Actes de la 11e conférence nationale sur les Applications Pratiques de l'Intelligence Artificielle

Nathalie Abadie, Ghislain Atemezing

AfIA

Association française
pour l'Intelligence Artificielle

APIA

11e conférence nationale sur les
Applications Pratiques de l'Intelligence Artificielle

PFIA 2025

Table des matières

Éditorial

11e conférence nationale sur les Applications Pratiques de l'Intelligence Artificielle

La onzième édition de la conférence nationale sur les Applications Pratiques de l'Intelligence Artificielle (APIA 2025) est hébergée par la plateforme PFIA, avec d'autres conférences francophones dans le domaine de l'intelligence artificielle (IA). Elle s'est tenue les 30 juin et 1er juillet 2025, à Dijon. Soutenue par le Collège Industriel de l'Association Française pour l'Intelligence Artificielle (AFIA), APIA est un lieu d'échanges entre les chercheurs académiques et les entreprises (industriels, entreprises de services, startups,...), pour partager leurs expériences dans le domaine de l'IA autour de cas d'usages concrets. L'objectif est double : démocratiser davantage l'utilisation de l'IA en contexte industriel et enrichir son potentiel applicatif pour les besoins des entreprises. Aussi la conférence s'intéresse-t-elle à tous les domaines de l'IA, tant que les méthodes proposées visent à résoudre des problèmes concrets, qu'ils soient industriels, sociétaux, économiques, politiques, environnementaux, patrimoniaux, culturels, etc.

Cette année la conférence APIA a reçu vingt-neuf soumissions d'articles. Vingt-et-un articles ont été acceptés, dont seize articles longs, quatre articles courts et un résumé d'article déjà publié à l'international. Chaque article a reçu deux relectures par des membres du comité de programme : un.e académique et un.e industriel.le. La conférence se tient sur deux jours, en ouverture de la plateforme PFIA. Elle comporte six sessions, sur des thèmes variées qui témoignent de la diversité des applications pratiques des différents domaines de l'IA :

- Jumeaux numériques et ingénierie des connaissances pour l'industrie.
- Réflexions sur l'IA générative.
- Interprétabilité, explicabilité.
- Grands modèles de langage.
- IA dans les systèmes embarqués.
- IA pour l'analyse de graphes, de textes et d'images.

Les deux dernières sessions sont organisées conjointement avec les Rencontres des Jeunes Chercheurs en Intelligence Artificielle (RJCIA) et regroupent des articles dont la première autrice ou le premier auteur est une jeune chercheuse ou un jeune chercheur (étudiant.e, doctorant.e, ou post-doctorant.e). Ces deux sessions ont été possibles grâce à une coordination avec Danai Symeonidou et l'aide du comité d'organisation de PFIA. Pour assurer une meilleure cohérence thématique des sessions, l'article *Repenser les collections patrimoniales par le prisme de l'IA* par Marion Charpier et Emmanuelle Bermès, est présenté dans une session de la journée thématique Humanités Numériques et IA (HNIA), mais demeure dans les actes de la conférence APIA à laquelle il a été soumis et accepté. Ceci a été décidé en accord avec les autrices et les responsables du comité de programme de la journée HNIA. Enfin, la plupart des sessions d'APIA 2025 accueillent au moins une présentation des partenaires Or et Platine de la plateforme PFIA :

- SKAIZen Group : Romain Alfred. *GDS : un réseau de neurones antagoniste générateur de messages SWIFT.*
- ARDANS : Patrick Prieur. *Intégration d'un système de gestion des connaissances conforme à l'ISO 30401, aux processus opérationnels existants d'une organisation.*
- SYARTEC : Eric Lebon et Hamid Ahaggach. *Analyse de données et apprentissage automatique pour la prise de décision intelligente dans le commerce automobile.*
- ENGIE : Guillaume Arbod, Ahmed Mabrouk. *Biométhane : optimisation par IA du plan d'incorporation des intrants.*
- Atol CD :

Nous les remercions pour leur soutien et leurs contributions à cette édition 2025.

Enfin, cette année, nous avons l'honneur d'accueillir Emiliano Lorini — directeur de recherche CNRS et co-responsable de l'équipe LILaC (Logic, Interaction, Language and Computation) à l'IRIT, Université Toulouse III – Paul Sabatier — dont la conférence invitée est intitulée *Raisonneurs Artificiels pour la Communication.*

Nous profitons de cet éditorial pour remercier chaleureusement les membres du comité de programme pour leur implication. Nous adressons également nos remerciements à l'ensemble des acteurs de la dynamique communauté APIA qui ont contribué au succès de cette édition avec leurs nombreuses soumissions, ainsi que le comité d'organisation de la plateforme PFIA 2025 qui a été d'un grand soutien et d'une grande efficacité.

Nathalie Abadie, Ghislain Atemezing

Comité de programme

Présidence

— Ghislain Atemezing (European Union Agency for Railways) ;
— Nathalie Abadie (LASTIG, Université Gustave Eiffel, IGN-ENSG).

Membres

— Florence Amardeilh (Elzeard) ;
— Fabien Amarger (Digitanie) ;
— Nicolas Audebert (LASTIG, Université Gustave Eiffel, IGN-ENSG) ;
— Nathalie Aussenac (IRIT, CNRS) ;
— Alain Berger (Ardans) ;
— Sandra Bringay (LIRMM) ;
— Xavier Briottet (Office National d'Etudes et de Recherches Aérospatiales) ;
— Stephan Brunessaux (Sensei Consult) ;
— Davide Buscaldi (LIPN, Université Paris 13, Sorbonne Paris Cité) ;
— Gaëtan Caillaut (Lingua Custodia) ;
— Bruno Carron (Airbus) ;
— Laurent Cervoni (Talan) ;
— Gaël de Chalendar (CEA LIST) ;
— Nicolas Chauvat (Logilab) ;
— Caroline Chopinaud (Hub France IA) ;
— Franck Cotton (Making Sense) ;
— Yves Demazeau (CNRS - LIG) ;
— Sylvie Despres (Laboratoire d'Informatique Médicale et de BIOinformatique) ;
— Gayo Diallo (ISPED & LABRI, Université de Bordeaux) ;
— Valentina Dragos (Office National d'Etudes et de Recherches Aérospatiales) ;
— Guillaume Dubuisson Duplessis (EDF) ;
— Françoise Fogelman-Soulié (Hub France IA) ;
— Nicolas Gonthier (LASTIG, Université Gustave Eiffel, IGN-ENSG) ;
— Céline Hudelot (Ecole Centrale Paris) ;
— Arnaud Lallouet (Huawei) ;
— Christine Largouët (Irisa - Agrocampus Ouest) ;
— Christelle Launois (Société Générale) ;
— Dino Lenco (INRAE) ;
— Dominique Lenne (Heudiasyc, Université de Technologie de Compiègne) ;
— Mustapha Lebbah (Université Paris-Saclay - UVSQ Versailles Campus) ;
— Sylvain Mahé (EDF R & D) ;
— Juliette Mattioli (Thales) ;
— Mathieu Roche (CIRAD, TETIS) ;
— Catherine Roussey (INRAE) ;
— Céline Rouveirol (LIPN, Université Paris 13) ;
— Brigitte Trousse (INRIA).

Session 1 : Jumeaux numériques et ingénierie des connaissances pour l'industrie

Contribution à la caractérisation de l'affordance d'un environnement de travail industriel : une approche basée sur l'apprentissage profond combinant données réelles et synthétiques.

Sarah Ouarab[1,2], David Garcia[1], Nicolas Ragot[3], Yohan Dupuis[4]

[1] CESI LINEACT, Lyon, France
[2] École Nationale Supérieure des Arts et Métiers, Paris, France
[3] CESI LINEACT, Rouen, France
[4] CESI LINEACT, Paris, France

souarab@cesi.fr

Résumé

Ce travail s'inscrit dans le cadre du projet "École De La Batterie", dont l'un des objectifs concerne l'optimisation de la conception des postes de travail manuel dans le but d'améliorer leur ergonomie. Nos travaux s'inscrivent dans cette démarche et visent à caractériser l'affordance des éléments de ces environnements avec lesquels les opérateurs interagissent (outils, composants, etc.) lors des opérations qu'ils réalisent. Un des verrous liés à cette problématique concerne la détection des éléments mobilisés par l'opérateur au cours de son activité pour aboutir à la caractérisation de leurs affordances. Les approches basées sur l'apprentissage profond fournissent de très bons résultats, mais nécessitent des bases de données d'apprentissage importantes. Dans le contexte industriel ces bases de connaissances labellisées n'existent pas ou sont en quantité très limitées pour ce type d'application. La méthode proposée repose sur un processus d'apprentissage automatique supervisé qui mobilise à la fois la génération de données réelles et synthétiques, qui sont respectivement issues de l'expérimentation et du jumeau numérique d'un poste de travail. Nous questionnons notamment la proportion de données réelles requises pour obtenir un modèle performant avec un effort de labellisation minimal, pour atteindre des performances de détection des outils cohérentes pour notre objectif de caractérisation des affordances du poste de travail.

Mots-clés

Affordance, environnement de travail, industrie, apprentissage profond, jumeau numérique, base de données réelle et synthétique.

Abstract

This work is part of the "École De La Batterie" project, one of whose objectives is to optimize the design of manual workstations in order to improve their ergonomics. Our contribution aligns with this goal by seeking to characterize the affordances of the elements in these environments (tools, components, workspace, etc.) with which operators interact during their tasks. One of the main challenges here is detecting the elements mobilized by the operator over the course of their work, in order to characterize their affordances. Deep learning–based approaches provide excellent results but require large training datasets. In industrial settings, however, such labeled datasets either do not exist or are available only in very limited quantities for this type of application. The proposed method relies on a supervised machine learning process that uses both real and synthetic data, derived from experiments and the digital twin of a workstation, respectively. In particular, we investigate the proportion of real data required to develop an effective model with minimal labelling effort, aiming to achieve consistent tool detection performance for our goal of characterising workstation affordances.

Keywords

Affordance, work environment, industry, deep learning, digital twin, real and synthetic dataset.

1 Introduction

Le concept d'affordance est né et a été popularisé en 1977 par Gibson [3], qui le définit comme les possibilités d'action ou d'utilisation offertes par un objet. En 1988, Norman reprend ce concept pour affirmer que l'affordance résulte de la relation entre les propriétés de l'objet et les capacités de l'agent [7]. En 2020, Simonian étend ce concept pour intégrer les inférences du sujet (l'humain dans notre cas) en situation réelle, en considérant que les affordances sont dépendantes de ce qu'il perçoit des propriétés de l'objet, pour agir [9]. Issu des champs disciplinaires de la psychologie cognitive et de la perception, le concept d'affordance a été progressivement investi par de nombreuses autres thématiques, notamment celles des sciences de l'ingénieur et de l'information et plus précisément celles de la robotique (sociale) et des interactions humains-systèmes [4]. Par ailleurs, la dynamique de l'industrie 5.0, qui repositionne l'humain au centre des systèmes de production, fait émerger de nombreux défis, notamment l'optimisation des inter-

actions entre les opérateurs et leur environnement de travail, ainsi que l'adaptation des outils et équipements aux besoins spécifiques des utilisateurs, ce qui implique une amélioration continue de l'ergonomie des postes de travail. Dans ce travail, nous cherchons à contribuer à cette problématique en caractérisant l'affordance des outils utilisés lors d'une tâche d'assemblage. Cela passe notamment par la détection des outils mobilisés par l'opérateur (cf. Figure 1). De nombreux travaux ont été menés sur les algorithmes d'apprentissage profond pour la détection d'objets. Cependant, ces méthodes étant particulièrement gourmandes en données d'entraînement, leur transposition au contexte industriel reste complexe en raison du manque de données annotées. Pour pallier ce problème, l'utilisation du jumeau numérique est intéressant, car il permet de générer des données synthétiques annotées massivement. Cependant, des modèles entraînés uniquement sur des données synthétiques n'offrent pas des performances acceptables [10]. Nos travaux visent à étudier une alternative qui repose sur des jeux de données d'entraînement mixtes puisqu'ils combinent des données réelles et synthétiques. En particulier, nous questionnons la proportion optimale de données réelles annotées nécessaires dans un jeu de données d'entraînement pour atteindre de bonnes performances en détection d'objets. Ici, le terme optimal renvoie au nombre minimal de données réelles nécessaire afin de minimiser le processus d'annotation tout en maximisant les performances du modèle. L'article est organisé de la façon suivante : nous présentons d'abord un bref état de l'art, suivi de notre approche méthodologique afin de créer les jeux de données et d'étudier les performances du modèle. Enfin, nous discutons des résultats obtenus avant de conclure et d'ouvrir sur des perspectives.

2 Travaux connexes

L'utilisation de données synthétiques pour entraîner des modèles d'apprentissage profond a montré son efficacité dans plusieurs domaines, par exemple pour le comptage de piétons [5] et celui de la classification et la détection de défauts dans l'acier [2]. Ces études démontrent que les données synthétiques peuvent combler les lacunes des ensembles de données réelles, notamment en cas de déséquilibre ou de manque de diversité. Une étude récente explore l'utilisation des données synthétiques pour l'entraînement des modèles de détection d'objets en milieu industriel [8]. L'étude évalue plusieurs proportions de données réelles et synthétiques pour l'entraînement des modèles de détection YOLOv8, et démontre qu'un jeu de données mixte de 890 échantillons, dont seulement 3% sont réels, permet d'atteindre des performances satisfaisantes. Enfin, l'étude de l'utilisation de jeux de données synthétiques pour l'estimation de la pose 3D d'objets industriels a été menée [6]. Elle confirme que les données synthétiques représentent une solution efficace face à la rareté des données réelles, tout en offrant une meilleure adaptation aux besoins spécifiques des cas d'usage. Ces travaux nous ont motivés à intégrer les données synthétiques à l'entraînement de notre modèle de détection, et à évaluer l'impact de cet ajout en fonction du ratio de données réelles utilisées. L'objectif est de déterminer le ratio minimal de données réelles qui permet d'obtenir des performances acceptables en termes de précision des détections. Cette approche vise à réduire la dépendance aux données réelles annotées manuellement, souvent coûteuses et limitées en quantité.

3 Méthode

Dans ce travail, nous nous intéressons à la caractérisation de l'affordance des outils utilisés par un opérateur humain sur un poste de travail, dans le cadre d'une tâche d'assemblage de composants. Cette caractérisation repose sur la détection des outils à l'aide de nos caméras, permettant ainsi le suivi en temps réel de leur utilisation. Notre approche repose sur l'apprentissage par transfert (*transfer learning*), une technique qui consiste à initialiser le modèle avec des poids pré-entraînés sur un grand ensemble de données (COCO pour YOLOv9) avant de l'adapter à notre tâche spécifique. L'entraînement a été réalisé sur 100 epoch avec une taille de lot (*batch size*) de 16 en utilisant les poids de YOLOv9 (Gelan-C).

3.1 Indicateurs de performance

Pour évaluer les performances du modèle, nous avons choisi d'utiliser les indicateurs mAP@0.5 (mAP : mean Average Precision) et mAP@0.5-95 et le f1-score, employés pour la détection d'objets :

- **mAP@0.5** : seuil de chevauchement (Intersection over Union, IoU) de 50% qui évalue la capacité du modèle à détecter et localiser les objets dans l'image.
- **mAP@0.5-0.95** : moyenne des mAP pour un intervalle de seuils d'IoU de [50%,$\cdots$,95%]. Ceci fournit une mesure plus précise de la détection et de la localisation des objets.
- **f1-score** : évalue la performance du modèle à distinguer les vrais positifs des faux positifs et des faux négatifs.

3.2 Collecte de données réelles

Des données réelles ont été collectées lors d'expérimentations sur un poste de travail manuel réel, où un opérateur réalise diverses actions avec les outils. Un ensemble de 200 images réelles a été enregistré. 80% servent à composer les bases de données d'entraînement et 20% sont dédiées à la base de données de test. L'annotation de ces images a été réalisée manuellement à l'aide de l'outil Roboflow, garantissant des annotations précises adaptées au format YOLO. Le temps d'annotation d'une image réelle est de l'ordre de 5 minutes sachant que celui-ci peut considérablement s'allonger en fonction de la complexité de la scène traitée et du nombre d'éléments à annoter.

3.3 Génération des données synthétiques

Les données synthétiques ont été générées avec Unity 3D et son module *Perception Package*, permettant la création automatisée de jeux de données annotés pour la vision par ordinateur. Afin de simuler un poste de travail manuel dans

des conditions réalistes, le Jumeau Numérique de l'Atelier Flexible de Production (JN-UFP), développé par l'équipe de recherche du campus CESI Rouen, a été exploité (cf. Figure 1-a). Ce jumeau numérique a été grandement amélioré dans le cadre du projet JENII [1], notamment son niveau de détail et la qualité de son rendu 3D hautement réaliste. Les outils modélisés au sein du JN-UFP correspondent aux cinq classes d'outils principales présentes sur le poste de travail réel, illustrées dans la Figure 1 : tournevis cruciforme, tournevis plat, clé Allen, clé plate et un jeu de clés Allen. La génération des données synthétiques a mobilisé des ressources de calcul modérées : environ 2 heures ont suffi pour produire 320 images sur une station dotée d'un processeur Intel Xeon W-2245 (3,9 GHz, 8 cœurs, 16 threads, 16,5 Mo de cache L3) et de 64 Go de RAM ainsi qu'une carte graphique NVIDIA Quadro RTX 6000.

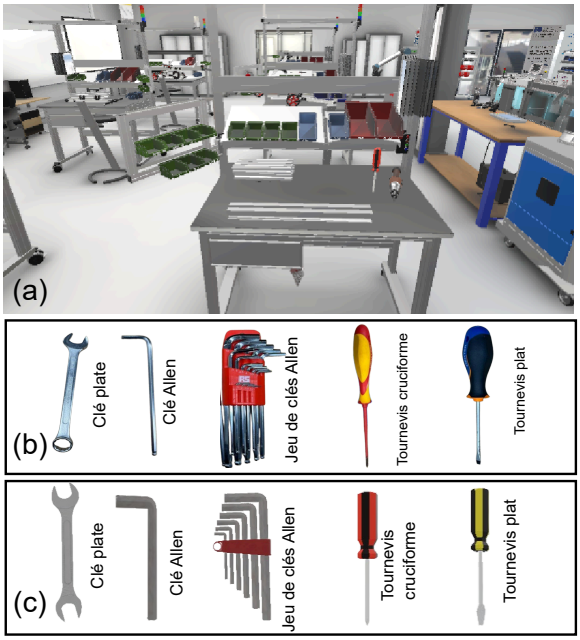


FIGURE 1 – (a) Vue d'un poste de travail du Jumeau numérique UFP (b) et (c) Vues des cinq classes d'outils mobilisés respectivement au sein du poste de travail réel et du poste de travail virtuel du JN-UFP.

4 Résultats et discussion

L'objectif de cette analyse est d'évaluer l'influence de la proportion des données réelles dans un jeu de données d'entraînement mixtes. Notre démarche vise à caractériser l'effort d'annotation de données réelles requis pour atteindre des performances acceptables pour l'application visée.

4.1 Performances obtenues avec un entraînement sur des données réelles

Pour rappel 200 images constituent le jeu de données réelles qui a été scindé en 160 images d'entrainement et 40 images pour le test. Ainsi le modèle entraîné sur ces données réelles a obtenu des performances prometteuses avec un mAP@0.5 de 0.943, un mAP@0.5-0.95 de 0.66 et un f1-score de 0.92. Ces résultats démontrent une bonne capacité de détection des objets avec un seuil d'IoU de 50%, bien que la précision diminue avec des seuils plus stricts. Néanmoins, un effort d'annotation important a été nécessaire pour labelliser les images. Dès lors, il est intéressant de questionner la proportion minimale de données réelles nécessaire dans le jeu de données d'entraînement en y ajoutant des données synthétiques issues du jumeau numérique.

4.2 Performances obtenues à partir de données mixtes

Nous avons créé onze bases de données d'entraînement contenant un nombre variable d'images synthétiques et réelles. La taille totale des différents jeux de données a été maintenue constante à 320 images. Le nombre d'images réelles dans les différents jeux de données d'entraînement a été progressivement augmenté jusqu'à atteindre un ratio de 50% d'images réelles, à savoir 160 images (totalité du nombre d'images réelles disponibles pour entraîner un modèle). La base de test est identique à celle utilisée pour l'entraînement sur le jeu de données réelles et commune à tous les jeux de données d'apprentissage mixtes : 40 images réelles (cf.4.1). Les résultats présentés Figure 2 montrent d'une part l'évolution des indicateurs mAP@0.5(mixtes), mAP@0.5-0.95(mixtes) et f1-score(mixtes) pour les modèles entraînés sur les bases de données mixtes, en fonction du pourcentage de données réelles (courbes pleines) et d'autre part l'évolution des indicateurs mAP@0.5(réelles), mAP@0.5-0.95(réelles) et f1-score(réelles) pour les modèles entraînés uniquement à partir des données réelles considérées pour chacune des bases de données mixtes (courbes en pointillés).

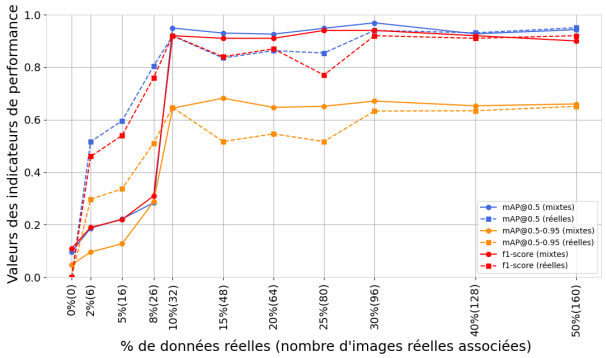


FIGURE 2 – Évolution des indicateurs mAP@0.5(mixtes), mAP@0.5-0.95(mixtes) et f1-score(mixtes) (courbes pleines) / Évolution des indicateurs mAP@0.5(réelles), mAP@0.5-0.95(réelles) et f1-score(réelles) (courbes en pointillés).

On observe que lorsque le jeu de données mixtes comporte moins de 10% de données réelles, les performances via les indicateurs mAP@0.5(mixtes), mAP@0.5-0.95(mixtes) et en f1-score(mixtes) se révèlent faibles. En revanche, dès que la proportion de données réelles est au delà de 10%,

les scores mAP@0.5(mixtes), mAP@0.5–0.95(mixtes) et f1-score(mixtes) augmentent drastiquement puis se stabilisent à des niveaux de performances beaucoup plus élevés, compris respectivement dans les intervalles [0.926, 0.969], [0.644, 0.682] et [0.90, 0.94]. On note également qu'entre 10% et 50%, la dispersion des valeurs autour de la valeur médiane, calculée pour chacun des indicateurs f1-score(mixtes), mAP@0.5(mixtes) et mAP@0.5-0.95(mixtes), varie peu. La meilleure performance est obtenue avec un ratio de 30% de données réelles : mAP@0.5(mixtes) de 0.969, mAP@0.5-0.95(mixtes) de 0.671 et un f1-score(mixtes) de 0.94. Si, désormais on s'intéresse aux évolutions des indicateurs mAP@0.5(réelles), mAP@0.5-0.95(réelles) et f1-score(réelles) comparativement à celles obtenues pour les indicateurs mAP@0.5(mixtes), mAP@0.5-0.95(mixtes) et f1-score(mixtes) on observe, sur la figure 2 trois comportements distinctifs. Pour l'intervalle [0%(0) à 10%(32)[, les performances des indicateurs (mixtes) sont très inférieures à celles des indicateurs (réelles). Ceci souligne que les données synthétiques associées aux données réelles dégradent ici globalement la performance des modèles. Pour l'intervalle [10%(32) à 30%(96)], les performances des indicateurs (mixtes) sont supérieures à celles des indicateurs (réelles). Les données synthétiques associées aux données réelles contribuent ici de manière significative aux performances des modèles sauf pour le jeu de données à 10% où seul la valeur de l'indicateur mAP@0.5(mixtes) se détache légèrement de celle de mAP@0.5(réelles). Pour l'intervalle]30% (96), 50% (160)], l'ajout de données synthétiques n'apporte plus d'amélioration significative sur les performances des modèles. En conclusion, les résultats obtenus pour l'intervalle de pourcentage de données réelles [10% 30%] mettent en évidence l'intérêt d'intégrer à un noyau minimal de données réelles des données synthétiques pour minimiser l'effort d'annotation tout en atteignant une performance acceptable pour le cas d'étude considéré.

5 Conclusion

Ce travail s'inscrit dans la perspective de contribuer à la caractérisation des affordances des éléments constitutifs d'un environnement de travail en développant un modèle de détection et de suivi de trajectoires d'outils basé sur l'apprentissage profond. Pour pallier le manque de données annotées en milieu industriel, nous avons combiné des données réelles issues d'expérimentations et des données synthétiques générées à partir d'un jumeau numérique. Nos expériences ont montré que l'ajout de données synthétiques à un noyau minimal de données réelles permet de franchir un palier en termes de performances. Dans le cadre de notre cas d'étude, cette proportion est comprise entre 10% et 30% pour la base d'entraînement. Les pistes d'amélioration envisagées portent sur l'optimisation de la contribution des données synthétiques aux performances du modèle, tout en visant à réduire la quantité de données réelles pour minimiser l'effort d'annotation. Selon nos analyses en cours, cela implique notamment l'amélioration du processus de génération des données virtuelles, par l'augmentation entre autres du nombre de points de vue et la réduction de l'écart de réalité entre les environnements réel et virtuel.

Remerciements

Ces travaux sont financés dans le cadre du projet Ecole de la Batterie (EDLB), opération soutenue par l'État dans le cadre de l'AMI « Compétences et Métiers d'Avenir » du Programme France 2030, opéré par la Caisse des Dépôts » (La Banque des Territoires).

Références

[1] ANR. Projet JENII. https ://anr.fr/ProjetIA-21-DMES-0006, 2025.

[2] Aleksei Boikov, Vladimir Payor, Roman Savelev, and Alexandr Kolesnikov. Synthetic data generation for steel defect detection and classification using deep learning. *Symmetry*, 13 :1176, 2021.

[3] James Jerome Gibson. *The Ecological Approach to Visual Perception*. Houghton Mifflin, 1979.

[4] Lorenzo Jamone, Emre Ugur, Angelo Cangelosi, Luciano Fadiga, Alexandre Bernardino, Justus Piater, and Jose Santos-Victor. Affordances in psychology, neuroscience, and robotics : A survey. *IEEE Transactions on Cognitive and Developmental Systems*, 10 :4–25, 2018.

[5] Hadi Keivan Ekbatani, Oriol Pujol, and Santi Segui. Synthetic data generation for deep learning in counting pedestrians. In *Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods*, page 318–323. SCITEPRESS - Science and Technology Publications, 2017.

[6] Aristide Laignel, Nicolas Ragot, Fabrice Duval, and Sarah Ouarab. *Synthetic Datasets for 6D Pose Estimation of Industrial Objects : Framework, Benchmark and Guidelines*, page 227–241. Springer Nature Switzerland, 2024.

[7] Donald A Norman. *The design of everyday things*. Bantam Doubleday Dell Publishing Group, 1990.

[8] Sarah Ouarab, Rémi Boutteau, Katerine Romeo, Christele Lecomte, Aristid Laignel, Nicolas Ragot, and Fabrice Duval. *Industrial Object Detection : Leveraging Synthetic Data for Training Deep Learning Models*, page 200–212. Springer Nature Switzerland, 2024.

[9] Stéphane Simonian, Rawad Chaker, and Jonathan Kaplan. Affordance en e-formation et régulation de l'apprentissage : une exploration dans un contexte d'études universitaires. *TransFormations : Recherche en éducation et formation des adultes*, 2020.

[10] T Zgheib, H Borges, V Feldman, T Guntz, C Di Loreto, O Desmaison, and F Corduant. Détection d'objets en temps réel : Entraînement de réseaux de neurones convolutifs sur images réelles et synthétiques. In *Conférence Nationale en Intelligence Artificielle*, 2022.

Utilisation d'intelligence artificielle pour la gestion autonome de bâtiments

Stéphane Reynaud[1,2], Anthony Dumas[1], Ana Roxin[2], Ludovic Journaux[3]

[1] B27-AI, Recherche & Développement

[2] Université Bourgogne Europe, Laboratoire d'Informatique de Bourgogne (LIB) UR 7534

[3] Institut Agro Dijon, Laboratoire d'Informatique de Bourgogne (LIB) UR 7534

sreynaud@b27.fr, adumas@b27.fr, ludovic.journaux@agro-dijon.fr, ana-maria.roxin@ube.fr

Résumé

Cette recherche explore l'intégration de l'intelligence artificielle (IA) dans les systèmes de gestion des bâtiments pour créer des bâtiments intelligents et autonomes, en combinant modèles numériques, données de capteurs et information d'usage. L'étude aborde les défis de l'efficacité énergétique, du comportement des usagers et de la conformité réglementaire. Les principales questions scientifiques incluent le développement de modèles de connaissances, la gestion de grands ensembles de données et l'autonomie de l'exploitation des bâtiments à travers des stratégies pilotées par l'IA, en utilisant des méthodes d'IA neuro-symbolique.

Mots-clés

Bâtiment, intelligence artificielle neuro-symbolique, jumeau numérique.

Abstract

This research explores the integration of artificial intelligence (AI) into building management systems to create intelligent and autonomous buildings, by combining digital models, sensor data, and usage information. The study addresses challenges in energy efficiency, user behavior, and regulatory compliance. Key scientific questions include the development of knowledge models, the management of large data sets, and the autonomy of building operations through AI-driven strategies, using neuro-symbolic AI methods.

Keywords

Buildings, Neuro-symbolic artificial intelligence (NSAI), Digital Twin (DT).

1 Introduction

L'intelligence artificielle (IA) change la façon dont nous gérons et exploitons les bâtiments, transformant les systèmes de gestion des bâtiments (« Building Management System », BMS) traditionnels en entités intelligentes capables d'optimiser leur fonctionnement, de prendre des décisions et d'effectuer des prédictions [1]. Cette évolution des BMS représente un changement de paradigme dans l'industrie de l'architecture, de l'ingénierie et de la construction (« Architecture, Engineering and Construction », AEC).

Aussi, dans le cadre d'une thèse CIFRE, l'entreprise B27-AI, bureau d'étude et d'ingénierie spécialisé dans le secteur de l'AEC, et le Laboratoire d'Informatique de Bourgogne (LIB) collabore pour permettre de coupler les données de la maquette numérique à des données capteurs et à des données d'usage, afin de transformer ces données en des connaissances supportant la prise de décision. Dans ce cadre, nous supposons la maquette numérique existante et disponible « telle que construite ». La finalité est de permettre à un bâtiment de se comporter de manière autonome.

En effet, les systèmes de BMS sont depuis longtemps la pierre angulaire de l'exploitation des bâtiments, permettant le contrôle centralisé de divers systèmes tels que le chauffage, la ventilation et la climatisation (« Heating, Ventilation, and air Conditioning », HVAC), l'éclairage, et la sécurité. Cependant, les bâtiments deviennent plus complexes, avec des demandes d'efficacité énergétique et de confort de leurs usagers en augmentation. Faire évoluer les solutions de BMS traditionnelles s'avère un véritable défi, à la fois métier et scientifique [2].

Or, l'intégration d'IA dans les BMS, notamment à travers l'utilisation d'algorithmes d'apprentissage automatique et l'analyse de vastes quantités de données provenant de capteurs et d'équipements divers [1], permet l'évolution vers une gestion plus autonome des bâtiments.

Malgré les nombreux défis existants, comme les coûts d'intégration, la faible qualité des données existantes et la grande hétérogénéité des systèmes [3], ces systèmes autonomes peuvent être capables de prendre des décisions complexes sans intervention humaine, comme :

1. Prédire et prévenir les pannes d'équipement avant qu'elles ne surviennent [4]
2. Ajuster dynamiquement les systèmes du bâtiment en fonction de leur occupation ou utilisation et de facteurs externes [5]
3. Optimiser la consommation d'énergie et réduire l'empreinte carbone [2]
4. Améliorer le confort des usagers grâce à des contrôles environnementaux personnalisés [6]

Ainsi, le verrou consiste en une approche pour des bâtiments autonomes, capables de s'adapter et de prendre des décisions par rapport aux interactions avec les usagers. Les bâtiments devenant des entités intelligentes capables d'effectuer les

opérations et la maintenance quotidiennes, les gestionnaires d'installations peuvent se concentrer sur la mise en œuvre de décisions complexes et stratégiques.

2 Contexte et problématiques métier

Cette section fournit le contexte et les problématiques métiers qui sous-tendent la nécessité de faire évoluer la gestion des bâtiments.

Enjeux environnementaux. Les bâtiments consomment d'importantes ressources, en particulier lors de leur construction, et il est estimé qu'ils sont responsables d'environ 40% de la consommation totale d'énergie dans les pays industrialisés [5]. Malgré les efforts visant à réduire l'impact environnemental en cours depuis les années 1970, le défi reste impératif face au changement climatique. Les technologies numériques offrent des possibilités de gérer plus efficacement le cycle de vie des bâtiments, permettant de prendre des décisions éclairées en matière de durabilité. Selon les objectifs de développement durable (« Sustainable Development Goals », SDG) de l'ONU, les futurs bâtiments doivent s'appuyer sur une énergie propre et promouvoir santé et bien-être dans des villes durables. Le Pacte vert pour l'Europe de l'Union Européenne (UE) vise un impact climatique nul d'ici 2050, avec des initiatives de rénovation et de numérisation des bâtiments. Le nouveau Bauhaus européen encourage les approches créatives et interdisciplinaires pour concevoir des espaces durables, inclusifs et agréables [7].

En réponse à ces défis, les normes réglementaires évoluent et intègrent des exigences strictes concernant l'empreinte carbone des bâtiments (e.g. le label « Bâtiments à Énergie Positive et Réduction Carbone » E+C-), le suivi des matériaux via des passeports numériques, et la certification des performances environnementales (e.g. Haute Qualité Environnementale ou HQE). Ces évolutions contrastent fortement avec l'inertie inhérente du secteur et le temps long des constructions, soulevant ainsi le défi majeur de l'adaptabilité du bâti.

Données liées au bâtiment. Les technologies numériques dans le secteur de l'AEC reposent historiquement sur la modélisation des informations du bâtiment (« Building Information Modeling », BIM) à travers toutes les phases du cycle de vie du bâtiment, de sa conception à sa démolition et son recyclage, en passant par sa construction, son exploitation et sa maintenance. Cependant, le BIM donne généralement une représentation statique des diverses phases, et l'arrivée de nouvelles technologies comme l'internet des objets (« Internet of Things », IoT) ont montré le besoin d'évoluer vers une représentation plus dynamique des bâtiments et une intégration en temps réel des données [8].

Afin de faciliter le partage d'information et de limiter l'hétérogénéité des solutions propriétaires, il a été créé l'openBIM, un processus collaboratif soutenu par l'organisation BuildingSMART International, mettant l'accent sur l'interopérabilité des données des projets de construction grâce à des normes ouvertes et des vocabulaires communs. « Industry Foundation Classes » (IFC) est le format d'échange standard openBIM pour l'échange de maquettes numériques de bâtiments et d'infrastructures dans le secteur de l'AEC, souvent caractérisé par des structures extrêmement complexes. Cette complexité a conduit à des défis dans la gestion des données IFC : bien qu'étant un vecteur important d'interopérabilité, l'IFC atteint ses limites lorsqu'il s'agit d'intégrer, traiter, manipuler, extraire et analyser des données provenant de maquettes numériques, de sources multiples et/ou de domaines adjacents [9].

Gestion des bâtiments axée sur les usages. Les bâtiments sont de plus en plus gérés de manière à optimiser non seulement la performance énergétique, mais également le confort global des usagers [2]. Les systèmes traditionnels basés sur des règles (e.g., les systèmes de gestion des bâtiments avec des points de consigne fixes) ne parviennent souvent pas à saisir la complexité et la variabilité des usages [10]. Les approches centrées sur les usages exploitent les données des capteurs (e.g., les mouvements, le CO_2, la température), les informations contextuelles (e.g. les horaires d'occupation) et les commentaires des usagers (par ex., les votes sur le confort thermique) [6].

Problématiques métiers. À partir du contexte présenté auparavant, 3 problématiques métiers ont été identifiées.

La première problématique métier (**PM1**) concerne l'enrichissement des représentations de la connaissance des bâtiments en intégrant les relations avec l'environnement et les usagers.

La seconde problématique métier (**PM2**) s'intéresse à la gestion des règles représentant des contraintes règlementaires, des comportements des usagers, ou des fonctionnements du bâtiment.

Enfin, la troisième problématique métier (**PM3**) consiste à interagir efficacement avec les bâtiments et à permettre l'adaptation dynamique du comportement de ces bâtiments en fonction de stratégies.

3 Problématiques scientifiques

Des objectifs et des problématiques métiers du projet de recherche découlent les problématiques scientifiques suivantes.

La première problématique scientifique (**PS1**), lié à la problématique métier PM1, porte sur la modélisation des connaissances intégrant les bâtiments, leurs systèmes et capteurs, leurs contextes environnementaux et règlementaires, et leurs relations avec les usagers. Pour construire cette représentation, il s'agira d'étudier les différentes ontologies de domaines existantes, de les complémenter au besoin, de les assembler et de combler les manques de données.

La problématique métier PM2 soulève la seconde problématique scientifique (**PS2**), qui concerne la définition, la structuration et l'interaction de règles liées à la gestion des systèmes de bâtiments, aux préférences et comportements des usagers et à des contraintes externes (e.g. règlementaire). Il sera nécessaire d'étudier l'utilisation de règles logiques exprimées par des humains ou construites par des ordinateurs, et de s'intéresser aux mécanismes de transformation, d'apprentissage et d'adaptation de ces règles.

Enfin, la troisième problématique scientifique (**PS3**)

concerne l'autonomie de fonctionnement du bâtiment et découle de la problématique métier PM3. Pour parvenir à établir un système de gestion autonome contrôlé par des stratégies et encadré par des contraintes, il sera nécessaire d'étudier les différentes approches permettant de bâtir un jumeau numérique du bâtiment particulièrement adapté au raisonnement.

Pour l'ensemble de ces trois problématiques scientifiques, les approches d'IA symboliques semblent insuffisantes pour obtenir les attendus de niveau d'adaptation au contexte, d'efficacité dans le traitement des grands ensembles de données, et de flexibilité dans la gestion du bâtiment. Par ailleurs, les approches d'IA statistiques semblent également insuffisantes pour les besoins d'interprétabilité, de transparence et d'explicabilité considérés.

En effet, en prenant l'exemple du comportement des usagers, qui est aléatoire, dépendant du contexte, et hétérogène, les règles symboliques (e.g., « si un usager est présent, régler la température à 22°C ») sont souvent trop rigides pour prendre en compte au changement et s'adapter au contexte. Et si les modèles statistiques basés sur les données peuvent être efficaces, ils sont souvent opaques, peuvent être sur/sous-ajustés, ou entrer en conflit avec les réglementations et les exigences de sécurité.

4 Contexte scientifique

Dans cette section, nous abordons les fondements théoriques et les technologies scientifiques essentielles pour adresser chaque problématique.

4.1 Structuration des connaissances (PS1)

La modélisation de la connaissance est centrale pour représenter les composants du bâtiment, ses usages et son environnement.

Une **ontologie** est, dans un domaine spécifique, une représentation formelle des connaissances, composée de la définition de ses concepts, de ses propriétés et de leurs relations. Ces relations sémantiques servent de fondations pour le raisonnement et l'inférence [11].

Une **base de connaissances** (« Knowledge Base », KB) est une collection de connaissances structurées à l'aide d'ontologies : les instances de la KB sont appelées connaissances assertionnelles ou ABox (« Assertional box ») et son ontologie la TBox (« Terminological box ») [11].

Un **graphe de connaissances** (« Knowledge Graph », KG) est une forme particulière de KB, qui relie des connaissances sous forme de graphe, mettant ainsi l'accent sur les relations entre différentes entités (e.g. des personnes, des lieux, des objets, des concepts) [12]. Un KG peut exploiter les principes des données liées (« Linked Data » [13]) pour relier des connaissances définies dans différentes ontologies.

4.2 Gestion de règles et contraintes (PS2)

L'objectif de PS2 est de formaliser les comportements attendus, contraintes et préférences sous forme de règles logiques, utilisables par les systèmes de gestion bâtiment.

Une **règle logique** est une expression logique prenant la forme d'une implication, combinant des antécédents avec des connecteurs logiques pour déduire des conséquents. Elles sont formulées au-dessus des ontologies afin de contraindre les connaissances qui y sont spécifiées, et les faire correspondre à un cas d'application ou une problématique spécifique. [14].

L'**IA symbolique** repose sur l'idée qu'un système intelligent peut être représenté par un modèle composé de symboles et de règles logiques explicites[15]. Ainsi, la connaissance est représentée de manière symbolique à l'aide d'un langage formel. Les raisonnements sont considérés comme des opérations formelles appliquées aux expressions et structures combinant les symboles du modèle.

4.3 Raisonnement et autonomie (PS3)

Le développement de la capacité d'autonomie du bâtiment repose sur des stratégies intelligentes d'adaptation, alimentées par des données réelles.

L'**IA statistique** se distingue de l'IA symbolique en privilégiant des approches basées sur les données et le calcul plutôt que sur les représentations symboliques [15]. La connaissance est encodée en données numériques et en probabilités, et traitée à l'aide de méthodes telles que l'apprentissage automatique et les algorithmes d'optimisation.

L'**IA neuro-symbolique** (« Neuro-symbolic artificial intelligence », NSAI) est un sous-domaine de l'intelligence artificielle qui combine les approches neuronales (i.e. statistiques) et symboliques [16]. L'IA neuro-symbolique utilise la logique formelle tirée du domaine de la représentation des connaissances et du raisonnement.

L'**apprentissage par renforcement** (« Reinforcement Learning », RL) est une catégorie d'algorithmes d'apprentissage automatique dans laquelle un agent intelligent apprend une politique optimale, en interagissant avec un environnement et en recevant des retours sous forme de récompenses ou de pénalités [17].

Un **jumeau numérique** (« Digital Twin », DT) est une « représentation numérique d'un objet observable, synchronisée avec celui-ci » [18]. Dans le cas d'un jumeau numérique de bâtiment (« Digital Building Twin », DBT), il s'agit d'un modèle virtuel du bâtiment et de ses composants, créé à partir de données et capteurs internes, permettant de surveiller et d'analyser ses performances en temps réel, et d'agir de manière éclairée sur sa forme physique [19].

Les **systèmes cyber-physiques** (« Cyber-Physical System », CPS) représentent une intégration de capacités de calcul et de processus physiques. Dans ces systèmes, la partie cybernétique (ordinateurs et réseaux) surveille et contrôle les processus physiques, avec des boucles de rétroaction [20].

5 État des lieux

Cette section analyse l'état actuel des recherches et des applications en lien avec les problématiques soulevées.

Apprentissage de règles logiques. L'extraction automatique de règles à partir de données utilise des approches neuro-symboliques [21] pour une meilleure précision [22] et une interprétabilité renforcée [23]. Il est possible d'utiliser ces approches dans le domaine de l'AEC, pour extraire automatiquement des règles, à partir de journaux de comportement des usagers ou de flux de

capteurs (e.g. température, humidité, CO_2, capteurs de mouvement, comptage d'occupation). Les KB, qui codent les contraintes de domaine, peuvent être intégrées aux règles apprises [24] et permettre d'établir une meilleure classification et détection d'éléments [25]. Appliquées au domaine de l'AEC, ces approches peuvent permettre d'intégrer des KB comportant, par exemple, les plages de confort thermique, les normes, et les préférences des usagers.

KG neuro-symboliques. Les KG peuvent être utilisés pour représenter plus finement les relations entre les zones du bâtiment, les usagers, les équipements et les éléments de contexte (e.g. météo, horaires d'occupation), en utilisant des ontologies adaptées aux divers domaines [26]. Les approches de raisonnement neuro-symbolique sur les KG [27] permettent notamment l'extraction et la synthèse de connaissances [28], l'apprentissage de règles logiques [29], et la vérification d'implications et de relations [30]. Dans le domaine de l'AEC, ces approches peuvent permettre de dériver des stratégies de contrôle basées sur les usages, ou pour détecter des anomalies (par ex., des données de capteurs contradictoires).

RL dans l'AEC. L'intégration de contraintes symboliques dans des méthodes de RL [17] peut permettre d'équilibrer les préférences spécifiques des usagers avec la consommation énergétique globale du bâtiment [31]. Les approches de RL peuvent également renforcer l'interprétabilité des décisions [32], consolidant la confiance des utilisateurs dans ces décisions [33]. Dans le domaine de l'AEC, ce type d'approche peut notamment être utiliser pour l'apprentissage des politiques de confort des usagers, de manière interprétable.

Intégration dans les CPS. Les CPS sont des systèmes qui reposent sur l'intégration transparente des composants physiques et des composants cybernétiques. Dans le domaine de l'AEC, les CPS sont peu étudiés, surtout dans la phase d'exploitation : les DBT leur sont préférés, car ils peuvent être vus comme une forme particulière de CPS, répondent davantage au besoin de simulation du secteur, et sont moins susceptibles à la faible densité de capteurs des bâtiments [34].

Intégration dans les DBT. Les DBT servent de répliques virtuelles pour l'intégration de données en temps réel et la modélisation prédictive. Les approches neuro-symboliques peuvent être intégrées dans les DBT pour simuler le comportement des usagers, adapter les règles de contrôle des bâtiments, puis les valider par rapport aux performances physiques du bâtiment [35], ou pour améliorer les interactions avec les utilisateurs [36]. La création, le positionnement et l'utilisation de capteurs virtuels dans les DBT peuvent améliorer la captation de données et compenser en partie la faible densité de capteurs physiques [37]. Les stratégies de maintenance réactives peuvent être transformer en approches proactives grâce à l'utilisation d'IA dans les DBT [38].

6 Contours de notre approche

Pour adresser l'ensemble des problèmes scientifiques exprimées, les technologies sémantiques constituent une approche solide et efficace. En utilisant des ontologies et des graphes de connaissances, ces technologies permettent de mieux structurer, lier et interpréter les données issues de sources diverses et de formats variés. Cependant, elles présentent des limitations, notamment en termes de standardisation et d'adaptation à des systèmes existants. L'adoption des technologies sémantiques nécessite également une approche rigoureuse pour garantir la cohérence et la qualité des données à travers les différents processus de conception et de construction. En outre, le passage à une gestion de données dynamiques impose des efforts supplémentaires en termes de mise à jour continue et de traitement en temps réel, ce qui peut engendrer des défis supplémentaires sur le plan de la performance et de l'extensibilité [39].

Les principales ontologies en lien avec le domaine de l'AEC ne permettent pas de capturer complètement, et avec la finesse attendue, les connaissances spécifiques de B27-AI. Une première approche va donc être de spécifier la connaissance métier de l'entreprise, en analysant, modifiant et alignant les ontologies existantes concernant les capteurs, la maquette numérique, ou encore le profil des usagers. Des travaux ont été démarrés pour créer une ontologie spécifique, alignée avec IFC [40]. Les travaux d'alignement seront poursuivis afin d'aligner l'ensemble des éléments de cette ontologie avec leurs équivalents tels qu'existants dans les ontologies du domaine i.e. BRICK [1], SAREF [2], BOT [3], SSN/SOSA [4]. L'idée est de permettre l'interopérabilité sémantique [41] entre ces vocabulaires en appliquant les principes de conception d'ontologies [13] et en répondant à la problématique scientifique PS1. Ceci permettra de constituer le vocabulaire métier B27-AI, qui sera la base de notre approche.

En plus des différentes ontologies, B27-AI a commencé à travailler sur la spécification d'ensembles de règles à destination de systèmes en lien avec le BMS, et prenant en compte les préférences des usagers (e.g. ouverture/fermeture de volets roulants en fonction de capteurs de luminosité et du profil de l'occupant de la pièce). Une seconde approche va donc être de spécifier sous forme logique, en suivant les termes de l'ontologie, différentes règles pour exprimer notamment les comportements des systèmes, les préférences des usagers, les habitudes d'usages, et les contraintes règlementaires. Il s'agira d'étudier également l'ajout de critères aux règles initiales, la sélection d'un ensemble de règles nécessaires et suffisantes par rapport à des critères, ou encore la priorisation entre les diverses règles. L'objectif est de parvenir à une classification de l'ensemble de ces règles par niveau de flexibilité et de granularité afin de répondre à la problématique scientifique PS2.

Grâces aux données d'usages et de capteurs couplées à celle de la maquette numérique, il est possible de créer un jumeau numérique du bâtiment. Une troisième approche va être l'ajout de connaissances, de règles et d'objectifs dans ce jumeau numérique puis de le doter de capacité de raisonnement et de prédiction. Il s'agira d'étudier la construction de stratégies de gestion du bâtiment, utilisant les règles existantes et en les adaptant en suivant les diverses

[1] https://brickschema.org/
[2] https://w3id.org/saref
[3] https://w3c-lbd-cg.github.io/bot/
[4] http://www.w3.org/ns/ssn/

contraintes (e.g. par RL neuro-symbolique). Il s'agira d'étudier également l'exploitation des connaissances en lien avec les stratégies de gestion pour permettre le comportement autonome du bâtiment, permettant ainsi de répondre à la problématique scientifique PS3.

7 Conclusion

L'intégration de l'IA dans la gestion des bâtiments représente une avancée significative vers l'automatisation et l'optimisation de leur exploitation. Grâce aux technologies telles que les jumeaux numériques, l'internet des objets (IoT), et les graphes de connaissances, les bâtiments peuvent évoluer pour devenir des entités intelligentes capables de gérer de manière autonome des processus complexes, en assimilant les comportements des usagers et les contraintes environnementales et règlementaires.
Toutefois, plusieurs défis demeurent, notamment en matière d'interopérabilité sémantique des données, de qualité des informations collectées, et d'intégration des systèmes existants. L'utilisation d'approches neuro-symboliques semble prometteuse pour surmonter ces obstacles. Les résultats de cette étude montrent qu'un cadre intégrant à la fois des connaissances et raisonnements symboliques, des règles logiques et des algorithmes d'apprentissage automatique pourrait répondre à ces défis.

Remerciements

Nous tenons à exprimer notre gratitude à B27-AI pour ses contributions financières et matérielles, et à l'Agence Nationale de la Recherche et de la Technologie (ANRT) pour sa subvention CIFRE.

8 Références

[1] S. K. Baduge *et al.*, « Artificial intelligence and smart vision for building and construction 4.0: Machine and deep learning methods and applications », *Automation in Construction*, vol. 141, p. 104440, sept. 2022, doi: 10.1016/j.autcon.2022.104440.

[2] F. Ferrada, A. I. Oliveira, J. Rosas, P. Macedo, R. Almeida, et L. M. Camarinha-Matos, « Development of a Conceptual Architecture for the Energy Management of Building Ecosystems », in *Collaborative Networks and Digital Transformation*, L. M. Camarinha-Matos, H. Afsarmanesh, et D. Antonelli, Éd., Cham: Springer International Publishing, 2019, p. 418-430. doi: 10.1007/978-3-030-28464-0_36.

[3] A. A. Khan, A. O. Bello, M. Arqam, et F. Ullah, « Integrating Building Information Modelling and Artificial Intelligence in Construction Projects: A Review of Challenges and Mitigation Strategies », *Technologies*, vol. 12, nº 10, Art. nº 10, oct. 2024, doi: 10.3390/technologies12100185.

[4] A. H. Gourabpasi et M. Nik-Bakht, « BIM-based automated fault detection and diagnostics of HVAC systems in commercial buildings », *J. Build. Eng.*, vol. 87, p. 109022, juin 2024, doi: 10.1016/j.jobe.2024.109022.

[5] A. Clausen *et al.*, « A digital twin framework for improving energy efficiency and occupant comfort in public and commercial buildings », *Energy Inform*, vol. 4, nº 2, p. 40, sept. 2021, doi: 10.1186/s42162-021-00153-9.

[6] R. Bortolini, R. Rodrigues, H. Alavi, L. Felix Dalla Vecchia, et N. Forcada, « Digital Twins' Applications for Building Energy Efficiency: A Review », *Energies*, vol. 15, nº 19, p. 7002, oct. 2022, doi: 10.3390/en15197002.

[7] G. Calcagno, A. Trombadore, G. Pierucci, et L. Montoni, « Untapping the Potential of the Digital Towards the Green Imperative: The Interdisciplinary BeXLab Experience », in *Technological Imagination in the Green and Digital Transition*, E. Arbizzani, E. Cangelli, C. Clemente, F. Cumo, F. Giofrè, A. M. Giovenale, M. Palme, et S. Paris, Éd., Cham: Springer International Publishing, 2023, p. 203-216. doi: 10.1007/978-3-031-29515-7_19.

[8] M. Deng, C. C. Menassa, et V. R. Kamat, « From Bim to Digital Twins: A Systematic Review of the Evolution of Intelligent Building Representations in the Aec-Fm Industry », *J. Inf. Technol. Constr.*, vol. 26, p. 58-83, 2021, doi: 10.36680/j.itcon.2021.005.

[9] C. Zhang, J. Beetz, et de Vries, « BimSPARQL: Domain-specific functional SPARQL extensions for querying RDF building data », *Semantic Web*, p. 1-27, août 2018, doi: 10.3233/SW-180297.

[10] S. Chávez-Feria, G. Giannakis, R. García-Castro, et M. Poveda-Villalón, « From obXML to the OP ontology: developing a semantic model for occupancy profile », présenté à LDAC, 2020.

[11] R. Brachman et H. Levesque, *Knowledge Representation and Reasoning*. in The Morgan Kaufmann Series in Artificial Intelligence. Morgan Kaufmann, 2004. doi: 10.1016/B978-1-55860-932-7.X5083-3.

[12] D. Fensel *et al.*, « Introduction: What Is a Knowledge Graph? », in *Knowledge Graphs: Methodology, Tools and Selected Use Cases*, D. Fensel, U. Şimşek, K. Angele, E. Huaman, E. Kärle, O. Panasiuk, I. Toma, J. Umbrich, et A. Wahler, Éd., Cham: Springer International Publishing, 2020, p. 1-10. doi: 10.1007/978-3-030-37439-6_1.

[13] T. Heath et C. Bizer, *Linked Data: Evolving the Web into a Global Data Space*. in Synthesis Lectures on Data, Semantics, and Knowledge. Cham: Springer International Publishing, 2011. doi: 10.1007/978-3-031-79432-2.

[14] J. H. Gallier, *Logic for Computer Science: Foundations of Automatic Theorem Proving, Second Edition*. Courier Dover Publications, 2015.

[15] M. Flasiński, *Introduction to Artificial Intelligence*. Springer, 2016.

[16] P. Hitzler, A. Eberhart, M. Ebrahimi, M. K. Sarker, et L. Zhou, « Neuro-symbolic approaches in artificial intelligence », *National Science Review*, vol. 9, nº 6, p. nwac035, juin 2022, doi: 10.1093/nsr/nwac035.

[17] H. Spieker, « Constraint-Guided Reinforcement Learning: Augmenting the Agent-Environment-

Interaction », in *2021 International Joint Conference on Neural Networks (IJCNN)*, juill. 2021, p. 1-8. doi: 10.1109/IJCNN52387.2021.9533996.

[18] ISO 23247-1, *Systèmes d'automatisation industrielle et intégration — Cadre technique de jumeau numérique dans un contexte de fabrication — Partie 1 : Vue d'ensemble et principes généraux*, 2021.

[19] A. Fuller, Z. Fan, C. Day, et C. Barlow, « Digital Twin: Enabling Technologies, Challenges and Open Research », *IEEE Access*, vol. 8, p. 108952-108971, 2020, doi: 10.1109/ACCESS.2020.2998358.

[20] E. A. Lee, « Cyber Physical Systems: Design Challenges », in *2008 11th IEEE International Symposium on Object and Component-Oriented Real-Time Distributed Computing (ISORC)*, mai 2008, p. 363-369. doi: 10.1109/ISORC.2008.25.

[21] R. Adamczak et K. Grabczewski, « A Hybrid Method for Extraction of Logical Rules From Data », juill. 2000.

[22] B. Wei et Z. Zhu, « Neural Symbolic Logical Rule Learner for Interpretable Learning », 21 août 2024, *arXiv*: arXiv:2408.11918. doi: 10.48550/arXiv.2408.11918.

[23] Z. Li, J. Guo, Y. Jiang, et X. Si, « Learning Reliable Logical Rules with SATNet », présenté à Thirty-seventh Conference on Neural Information Processing Systems, nov. 2023.

[24] M. Xu *et al.*, « Reasoning based on symbolic and parametric knowledge bases: a survey », 2 janvier 2025, *arXiv*: arXiv:2501.01030. doi: 10.48550/arXiv.2501.01030.

[25] J. Hong et T. P. Pavlic, « An Insect-Inspired Randomly, Weighted Neural Network with Random Fourier Features For Neuro-Symbolic Relational Learning », 11 septembre 2021, *arXiv*: arXiv:2109.06663. doi: 10.48550/arXiv.2109.06663.

[26] F. Lygerakis, N. Kampelis, et D. Kolokotsa, « Knowledge Graphs' Ontologies and Applications for Energy Efficiency in Buildings: A Review », *Energies*, vol. 15, nº 20, Art. nº 20, janv. 2022, doi: 10.3390/en15207520.

[27] W. Hua et Y. Zhang, « System 1 + System 2 = Better World: Neural-Symbolic Chain of Logic Reasoning », in *Findings of the Association for Computational Linguistics: EMNLP 2022*, Y. Goldberg, Z. Kozareva, et Y. Zhang, Éd., Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, déc. 2022, p. 601-612. doi: 10.18653/v1/2022.findings-emnlp.42.

[28] L. Liu, Z. Wang, et H. Tong, « Neural-Symbolic Reasoning over Knowledge Graphs: A Survey from a Query Perspective », 30 novembre 2024, *arXiv*: arXiv:2412.10390. doi: 10.48550/arXiv.2412.10390.

[29] Y. Wei, H. Li, G. Xin, Y. Wang, et B. Wang, « An original model for multi-target learning of logical rules for knowledge graph reasoning », 15 juillet 2022, *arXiv*: arXiv:2112.06189. doi: 10.48550/arXiv.2112.06189.

[30] Y. Xie, Z. Xu, M. S. Kankanhalli, K. S. Meel, et H. Soh, « Embedding Symbolic Knowledge into Deep Networks », in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2019.

[31] P. Graf et P. Emami, « Three Pathways to Neurosymbolic Reinforcement Learning with Interpretable Model and Policy Networks », 7 février 2024, *arXiv*: arXiv:2402.05307. doi: 10.48550/arXiv.2402.05307.

[32] S. Milani *et al.*, « MAVIPER: Learning Decision Tree Policies for Interpretable Multi-Agent Reinforcement Learning », 11 juillet 2022, *arXiv*: arXiv:2205.12449. doi: 10.48550/arXiv.2205.12449.

[33] E. M. Kenny, M. Tucker, et J. Shah, « Towards Interpretable Deep Reinforcement Learning with Human-Friendly Prototypes », présenté à The Eleventh International Conference on Learning Representations, sept. 2022.

[34] B. Zhang, B. Yang, C. Wang, Z. Wang, B. Liu, et T. Fang, « Computer Vision-Based Construction Process Sensing for Cyber–Physical Systems: A Review », *Sensors*, vol. 21, nº 16, Art. nº 16, janv. 2021, doi: 10.3390/s21165468.

[35] S. B. Hakim, M. Adil, A. Velasquez, et H. H. Song, « ANSR-DT: An Adaptive Neuro-Symbolic Learning and Reasoning Framework for Digital Twins », 15 janvier 2025, *arXiv*: arXiv:2501.08561. doi: 10.48550/arXiv.2501.08561.

[36] A. Siyaev, D. Valiev, et G.-S. Jo, « Interaction with Industrial Digital Twin Using Neuro-Symbolic Reasoning », *Sensors*, vol. 23, nº 3, Art. nº 3, janv. 2023, doi: 10.3390/s23031729.

[37] S. Yoon, Y. Choi, et J. Koo, « In situ virtual sensors in building digital twins: framework and methodology », *J. Ind. Inf. Integr.*, vol. 36, p. 100532, déc. 2023, doi: 10.1016/j.jii.2023.100532.

[38] S. Agostinelli, « COGNIBUILD: Cognitive Digital Twin Framework for Advanced Building Management and Predictive Maintenance », in *Technological Imagination in the Green and Digital Transition*, E. Arbizzani, E. Cangelli, C. Clemente, F. Cumo, F. Giofrè, A. M. Giovenale, M. Palme, et S. Paris, Éd., Cham: Springer International Publishing, 2023, p. 69-78. doi: 10.1007/978-3-031-29515-7_8.

[39] C. Jones, « Semantic Technologies in Knowledge Management », *European Journal of Information and Knowledge Management*, vol. 3, nº 1, Art. nº 1, mars 2024, doi: 10.47941/ejikm.1750.

[40] S. Reynaud, A. Dumas, et A. Roxin, « Knowledge representation for neuro-symbolic digital building twin querying », in *The Second Workshop on AI for Digital Twins and Cyber-Physical Applications*, in CEUR Workshop Proceedings, vol. 3807. Jeju - South Korea, August: CEUR, août 2024, p. 48-75.

[41] A. Roxin, W. Abdou, et W. Derigent, « Interoperable Digital Building Twins Through Communicating Materials and Semantic BIM », *SN COMPUT. SCI.*, vol. 3, nº 1, p. 23, oct. 2021, doi: 10.1007/s42979-021-00860-w.

Vers une prédiction optimisée : réduction de dimension, mécanisme d'attention et sélection dynamique au service d'un jumeau numérique

Bruno PEREZ[1], Imad MOURTAJI[1], Imen ABIDI[1]

[1] Akkodis Research , AKR. Tour Coeur Défense, 92400 Courbevoie France.

bruno.perez@akkodis.com

Résumé

Lors de nos premiers travaux, nous avons couplé un système multi-agents (SMA) à un apprentissage par renforcement (AR) pour optimiser la prédiction des évènements indésirables associés aux soins (EIAS). Ce modèle s'appuie sur un jumeau numérique du bloc opératoire, synchronisé avec le jumeau physique, afin d'alerter en temps réel lors d'un EIAS. Dans ce papier, nous présentons l'évolution de notre approche, qui ne se limite plus à la génération d'alertes, mais anticipe désormais l'évolution globale de l'environnement. Pour ce faire, nous avons intégré des séries temporelles et un réseau de neurones LSTM pour mieux modéliser les dépendances temporelles. Cette solution permet d'alerter et de prédire le risque d'EIAS à court, moyen ou long terme, renforçant ainsi la sécurité des soins. Dans une perspective d'analyse des interactions, nous avons pu observer comment l'accumulation de signaux faibles et l'escalade de certains comportements collectifs entre agents préfigurent la survenue d'un EIAS. Ces observations ouvrent la voie à de nouvelles stratégies préventives, en adaptant les réponses du système aux dynamiques interactionnelles identifiées.

Mots-clés

SMA, LSTM, ST, EIAS, Propagation d'incertitude

Abstract

In our initial work, we coupled a multi-agent system (MAS) with reinforcement learning (RL) to optimize the prediction of adverse events associated with care (AEAC). This model relies on a digital twin of the operating room, synchronized with the physical twin, to alert in real time when an AEAC occurs. In this paper, we present the evolution of our approach, which is no longer limited to generating alerts, but now anticipates the global evolution of the environment. To achieve this, we have integrated time series and an LSTM neural network to better model temporal dependencies. This solution makes it possible to alert and predict the risk of AEAC in the short, medium or long term, thereby reinforcing the safety of care. From an interaction analysis perspective, we were able to observe how the accumulation of weak signals and the escalation of certain collective behaviors between agents foreshadow the occurrence of an AEAC. These observations pave the way for new preventive strategies, by adapting system responses to the interactional dynamics identified.

Keywords

MAS, LSTM, TS, AEAC, Uncertainty Propagation

1 Introduction

Les opérations chirurgicales évoluent dans un environnement médical en constante innovation, offrant des procédures diversifiées mais augmentant la complexité des risques. La sécurité des patients dépend d'un cadre technique sophistiqué et de la collaboration entre disciplines. Les protocoles standardisés, tels que ceux de l'OMS, améliorent la gestion des risques, mais les évènements indésirables graves restent préoccupants, avec 2385 cas signalés en 2022. Face à ce constat, nous avons dans des premiers travaux de recherche développé un jumeau numérique du bloc opératoire nommé PRIA (Prédiction des Risques à l'aide de l'Intelligence Artificielle) intégrant un système multi-agents (SMA) et un modèle d'apprentissage continu par renforcement (AR) [10]. L'objectif était de détecter des évènements indésirables associés aux soins (EIAS) afin de générer des alertes en limitant au maximum les biais (faux négatifs, faux positifs) grâce au couplage SMA/AR. Dans ce papier, nous exposons l'intégration de la prédictivité à notre architecture. Il s'agit pour nous d'agréger à notre architecture un paradigme capable de déterminer l'évolution de notre système dans notre contexte non déterministe. Nous avons orienté nos choix sur des séries temporelles enrichies par des algorithmes capables d'évoluer dans des contextes non forcément continus. Nous présentons dans la suite de cet exposé un état de l'art condensé mais représentatif des grandes tendances. Nous décrivons ensuite nos objectifs de recherche en terme de modélisation prédictive dans un contexte connecté de simulation. La présentation de nos premiers résultats suivie d'une discussion viennent clôturer ce papier.

2 Travaux antérieurs

La modélisation des risques en milieu de santé est largement répandue, car elle constitue une réponse essentielle aux enjeux de santé publique. Face à la complexité des environnements médicaux, notamment en bloc opératoire, plusieurs approches et paradigmes sont mis en œuvre afin d'anticiper la survenue d'évènements indésirables et d'améliorer la sécurité des patients. Ces modèles reposent sur des méthodologies variées, allant des analyses probabilistes aux systèmes d'intelligence artificielle, chacune présentant ses avantages et ses limites.

Nos précédents travaux portaient sur un générateur d'alertes, matérialisé par un SMA couplé à un AR, permettant ainsi de limiter les fausses alertes [10]. Notre état de l'art décrivait les systèmes multi-agents, l'apprentissage possible d'un SMA et, enfin, les jumeaux numériques dans le domaine de la santé. Dans cette section, nous abordons les travaux connexes à nos recherches actuelles, qui portent sur la prédictivité de la survenue d'un évènement indésirable.

2.1 Prédictivité de l'évolution d'un système complexe non déterministe

Au-delà de la génération optimisée d'une alerte dans un bloc opératoire, nous souhaitons prédire l'évolution de ces systèmes complexes. Les séries temporelles sont largement utilisées pour modéliser des systèmes dynamiques où les observations sont séquentielles et dépendent du temps [5, 7].

Les modèles ARIMA (AutoRegressive Integrated Moving Average) sont classiques pour les prévisions linéaires, mais ils montrent leurs limites face à des systèmes non linéaires et complexes [3]. Pour pallier ces insuffisances, les réseaux de neurones récurrents (RNN), et plus particulièrement les long short-term memory (LSTM), se sont imposés comme des alternatives efficaces. Ces derniers sont capables de capturer des dépendances temporelles longues et complexes, ce qui les rend particulièrement adaptés aux données séquentielles [6].

Cependant, les LSTM ne sont pas les seuls modèles à exploiter les données temporelles. Les réseaux de neurones convolutifs (CNN) peuvent également être utilisés pour extraire des caractéristiques temporelles lorsqu'ils sont appliqués à des séries temporelles. Bien qu'ils soient souvent utilisés en complément des LSTM pour améliorer la précision des prédictions, ils montrent des capacités intéressantes pour traiter des données structurées de manière spatiale [12, 1, 9]. Ici, une structure spatiale désigne l'organisation locale des données, où des points proches dans le temps sont considérés comme voisins, à la manière des pixels adjacents dans une image.

En outre, l'intégration de modèles statistiques traditionnels avec des techniques d'apprentissage profond a montré des résultats prometteurs. Par exemple, combiner ARIMA avec LSTM permet de tirer parti des forces des deux approches pour améliorer la précision des prévisions [11]. Cette approche hybride permet de mieux capturer à la fois les tendances linéaires et les dynamiques non linéaires des systèmes complexes.

Malgré leurs avantages, ces modèles présentent plusieurs limites. La complexité computationnelle des modèles basés sur l'apprentissage profond, comme les LSTM, nécessite des ressources importantes et des temps d'entraînement longs. De plus, l'interprétabilité de ces modèles de type boîte noire reste un défi, ce qui peut être un obstacle dans des domaines sensibles comme la santé. Enfin, la performance des modèles dépend fortement de la qualité et de la quantité des données disponibles. Les données manquantes ou bruitées peuvent affecter négativement les prédictions, rendant ainsi crucial le prétraitement des données.

2.2 Plateformes et solutions de modélisation des risques au bloc opératoire

Plusieurs plateformes ont été développées pour améliorer la sécurité et gérer les risques au sein des blocs opératoires. Parmi celles-ci :

- SIM-PRO-BLOC [8] : Cette plateforme de simulation professionnelle est le fruit d'une collaboration entre différents professionnels de santé, visant à améliorer la qualité des soins et la prévention des risques au bloc opératoire.
- 3D Virtual Operating Room [4] : Il s'agit d'un serious game multilingue en 3D conçu pour former et entraîner les professionnels à la gestion des risques et à la prévention des évènements indésirables graves.
- Caresyntax [2] : Cette plateforme connecte la salle d'opération au système d'information hospitalier, permettant une gestion évolutive des vidéos et des données pour améliorer les performances et la sécurité au bloc opératoire.

Bien que ces solutions apportent des avancées significatives, elles présentent certaines limites. Par exemple, les plateformes de simulation, bien qu'efficaces pour la formation, peuvent ne pas refléter toutes les complexités des situations réelles. De plus, l'intégration de systèmes comme Caresyntax dans les infrastructures hospitalières existantes peut s'avérer coûteuse et techniquement complexe. Enfin, la dépendance à des bases de données et à des systèmes informatiques soulève des questions concernant la confidentialité des données et la cybersécurité.

2.3 Nos contributions en réponse aux verrous et limites identifiés

Pour surmonter les limites et verrous précédemment identifiés, notre approche se distingue par l'utilisation de séries temporelles modélisées à l'aide d'un LSTM, et enrichies par des techniques innovantes de traitement des données. Nous proposons d'explorer l'utilisation de méthodes de réduction de dimensionnalité, telles que l'analyse en composantes principales (ACP), pour prétraiter les données et ré-

duire leur complexité avant de les introduire dans le modèle LSTM.

De plus, nous proposons de mettre en œuvre un mécanisme de sélection dynamique, où le système compare en temps réel les erreurs absolues moyennes (EAM) de chaque modèle pour chaque variable et choisit automatiquement celui qui minimise l'erreur. En complément, nous envisageons d'incorporer des mécanismes d'attention afin de permettre au modèle de se concentrer sur les parties les plus pertinentes des séries temporelles, améliorant ainsi la précision des prévisions.

En parallèle, nous mettrons en œuvre une stratégie de propagation d'incertitude qui permettra de quantifier et de diffuser l'incertitude inhérente aux données et aux prédictions à travers l'ensemble du modèle. Cette approche innovante vise à :

— Mesurer la fiabilité des prédictions en intégrant explicitement l'incertitude à chaque étape du traitement.
— Identifier les sources d'erreur potentielles, facilitant ainsi l'analyse de la robustesse du modèle dans des environnements complexes.
— Améliorer l'interprétabilité des résultats en fournissant des indicateurs quantitatifs de confiance, essentiels pour la prise de décision dans des contextes opérationnels critiques.

Cette double contribution, combinant une modélisation avancée avec LSTM et mécanismes d'attention, à une approche de propagation d'incertitude, offre une solution complète pour optimiser la prédictivité et la transparence des prévisions dans des environnements synchronisés.

3 Notre architecture

Dans nos premiers travaux, le couplage SMA/RL, combinant les Systèmes Multi-Agents et l'apprentissage par renforcement (AR ou plus communément RL) est une contribution novatrice à la création de jumeaux numériques, une capacité avancée pour simuler de manière approfondie la complexité des environnements réels. L'intérêt substantiel de cette approche se situe à la fois sur le plan théorique, grâce à la formalisation des interactions entre agents autonomes, et sur le plan pratique, en permettant la simulation avancée de scénarios complexes dans divers secteurs. Les premiers résultats ont permis de vérifier l'efficacité de cette architecture dans la mise en œuvre de notre générateur d'alertes, en limitant les fausses alertes (faux positifs et faux négatifs).

Nos travaux de recherche actuels sont axés sur la prédictivité de l'évolution de notre système, en se concentrant sur l'estimation de la variation probable du SMA entre deux synchronisations du bloc opératoire avec le SMA. Dans cette optique, nous avons enrichi notre architecture en y intégrant un modèle ARIMA pour l'analyse des séries temporelles, un réseau LSTM pour l'exploitation de ces mêmes séries, ainsi qu'une méthode de propagation d'incertitude.

3.1 Modèle ARIMA

Le modèle ARIMA (AutoRegressive Integrated Moving Average) est une méthode classique de modélisation des séries temporelles qui combine trois composantes essentielles : la partie autorégressive (AR), la différenciation (I) pour rendre la série stationnaire et la partie de moyenne mobile (MA). Formellement, un modèle ARIMA(p, d, q) s'exprime par :

$$\phi(L)(1-L)^d x_t = \theta(L)\epsilon_t,$$

où : - x_t désigne la valeur de la série à l'instant t, - L est l'opérateur de retard, défini par $Lx_t = x_{t-1}$, - ϵ_t représente un bruit blanc.
Les polynômes $\phi(L)$ et $\theta(L)$ sont donnés par :

$$\phi(L) = 1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p,$$

$$\theta(L) = 1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q.$$

La composante intégrée, notée $(1-L)^d$, permet de différencier la série d fois afin d'éliminer toute tendance non stationnaire. En combinant ces trois éléments, le modèle ARIMA est capable de capturer à la fois les dépendances temporelles et les effets des chocs aléatoires, ce qui en fait un outil puissant pour la prévision des séries temporelles.

3.2 LSTM, mécanismes d'attention et réduction de dimension

Afin de tirer profit des capacités de mémorisation du LSTM pour extraire des caractéristiques pertinentes tout en réduisant la dimensionnalité du problème, nous proposons d'appliquer ce modèle aux données du SMA. Soit une variable d'état $x(t) \in \mathbb{R}^n$ représentant les mesures du SMA à différents instants t. Notre objectif est d'estimer la variation ΔSMA entre deux instants de synchronisation, que nous formulons comme une fonction f :

$$\Delta SMA = f\big(x(t),\, x(t-1),\, \ldots,\, x(t-T)\big)$$

où T représente la taille de la fenêtre temporelle considérée. Pour améliorer la capacité prédictive et réduire la complexité des données, nous appliquons une réduction de dimensionnalité via l'analyse en composantes principales. Soit $X \in \mathbb{R}^{T\times n}$ la matrice des observations sur la fenêtre T, nous cherchons une transformation linéaire définie par :

$$Y = XW$$

avec $W \in R^{n\times m}$ (où $m < n$) constitué des vecteurs propres associés aux m plus grandes valeurs propres de la covariance de X. Cette étape permet de prétraiter les données et de réduire leur dimensionnalité tout en préservant l'essentiel de la variance.
Les données ainsi prétraitées sont ensuite introduites dans un modèle LSTM, dont le fonctionnement interne se résume par les équations suivantes :

$$
\begin{aligned}
i_t &= \sigma\Big(W_{xi}\, y_t + W_{hi}\, h_{t-1} + b_i\Big),\\
f_t &= \sigma\Big(W_{xf}\, y_t + W_{hf}\, h_{t-1} + b_f\Big),\\
o_t &= \sigma\Big(W_{xo}\, y_t + W_{ho}\, h_{t-1} + b_o\Big),\\
\tilde{c}_t &= \tanh\Big(W_{xc}\, y_t + W_{hc}\, h_{t-1} + b_c\Big),\\
c_t &= f_t \odot c_{t-1} + i_t \odot \tilde{c}_t,\\
h_t &= o_t \odot \tanh(c_t),
\end{aligned}
$$

où i_t, f_t et o_t représentent respectivement les portes d'entrée, d'oubli et de sortie, c_t est l'état de la cellule, h_t la sortie et $\odot$ désigne la multiplication élément par élément.
Afin de renforcer la précision des prévisions, nous envisageons également l'intégration d'un mécanisme d'attention. Ce mécanisme permet au modèle de pondérer différemment les contributions de chaque instant de la séquence en définissant des poids $\alpha_{t,i}$ calculés par :

$$\alpha_{t,i} = \frac{\exp(e_{t,i})}{\sum_{j=1}^{T} \exp(e_{t,j})},\ avec\, e_{t,i} = \phi\big(h_t,\, y_i\big),$$

où ϕ est une fonction de score (par exemple, une fonction linéaire ou basée sur le produit scalaire). Le vecteur de contexte c_t est ensuite obtenu par :

$$c_t = \sum_{i=1}^{T} \alpha_{t,i}\, y_i,$$

et sert à affiner la prédiction finale, qui s'exprime par :

$$\Delta SMA = g\big(h_t,\, c_t\big),$$

où g est une fonction de combinaison, par exemple une couche entièrement connectée.
Ainsi, notre contribution réside dans la combinaison de techniques de réduction de dimensionnalité (ACP), de modélisation temporelle par LSTM et d'un mécanisme d'attention, visant à optimiser l'estimation de la variation du SMA dans des environnements synchronisés. Cette approche intégrée permet de réduire la complexité des données tout en capturant efficacement les dynamiques temporelles et en focalisant l'apprentissage sur les informations les plus pertinentes pour la prédiction. Nous abordons à présent la stratégie appliquées à la propagation d'incertitudes.

3.3 Propagation d'incertitudes

Dans notre approche, la propagation d'incertitude repose sur une méthode d'approximation linéaire fondée sur le développement de Taylor. Soit une fonction

$$y = f(x_1, x_2, \ldots, x_n)$$

où chaque variable x_i est mesurée avec une incertitude associée σ_{x_i}. En supposant l'indépendance des variables, l'incertitude totale sur y est alors évaluée par la formule classique de propagation d'erreur :

$$\sigma_y^2 = \sum_{i=1}^{n} \left(\frac{\partial f}{\partial x_i}\right)^2 \sigma_{x_i}^2.$$

Cette approche consiste à linéariser f autour des valeurs moyennes des variables d'entrée, de sorte que le coefficient de sensibilité $\frac{\partial f}{\partial x_i}$ quantifie l'impact de l'incertitude sur x_i sur la sortie y. Pour des systèmes non linéaires ou comportant des interactions complexes, cette méthode peut être complétée par des techniques de simulation Monte Carlo, offrant ainsi une estimation plus robuste de la distribution des incertitudes. Ce cadre formel permet de mesurer précisément la fiabilité des prédictions et d'identifier les sources dominantes d'erreur, contribuant ainsi à une meilleure compréhension et interprétation des résultats obtenus.

La Figure 1 décrit notre architecture dans sa globalité. Le jumeau physique se synchronise avec le jumeau numérique sachant que les échanges entre le SMA et le jumeau numérique se fait via une base pivot qui communique avec la couche interaction (côté SMA). L'apprentissage par renforcement est quant à lui connecté à la couche d'apprentissage du SMA. Grâce à cet apprentissage, les agents du système peuvent apprendre en continu à partir de l'évolution du système lui-même, ce qui permet de prédire son évolution future et d'optimiser les décisions en temps réel. Cette approche est particulièrement utile dans notre contexte, où les systèmes en question sont en constante évolution et nécessitent une adaptation continue pour fonctionner de manière optimale. La brique prédictivité est le module qui correspond à nos travaux actuels que nous présentons dans ce papier. Il s'agit de prédire l'évolution de l'état du système entre deux synchronisations. Le choix des séries temporelles couplés à un LSTM auquel s'agrège la propagation d'incertitude vient enrichir notre outil destiné à l'anticipation de la survenue d'un EIAS. Comme illustré dans la figure 2, cette approche combine plusieurs étapes clés : réduction de dimensionnalité pour simplifier les données d'entrée, modélisation dynamique à l'aide de LSTM et mécanismes d'attention pour se concentrer sur les parties les plus pertinentes des séries temporelles. En parallèle, la propagation d'incertitude permet de quantifier et diffuser l'incertitude à travers le modèle, améliorant ainsi la fiabilité et l'interprétabilité des prédictions.
Nos premières expérimentations et résultats sont présentés dans la section qui suit.

4 Expérimentations, résultats et discussion

Le protocole expérimental consiste à comparer trois approches de prévision sur des séries temporelles réelles issues d'un fichier CSV contenant les variables *FC* (fréquence cardiaque), *Temp* (température), *Cholesterol* et *Glycemie*. Le millier de données généré par le jumeau physique virtuel (voir nos premiers travaux [10]) est d'abord chargé, puis ensuite réparti en deux ensembles : 80 % pour l'entraînement et 20 % pour le test. Sur la base de l'ensemble d'entraînement, trois modèles sont ajustés :

— Un modèle ARIMA classique, adapté aux tendances linéaires,

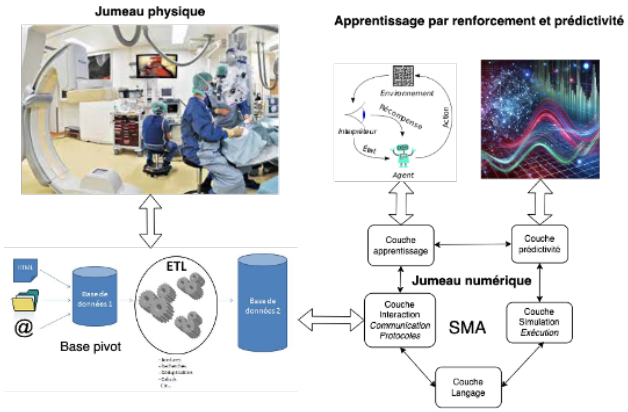


FIGURE 1 – Architecture globale

— Un modèle LSTM, capable de capturer des non-linéarités,
— Un modèle hybride qui corrige la prévision ARIMA à l'aide d'un LSTM appliqué sur les résidus.

Pour chacun de ces modèles, la performance est évaluée sur l'ensemble test à l'aide de l'erreur absolue moyenne. Par ailleurs, une propagation d'incertitude est réalisée pour le modèle hybride en utilisant l'écart-type des résidus. Enfin, un mécanisme de sélection dynamique compare les EAM et choisit le modèle ayant la meilleure performance pour produire la prévision finale.

4.1 Premiers résultats

Les premiers résultats montrent que pour la plupart des variables, ARIMA et le modèle hybride obtiennent des erreurs quasi nulles, tandis que le LSTM présente des erreurs plus élevées. Toutefois, pour la variable *FC*, le LSTM affiche une EAM légèrement inférieure à celle d'ARIMA et du modèle hybride, bien que l'incertitude associée soit élevée.
Nous pouvons résumer ces premiers résultats dans le tableau 1 :

Variable	FC	Temp	Cholesterol	Glycemie
ARIMA_EAM	17.48	0.03	0.03	0.03
LSTM_EAM	15.97	14.06	7.94	4.71
Hybrid_EAM	19.51	0.03	0.03	0.03
Incertitude_EAM	16.23	0.00	0.00	0.00

Tableau 1 – Comparaison des EAM des modèles

Pour *Temp*, *Cholesterol* et *Glycemie*, ARIMA et le modèle hybride obtiennent des performances quasi parfaites (EAM $\approx$ 0.03), suggérant que ces séries suivent des tendances linéaires bien capturées par des modèles statistiques classiques.
En revanche, pour *FC*, le LSTM présente une performance légèrement meilleure (EAM $\approx$ 15.97) que ARIMA et le modèle hybride, ce qui indique une possible présence de non-linéarités dans cette variable.
La propagation d'incertitude (nous le rappelons) consiste à combiner les incertitudes de chaque variable pour obtenir

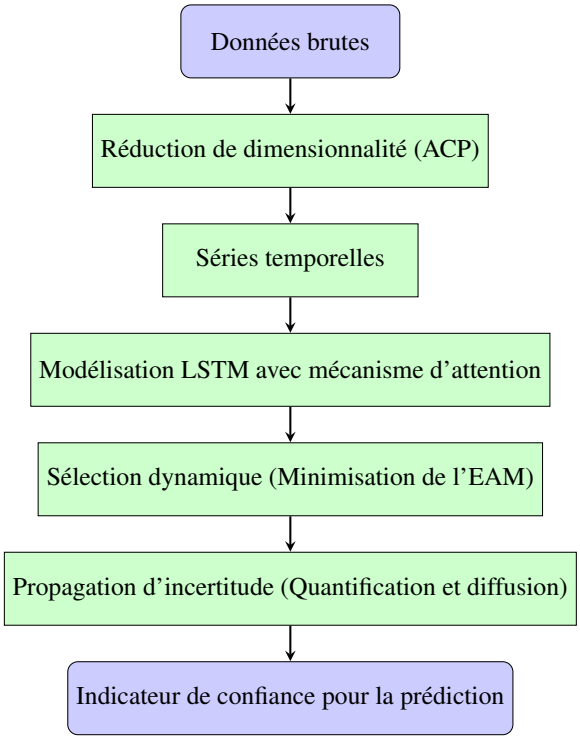


FIGURE 2 – Architecture globale du processus de prédiction

une estimation globale de la fiabilité des prédictions. Cependant, dans notre cas, l'incertitude, exprimée via l'écart-type et les intervalles, est significative uniquement pour la variable *FC*, tandis qu'elle est quasi nulle pour les autres variables. Cette disparité crée un déséquilibre dans la quantification de l'incertitude globale : la contribution de *FC* domine entièrement le calcul, ce qui fausse la représentation de la fiabilité du modèle dans son ensemble.
En d'autres termes, la propagation d'incertitude n'est pas exploitable dans notre contexte car l'incertitude élevée sur *FC* ne permet pas d'obtenir une mesure homogène et robuste de l'incertitude globale. Le faible niveau d'incertitude pour les autres variables ne compense pas ce déséquilibre, rendant l'indicateur final peu pertinent pour guider les décisions ou affiner le modèle. Ainsi, la méthode de propagation d'incertitude, dans ces conditions, n'apporte pas d'information utile par rapport aux résultats obtenus.
Afin d'optimiser le choix du modèle en temps réel, le code a été modifié pour intégrer un mécanisme d'attention et de sélection dynamique.

4.2 Résultats avec réduction de dimension, sélection dynamique et mécanismes d'attention

Dans cette nouvelle approche, nous appliquons la réduction de dimension pour le modèle LSTM. Nous comparons ensuite les EAM obtenues par les trois approches pour chaque variable et choisissons automatiquement le modèle qui minimise l'erreur. De plus, nous intégrons un mécanisme d'at-

tention afin de permettre au modèle de se concentrer sur les parties les plus pertinentes des séries temporelles, améliorant ainsi la précision des prévisions. Cette amélioration permet une adaptation automatique du système aux caractéristiques spécifiques de chaque série, en exploitant la capacité d'ARIMA à modéliser des tendances linéaires et la force du LSTM dans le traitement de non-linéarités, tout en optimisant l'apprentissage grâce à l'attention.
Le tableau 2 résume les performances obtenues :

Variable	FC	Temp	Cholesterol	Glycemie
ARIMA_EAM	17.48	0.0318	0.0318	0.0318
LSTM_EAM	16.32	6.6095	5.5443	4.3904
Hybrid_EAM	17.82	0.0316	0.0316	0.0316
Incertitude_EAM	16.23	0.00	0.00	0.00
Best Model	LSTM	Hybrid	Hybrid	Hybrid

Tableau 2 – Comparaison des performances des modèles

— Pour *FC*, le mécanisme de sélection dynamique opte pour le modèle LSTM, ce qui est cohérent avec une performance légèrement meilleure pour cette variable.
— Pour *Temp*, *Cholesterol* et *Glycemie*, le modèle hybride est sélectionné, offrant des erreurs comparables à celles d'ARIMA (EAM ≈ 0.0316–0.0318), confirmant que ces séries sont bien prédites par des modèles linéaires ou hybrides.
— Les intervalles d'incertitude sont identiques à nos premiers résultats et confirment la non-pertinence quant à l'utilisation de la propagation d'incertitude.

Le premier graphique (figure 3) présente les EAM obtenues pour chaque modèle sur les quatre variables (*FC*, *Temp*, *Cholesterol* et *Glycemie*) avant l'intégration des mécanismes d'attention, de sélection dynamique et de réduction de dimension. On constate que, pour *Temp*, *Cholesterol* et *Glycemie*, les modèles ARIMA et Hybrid obtiennent des EAM quasi nulles (0,03), alors que pour *FC* le LSTM affiche une erreur légèrement inférieure (15,97) comparée à ARIMA (17,48) et Hybrid (19,51).

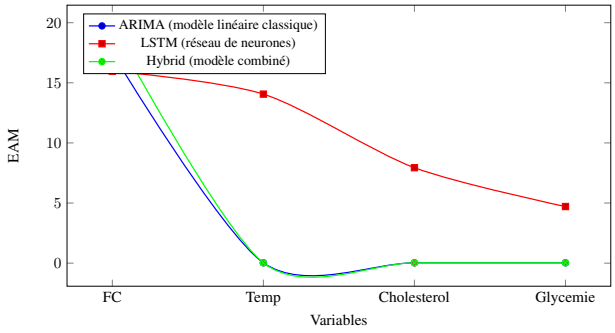


FIGURE 3 – Résultats sans intégration des principes d'attention de sélection dynamique et de réduction de dimension

Le second graphique (figure 4) regroupe les résultats obtenus après intégration des mécanismes d'attention, de sélection dynamique et de réduction de dimension. Pour chaque variable, le modèle final est choisi en fonction de l'EAM la plus faible. Ainsi, pour *FC* le LSTM est retenu (EAM ≈ 16,32) tandis que pour *Temp*, *Cholesterol* et *Glycemie*, le modèle Hybrid est sélectionné (avec des EAM très faibles, autour de 0,0316 à 0,0318). Cette approche dynamique permet d'adapter la méthode à la nature spécifique de chaque série en tirant parti des mécanismes d'attention, qui mettent en évidence les informations les plus pertinentes et renforcent la capacité du modèle à capturer des structures complexes.

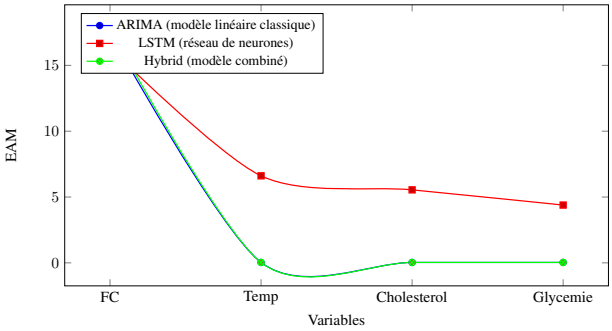


FIGURE 4 – Résultats avec intégration des principes d'attention de sélection dynamique et de réduction de dimension

Les premiers résultats montrent que pour trois variables (*Temp*, *Cholesterol* et *Glycemie*), les modèles ARIMA et Hybrid obtiennent des EAM extrêmement faibles, indiquant une dynamique linéaire bien capturée par ces approches. En revanche, pour la variable *FC*, le LSTM présente une légère supériorité avec une EAM inférieure (15,97 vs 17,48 pour ARIMA et 19,51 pour Hybrid), suggérant des non-linéarités dans cette série.
Après l'implémentation du mécanisme de sélection dynamique, le système choisit automatiquement le LSTM pour *FC* et le modèle Hybrid pour les autres variables, garantissant ainsi la meilleure performance possible en fonction de la nature des données. De plus, l'intégration d'un mécanisme d'attention a permis d'améliorer la prise en compte des dépendances temporelles, en mettant en avant les séquences les plus pertinentes pour la prédiction. Parallèlement, une réduction de dimension a été appliquée afin d'éliminer les redondances dans les variables explicatives, réduisant ainsi la complexité des modèles et améliorant leur généralisation. Ces ajustements permettent d'optimiser la prévision en temps réel en adaptant à la fois le choix du modèle et la pertinence des informations utilisées.
En conclusion, l'intégration d'un mécanisme de sélection dynamique des modèles, couplé à l'attention et à la réduction de dimension, a permis d'adapter en temps réel la méthode de prévision à la nature des données tout en améliorant la robustesse des prédictions. Pour des variables présentant des dynamiques linéaires (Temp, Cholesterol, Glycemie), ARIMA et le modèle hybride se montrent très performants, tandis que pour la variable *FC*, qui semble contenir des non-linéarités, le LSTM offre de meilleurs résul-

tats grâce à sa capacité à capturer des relations complexes. L'attention renforce encore cette différenciation en pondérant intelligemment les séquences influentes, tandis que la réduction de dimension améliore l'efficacité computationnelle et la stabilité des modèles. Cette approche modulaire et adaptative constitue une avancée prometteuse pour la prévision en temps réel, tout en ouvrant la voie à des améliorations ultérieures, notamment en matière d'estimation de l'incertitude et d'optimisation des architectures neuronales.

5 Conclusion

Dans le cadre de nos travaux de recherche, inscrits dans la continuité de PRIA (jumeau numérique du bloc opératoire), nous avons intégré la brique de prédictivité à notre architecture. L'objectif n'est plus seulement de générer des alertes en temps réel, mais également de prédire la survenue d'un EIAS entre deux synchronisations du jumeau numérique avec le bloc opératoire. Pour ce faire, nous avons exploité des données extraites du jumeau physique afin d'éprouver nos modèles de prévision ; à savoir ARIMA, LSTM et un modèle hybride combinant ARIMA et LSTM, auquel est associée une composante de propagation d'incertitude. Cette approche vise à adapter le choix du modèle à la dynamique spécifique de chaque variable.

Les premiers résultats expérimentaux, obtenus sans mécanismes d'attention, de sélection dynamique et de réduction de dimension, montrent que le choix du modèle optimal dépend fortement du comportement des variables. Pour trois variables (*Temp*, *Cholesterol*, *Glycemie*), dont la dynamique s'avère quasi linéaire, ARIMA et le modèle hybride affichent d'excellentes performances (EAM $\approx$ 0,03). En revanche, pour la variable *FC*, dont le comportement semble plus complexe et non linéaire, le LSTM présente une EAM légèrement inférieure (15,97) comparé à ARIMA (17,48) et au modèle hybride (19,51).

Ces résultats indiquent que, pour les variables présentant une dynamique linéaire, les approches classiques (ARIMA) ou leur hybridation avec un LSTM offrent des prédictions très précises, alors que pour *FC*, la capacité du LSTM à capturer des non-linéarités lui confère un léger avantage.

Suite à l'intégration d'un mécanisme de sélection dynamique, le système compare en temps réel les EAM de chaque modèle pour chaque variable et choisit automatiquement celui qui minimise l'erreur. En outre, l'ajout de mécanismes d'attention et de réduction de dimension permet au modèle de se concentrer sur les informations les plus pertinentes des séries temporelles, renforçant ainsi la précision des prévisions. On constate que, grâce à cette approche dynamique, le système opte pour le LSTM pour *FC* (confirmant la présence de non-linéarités) et pour le modèle hybride pour les autres variables linéaires. L'automatisation de la sélection, couplée à l'attention, permet ainsi d'optimiser les prévisions en fonction de la nature de chaque série.

Du point de vue de la visualisation, nous avons représenté ces résultats sous forme de courbes dans deux graphiques distincts. Le premier graphique illustre les premiers résultats obtenus sans sélection dynamique, tandis que le second met en évidence les performances après implémentation du mécanisme adaptatif. Ces courbes démontrent clairement l'intérêt de l'hybridation dynamique, qui permet de réduire l'erreur globale en sélectionnant le modèle le plus adapté à chaque variable.

En conclusion, l'intégration de la réduction de dimension, d'un mécanisme d'attention et la sélection dynamique constitue une avancée encourageante dans notre démarche de prédictivité. En adaptant le choix du modèle aux caractéristiques spécifiques de chaque variable, nous parvenons à optimiser les performances de prévision. Néanmoins, la composante de propagation d'incertitude nécessite encore des ajustements, car les résultats actuels ne sont pas suffisamment probants pour une exploitation opérationnelle. En termes de perspectives, nous envisageons un déploiement en conditions réelles, ainsi qu'un approfondissement de l'ajustement de la propagation d'incertitude, afin d'intégrer de manière fiable les marges d'erreur dans notre système prédictif. Cette démarche représente une étape cruciale pour améliorer la synchronisation entre le jumeau numérique et le bloc opératoire, garantissant ainsi une meilleure anticipation des évènements indésirables.

Références

[1] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv :1803.01271*, 2018.

[2] L Bourgeois, J Delode, S Kirche, A Massei, V Moreno, X Payet-Burin, et al. Bloc opératoire : état de l'art des technologies biomédicales. *IRBM News*, 40(4) :117–156, 2019.

[3] George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis : forecasting and control*. John Wiley & Sons, 2015.

[4] Gilles Devreux. *Le rôle des comportements informationnels dans la prise de conscience de la situation : usage dans le serious game 3D Virtual Operating Room*. PhD thesis, Université Toulouse le Mirail-Toulouse II, 2015.

[5] James D. Hamilton. *Time Series Analysis*. Princeton University Press, Princeton, NJ, 1994.

[6] S Hochreiter. Long short-term memory. *Neural Computation MIT-Press*, 1997.

[7] Bryan Lim and Stefan Zohren. Time-series forecasting with deep learning : a survey. *Philosophical Transactions of the Royal Society A*, 379(2194) :20200209, 2021.

[8] Didier Maocec. Sim-pro-bloc : Se former autrement. *Manipulateur d'imagerie médicale et de radiothérapie (342, septembre)*, pages 20–25, 2024.

[9] Boris N Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. N-beats : Neural basis expansion analysis for interpretable time series forecasting.

In *International Conference on Learning Representations (ICLR)*, 2020.

[10] B Perez. Amélioration de la sécurité chirurgicale avec un jumeau numérique prédictif : le rôle des systèmes multi-agents et de l'apprentissage par renforcement. 2024.

[11] G Peter Zhang. Time series forecasting using a hybrid arima and neural network model. *Neurocomputing*, 50 :159–175, 2003.

[12] Guoqiang Zhang, B Eddy Patuwo, and Michael Y Hu. Forecasting with artificial neural networks : : The state of the art. *International journal of forecasting*, 14(1) :35–62, 1998.

Ingénierie de la Connaissance et Management de la Connaissance au service de l'efficience de la « Mémoire d'Entreprise »

Alain Berger[1] & Patrick Prieur[1]

[1] Ardans SAS,
6 rue Jean Pierre Timbaud, « *Le Campus* » Bâtiment B1, 78180 Montigny-le-Bretonneux, France
{aberger, pprieur}@ardans.fr • https://www.ardans.fr

30 juin 2025

Résumé

En 2024, l'Observatoire B2V des Mémoires® s'est emparé de la question de la « mémoire de l'entreprise » et a lancé des actions concrètes pour pointer ce sujet dans les sphères managériales. Membre de son Conseil Scientifique, le Pr Jean-Gabriel Ganascia nous a invités à éclairer cette réflexion par notre vision industrielle des démarches en cours et de l'apport de l'Intelligence Artificielle (IA) dans le Management de la Connaissance (KM), la question étant ainsi formulée : « Comment positionner le « Management de la Connaissance » dans cette réflexion à propos de la « Mémoire d'Entreprise » ? » L'article présente dans une première partie la démarche considérée comme pionnière, singulière et exemplaire du Commissariat à l'Énergie Atomique et aux Énergies Alternatives. Dans un deuxième temps, l'article présente un regard sur l'arrivée de la norme ISO30401 *et les exigences attendues sur les « Systèmes de Management de la Connaissance » (SKM). La contribution de l'IA déjà significative et la venue sur les grands modèles de langage sont abordés avec la prudence industrielle qui s'impose. La conviction des auteurs est claire : l'ingénierie de la connaissance et le management de la connaissance sont au service de l'efficience de la « Mémoire d'Entreprise ».*

Mots-clés

Management de la Connaissance, Knowledge Management, KM, Mémoire d'Entreprise, Capitalisation et Exploitation des Connaissances, PARNASSE, ISO30401, *KB_ Scope®, Observatoire B2V des Mémoires®, Intelligence Artificielle, IA, Connaissance clé, Connaissance cruciale, Ingénierie de la Connaissance, Transfert de Connaissance d'Expert, EKT,* COGNICOACH, *Système de Knowledge Management, SKM, Grands Modèles de Langage, LLM.*

Abstract

In 2024, the B2V Memory Observatory® took up the issue of "corporate memory" and launched concrete actions to bring this subject to the forefront of management. Prof. Jean-Gabriel Ganascia, a member of the Scientific Advisory Board, invited us to shed light on the subject, based on our industrial vision of current approaches and the contribution of artificial intelligence to KM : "How can we position Knowledge Management in this reflection on Corporate Memory ? In the first part of the article, we present the pioneering, singular and exemplary approach of the Commissariat à l'Énergie Atomique et aux Énergies Alternatives. The second part looks at the arrival of the ISO30401 standard and the expected requirements for KMS or "Knowledge Management Systems". The already significant contribution of Artificial Intelligence and the arrival of major language models are discussed, with the necessary industrial limits of caution. The author's conviction is clear : knowledge engineering and knowledge management are at the service of the efficiency of the "Corporate Memory".

Keywords

Knowledge Management, Corporate Memory, Knowledge Capitalisation & Exploitation, PARNASSE, *KB_ Scope®, Observatoire B2V des Mémoires®, Artificial Intelligence, Key knowledge, Crucial knowledge,* ISO30401, *Knowledge Engineering, Expert Knowledge Transfer,* COGNICOACH, *Knowledge Management System, Large Language Models.*

1 Avant-Propos

Quand début 2024, l'Observatoire B2V des Mémoires® [1] affirme que « *Sans mémoire, pas d'avenir* », la question de la *« Mémoire d'Entreprise »* est au cœur du sujet. Son Conseil Scientifique valide l'engagement pris par la Direction Générale d'instruire ce sujet stratégique pour les entreprises et leurs salariés. Le Pr Jean-Gabriel Ganascia nous a alors invités à éclairer cette réflexion par notre vision industrielle, justifiée par vingt-cinq années de réalisation de telles opérations et, complétée par notre connaissance de l'intelligence artificielle : ceci afin d'illustrer son apport dans le *« Management de la Connaissance »* (ou Knowledge Management *i.e.* KM). La question était donc initialement ainsi formulée :

1. https://www.observatoireb2vdesmemoires.fr est le fonds de dotation créé en avril 2013 par le Groupe de protection sociale B2V. Il constitue son laboratoire social et sociétal sur la « *mémoire* ».

« *Comment positionner le « Management de la Connaissance » dans cette réflexion à propos de la « Mémoire d'Entreprise » ?* » Il convient de rappeler que l'action de l'Observatoire B2V des Mémoires® s'est déjà concrétisée par un premier sondage d'opinion [2], par une conférence à Lille le 7 juin 2024 [3] par un enseignement sous la forme d'un Certificat de formation Continue « *Mémoire de l'Entreprise* » [4] avec l'Université Paris Dauphine-PSL et la Fondation Maison de Salins, et par une première action mémorielle auprès de retraités chez un industriel.

Notre intervention lors du Certificat a mis en lumière toute la problématique sur l'apport du KM dans la *« Mémoire d'Entreprise »*, d'où cette proposition de clarification de notre perception.

2 Introduction

Comme tout organisme vivant, l'entreprise se dote naturellement d'une mémoire. Elle construit des documents, des procédures, des archives tant pour son existence administrative que pour exercer ses activités métiers. La mise en place de systèmes d'information pour les différentes fonctions de soutien ou de production fait que si l'enjeu de l'efficience collective est adressé par ce support technologique informatique, la pertinence de la justification « *métier* » reste dans la tête des humains qui font montre de discernement et surtout d'expertise.

Comment pérenniser ces savoirs, comment les expliciter, comment les transmettre, comment les exploiter ? Dans les systèmes de management, la norme ISO9001:2015 qui traite de la qualité est depuis 2018 consolidée par la norme ISO30401:2018 qui est consacrée au système de management de la connaissance.

A ceux qui souhaitent se doter d'une solution « base de connaissance » augmentée par une intelligence artificielle, il convient de leur recommander de débuter par se doter d'un « *Système de Management de la Connaissance* » (ou SKM pour « *Système de Knowledge Management* ») autour d'une équipe dédiée ce qui est rarement le cas. Un système informatique même doté d'IA ne saurait pas pallier un tel déficit organisationnel.

S'il faut donner du temps au temps pour une telle mise en place, les différentes étapes qui seront réalisées consolideront tant la maîtrise des processus métier que celles des compétences nécessaires pour identifier les savoirs clés et pérenniser les connaissances cruciales. Un tel SKM renforce l'identité culturelle de l'organisme, améliore la qualité des échanges par un langage commun partagé, accélère l'intégration de nouvelles ressources humaines, consolide la qualité des produits ou services rendus, et appuie la R&D pour anticiper les innovations. C'est certes un long chemin, cependant il procure en général un résultat particulièrement fructueux pour ceux qui l'ont emprunté.

Nous observerons que le *« Management de la Connaissance »* est bien devenu la clé stratégique de la réflexion sur l'apport de la *« Mémoire d'Entreprise »*.

L'article introduit la démarche conduite au sein du Commissariat à l'Énergie Atomique et aux Énergies Alternatives (CEA) qui se révèle aujourd'hui comme pionnière, singulière et exemplaire.

Dans un deuxième temps, un regard est posé sur l'apport de la norme ISO30401, et les exigences attendues sur les SKM en particulier par leurs utilisateurs (cf. figure 3a) (voir § 4.4).

A la contribution préliminaire de l'Intelligence Artificielle déjà significative, la venue des grands modèles de langage est abordée avec les limites industrielles de prudence qui s'imposent : halte aux ultracrépidarianistes [5] [12].

3 L'exemple remarquable du CEA

Si l'on remonte aux années 1980, premier âge d'or de l'IA, de nombreuses sociétés de tous les domaines (automobile, aéronautique, sidérurgie, transport, banque, énergie, l'agroalimentaire, etc.) disposaient d'une cellule ou d'un groupe en IA dont l'objectif était de « *gérer la connaissance* ». Une quarantaine d'année plus tard, la plupart de ces unités ont disparu. Depuis le réveil de l'IA par les technologies connexionnistes (Machine learning, deep learning) de nouveaux groupes se sont créés et une accélération est notable depuis 2022 avec les grands modèles de langage : serait-ce un feu de paille généré par un effet de mode ?

L'exemple du CEA est singulier car il est ancré dans le temps avec une certaine anticipation et modernité. Il paraît important de préciser que le choix de cet organisme est aussi une preuve que les structures dites étatiques ont mis en place des mécanismes remarquables qui durent dans le temps indépendamment des évolutions politiques à la tête de l'état ; les intérêts fondamentaux de la Nation [6] et une « *Mémoire de la Nation* » est alors instaurée.

Voici comment la mission initiale de cet EPIC [7] de développer de nouvelles connaissances scientifiques et transfère des innovations technologiques auprès du monde industriel, a pris un nouvel essor ces trente dernières années.

Dans le prolongement de la signature du traité d'interdiction des essais nucléaires décidé en 1996, le Président de la République, Jacques Chirac, en visite au Commissariat à l'Énergie atomique-Direction des applications militaires (CEA DAM), au Centre de Bruyères-le-Châtel, le 7 sep-

2. Sondage réalisé par l'institut IFOP sur le sujet « *mémoire de l'entreprise* » en mars 2024 auprès de 1000 cadres français : https://www.observatoireb2vdesmemoires.fr/sondage-dopinion

3. Conférence avec le Medef des Hauts de France et l'Institut Choiseul présentant la démarche en cours et les premiers résultats obtenus https://www.observatoireb2vdesmemoires.fr/lobservatoire/memoire-de-lentreprise/conference-lentreprise-en-memoires

4. https://executive-education.dauphine.psl.eu/formations/certificat/memoire-entreprise

5. « *Sutor, ne supra crepidam* », littéralement, le cordonnier (*sutor*), pas plus haut que la sandale (*crepidam*). Rapportée par Pline l'ancien dans son Histoire naturelle, cette sentence latine signifie que, « *de ce qui va au-delà de son métier, et que l'on ignore, on ne devrait parler* ».

6. Article 410-1 du Code pénal https://www.legifrance.gouv.fr/codes/article_lc/LEGIARTI000006418343.

7. EPIC pour Établissement Public Industriel et Commercial : c'est une personne morale de droit public gérant un service public spécialisé, distincte de l'État, et des collectivités territoriales, mais rattachée à eux.

tembre 2006, déclarait [8] qu'il « *assignait au CEA, et notamment à la DAM, la mission afin qu'il continue à assurer la crédibilité de notre dissuasion et à participer au développement et au rayonnement mondial de la science française* » ; c'est un message clair où l'institution doit rester à l'état de l'art.

3.1 Le préalable incontournable

Si cette démarche peut sembler être une tautologie, il faut considérer deux notions fondamentales liée à l'IGI1300 [22] qui définit les exigences de sécurité des systèmes d'information amenés à traiter des informations ou supports classifiés. Celle du « *droit à en connaître* » et celle de la « *protection des personnes physiques* » dont la conséquence est une distribution sur le territoire des visions partielles du sujet à résoudre collectivement. A ce point, les actions peuvent être planifiées par une équipe dédiée maîtrisant les outils de capitalisation [9] de cette mémoire, et ce, afin de garantir la mission assignée au CEA.

3.2 La traduction de la mission

Après avoir réalisé un programme de simulation numérique qui garantisse la fiabilité et la sûreté des armes, la question de la pérennité du savoir des scientifiques impliqués se posait alors. Il convenait aussi de se doter de l'outil scientifique qui allait valider les phénomènes de la physique fondamentale prévus par des expérimentations en laboratoire : c'est la finalité du Laser MégaJoule.

3.3 De l'esprit humain à l'explicitation et la modélisation à des fins de pérennisation pour l'exploitation et la transmission

Le programme « Capitalisation et Exploitation des Connaissances » (ou CEC) se voyait confié dès avril 1996 la mission de capitaliser, pérenniser et transmettre les connaissances initialement acquises lors de ces essais, puis, de préserver les savoirs critiques ou cruciaux avant le départ d'experts, et enfin, lorsque les expérimentations d'un projet se heurtent à un mur technologique, de conserver tous les acquis et retours d'expérience à la suspension de ces travaux de recherche. Bien évidemment, cette dimension stratégique de gestion de la connaissance avait été anticipée et confirmée bien avant par la Direction du CEA, dès 1994, comme une directive de son manuel qualité [24] et des publications reconnues [21, 16, 11].

3.4 Une mémoire d'entreprise pour une double efficience

Cet exemple est particulièrement riche car il s'agit pour un tel organisme de se doter de l'outil de mémoire qui lui confère une double efficience technique et économique ; en disposant d'un moyen d'éviter de refaire deux fois la même étude et la même expérimentation, et donc, d'éviter de perdre du temps en étant sûr de consacrer l'argent de son budget à aller plus avant dans la recherche.

3.5 Une mémoire d'entreprise multi-facettes

Cet outil de mémoire destiné à soutenir la recherche, est multi-facettes dans la mesure où il se traduit par différents types d'objectifs parmi lesquels :

- « *Agréger* » : dans des codes de calculs les fruits des expériences analysées et de leur modélisation associées.
- « *Numériser* » : des documents (avant qu'il ne s'effacent), les référencer et les archiver selon les règles.
- « *Filmer* » : les gestes métiers appropriés dans les opérations manuelles.
- « *Recueillir et expliciter* » : les retours d'expériences, les savoirs, les expertises des sachants dans des recueil de connaissance avant qu'ils ne quittent leurs fonctions.
- « *Pérenniser* » : la connaissance acquise jusqu'à l'achèvement du programme qu'il soit parfaitement atteint ou suspendu dans l'attente d'une évolution de l'état de l'art.

3.6 La cartographie partagée pour analyser la nature de cette mémoire

Pour le CEA comme pour d'autres organismes, ces travaux ont permis d'approcher une nouvelle vision de cette mémoire du savoir : une vision « méta » traduite sur une carte (cf. figure 1b). Ces cartes sont des graphes où les noeuds représentent les éléments de connaissances et les arêtes (voire arcs) composent les liens entre des éléments en relation.
L'exemple de cartographie [10] est celui d'un processus de l'activité métier qui consomme k ressources cognitives qui sont détenues par x acteurs répartis sur y sites (cf. figure 1a) [25]. Cette prise de conscience révèle les sujets qui sont prioritaires parmi ceux qui sont à prendre en compte. On parle de **connaissances clés** - celles pour la réalisation des activités qui constituent le savoir-faire métier et sont l'élément différenciant vis-à-vis de la concurrence - qui peuvent devenir des **connaissances cruciales** [1] - celles sans lesquelles les problèmes essentiels de l'entreprise n'ont pas de solution - en particulier lorsque le maintien de ce savoir n'est pas garanti à un horizon défini.

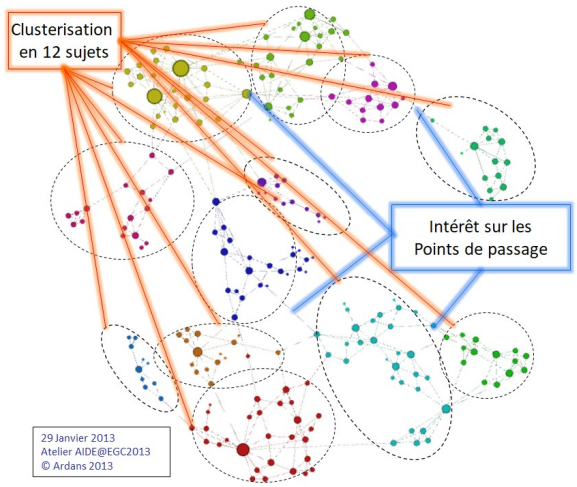
3.7 L'émergence d'une stratégie d'organisation manageant cette mémoire

Les actions détectées sont alors ordonnées selon des critères bien établis : il est clair qu'une stratégie de management de la pérennisation de la connaissance émerge et se déroule. Ceci a nécessité une réflexion sur l'organisation à mettre en place pour réaliser ces cartographies, les consolider, les agréger, les analyser pour proposer la stratégie pertinente fondée sur ce processus d'élicitation d'une solution

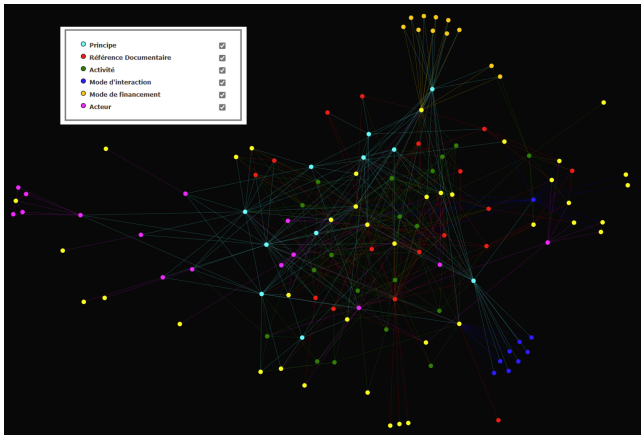
8. https://www.vie-publique.fr/discours/163298-declaration-de-m-jacques-chirac-president-de-la-republique-sur-la-dis

9. Capitaliser sur les connaissances de l'entreprise, c'est considérer les connaissances utilisées et produites par l'entreprise comme un ensemble de richesses constituant un capital, et en tirer des intérêts contribuant à augmenter la valeur de ce capital [14]

10. AKM ou Ardans Knowledge Maker® v2025 est l'implantation de la méthode Ardans [17].



(a) Les « *clusters* » du graphe de l'expertise modélisée avec AKM [25]



(b) Le graphe généré par le KB_ Scope® d'AKM : une cartographie "3D" de l'expertise

FIGURE 1 – L'analyse du graphe de la base d'expertise fruit de l'ingénierie de la connaissance

optimale pour toute l'entité. L'application EDIFICE [11] accompagne les membres du projet CEC à piloter cette activité modélisée selon le triptyque « *Activité Métier, Connaissance, Objet Cognitif* » en liant les aspects organisationnels (unité, programme), process (REX, référentiel) et cognitifs (documents, experts) [7].

4 La norme pour traduire le SKM

Le savoir individuel a été échangé, montré, expliqué (ou Socialisé) avant d'être traduit, formalisé, explicité pour être transmis (ou Externalisés) à une communauté choisie. Celle-ci va alors le structurer et l'intégrer (ou le Combiner) pour générer de nouveaux éléments qui appropriés (ou Internalisés) vont constituer de nouveaux savoirs individuels qui à leur tour… On retrouve ce qui a été identifié et qualifié de SECI par Nonaka [20].
Depuis 2018, la question s'est ainsi naturellement replacée vers la question du SKM ou Système de Management de la Connaissance. Les organismes confrontés au §7.1.6 de la norme ISO9001:2015, ont pu apprécier la retranscription de la question du management de la connaissance organisationnelle dans la norme ISO30401:2018 [23].

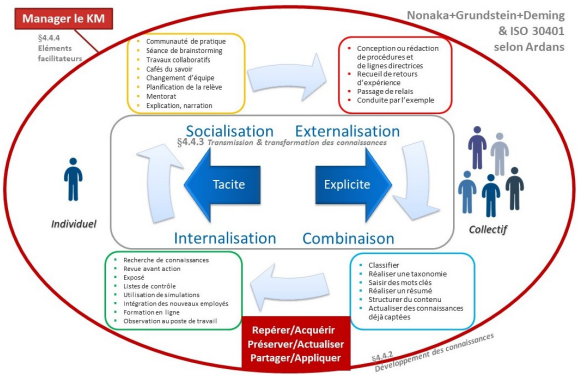
Il ne s'agit pas d'une révolution mais d'une consolidation où coexistent les travaux de Nonaka (cf. supra), de Grundstein [15] et de Deming [12] !

L'apport majeur de cette norme est de poser les bases d'un langage commun sur les exigences attendues pour qu'un organisme puisse se prévaloir de l'implantation d'un Système de Management de la Connaissance.

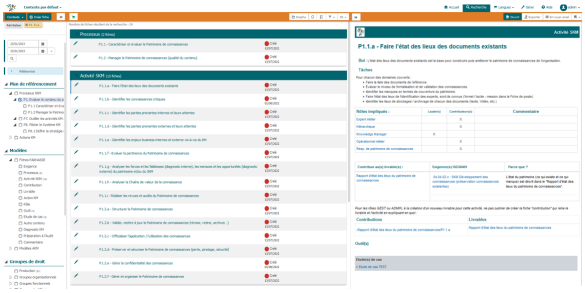
4.1 Le regard processus pour s'approprier et se mesurer à la norme

La question du management de la connaissance est *in fine* une question de progrès pour l'espèce humaine. Le passage de l'oralité à l'écrit ne s'est pas passé en une génération. De la même façon que l'échange oral est l'objet d'une langue avec ses propres codifications, celle de la représentation écrite a été l'objet d'avancée technique de la pierre, au papyrus, au papier, du pictogramme, idéogramme, hiéroglyphe, à une écriture syllabique puis alphabétique avec des caractères… la formalisation étant une codification et la lecture un décryptage. A l'heure du numérique, les sujets se sont décalés mais sont de même nature : comment représenter une image ? En point, en vecteur ? en discret ou en continu ? et la couleur ? Et le format…lequel sera celui de référence dans 10 ans ou dans un siècle ? Idem pour l'animation de l'image, la vidéo, le son, le texte… quel est le bon format qui va s'avérer pérenne ? La question est bien de savoir comment un organisme s'approprie la bonne organisation pour faire grandir le savoir contenu dans sa « mémoire collective d'entreprise », partagée par ses collaborateurs et sécurisée par rapport à des éventuelles agressions hostiles. Si la réponse est « ce n'est pas avec un outil informatique », il est clair aussi que l'outil informatique accompagne une démarche de management de la connaissance afin que l'utilisateur habilité puisse consulter, contribuer, questionner ses pairs, actualiser les contenus de son domaine de compétence. La prise en compte de la culture et du métier est essentielle pour que le dispositif puisse se fondre dans le quotidien. Il se révèle enfin indispensable de mettre en place la gouvernance qui s'impose et qui dispose des moyens adéquats, les comités qui modèrent, le langage commun qui est partagé dans le métier, et les processus qui concourent à la bonne hygiène de la vie du Système de Management de la Connaissance. La mémoire de l'entreprise est définitivement de la responsabilité des humains qui y collaborent à commencer par les responsables qui la dirigent.

11. Entrepôt de Données Intégrées et FIabilisées pour le pilotage de la gestion de ConnaissancE : méthode éponyme au logiciel du CEA.



(a) L'ISO30401 reprend Nonaka, Grundstein et Deming selon Ardans [3]



(b) PARNASSE guide le knowledge manager pour visualiser le processus SKM de son organisme

FIGURE 2 – L'ISO30401:2018 & PARNASSE aident à positionner la *« Mémoire d'Entreprise »*

4.2 Le Portail pour manager en connaissance le KM

Il convient de citer à ce point l'excellente initiative de l'association « Club Gestion des connaissances » [9] qui contribue à l'établissement de la norme ISO30401 pour la France (via l'Afnor) depuis son origine, l'a traduite avec une vision processus (cf. §4.3) au sein d'une méthode et d'un outil : PARNASSE [13] . L'idée est de rendre audible la norme, d'aider celui ou celle qui aura le rôle de Knowledge Manager par la mise à disposition d'un outil qui clarifie à quoi ressemble le KM dans son organisation.

De la même façon qu'une expertise peut s'illustrer sous forme du graphe (2D/ 3D) de la base de connaissance (cf. figure 1b), dans PARNASSE , le système KM se décline sous forme de 8 processus et 20 activités (cf. §4.3). Avec une telle modélisation, le Knowledge Manager dispose de l'outil pour maîtriser et manager en parfaite connaissance le SKM (cf. figure 2b) et ainsi le processus de mémoire de son entreprise.

12. La roue de Deming est une représentation graphique de la méthode d'amélioration continue des processus et de gestion de la qualité dite PDCA (Plan-Do-Check-Act)
13. PARNASSE pour Portail Associant la Référence Normative avec un Référentiel Structuré d'Entreprise implante la méthode du Club GC.

4.3 Les Processus du SKM de référence décrits dans PARNASSE

PARNASSE : **Les processus du SKM de référence** du Club Gestion des Connaissances

Processus 1. Évaluer le contenu du patrimoine et le gérer

- P 1.1. Caractériser et évaluer le Patrimoine de connaissances
- P 1.2. Manager le Patrimoine de connaissances (qualité du contenu)

Processus 2. Faire vivre le patrimoine de connaissances et garantir son application

- P 2.1. Formaliser et mettre à disposition les connaissances
- P 2.2. Garantir l'application des connaissances
- P 2.3. Recenser les connaissances utiles à l'Organisation
- P 2.4. Gérer les Communautés de savoir et gérer l'expertise

Processus 3. Gérer et piloter les dispositifs d'acquisition de connaissances

- P 3.1. Processus RH - Recenser le besoin en formations nécessaires à l'activité (actuelle et future)
- P 3.2. Processus RH - Gérer et piloter l'apprentissage individuel (MOOC, e-learning, Coaching, etc.)
- P 3.3. Gérer et piloter l'apprentissage en interaction collective (groupes d'expertises, séminaires, communautés d'apprentissage, etc.)
- P 3.4. Définir les besoins en recrutement en lien avec les connaissances critiques de l'Organisation
- P 3.5. Processus RH - Gérer et piloter la construction des formations et solutions d'apprentissage

Processus 4. Soutenir les dispositifs de créativité et d'innovation

- P 4.1. Soutenir les activités de créativité
- P 4.2. Soutenir l'activité d'innovation
- P 4.3. Faire le bilan des connaissances acquises au cours des activités d'innovation / créativité

Processus 5. Soutenir les processus opérationnels

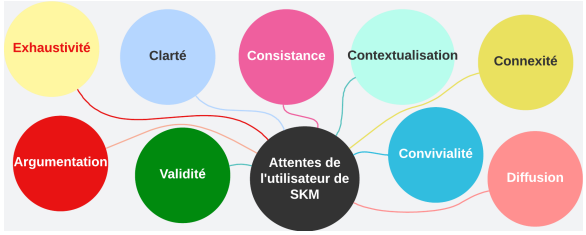
Processus 6. Transformer l'information externe en connaissance utile pour l'organisation

Processus 7. Outiller les activités KM

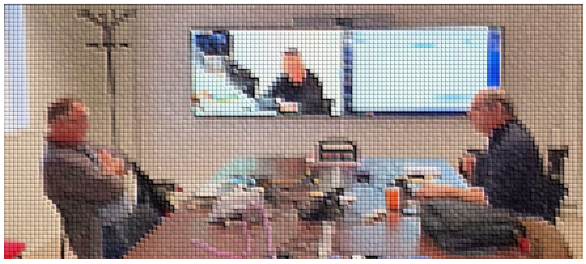
- P 7.1. Interagir avec les outils d'IA

Processus 8. Piloter le Système KM

- P 8.1. Définir la stratégie et les objectifs KM
- P 8.2. Construire le plan KM accepté par la direction de l'Organisation
- P 8.3. Évaluer le Système KM : les audits
- P 8.4. Superviser le Système KM : processus de décision, revues de pilotage, tableaux de bord des indicateurs, ressources humaines et matérielles, niveau de compétence, etc.
- P 8.5. Organiser et conduire les actions de mise en place et d'amélioration du Système KM : sensibiliser, communiquer, mobiliser les acteurs, conduire les actions, etc.



(a) Les exigences d'un utilisateur de SKM



(b) La session de COGNICOACH

FIGURE 3 – L'ingénieur de la connaissance médiateur entre expert, receveur et usager

4.4 L'expression des exigences d'un utilisateur par rapport au SKM

Pour être plébiscité par les acteurs d'une organisation, la base de connaissance ou le Système de Management de la Connaissance doit satisfaire aux attentes des acteurs [3] (cf. figure 3a) dont notamment :

- « *Exhaustivité* » : il convient que la connaissance soit exhaustive sur le périmètre sur laquelle elle s'applique ; un nouvel utilisateur débute toujours par un test et afin d'obtenir sa confiance le système devra lui retourner la réponse pertinente !
- « *Clarté* » : les contenus sont clairs, dénués de toute ambiguïté, cela pour faciliter l'adhésion, l'appropriation et le bon usage par l'utilisateur.
- « *Consistance* » : les résultats de « *navigation* » pour obtenir les contenus sont cohérents, réplicables, non-contradictoires ; cette stabilité rassure l'utilisateur.
- « *Contextualisation* » : il est fondamental de bien décrire le contexte dans lequel cette connaissance est valide pour être exploitée en toute sérénité.
- « *Connexité* » : l'élément de connaissance consulté est au cœur d'un réseau (implicitement sémantique) d'éléments de connaissance au sein desquels il est positionné dans une représentation cartographique multidimensionnelle : un réseau précieux pour évaluer la qualité de la base comme son homogénéité, ses relations, ses trous, ses densités.
- « *Diffusion* » : la connaissance est un actif précieux et est restreinte à ceux habilités à en connaître, celui qui en bénéficie doit savoir le mesurer.
- « *Convivialité* » : plus que jamais l'ergonomie d'un système à base de connaissance moderne doit être d'une ergonomie intuitive et fluide et démontrer qu'elle offre à l'usager un retour sur investissement sans pareil.
- « *Validité* » : la connaissance est vivante, comme elle s'affine dans le temps, elle est intrinsèquement « non monotone » et doit être datée.
- « *Argumentation* » : les contenus sont argumentés et disposent des niveaux de preuve nécessaires pour une bonne appropriation par le lecteur.

Le processus P 7.1 « *Interagir avec les outils d'IA* » (cf. §4.3) doit donc prendre en compte tous ces éléments implicitement attendus d'éthique et de confiance afin de satisfaire pleinement l'utilisateur du SKM.

5 L'apport de l'intelligence artificielle

Le « *Management de la Connaissance* » devant être pratiqué par les ingénieurs de la connaissance [2] s'inscrit directement dans la branche « *Connaissance* » de l'intelligence artificielle telle que nativement définie à Dartmouth [18] lors du séminaire fondateur de l'IA. Si les systèmes experts et les systèmes à base de connaissance « *d'antan* » ne fonctionnent pas de la même façon que les bases de connaissance actuelles [14], ces dispositifs nécessitent tous de colliger de la connaissance, de l'éliciter et de suivre un processus pour faire qu'elle soit validée par l'expert donneur avant d'être mise à disposition pour être actualisée dans le futur. Quand la validation des règles des systèmes experts était très complexe (maîtrise du déclenchement de la règle ou validation du système général), la validation des bases de connaissance actuelles est rendue plus abordable même si elle requiert le même soin, la même attention et la même précision. Nous aurons bien noté au passage que ces bases actuelles dérivent de « *concepts de l'IA* ».

5.1 La contribution dans l'abstraction et la représentation

L'outil Ardans Knowledge Maker® reconnu en France comme le référent dans les démarches de capitalisation d'expertise intègre des notions industrialisées issues des techniques et « *concepts de l'IA* ». C'est une « *solution hybride* » car l'outil exploite ces dispositifs :

- « *Taxonomie* » : la constitution d'arborescences de concepts classifiés pour décrire le langage du métier est un concept très précieux : cela aide le novice à comprendre cette hiérarchie de termes, à les positionner les uns par rapports aux autres, cela aide à décrire les environnements qu'ils soient physiques (comme dans l'ingénierie système) ou fonctionnels, que cela soit des contextes de travail ou des notions de priorité de droit ; L'arbre de Porphyre (cf. figure 4) est l'ancêtre

14. Les connaissances étaient écrites sous format de règles pour qu'un moteur d'inférence raisonne selon des faits établis, quand les connaissances sont aujourd'hui décrites dans des objets comprenant des champs de texte ou de valeur avec des liages entre les objets.

de cette représentation. Cette mise en avant de concept sert aussi l'indexation de la base de connaissance.

- « *Objet* » : la représentation des connaissances est très friande de l'usage de ce concept ; que l'on parle de Classe, Attribut, et Instance, ou de Modèle, Rubrique, et Fiche, il s'agit de peu ou prou de la même chose !
- « *Héritage* » : l'héritage est une notion essentielle de la programmation orientée objet qui permet de définir une nouvelle classe à partir de classes existantes ; idem pour les modèles qui peuvent hériter de modèles.
- « *Ontologie* » : le liage entre les éléments qui existent dans les bases de connaissance est extrêmement précieux. Il s'agit d'un véritable « *graphe sémantique* » qui est élaboré au fil de l'eau. Il est autant utilisé pour l'élaboration que pour la consultation de la base de connaissance [25].
- « *Apprentissage* » : l'indexation n'est pas que syntaxique, elle est aussi sémantique. elle se réalise par un apprentissage sur les contenus validés de la base de connaissance pour appuyer l'utilisateur en consultation comme en contribution [6].
- « *Hypertexte, url, web* » : le succès initial de l'IA est sa capacité de tissage du lien qui rend possible la navigation dans l'ontologie [13] si implicite à nos systèmes actuels qu'il serait indécent d'omettre de le citer !

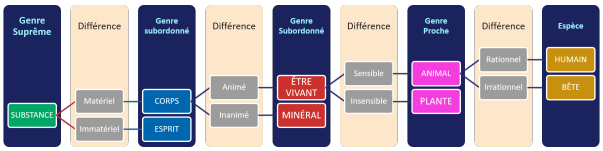


FIGURE 4 – L'arbre de Porphyre (234-305) : à chaque étape un genre se différencie du précédent

5.2 Le process COGNICOACH en garantie du transfert de connaissance

Parmi les actions de *« Mémoire d'Entreprise »* on remarque la pertinence de l'« *EKT* » pour Transfert de Connaissance d'Expert (*i.e. Expert Knowledge Transfer*). Nous avons évoqué supra le processus long qui concerne l'explicitation de la connaissance avec en particulier le travail méticuleux et précis de la validation d'une base de connaissance : la photographie (cf. figure 3b) illustre ici concrètement à quoi ressemble les deux objectifs d'un transfert, avec d'une part de validation de la connaissance de l'expert "donneur", et d'autre part la bonne appropriation du "receveur" : on définit ce process « COGNICOACH ».

L'ingénieur de la connaissance échange avec l'expert face à lui sur l'explicitation qui a été traduite dans la base de connaissance. Ce travail est présenté sur l'écran de l'ordinateur qui est retransmis simultanément via la vidéoconférence à l'expert receveur (ici à 1350 km).

On observe ainsi que la retranscription de cet « *élément de connaissance* » est soumis à l'expert pour valider la fidèlité de l'élicitation de son point de vue. Ce contenu est alors proposé au regard du receveur, qui étant déjà un spécialiste du domaine va naturellement « *stresser* » ce contenu. De cette réaction, un ajustement est réalisé si nécessaire afin de consolider l'appropriation par le receveur. Ce « COGNICOACH » vise à garantir l'appropriation du receveur et vérifie notamment les aspects « *Clarté* », « *Contextualisation* », « *Argumentation* »…(cf. §4.4).

Dans certains cas, un complément nécessitant un nouvel élément de connaissance est produit et ajouté : la formulation de Davenport [10] a été justement transcendée en « *Transfert de connaissances = Transmission + Absorption & Utilisation + Enrichissement* » [4] [15].

En effet, un transfert de connaissance n'est parfaitement réalisé que lorsque le receveur va au-delà de ce qui lui a été transmis, en contribuant, à son tour, à l'enrichissement du patrimoine de son organisation. Ce mode de construction « COGNICOACH » garantit une excellente confiance dans le système qui gère cette mémoire : à charge de l'organisation de maintenir la bonne actualisation du travail consciencieux initialement produit.

5.3 L'interrogation sur l'exploitabilité en confiance des LLM pour cette *« Mémoire d'Entreprise »*

L'arrivée des grands modèles de langage (en anglais « *Large Language Models* » ou LLM) est en train de chahuter la question de l'exploitation des outils informatiques pour interroger de grands volumes de données et de textes en particulier. Le sujet sur lequel il n'y a pas de question est que les bases de connaissance qui contiennent une mémoire stratégique de l'entreprise ne sont pas connectées à l'extérieur pour des questions triviales de sécurité. La problématique concerne donc le volume de "contenu" qui est nécessaire pour avoir la capacité à disposer d'un apprentissage efficient afin d'avoir des réponses « *pertinentes* » et celui à partir duquel les « *hallucinations* » vont insidieusement apparaître. Aujourd'hui, nous observons sur nos bases qu'une progression dans le séquencement des étapes suivantes délivre des résultats prometteurs selon Moris [19] :

1. Disposer de l'ontologie et la modélisation pertinente.
2. Fonder un corpus de connaissance validé significatif.
3. Produire une indexation sémantique sur ce corpus.
4. Exploiter 1, 2 et 3 pour exécuter le principe du RAG (Retrieval-Augmented Generation [16]) en produisant un « *Prompt Augmenté* » afin que le LLM génère une réponse pertinente, reproductible et donc consistante.

Ce dernier point est essentiel pour conserver la confiance dans le dispositif (cf. figure 5). On note que pour l'utilisateur lambda les impératifs suivants sont implicitement respectés :

15. A la version **Knowledge Transfer = Transmission + Absorption & Use** est ajouté **+ Enrichment**

16. En français « *Génération Augmentée de Récupération* ». Dans le Bulletin n°123 de l'AfIA, le RAG est défini comme une technique de traitement du langage naturel considérée comme un sur-ensemble du LLM. L'objectif est de prendre en compte « des règles ou des faits plus récents et plus fiables » afin de gommer le côté statistique des LLM.

- « *Justification* » : il peut disposer de la « *justification* » de cette réponse avec la ou les éléments de connaissance qui ont permis de répondre à la requête.
- « *Maîtrise* » : le système s'abstient de générer des extrapolations si la base de connaissance est vide sur le sujet du questionnement.
- « *Consistance* » : le système garantit le fait de rester dans le périmètre maîtrisé par la base et donc n'est pas victime d'hallucinations en mélangeant allègrement des notions présentes mais décontextualisées.

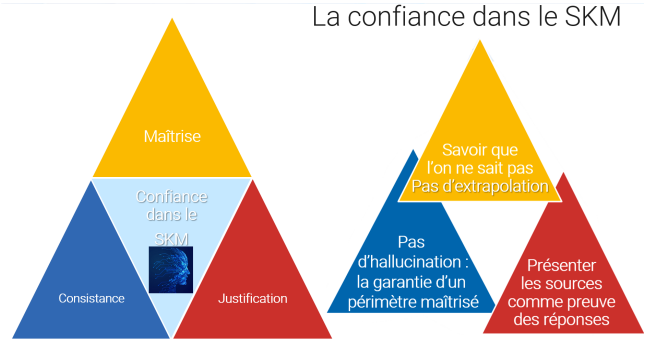


FIGURE 5 – Acquérir la confiance de l'utilisateur

Néanmoins la spécification du RAG comme sa validation sont très ésotériques et subtiles voire métastables à l'usage. Le danger réside dans le fait que l'utilisateur soit fasciné par l'élégance de la réponse et qu'aveuglé, il ne soit pas en mesure de discerner qu'il ne s'agisse que d'une assertion fantasque, magnifiquement sublimée par une écriture délicieusement spécieuse.

6 La mémoire de l'entreprise utile valorisée par le KM

Si Friedrich Nietzsche affirme que « *le futur appartient à celui qui a la plus longue mémoire* », il faut admettre que l'entreprise a un véritable enjeu patrimonial à savoir valoriser sa mémoire collective. Les processus métiers, qui évoluent pour suivre les changements réglementaires ou législatifs, pour anticiper les mouvements des marchés qui les composent, pour enrichir l'offre en continu pour satisfaire la clientèle, doivent être parfaitement appropriés par les personnels concernés. Ils doivent aussi disposer de l'entière adhésion de tous les collaborateurs qui contribuent aux produits ou aux services considérés. Le problème est que la transmission d'un savoir et de son assimilation pour que celui-ci devienne une compétence réelle, nécessitent du temps. Dans le questionnement de ce qui doit (est utile à) être mémorisé pour l'entreprise, il y a :

- « *Infobésité ou Pertinence* » : l'accumulation de contenu dans le temps risque de noyer le collaborateur par des informations superfétatoires (car prochainement obsolètes). Par exemple, un obstacle qui bloque l'innovation aujourd'hui ne sera-t-il pas contourné demain ? Mémoriser une telle situation serait-il potentiellement profitable à l'entreprise ?
- « *Bruit ou Intérêt* » : un contenu n'a de valeur ajoutée que s'il facilite le discernement du collaborateur. Il doit faciliter les choix et les justifications du collaborateur par sa pertinence et sa complétude (absence de silence), et ne pas perturber la réflexion de ce dernier par la proposition d'éléments inappropriés.
- « *Transverse ou Verticale* » : la complexité comme la richesse des objets manipulés dans les organisations font que, à la connaissance verticale d'un expert, s'ajoute une connaissance transversale d'un architecte, voire une expérience en profondeur de la vie de l'objet en exploitation par le mainteneur. Comment garantir la bonne cohérence globale ?
- « *Lièvre ou Tortue* » : à la vitesse d'exécution proposée par les RAG+LLM & Co, il convient d'apprécier celle de la sédimentation de l'humain pour l'élicitation et la justification de la connaissance auprès des pairs.

La question de la frugalité est aussi très prégnante dans les organisations. En tous les cas le « *Management de la Connaissance* » se doit d'être force de discernement pour l'organisation et l'« *Ingénierie de la Connaissance* » est un outil précieux pour se concentrer sur l'essence de la connaissance (cf. figure 6).

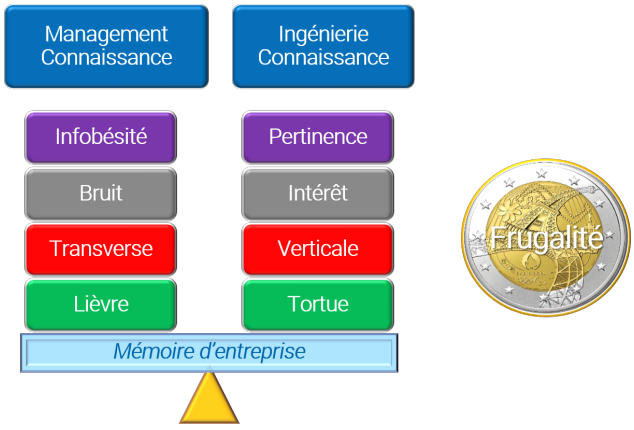


FIGURE 6 – Les enjeux du KM et l'IA en entreprise

7 Conclusion

La synthèse du cas CEA DAM illustre l'exemplarité de cette démarche de mise en place d'une mémoire avec une pérennisation du savoir à plusieurs niveaux : celui de la stratégie et de la finalité de cette mémoire d'entreprise, celui de la maîtrise du processus métier dans le contexte de sécurité attendu, celui de l'organisation pour détecter, acquérir, pérenniser, exploiter, actualiser les acquis, avec celui essentiel de la mise en place d'une équipe dotée des ressources pour éliciter et faire vivre cet actif patrimonial.
Acquérir une vision holistique de la connaissance dans certaines organisations est une gageure aujourd'hui. Il va falloir s'attacher à rendre cette mémoire vivante et accessible dans tous les sens du terme. On se gardera de croire encore pendant quelque temps que les moteurs fondés sur les

LLM résoudront le sujet en un clic : la *« Mémoire d'Entreprise »* est définitivement un sujet stratégique. Elle ne se construira que collectivement, intégrant la culture et l'ADN de l'organisme pour s'appuyer sur les acquis présents ou passés afin de mieux préparer l'avenir.
Il reste que l'ingénierie de la connaissance fille de l'« *IA Symbolique* » est une clé de l'IA qui dans son usage pour le management de la connaissance est d'une source d'efficience certaine pour concevoir et faire grandir la *« Mémoire d'Entreprise »*.

8 Postface

Cet article a été débattu lors des ateliers DAHLIA [17] et l'atelier KM-IA [18] qui se sont déroulés le 28 janvier 2025 à Strasbourg lors de la conférence EGC2025[5]. Concomitamment l'Institut Choiseul et l'Observatoire B2V des Mémoires® présentaient à Paris les résultats de leur étude [8] sur la « *Mémoire des entreprises* ».

Références

[1] Aline Belloni, Alain Berger, and Jean-Pierre Cotton. Cibler une action de gestion des connaissances appropriée dans un cadre industriel : retour d'expérience d'Ardans. In Sandra Bringay, editor, *3ème Conférence Nationale sur les Applications Pratiques de l'Intelligence Artificielle, APIA'2017*, volume https://lc.cx/bRYKO1, pages 35–43, Caen, 3-4 VII 2017.

[2] Alain Berger. Évolution dans l'industrie du métier d'ingénieur cogniticien ou d'ingénieur de la connaissance entre 1985 et 2015. In *1st Conférence Nationale sur les Applications Pratiques de l'Intelligence Artificielle (APIA'2015) at the Plate-forme Intelligence Artificielle*, volume https://lc.cx/YdDZMv, pages 23–33, Rennes, VII 2015.

[3] Alain Berger. Regard sur l'ingénierie de la connaissance face à l'ISO30401. In *34èmes Journées francophones d'Ingénierie des Connaissances (IC'2023)*, volume https://lc.cx/Hk8mDx, Strasbourg, VII 2023. Plate-Forme Intelligence Artificielle (PFIA).

[4] Alain Berger, Sébastien Boblet, Thierry Cartié, Jean-Pierre Cotton, and François Vexler. Implanter une approche hybride dans une démarche d'ingénierie de la connaissance pour manager les avis techniques relatifs au retour d'expérience d'exploitation d'un équipement sensible complexe. In *35èmes Journées francophones d'Ingénierie des Connaissances (IC'2024)* https://lc.cx/d0V0WQ. PFIA par AFIA et L3i La Rochelle Univ., VII 2024.

[5] Alain Berger and Patrick Prieur. Le « management de la connaissance » : la clé stratégique de la réflexion sur l'apport de la « mémoire d'entreprise ». In *25èmes Journées Francophones Extraction et Gestion des Connaissances, EGC'2025*, volume https://lc.cx/qEClSI, Stasbourg, I 2025.

[6] Alain Berger, François Vexler, Corentin Mary, and Jean-Pierre Cotton. Réflexion sur le choix d'un classifieur sémantique destiné à aider le cogniticien dans l'élaboration d'une base de connaissance et la garantie de sa consistance dans le temps. In *6ème Conférence Nationale sur les Applications Pratiques de l'Intelligence Artificielle, APIA 2020*, volume https://lc.cx/9m7WwU, pages 67–74, Angers, VII 2020.

[7] J.P. Bouchez. *L'entreprise à l'ère du digital - Les nouvelles pratiques collaboratives*. De Boeck, 1996.

[8] Institut Choiseul and Observatoire B2V. *La mémoire des entreprises*. Etudes Choiseul, Paris, I 2025.

[9] Patrick Coustillière. L'Ingénierie Système, un outil pour le KM Manager ? In *Club Gestion des Connaissances*, volume https://lc.cx/O3wTat, V 2022.

[10] Thomas Davenport and Laurence Prusak. *Working Knowledge : How Organizations Manage what They Know*, volume https://lc.cx/fPR85V of *EBSCO eBook Coll.* Harvard Business School, 1998.

[11] JL. Ermine, M. Chaillot, P. Bigeon, B. Charenton, and D. Malavielle. MKSM a method for knowledge management. In *International Symposium on the Management of Industrial & Corporate Knowledge (ISMICK'96)*, pages 288–302, Rotterdam, NL, 1996.

[12] Jean-Gabriel Ganascia. Le cordonnier, l'ultracrépidarianiste et chatgpt. In *Sciences et Avenir*, volume https://lc.cx/nyFPXU, Mai 2023.

[13] Jean-Gabriel Ganascia. *L'I.A. expliquée aux humains*. Edition du Seuil, Paris, IX 2024.

[14] Michel Grundstein. La capitalisation des connaissances de l'entreprise, système de production des connaissances. In *Colloque L'Entreprise Apprenante et les Sciences de la Complexité*, Aix-en-Pce, Mai 1995. Jeanne Mallet : L'organisation Apprenante. Faire, chercher, comprendre.

[15] Michel Grundstein. De la capitalisation des connaissances au management des connaissances dans l'entreprise, les fondamentaux du knowledge management. In *Management des connaissances en entreprise*, pages 25–54. Economics Papers from University Paris Dauphine, IV 2003.

[16] Pierre Malvache and Patrick Prieur. Mastering Corporate Experience with the REX Method, Management of Industrial and Corporate Memory. In *International Symposium on the Management of Industrial and Corporate Knowledge(ISMICK'93)*, pages 33–41, Compiègne, June 1993.

17. Site de DAHLIA : DigitAl Humanities and cuLtural herItAge : data and knowledge management and analysis :https://dahlia.egc.asso.fr/atelierDAHLIA-EGC2025.html

18. https://km-ia.sciencesconf.org/ et KM-IA pour « *Gestion des connaissances tacites en entreprise : réflexions, retours d'expériences, bonnes pratiques et mauvaises surprises de l'intelligence artificielle* »

[17] Pierre Mariot, Christine Golbreich, Jean-Pierre Cotton, and Alain Berger. Méthode, Modèle et Outil Ardans de capitalisation des connaissances. In *RNTI E12 Modélisation des Connaissances*, volume https://lc.cx/grwT3F, pages 187–206, 2007.

[18] John McCarthy, Marvin Minsky, Nathaniel Rochester, and Claude Shannon. A proposal for the dartmouth summer research project on artificial intelligence. In https://lc.cx/hzjF0Q, 1955.

[19] Elise Moris. Sémantique et LLM dans AKM. In *8ème édition d'Ardans Users'Group Meeting (AUGM2024)*. Ardans, Paris-Saclay, X 2024.

[20] Ikujiro Nonaka and Hirotaka Takeuchi. *The Knowledge-Creating Company : How Japanese Companies Create the Dynamics of Innovation*. Oxford University Press, 1995.

[21] Jean-Michel Penalva. Sagace : une représentation des connaissances pour la supervision de procédés continus. In https://lc.cx/6GA_8U, Juin 1990.

[22] Premier-Ministre. Instruction générale interministérielle 1300 sur la protection du secret de la défense nationale. In https://lc.cx/owM0sM, Paris, 11 août 2021. JORF n°0185.

[23] ISO Central Secretary. Knowledge management systems — requirements iso30401 :2018. In https://lc.cx/TWDeFT, International Organization for Standardization. Geneva, CH, 2018.

[24] Jean-Marie van Craeynest, Jean-Louis Ermine, and Christophe Chagnot. Capitalisation des connaissances dans le cadre d'un transfert industriel. In *IC'97, Ingénierie des Connaissances*, Roscoff, May 1997.

[25] François Vexler, Alain Berger, Jean-Pierre Cotton, and Aline Belloni. Eléments d'appréciation et d'analyse d'une base de connaissance : l'expérience industrielle d'Ardans. In *Actes Atelier AIDE à EGC'2013, 13ème Conférence Francophone sur l'Extraction et la Gestion des Connaissances*, volume https://lc.cx/z4taz0, pages 59–72, Toulouse, 2013.

Session 2 : Réflexions sur l'IA générative

Au-delà des discours, l'IA générative à l'épreuve des usages réels en entreprise

Robin Héron[1], Myriam Fréjus[1]

[1] EDF Recherche & Développement, SEQUOIA

robinrheron@gmail.com ; myriam.frejus@edf.fr

Résumé

L'Intelligence Artificielle Générative (IAG) prend une place centrale dans le débat public et les stratégies d'entreprise, notamment avec l'idée d'améliorer la productivité et l'efficience. Cependant il existe encore peu de travaux qualitatifs sur l'appropriation de ces dispositifs et les usages réels des utilisateurs. Nous présentons une étude réalisée dans le cadre d'une expérimentation en entreprise d'une IAG intégrée aux outils informatiques. Nous en décrivons le début d'appropriation et comment les usages s'instancient (finalités et freins) dans des situations réelles de travail.

Mots-clés

IAG, appropriation, situations réelles de travail, activité, ergonomie

Abstract

Generative Artificial Intelligence (GenAI) is making its way into the public debate and into corporate strategies, particularly with the idea of improving productivity and efficiency. However, there is still few qualitative studies on the appropriation of these devices and the actual uses made of them by users. We present a study carried out as part of a corporate experiment with a GenAI integrated with IT tools. We describe how the system is first adopted and how it is used (purposes and obstacles) in real work situations.

Keywords

GenAI, appropriation, real work situations, activity, ergonomics

1 Introduction

Le développement de systèmes à base d'IA a permis le transfert de tâches associées à la pensée humaine vers la machine. Avec l'IA dite « générative » (IAG), un nouvel élan a emporté l'opinion publique et, comme le titrait la Une du Times du 27 février 2023, une « course à l'armement » a lieu pour les grandes entreprises. Dans ce contexte, de nombreux rapports, particulièrement issus de la littérature « grise », font état des gains en termes de productivité, d'efficience et plus rarement de qualité lorsque les salariés utilisent des systèmes d'IAG. Les travaux en situations sont encore rares et les résultats rapportés sont centrés sur des indicateurs quantitatifs [7] offrant tout de même un éclairage sur l'IAG au travail. Dans ce contexte, il importe de préciser comment l'intégration d'un outil à base d'IAG dans une organisation sociotechnique en place transforme l'activité des salariés et de quelle nature est sa valeur ajoutée.

1.1 Travaux antérieurs

1.1.1 L'impact de l'IA : dépasser la vision quantitative et substitutive

La question de l'impact de l'IA sur les emplois est complexe, comme le montrent les controverses suite aux prédictions de Frey & Osborne[16] [31]. Bien que d'un côté l'automatisation de tâches conduise à la perte de certains emplois, de l'autre côté cela conduit à l'augmentation de la demande en travailleurs pour des tâches non automatisables du fait de l'augmentation de la productivité notamment [1]. Cela génère aussi de nouveaux métiers, comme par exemple pour la supervision / alimentation des IA [20].
Ainsi, penser l'impact des IA en termes de remplacement d'emplois ou de substitution de certaines tâches par la machine s'avère insuffisant.

1.1.2 Valoriser la complémentarité Humain-IA

Penser le rapport Humain-IA en termes de complémentarité offre une vision alternative qui permettrait de dépasser les limites de la substitution. Cette vision s'inscrit dans une perspective optimiste espérant de meilleures conditions de travail et de performance.
En ce sens, il apparaît que des outils à base d'IA peuvent soutenir le travail humain en apportant des feedbacks pertinents [43] ; en étant un assistant digital [29], en améliorant les prises de décision [8, 31, 44] et la relation de service par exemple [22, 37].
L'IA a été particulièrement développée et étudiée dans le milieu médical, et plus particulièrement la radiologie. La complémentarité Humain-IA est privilégiée pour laisser au médecin la responsabilité du diagnostic. Ces systèmes lui permettent une réévaluation active de son diagnostic en questionnant point par point les suggestions de la machine [25]. Pour autant cela n'est pas systématique notamment du fait de l'opacité du fonctionnement des systèmes proposant simplement une liste de solutions [2, 25]. En outre, le filtrage proposé par les systèmes peut devenir un obstacle pour les médecins en formation qui ne peuvent développer leur regard [2]. De plus, l'impossibilité de l'IA d'accéder à l'ensemble du contexte et aux connaissances de l'humain « entrave l'alliance » (ibid, p. 45.). On retrouve ici une limite connue en interaction Humain-Machine dans les situations d'automatisation : l'asymétrie entre l'humain et la machine [40]. L'impossibilité pour la machine de connaître l'ensemble des éléments de contexte de l'humain fait que ce dernier doit pouvoir garder/reprendre le contrôle et que la machine doit lui rendre des comptes sur ses actions et connaissances.
On retrouve avec l'IA les limites inhérentes à toute mise en œuvre d'une automatisation telles que décrites par Bainbridge [3]. En contexte professionnel, l'automatisation peut conduire à une trop grande confiance dans le système et donc à son utilisation dans des contextes inappropriés par les employés et à l'inverse,

à une absence d'usage là où il serait pertinent. Il se peut également que la hiérarchie intègre ces systèmes sans considérer les impacts sur le travail [38].
L'intégration d'un nouvel outil dans une situation de travail constitue une transformation de cette situation. Si les effets de l'IA ou de l'automatisation ont pu être décrits, ceux de l'IAG restent peu connus.

1.1.3 L'intégration de l'IAG en entreprise : des effets à documenter

Le déploiement de l'IAG étant récent, les études de cas sont encore rares. Brynjolfsson, Raymond et Li [7], dans un travail en cours, rapportent le cas du déploiement d'IAG dans un service clients. Dans leur étude, les conseillers d'une entreprise américaine se voient attribuer un assistant virtuel qui analyse les échanges par chat avec les clients et conserve l'historique dans le temps. Le système suggère ainsi des réponses au cours des échanges clients. Les employés choisissent ce qu'ils intègrent ou non dans leurs réponses.
Cette étude offre une vision ancrée dans le déploiement réel d'une solution à base d'IAG à une échelle macro à l'aide d'indicateurs tels que le « temps de gestion moyen », le « taux de résolution », et la « satisfaction des clients ».
Ils décrivent des résultats intéressants à considérer lors du déploiement de tels outils. Ils mettent notamment en évidence que les réponses suggérées par l'IA sont similaires à celles des employés expérimentés et qu'avec le temps le style des novices utilisant le plus l'IA s'en rapprochait. D'autres auteurs rapportent des résultats similaires : les bénéfices en termes de qualité de production et de performance sont plus grands pour des personnes moins qualifiées ou moins compétentes pour une tâche donnée, en comparaison à des experts [12, 36].
Brynjolfsson et al. (ibid) considèrent donc que l'IAG permettait aux salariés d'apprendre la bonne façon de faire. Cependant, cela suggère également que l'IA conduit à une perte de diversité dans les réponses. En effet, plusieurs auteurs soulignent que si les productions textuelles sont jugées de meilleure qualité, la diversité des productions s'appauvrit et les textes convergent [15, 36, 45]. Les effets positifs sur le ressenti client pourraient à terme être négatifs face à la perte de caractère humain, organique et imparfait des réponses des conseillers.

L'IAG semble donc pleine de promesses quant à l'amélioration des productions, le développement des compétences et cela peut se traduire en gains pour les entreprises. La majorité des auteurs s'intéresse aux résultats du travail. Le travail avec une IAG lui n'est pas décrit, on ne sait pas ce que font réellement les personnes avec ces outils. Cependant, comme toute modification de l'environnement de travail, cela implique de la part des salariés des adaptations de leur mode de fonctionnement, et cela ne va pas forcément de soi. Par exemple, Gras Gentiletti [19] montre que des experts métiers se retrouvent à alimenter en contenu les chatbots censés alléger leur charge de travail.
Suh et al. [41] montrent comment l'introduction d'un outil d'IAG dans le processus de composition transformait le travail à réaliser et la façon dont les musiciens l'abordaient. Ils soulignent les transformations du travail de compositeur vers celui d'un producteur de musique ou de conseiller, réalisant du patchwork d'œuvres. En ce sens, les compositeurs se sentaient plus créatifs et engagés dans le travail collaboratif sans l'IA.

Pour les personnes, les bénéfices en termes d'augmentation de la productivité et de la confiance en ses capacités sont précieux. Pour autant, cela est associé à une transformation de leur activité de travail. Ainsi, il est préférable d'avoir un accompagnement relatif à l'utilisation de l'IAG pour en maximiser les bénéfices et maîtriser les risques comme le sentiment de perte de responsabilité ou de diversité des écrits [13, 23, 27, 29].

1.2 L'analyse de l'activité : un regard anthropocentré au service de l'amélioration des situations sociotechniques

Les promesses d'amélioration faites lors de l'intégration de nouveaux dispositifs ne sont pas toujours tenues du fait d'une mauvaise ou non prise en compte du travail quotidien [6, 19]. Ceci s'explique par une approche principalement technique des systèmes conçus pour être utilisés ou pour fonctionner en parallèle de l'activité humaine. Les conceptions techno-centrées [39] peuvent limiter l'appropriation de ces dispositifs et/ou la rendre peu aisée.
L'insertion de nouveaux dispositifs impacte l'ensemble de la situation sociotechnique : leur appropriation est difficile à anticiper et les utilisateurs concernés ont toujours du mal à se représenter leurs besoins et les apports possibles de ces nouveaux outils proposés, et ce, d'autant plus que l'apport de l'IA est façonné par des imaginaires portés par des récits dominants [17, 19]. Ainsi, tant les futurs utilisateurs, les métiers, que les concepteurs peuvent être amenés à envisager des fonctionnalités éloignées des activités réelles des futurs utilisateurs.
De nombreux auteurs appellent à une conception des systèmes à base d'IA centrée sur l'Humain [20, 46, 47] pour pouvoir combiner les intelligences humaines et artificielles [14].

Notre point de vue d'ergonomes sur le travail consiste en premier lieu à distinguer la tâche et l'activité. La première correspondant au but à atteindre dans des conditions déterminées hors contexte et souvent prescrites, alors que la seconde considère tout ce qui est mis en œuvre en situation par la personne pour réaliser une tâche [26]. Cette distinction est fondamentale car elle permet, notamment dans le cas présent, de dépasser une définition de l'usage possible de l'IA en termes de tâches décontextualisées (par exemple, résumer un texte) et transposables indéfiniment sans considération pour leur contexte de réalisation. L'activité, permet une approche systémique du travail où les caractéristiques de la tâche, de son environnement plus large et de la personne conduisent cette dernière à réaliser tout un ensemble d'actions, s'adaptant à la variabilité et la diversité des

situations. Plus simplement, la rédaction d'un rapport n'impliquera pas la même activité pour un juriste ou un data scientist, tant du point de vue du contexte, des compétences ou encore des objectifs. De même, que l'activité d'un salarié sera différente pour une même tâche en fonction d'éléments contextuels. Ainsi, une approche qualitative et située de l'usage des technologies plaçant les processus d'appropriation au centre de la valeur des dispositifs technologiques nous apparaît nécessaire [6, 39].

2 Méthode

2.1 Contexte et démarche

Dans l'optique de déterminer les orientations futures en matière d'outils à base d'IA générative, de nombreuses expérimentations ont démarré courant 2024 au sein d'une grande entreprise française. Notre recherche s'inscrit dans une de ces expérimentations internes.
Trois cents licences d'un outil IAG intégré aux applications bureautiques ont été attribuées à des « expérimentateurs » de différents services et métiers. De juin à octobre, les employés sélectionnés ont pu utiliser l'outil à leur guise. En parallèle, un accompagnement a été mis en place : l'entreprise a proposé un ensemble de documents et ateliers (webinaires) afin d'aider ou de montrer les usages possibles de ce nouveau système. Une liste de « cas d'usages transverses » a été définie. La notion de « cas d'usage » vient de l'ingénierie et du développement logiciel. Les « cas d'usage » correspondent à un usage possible et permettent de préciser les exigences fonctionnelles d'un système et comment les utilisateurs interagissent avec ce système. Ils servent de fil conducteur pour la discussion entre les différentes parties d'un projet de conception [21]. Les cas d'usage prédéfinis pour l'IAG servent de référence pour les ateliers. En effet, ceux-ci reprenaient ces cas d'usage et proposaient logiciel par logiciel les prompts à y associer– . En plus d'avoir une entrée par logiciel et non par finalité, les cas d'usage présentés, bien que proposant parfois un contexte (par exemple « vous venez de recevoir un long document et n'avez pas le temps de le lire avant votre prochaine réunion. »), s'arrêtent à la proposition d'un prompt à utiliser ou un bouton à cliquer. Ainsi, toute la complexité de l'activité et de ses finalités n'est pas envisagée. Un objectif = un prompt. Les « expérimentateurs » sont évidemment invités à affiner les réponses à l'aide de prompts successifs ou d'utilisation des suggestions de l'outil, mais c'est le prompt qui permet d'atteindre l'objectif.

De plus, les « expérimentateurs » étaient invités par les organisateurs de ce test à répondre à deux questionnaires (en août et en octobre), afin de déterminer les gains de productivité offerts par l'outil et ainsi d'orienter les choix d'achat de licences supplémentaires.
Avec une démarche qualitative inscrite dans l'analyse de l'activité, des participants ont été recrutés parmi les 300 « expérimentateurs ». Nous les avons contactés par e-mail ou par la messagerie d'entreprise. Le premier contact visait à présenter la méthode de l'étude et ce qu'ils auraient à faire s'ils ou elles acceptaient de participer.
Neuf employés d'entités variées ont accepté. Comme l'ensemble des expérimentateurs, ceux-ci avaient des profils principalement orientés vers l'innovation et le numérique. De fait, la plupart avaient de bonnes connaissances en matière d'outils informatiques. Trois de nos participants sont informaticiens. (Tableau 1)

Tableau 1 – Participants de l'étude

Employé	Genre	Fonction
e1	F	Référente outils collaboratifs
e2	F	Responsable transition numérique
e3	F	Data analyste en système d'information
e4	M	Responsable numérisation
e5	M	Manageur des processus d'innovation
e6	F	Cheffe de projet RH
e7	M	Chargé de Mission en transition numérique
e8	M	Chargé de Mission opérationnel
e9	M	Manageur en système d'information

2.2 Carnet de bord et entretien

Nous avons réalisé des entretiens de remise en situation à partir de traces de l'activité [32, 42]. Les traces utilisées en entretiens étaient recueillies par les participants eux-mêmes, dans un carnet de bord durant 1 à 2 mois.
Le carnet était transmis au format Word et les participants étaient informés qu'ils pouvaient le compléter directement sur leur ordinateur ou l'imprimer s'ils le souhaitaient. La consigne invitait les participants à tracer des moments d'usages spécifiques où l'IA a particulièrement bien fonctionné ou au contraire, des moments où il fut difficile, voire impossible, d'arriver au résultat escompté.
La première page présentait des informations sur notre recherche et son déroulement. Puis le carnet était composé de plusieurs fiches vierges. Chaque fiche était structurée à l'identique avec 4 blocs : le recto comprenait trois blocs (contexte, déroulement, évaluation) et le verso comprenait un bloc vierge pour y insérer des captures d'écrans ou des photographies. La première fiche était remplie avec un exemple pour aider les participants.
Ils avaient également la possibilité de tracer leurs usages sans utiliser le carnet de bord s'ils le souhaitaient. L'objectif de ces carnets de bord était d'obtenir suffisamment de traces structurées de l'activité de travail des employés lors de l'utilisation de l'outil basé sur l'IAG afin d'approfondir cela en entretien.

Les entretiens étaient organisés en trois parties. La première abordait des questions relatives à l'ancienneté des participants, leur métier, leur poste et leur niveau estimé de compétence/connaissances en informatique et IA. La deuxième partie permettait de comprendre le poste et l'activité de travail globale du ou de la participante et la dernière revenait sur les usages tracés dans le carnet de bord, ainsi que sur des usages supplémentaires non tracés. En plus des carnets de bord, d'autre traces de leur activité étaient utilisées (ex. documents produits, historique d'interaction, etc.).

2.3 Analyses

Tous les entretiens ont été retranscrits dans leur intégralité. Puis nous avons procédé à une analyse thématique des données : à la relecture des transcriptions, des passages ont été sélectionnés et des catégories ont été créées. Pour chaque entretien, les passages sélectionnés étaient soit ajoutés aux catégories préexistantes soit une nouvelle catégorie fut créée.

Du fait de notre centration sur des usages réels par la remise en situation des participants en contexte, certains passages des entretiens relatent des pans détaillés de l'activité des participants. Pour ces passages, nous avons procédé à une seconde analyse centrée sur les objectifs poursuivis par les participants et le déroulement des faits rapportés. Au total, 56 usages, dont la finalité était clairement identifiable, ont été rapportés par les participants. Dans un premier temps nous les avons catégorisés selon leur complexité et leur destination : les usages simples (i.e., l'IA est utilisée pour des tâches spécifiques simples pour atteindre une finalité) ; les usages complexes (i.e., l'employé essaye d'utiliser l'IA pour un cas d'usage métier spécifique, la finalité est plus lointaine et il s'agit pour l'heure de tests) ; les usages pour des collègues (i.e., les utilisations de l'IA dans le cadre de l'expérimentation en entreprise réalisées pour le compte d'autres employés n'ayant pas la licence et donc dissociées de l'activité de la personne utilisant l'outil).
Dans un second temps, nous avons catégorisé les usages en fonction des utilisations de l'IAG. Dix catégories, se rapprochant des cas d'usage proposés par l'entreprise (cf. 2.1.), ont émergé : Produire un document ; Rédiger un courriel ; Générer une image, ;Rechercher des informations Web ; Rechercher des informations internes ; Rechercher des messages ; Synthétiser un document ; Faire un compte rendu de réunion ; Produire du code informatique ; Autre. Puis pour chaque usage rapporté, en plus des catégorisations précédentes, nous avons relevé les objectifs poursuivis par les employés, les détails du contexte et les éventuelles difficultés rencontrées.

3 Résultats

3.1 Les usages réels de l'IAG en entreprise

Notre analyse nous a permis de faire émerger différents usages de l'outil IAG. Les dix catégories ressorties en analyse sont regroupées en quatre supra-catégories : concevoir des documents, rechercher des informations, résumer des informations, autres. Nous allons maintenant détailler les usages réels, en considérant les objectifs poursuivis par les salariés et comment l'IAG est mobilisée différemment en contexte.

3.1.1 Concevoir des documents

Dans l'ensemble des usages rapportés, seize concernaient la conception d'un document. Les personnes vont ainsi utiliser l'IAG pour les aider dans la rédaction de courriels (3), de supports de présentation (5), de tableurs (2), de documents textuels divers (guidelines métier, courrier de réponse à un prestataire, etc.) (3), ou d'images (3).

Concevoir un document est plus complexe que produire un prompt exprimant le résultat attendu. Il est rare que l'utilisateur se contente de demander au système de produire un document à partir d'un prompt seul.
Il s'agit plus souvent de transformer ou restructurer des données ou des documents que l'utilisateur fournit en entrée au système et cela peut s'inscrire à différents moments dans la conception d'un document. En début de conception, l'IAG va être sollicitée pour : proposer une structure de présentation (2) ou de document textuel (1). Au cours du processus il peut s'agir de fournir un document et de demander de modifier le format des informations fournies (e.g., passer d'un PDF à un tableau Excel) (5). Les salariés peuvent également utiliser l'IAG vers la fin de la conception de leur document pour revoir des formulations (1), traduire le texte dans une autre langue (1) ou concevoir des images ou icônes afin d'illustrer le document (2).
Au-delà de ce qui est demandé à l'outil, les finalités poursuivies par les utilisateurs sont nombreuses et ne se limitent pas à la conception dudit document. Ainsi, il peut s'agir d'illustrer une présentation, d'extraire des données pour faciliter leur traitement ultérieur, d'améliorer un processus d'évaluation en produisant des guidelines. Dans certains cas, il peut s'agir de tester les limites du système ou son utilité dans des cas métiers.

De plus, nous identifions des cas complexes dans lesquels la production du document souhaité s'intègre dans une suite d'actions demandées à l'IA, comme la recherche d'informations spécifiques en amont. Cela peut s'inscrire dans l'exécution d'un prompt complexe et des approfondissements successifs. Par exemple une utilisatrice (e1) rapporte comment elle a dans un premier temps recherché de l'information sur les liens entre sédentarité et risques pour la santé. Pour cela, elle a utilisé un prompt structuré pour obtenir le plus de détails possibles suivi de prompts pour approfondir la réponse « *Dans un tout premier temps je n'avais pas de chiffres donc je lui ai demandé les chiffres parce que c'était parlant* ». S'agissant d'un courriel à destination d'anglophones également, elle a aussi demandé une traduction du texte. Par la suite elle a mis en forme le texte français et anglais dans un courriel de communication interne. « *Ensuite, nous on reprend le corps du mail mais après il faut toujours qu'on indique que c'est à destination de tout le monde que c'est de la part de la COM et de l'équipe collaborative, que ça s'adresse aux anglophones et francophones* ».
Cela peut également être réalisé par une série d'actions où différents prompts visent des objectifs intermédiaires différents. Un utilisateur (e8) souhaitait répondre à une question technique concernant une étude de danger. Pour obtenir une réponse, l'utilisateur a réalisé une série d'actions et de prompts : ajouter l'étude de danger sur le serveur « *Le problème, c'est qu'aujourd'hui nous, on travaille toujours sur des bases de, des serveurs, les*

anciens serveurs quoi » ; demander des résumés du document pour s'assurer de sa bonne prise en compte « *Il résume celui-là, celui-là puis après faut construire la réponse* » ; demander la rédaction de la réponse. La réponse n'étant pas adaptée à la posture de l'entreprise dans ce type de cas « *La probabilité de réponse c'est dire "Bah oui, vous avez raison d'étudier ça sous cet angle-là", techniquement, il a raison. Par contre, avec notre posture réglementaire, on n'a pas le droit.* », l'employé, après avoir demandé d'anciennes réponses à fournir à l'outil, doit réitérer les étapes décrites précédemment.

Difficultés
La conception d'un document regroupe différents usages de l'IAG, ainsi différents types de difficultés émergent. Les réponses à des courriels ne prennent pas toujours en compte la relation entre les interlocuteurs et se trouvent être souvent trop formelles vis-à-vis de la situation « *Il te fait un truc, comme s'il parlait au Premier ministre* » (e5). Son ignorance des codes sociaux ou, comme l'a montré le cas rapporté plus haut sur l'étude de danger (e8), des postures de l'entreprise, limite les employés dans leurs usages et leur rajoute des tâches.
Pour ce qui est de la création de supports de présentation, il semble que l'outil ne convainc pas « *Je lui ai demandé de me faire un PowerPoint de mon document Word. Alors il m'a sorti un de 46 slides pour 11 pages* » (e2), ou ne fonctionne tout simplement pas « *je voulais qu'il me fasse un résumé de mon appel d'offres sur 4-6 slides. Il est jamais arrivé à le faire* ».
Concernant les images, la première difficulté consiste à décrire précisément l'image que l'on souhaite. De plus, il apparaît impossible ou du moins difficile de supprimer des éléments de l'image générée « Au début, j'ai mis quelque chose comme "*je voudrais une image avec des fleurs et des oiseaux, mais sans arbre […] Et en fait, dès lors que t'avais le mot arbre dedans, il y était.*" » (e1).

3.1.2 Rechercher des informations

Au total, nous notons vingt-cinq cas de recherche impliquant l'outil à base d'IA générative.
Il est possible de différencier trois types de recherche en fonction de la provenance de l'information : recherche web (9), recherche dans les bases de données entreprise (6), recherche dans les messages (10).
Pour autant, comme nous le verrons, il est nécessaire de préciser que le réel montre une porosité entre ces formes de recherches.

Les usages associés à des recherches web sont larges. Il peut s'agir de trouver des informations sur une personnalité publique ou sur un événement d'actualité (2), d'explorer un thème et ainsi de réaliser des benchmarks (2), de trouver et comprendre des réglementations ou jurisprudences (2), ou encore de trouver du contenu pour préparer un webinaire (1). Nous notons également le cas d'une recherche web sur une question propre à l'entreprise mais introuvable dans la recherche interne. Dernièrement une utilisatrice recherchait une méthode pour résoudre un cas métier pour lequel elle doit développer une solution et a trouvé un article scientifique. Pour ces cas de recherche web, les participants exposent comment l'IAG vient en remplacement d'un moteur de recherche classique, avec l'avantage d'évaluer un ensemble de possibilités et de proposer un résumé de plusieurs sources.

Les recherches à partir de bases de données internes sont de trois types : rechercher des informations dans un document particulier ou une série de documents identiques (2) ; rechercher par curiosité/intérêt un projet spécifique dont l'employé a entendu parler (2) ; effectuer une recherche générique sur un thème ou un sujet donné (2).

Concernant l'utilisation de l'IAG pour retrouver des courriels ou des messages, différentes finalités apparaissent. Les salariés cherchent à comprendre un ensemble d'échanges par email ou messagerie instantanée afin d'évaluer la situation (3) ; à se remémorer les échanges avec une personne spécifique afin de relancer la personne ou préparer une réunion (4) ; à s'informer à leur retour de congé (1) ; ou encore à explorer l'état d'avancement sur un sujet particulier (2) : par exemple un utilisateur a récupéré des échanges avec des partenaires industriels pour aider un nouveau collègue à reprendre le dossier. Les recherches dans les messages ne sont pas toujours de simples recherches de contenu et visent à obtenir une synthèse de plusieurs messages (cf. 3.1.3)

Un cas particulier de la recherche d'informations concerne la programmation informatique. Un salarié (e4) nous rapporte deux cas d'aide à la programmation. Dans un premier, il formule son objectif « confectionner une base graphe sur des données internes » Il a donc décrit ce qu'il souhaite et fourni les documents en entrée pour obtenir des bouts de codes à utiliser. Dans un autre cas, il souhaite trouver comment corriger une erreur dans son code. Il a alors simplement copié-collé l'erreur dans l'outil qui lui a fourni une solution. Pour autant la réponse trop verbeuse de l'outil l'a conduit à penser que l'IA n'avait pas compris sa demande.

Difficultés
Outre le format de réponse de l'outil, le principal problème associé à la recherche d'informations est lié à l'accès aux données. Du fait des systèmes d'informations de l'entreprise et de l'intégration de cet outil IAG, ce dernier n'a pas accès à l'intégralité des bases. « *Les échanges de mails que je cherchais se trouvaient dans la BAL [commune] et non pas dans ma boite de réception à moi et donc c'est vrai que j'ai été un petit peu frustrée* » (e1). Plus grave, l'outil, en plus d'être comme toute IAG, sujet aux hallucinations, ne donne pas toujours les bonnes références documentaires lorsqu'il rapporte les résultats d'une recherche « *Et après le gros problème qu'il y a, c'est les références qu'il me donne. Elles sont pas bonnes ou des fois il me fait des résumés. Il donne du contenu qui est pas lié à la référence qu'il m'a donnée*. » (e4)

3.1.3 Résumer de l'information

Résumer à l'aide de l'IAG est une fonctionnalité particulièrement mise en avant, notamment du fait de sa

capacité de travailler des données textuelles.
Dans l'entreprise, la réalisation de comptes rendus de réunion a largement été poussée et utilisée (6).
Avec le travail à distance et le télétravail, les opportunités pour réaliser des réunions en visioconférence ne manquent pas. Tous nos participants ont ainsi pu tester cet usage. Pour certains, cela est effectué par l'action IAG « Résumer » ou à l'aide d'un prompt demandant une synthèse plus ou moins précise, qui sera affinée par des prompts complémentaires. « *Je fais enregistrement, transcription, puis après je fais 2 - 3 tests. " Qu'est ce qui a été dit ? Fais-moi un résumé comme-ci, un résumé comme ça… "* » (e8)
Pour d'autres, il s'agira de demander à l'IAG de structurer (reformuler) des notes prises manuellement. Dans les deux cas, l'objectif visé est d'obtenir un compte rendu qui sera partageable à des collègues. « *Je vais avoir mon téléphone donc je vais faire des prises de notes. Voilà pendant les réunions avec mon téléphone, donc je note que voilà c'est pas du tout organisé […] c'est pas du tout textuel. Enfin voilà. Et pour le coup, ouais, il me bluffe aussi à pouvoir rédiger de façon assez optimisée quand même, les quelques points de texte. […] Tu sélectionnes ce que tu veux réécrire. [cliquer sur] "[Outil IAG]", [cliquer sur]"Réécrire*"» (e9).

Autrement résumer des documents peut viser à aider à la lecture (3) d'une thèse ou d'une procédure complexe par exemple, ou peut directement participer à la rédaction d'une note en produisant la synthèse à ajouter en début de document (2). On retrouve de l'interdépendance entre les cas d'usages comme nous l'évoquions concernant la recherche dans les messages : les utilisateurs demandent un résumé des courriels reçus, pour avoir un aperçu de ce qu'il y a à traiter au retour de congés (1), pour préparer une réunion sur les sujets des derniers échanges (2), pour être aidé dans la compréhension d'un échange de nombreux courriels et agir en conséquence (2).

Difficultés
Plusieurs limites émergent cependant vis-à-vis de la synthèse d'informations avec l'outil IAG utilisé. Pour ce qui est des comptes-rendus de réunion directement produits par l'outil, il est nécessaire de démarrer l'enregistrement et la transcription. De plus, pour tout type de résumé l'outil est en difficulté pour les textes longs. Ainsi les utilisateurs demandent un résumé intermédiaire pour les longues réunions « *Il galère un peu sur des réunions qui étaient trop longues […] c'est pour ça que je le lançais en cours de route et à la fin je lance le prompt qui avait été suggéré* » (e2) ou découpent leur demande partie par partie pour les documents « *Du coup je segmentais les documents quand j'ai les gros trucs que je voulais*. » (e8). Enfin, les utilisateurs pointent le manque de précision de ces résumés.

3.1.4 D'autres formes d'usages

Certains usages rapportés par les participants n'ont pas pu être classés convenablement dans les catégories précitées. Parmi ces cas se trouvent deux usages complexes de l'IAG. Les employés cherchaient à utiliser l'IAG pour des cas métiers spécifiques pour lesquels cet outil transverse n'a pas été pensé. Une utilisatrice, data scientist concevant des outils à base d'IA commandités par d'autres métiers, a voulu voir comment cet outil pouvait répondre à un cas d'analyse d'images : il s'agissait d'identifier certaines caractéristiques des sols en amont des constructions « *En fait, comme il génère des images, je pensais qu'il pouvait aussi analyser des images. […] on va essayer, on sait jamais.* » (e3). Elle souhaitait ainsi éviter le développement inutile d'une application. Cependant, cet outil n'est pas en mesure d'analyser les images.
Une autre employée travaillant sur les outils numériques à disposition des ressources humaines a, quant à elle, testé la capacité de cet outil IAG à répondre à des questions techniques RH. Pour cela, après avoir extrait et anonymisé de vraies questions, elle a fourni en entrée des documents internes auxquels l'outil n'a pas accès et a posé ses questions en demandant des réponses structurées. « *Alors, ce qui est embêtant, c'est [que l'outil] a pas accès [à ces données]. Et en fait, c'est la Bible pour toutes ces réponses réglementaires.* » (e6). Au-delà de ce problème, une fois les documents récupérés l'outil ne les prenait pas toujours en référence ce qui rendait la tâche difficile.

Avec le nombre limité de licences certains employés se voient attribuer un rôle de testeur de « cas d'usage métier ». Dans ce contexte, en fonction des directions, il leur incombe de tester une liste d'usages définis en comité de pilotage. Dans d'autres, ils sont chargés d'inviter leurs collègues à leur adresser des demandes de tests, c'est le cas de la participante e6.
Ainsi, dans ces tests, le contexte est décorrélé de l'usage, conduisant à des biais dans la façon dont les « expérimentateurs » se saisissent de l'outil. En utilisant l'outil sans connaître les finalités et le contexte propre au métier concerné, certaines informations leur manquent « *Tout ce qui était en rapport a été détecté par l'outil. Par contre, il y a des choses qu'ils rajoutent manuellement, donc qui n'étaient pas dans le rapport que je peux pas connaître*. » (e9), et il est difficile pour eux d'évaluer la qualité et la pertinence des réponses de l'outil *« Non [je n'évalue pas la fiabilité], alors en fait en fait la personne avec qui je travaille, c'est une experte senior qui maîtrise énormément de sujets RH, réglementaires. Donc, tous les sujets qu'elle m'a donné à tester, elle les connaît parfaitement en fait*. » (e6).

Bien que l'utilisation de l'outil IAG pour autrui soit particulière, cela soulève la question du travail collectif avec ce type d'outil. Notre étude n'a permis d'identifier qu'un cas d'utilisation collaborative de l'IAG. Un groupe en charge des outils numériques et collaboratifs d'une unité travaillait à produire un quizz sur la sûreté en entreprise. Pour cela, une personne affichait son écran et l'outil était utilisé collectivement pour résumer des documents internes sur ces questions, puis pour générer des séries de questions sur différents thèmes. « *On était 4 en tout. On était dans une salle de réunion, j'ai demandé à l'une des personnes d'afficher son écran* » (e1). Ce travail était aussi l'occasion de former en contexte des

collègues à l'utilisation de ce nouvel outil.

3.2 Les attentes et le réel des apports de l'IAG en entreprise

La communication interne promettait des gains de temps, de productivité et de créativité, en mettant principalement les deux premiers en avant. Du moins c'est ce que les participants gardent en tête (7 participants sur 9).
Cependant, les gains réels associés à l'utilisation de l'outil, s'il y en a, ne sont pas évidents à quantifier « *Si tu me demandes combien de temps gagné avec ça, je suis incapable de le dire. Matérialiser le temps gagné, c'est impossible.* » (e9). Et ces attentes peuvent conduire à quelques déceptions « *Au début, j'ai essayé de voir en fait, je me suis dit que ça va beaucoup m'aider et que je vais gagner beaucoup de temps et en fait non*. » (e5). Les utilisateurs imputent cet écart notamment à la variabilité des résultats de l'outil qui impose de vérifier les résultats et de s'adapter en conséquence (6/9).

Les gains se trouvent peut-être ailleurs. Il peut s'agir de la qualité du travail par exemple comme le souligne un participant « *Je pense que c'est un gain de qualité. De compréhension peut-être aussi, de facilité de compréhension des personnes avec qui je travaille enfin, et cetera.* » (e9). L'outil offre également des possibilités nouvelles. Ainsi, des participants rapportent une amélioration de la qualité des écrits pour ceux qui ne se considèrent pas comme littéraires, la possibilité d'obtenir des informations plus larges qu'une recherche classique dans ses messages (e.g., des éléments sur le contexte, sur les interlocuteurs, obtenir un résumé), ou encore la découverte d'accès à des données internes utiles.

Pour évaluer la valeur ajoutée de l'outil, les employés approchent certains des usages comme des tests opportunistes (4/9). « *Je me suis prise au jeu de tout tester sur [l'outil] pour voir ce que ça donnait*. » (e2). En effet, plusieurs des usages rapportés sont explicitement des tests de l'IA dès le début et certains deviennent des tests suite à une erreur ou une hallucination pour déterminer les limites de l'outil. « *De mon expérience, c'est plus fiable en termes de restitution de contenu que dans [le logiciel de traitement de texte], ça va très bien rédiger, c'est-à-dire va faire des belles phrases, des voilà des belles articulations dans les phrases, et cetera. Mais le contenu n'est pas forcément très fiable. Et donc je me suis dit que j'allais le faire plutôt dans l'[interface principale]*. » (e6).
Cela permet également de développer la confiance dans le système (2/9) dans un contexte où la responsabilité des productions de l'outil leur est attribuée.
Finalement, les employés, par le test d'usages métiers plus complexes, même en perdant du temps sur le moment, pensent aux gains futurs en termes de temps, de simplicité et de qualité que pourra amener l'outil. Cette expérimentation était ainsi une opportunité de tester l'outil pour apporter des solutions aux différents métiers (7/9).

4 Discussion : des usages réels différents des attendus avec l'IAG

Nos résultats mettent en évidence (1) comment différents usages impliquant un outil IAG en entreprise prennent forment et soutiennent des finalités différentes en fonction des situations ; (2) les apports de l'IAG en contexte professionnel ; (3) comment l'appropriation en entreprise, par une expérimentation, guide les attentes et les usages.

4.1.1 Un usage ne revient pas à exécuter un prompt

Lorsqu'il est question d'usage réel de l'IAG, les entreprises évoquent des « cas d'usage ». Cependant, du fait de leur centration sur l'interaction avec le système, ces cas d'usage ne rendent pas compte des formes diverses que peut prendre une activité finalisée.

Comme nous l'avons vu, la catégorisation des usages de l'IAG basée sur les actions réalisées n'est pas aisée, et parfois poreuse. Bien souvent, l'objectif visé par l'utilisateur implique l'utilisation de l'outil pour différentes tâches intermédiaires. Par exemple, lorsqu'il s'agit de concevoir un document textuel, l'employé peut être amené à effectuer des recherches dans un premier temps sur la base desquelles il est demandé à l'outil de produire une structure de document ou un paragraphe. L'outil peut également être utilisé simplement pour reformuler un passage ou traduire le texte dans une autre langue. La génération d'une réponse par l'IA n'indique pas l'atteinte de l'objectif, il reste à la charge de l'utilisateur un ensemble de tâche à réaliser, notamment du fait de l'asymétrie de la relation Humain-Machine [3, 34]. Par exemple, l'utilisateur peut encore être amené à assembler les différents morceaux qu'il a produit à l'aide de l'outil [35].
Il y a plusieurs intentions derrière une tâche et donc une interconnexion des cas d'usages, cela rompt avec l'idée que les cas d'usage peuvent être pensés indépendamment et indifféremment, en dehors d'une activité.

Cela met en évidence les limites d'une interaction pensée exclusivement par prompt. Le prompting est souvent pensé comme le futur de l'interaction entre Humains et IAG car supposé aisé et naturel puisque renvoyant au langage tel que maîtrisé par les humains. Cependant, prompter n'est pas parler en langage naturel. Les interactions entre humains supposent une collaboration faite de plusieurs échanges qui permettent d'arriver à un modèle mental partagé de la conversation et des intentions de l'autre locuteur, aidé par des indices contextuels variés (non) verbaux, historiquement et socialement construits [10, 18].
Or, dans le cas de l'IAG, bien poser la question suppose d'être capable *a priori* de formuler son besoin, de l'exprimer dans des termes compris par la machine tout en se figurant les possibilités de la technologie, ce qui n'est pas accessible à tous. Le fait que de légères variations dans le prompt, qui seraient non significatives pour un interlocuteur humain, résultent en des changements majeurs dans le comportement du modèle rappelle que les prompts ne sont pas des interfaces de langage naturel. [48]

4.1.2 La valeur ajoutée est protéiforme et reste à évaluer

Les gains associés à l'utilisation de l'IAG ne sont pas toujours évidents à identifier par les employés. Notamment lorsqu'il s'agit du temps gagné (principal gain attendu), ce temps gagné est difficile à estimer. Les employés évoquent alors d'autres types de gains qualitatifs concernant leur travail. Ainsi, nos résultats vont dans le sens de la littérature soulignant l'amélioration de la qualité des productions [12, 36].

Nos résultats soulignent également comment l'outil permet dans certaines situations de faciliter la compréhension de documents, d'ouvrir des accès à des données jusqu'alors inconnus ou encore donner plus de détails qu'initialement demandés lors d'une recherche dans les messages. Ce type d'apport permet de soutenir l'activité de personnes en élargissant leur marge de manœuvre situationnelle – « la possibilité pour l'opérateur, dans une situation précise, d'élaborer un mode opératoire efficient » p.15 [11] – dans leur approche des situations en vue d'atteindre leurs objectifs.

4.1.3 Les usages sont orientés par les attendus de l'entreprise et les discours

Le cadre proposé par Bauchet et al. [4] propose une vision évolutive des questions d'acceptabilité et d'appropriation des technologies numériques en contexte professionnel (plus particulièrement l'éducation). En combinant plusieurs modèles, les auteurs conçoivent l'acceptabilité, l'acceptation, l'adoption et l'appropriation comme les parties d'un processus dans le temps démarrant en amont de l'intégration de l'outil du fait des représentations préexistantes influençant l'acceptabilité. Ainsi, ils soulignent comment les représentations façonnent les attitudes vis-à-vis de l'outil et, dans la suite des usages réels, participent de l'appropriation.

Dans le cas de l'appropriation à l'échelle d'une entreprise, il est possible de considérer comment l'appropriation de l'outil débute par le cadrage par l'entreprise des représentations et des usages comme nous l'avons vu avec les discours, les ateliers et la prise en charge des tests par certaines directions.

Dans une certaine mesure, les représentations dont disposaient les participants en amont de l'expérimentation, mêlées au cadre donné par l'entreprise, ont participé à guider leurs attentes et leurs usages de l'outil.

5 Conclusion

Nielsen [34] estimait que l'IAG, avec les interfaces tchat, amène un nouveau paradigme d'interaction basé sur la spécification du résultat attendu selon le but visé et l'intention de l'utilisateur. Contrairement à l'utilisation de logiciels dans lesquels l'utilisateur dit à l'ordinateur quoi faire, ici il exprime le résultat qu'il souhaite. Cependant, nos résultats soulignent comment les usages réels de l'IAG débordent du concept de cas d'usage notamment parce que ce dernier se centre, du moins dans l'entreprise de notre étude, sur les réponses du système comme finalités. Les finalités sont multiples et l'utilisation du système pour une même finalité peut être faite à différents instants et de différentes façons. L'utilisateur prend en charge ce qu'il faut faire et les étapes intermédiaires. L'IAG n'est appelée que pour une ou plusieurs de ces étapes.

Comme nous l'évoquions en introduction, avec le déploiement de nouveaux outils, l'activité des salariés est transformée. Pour identifier ces transformations et les accompagner, il est nécessaire de préciser comment l'intégration d'un tel outil est faite dans une organisation sociotechnique en place. C'est par une approche centrée humain qu'il est possible de mettre en évidence la richesse du réel pour la prendre en compte dans la conception. Cette étude nous a permis d'identifier de nombreuses limites à l'expérimentation interne, comme le choix des participants (demande motivée par le test d'un cas d'usage métier à forte valeur ajoutée, conduisant notamment aux utilisations pour autrui), le cadrage de l'exploration de l'outil (documentation et ateliers orientés prompt),et le temps limité pour utiliser l'outil et l'intégrer ou non à ses activités (relativement court avec une grande période de vacances).

Nous proposons trois préconisations pour la mise en place de ce type d'expérimentations ou pour le déploiement d'IAG en entreprise :

- Une approche située de l'acceptation et de l'appropriation suppose de ne pas étudier les outils de façon isolée et décontextualisée. Il faut mettre l'outil à l'épreuve de son contexte d'usage réel et dans le temps pour évaluer concrètement les apports et limites vis-à-vis de l'activité des employés [5]. Il serait pertinent de mener des études centrées sur les usages réels au long cours, en limitant par exemple le déploiement de l'outil à quelques équipes pour identifier les transformations du travail individuel et collectif.
- Une expérimentation suppose la mise en situation d'usages réels par les utilisateurs finaux. L'accompagnement des salariés devrait être, non pas sur les prompts à utiliser dans un premier temps, mais proposer une approche plus systémique en mettant en avant les liens entre les différentes possibilités offertes par l'outil. C'est-à-dire mettre évidence comment les salariés peuvent utiliser l'outil pour différentes tâches et dans différents logiciels afin d'atteindre un but spécifique. Par exemple, comme nous l'avons vu, pour rédiger une newsletter, il est possible d'utiliser l'outil pour réaliser une recherche sur le web, les informations trouvées peuvent ensuite être résumées et retravaillées pour constituer le corps du texte. Puis l'outil peut servir à obtenir une traduction. En parallèle de cela, la salariée vérifie les informations obtenues, redemande des précisions si nécessaire, vérifie la traduction et mets en forme le texte avant l'envoi. L'ensemble de ces actions visant un même objectif.

  De plus, une formation préalable n'est pas suffisante. Il semble important de mettre en place un

accompagnement sous forme d'aide en cours d'usage. Des experts, pourraient ainsi aider les salariés à atteindre des finalités précises et ainsi penser l'activité de façon globale et plus opérationnelle.

- Les modalités d'interaction avec l'IAG méritent d'être repensées du point de vue de l'humain, pour une réelle interaction Humain-IAG. Morris [33] appelle à dépasser le prompt comme seul moyen d'interaction avec les IAG pour que celles-ci deviennent utiles, utilisables et sûres pour les utilisateurs finaux. Les menus et formalismes constituent des propositions saillantes (i.e., des affordances, [35]) qui montrent les possibilités à l'utilisateur et facilitent l'apprentissage dans le temps. Les futures interfaces d'IAG devraient intégrer des stratégies d'accompagnement adaptatif. Il est erroné de penser que la question de l'interaction disparaîtra avec l'émergence d'une intelligence artificielle générale de grande capacité : plus les modèles auront de capacités, plus les composants de l'interaction auront d'importance. Former les utilisateurs au prompting n'est pas la solution et démontre l'absence de naturel de l'interaction avec la machine. De plus, des aides contextuelles au prompt proposées par la machine ne semblent pas être suffisantes [9]. Les systèmes devraient proposer un échange, une réelle interaction avec les utilisateurs, pour les aider à apprendre, à affiner leurs choix, et pour leur faire savoir ce dont le système a besoin pour fournir de meilleures réponses [34].

Références

[1] D. Acemoglu, & P. Restrepo. Secular stagnation? The effect of aging on economic growth in the age of automation. *American Economic Review*, 107(5), 174-179, 2017

[2] G. Anichini & B. Geffroy. L'intelligence artificielle à l'épreuve des savoirs tacites. Analyse des pratiques d'utilisation d'un outil d'aide à la détection en radiologie. *Sciences sociales et santé*, 39(2), 43-69,2021.

[3] L. Bainbridge. Ironies of automation. *Automatica*, 16(9), 775-779, 1983.

[4] C. Bauchet, B. Hubert, & J. Dinet. Entre acceptabilité et appropriation des outils numériques intégrés dans le système éducatif : Le modèle des 4A. *13ème colloque international RIPSYDEVE*, pp. 158-161, 2020.

[5] M. E. Bobillier Chaumon. L'acceptation située des technologies dans et par l'activité : premiers étayages pour une clinique de l'usage. *Psychologie du Travail et des Organisations*, 22(1), 4-21, 2016.

[6] M. E. Bobillier Chaumon. Du rôle des TIC dans la transformation digitale de l'activité et de la santé au travail. Mieux travailler à l'ère du numérique : définir les enjeux et soutenir l'action. *La revue des conditions de travail ANACT 6, pp. 16-24*, 2017.

[7] E. Brynjolfsson, D. Li, & L. R. Raymond. Generative AI at work (No. w31161). *National Bureau of Economic Research*, 2023.

[8] E. van den Broek, A. Sergeeva, & M. Huysman. When the Machine Meets the Expert: An Ethnography of Developing AI for Hiring. *MIS quarterly*, 45(3), 2021.

[9] C. Chen, S. Lee. E. Jan, & S. S. Sundar. Is Your Prompt Detailed Enough? Exploring the Effects of Prompt Coaching on Users' Perceptions, Engagement, and Trust in Text-to-Image Generative AI Tools. *Proceedings of the Second International Symposium on Trustworthy Autonomous Systems*, pp. 1-12, 2024.

[10] H. H, Clark, & S. E., Brennan. Grounding in communication. In L. B. Resnick, J. M. Levine, & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 127–149). American Psychological Association, 1991

[11] F. Coutarel, S. Caroly, N. Vézina, & F. Daniellou. Marge de manœuvre situationnelle et pouvoir d'agir : des concepts à l'intervention ergonomique. *Le travail humain*, 78(1), 9-29, 2025.

[12] J. H. Choi, & D. Schwarcz. AI Assistance in Legal Analysis: An empirical study. *SSRN*, 4539836, 2023

[13] T. K. Chiu, B. L. Moorhouse, C. S. Chai, & M. Ismailov. Teacher support and student motivation to learn with Artificial Intelligence (AI) based chatbot. *Interactive Learning Environments*, 32(7), 3240-3256, 2024.

[14] D. Dellermann, P. Ebel, M. Söllner, & J. M. Leimeister. Hybrid intelligence. *Business & Information Systems Engineering*, 61(5), 637-643, 2019.

[15] A. R. Dosh, & O.P. Hauser. Generative artificial intelligence enhances creativity but reduces the diversity of novel content. *SSRN*, 4535536, 2023.

[16] C. B. Frey, & M. A. Osborne. The future of employment: How susceptible are jobs to computerisation?. *Technological forecasting and social change*, 114, 254-280, 2017

[17] T. Gamkrelidze, F. Barcellini, & M. Zouinar. Intelligence Artificielle dans les activités professionnelles: quelles visions des acteurs concernés ? *In 55e congrès de la SELF. L'activité et ses frontières–Penser et agir sur les transformations de nos sociétés*, 2021.

[18] S. Garrod, & M. J. Pickering. Joint action, interactive alignment, and dialog. *Topics in Cognitive Science*, 1(2), 292-304, 2009.

[19] M. Gras Gentiletti. *Soutenir la dimension constructive de l'activité instrumentée par des dispositifs techniques à base d'intelligence artificielle*. Thèse : Université Paris 8. 2022

[20] K. Inkpen, S. Chancellor, M. De Choudhury, M. L. Veale, & E. P. S. Baumer. (2019). Where is the Human? Bridging the Gap Between AI and HCI. *CHI'19 Extended Abstracts*, 2019.

[21] I. Jacobson, M. Christerson, P. Jonsson, G. Overgard, Object-Oriented software engineering: A use case driven approach, Addison Wesley Professional, 1992.

[22] T. Kim, H. Jo, Y. Yhee, & C. Koo, C. Robots, artificial intelligence, and service automation (RAISA) in hospitality: sentiment analysis of YouTube streaming data. *Electronic Markets*, 32(1), 259-275, 2022.

[23] Y. Kotturi, A. Anderson, G. Ford, M. Skirpan, & J. P. Bigham. Deconstructing the veneer of simplicity: Co-designing introductory generative AI workshops with local entrepreneurs. *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems,* pp. 1-16), 2024.

[24] T. Kraljic & M. Lahav. From Prompt Engineering to Collaborating: A Human-Centered Approach to AI Interfaces. *Interactions*, Volume 31, Number 3, Pages 30-35, 2024.S. Lebovitz, H. Lifshitz-Assaf, & N. Levina. To engage or not to engage with AI for critical judgments: How professionals deal with opacity when using AI for medical diagnosis. *Organization science*, 33(1), 126-148, 2022.

[25] J. Leplat. L'analyse psychologique du travail. *European review of applied psychology*, 54(2), 101-108, 2004.

[26] N. Li, H. Zhou, W. Deng, J. Liu, F. Liu, & K. Mikel-Hong. When Advanced AI Isn't Enough: Human Factors as Drivers of Success in Generative AI-Human Collaborations, *SSRN* 4738829, 2024.

[27] A. Malik, P. Budhwar, C. Patel, & N. R. Srikanth, N. R. May the bots be with you! Delivering HR cost-effectiveness and individualised employee experiences in an MNE. *Artificial Intelligence and International HRM*, pp. 83-113, Routledge, 2023.

[28] A. Mahdavi Goloujeh, A. Sullivan, & B. Magerko. Is It AI or Is It Me? Understanding Users' Prompt Journey with Text-to-Image Generative AI Tools. *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems,* pp. 1-13, 2024

[29] D. Méda. Sens et avenir du travail en Europe. *The Future of Work*, 29, 2016.

[30] A. Meijer, L. Lorenz, & M. Wessels. Algorithmization of Bureaucratic Organizations. *Public administration Review*, 81(5), 837 - 846, 2021.

[31] V. Mollo, & P. Falzon. Auto-and allo-confrontation as tools for reflective activities. *Applied ergonomics*, 35(6), 531-540, 2004.

[32] M. R. Morris. Prompting considered harmful. *Communications of the ACM*, 67(12), 28-30, 2024.

[33] J. Nielsen. AI Is First New UI Paradigm in 60 Years. 2023. Récupéré à : https://www.uxtigers.com/post/ai-new-ui-paradigm

[34] D. A. Norman. Affordance, conventions, and design. *interactions*, 6(3), 38-43, 1999.

[35] S. Noy, & W Zhang. Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654), 187-192, 2023.

[36] F. Pemer. Enacting professional service work in times of digitalization and potential disruption. *Journal of Service Research*, 24(2), 249-268, 2021.

[37] R. Parasuraman, & V. Riley. Humans and automation: Use, misuse, disuse, abuse. *Human factors*, 39(2), 230-253, 1997

[38] P. Rabardel. *Les hommes et les technologies; approche cognitive des instruments contemporains*. Armand Colin, 1995.

[39] L. A. Suchman. *Plans and situated actions: The problem of human-machine communication.* Cambridge university press, 1987.

[40] M. Suh, E. Youngblom, M. Terry, & C. J. Cai. AI as social glue: uncovering the roles of deep generative AI during social music composition. *Proceedings of the 2021 CHI conference on human factors in computing systems*, pp. 1-11, 2021

[41] J. Theureau. Les entretiens d'autoconfrontation et de remise en situation par les traces matérielles et le programme de recherche « cours d'action ». *Revue d'anthropologie des connaissances*, 4(2), 2010.

[42] S. Tong, N. Jia, X. Luo, & Z. Fang. The Janus face of artificial intelligence feedback: Deployment versus disclosure effects on employee performance. *Strategic Management Journal*, 42(9), 1600-1631, 2021.

[43] C. Trocin, I. V. Hovland, P. Mikalef,& C. Dremel. How Artificial Intelligence affords digital innovation: A cross-case analysis of Scandinavian companies. *Technological Forecasting and Social Change*, 173, 121081, 2021.

[44] M. Torricelli, M. Martino, A. Baronchelli, & L. M. Aiello. The role of interface design on prompt-mediated creativity in Generative AI. *Proceedings of the 16th ACM Web Science Conference*, pp. 235-240, 2024.

[45] W. Xu. Toward human-centered AI: A perspective from human-computer interaction. *Interactions*, 26(4), 42-46, 2019.

[46] Q. Yang, A. Steinfeld, C. Rosé, & J. Ztimmerman. Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design. *Proceedings of the 2020 CHI conference on human factors in computing systems*, pp. 1-13, 2020.

[47] J. D. Zamfirescu-Pereira, R. Y., Wong, B. Hartmann, & Q. Yang. Why Johnny can't prompt: how non-AI experts try (and fail) to design LLM prompts. *Proceedings of the 2023 CHI conference on human factors in computing systems*, pp. 1-21, 2023.

Des fonctionnalités à coût maîtrisé ? Le modèle A-U appliqué à l'IA générative.

Robert Viseur[1]

[1] UMONS, FWEG, Service TIC

robert.viseur@umons.ac.be

Résumé

L'essor des IA génératives (IAG) donne lieu à des projections alarmistes quant à leur impact environnemental. Ce risque est-il correctement évalué ? L'innovation tend à suivre un cycle où les efforts se concentrent sur le produit puis son processus de production et enfin sa simplification (modèle A-U). Ce cycle s'applique-t-il également aux IAG ? L'analyse des premières IA basées sur le deep learning montre le potentiel lié aux optimisations matérielles et logicielles. Ce type d'optimisation est-il mis en œuvre dans le cas des IAG ? Les mesures d'optimisation peuvent porter sur le matériel mais aussi sur les données ou sur les modèles eux-mêmes. Compte tenu de la forte pression sur les coûts, la tendance actuelle pencherait davantage vers une croissance maîtrisée des coûts financiers et environnementaux, du fait notamment des économies d'échelle. La recherche met en avant l'importance de l'optimisation des coûts d'inférence pour l'atteinte de la rentabilité des grands chatbots internationaux. Une perspective se dégage par ailleurs en matière de développement collaboratif des IAG sur le principe des fondations open-sources.

Mots-clés

Intelligence artificielle générative, grand modèle de langage, sobriété numérique.

Abstract

The rise of generative AI (GAI) has led to alarmist projections regarding its environmental impact. Is this risk being properly assessed? Innovation tends to follow a cycle where efforts initially focus on the product, then on its production process, and finally on its simplification (A-U model). Does this cycle also apply to GAI? An analysis of early deep learning-based AI systems highlights the potential of hardware and software optimisations. Are such optimisations being implemented in the case of GAI? Optimisation measures can target hardware but also data or the models themselves. Given the strong cost pressures, the current trend seems to favour a controlled growth in financial and environmental costs, particularly due to economies of scale. Research highlights the importance of optimising inference costs to achieve profitability for major international chatbots. Another emerging perspective concerns the collaborative development of GAI, based on the principles of open-source foundations.

Keywords

Generative artificial intelligence, large language model, digital sobriety.

1 Contexte

Le développement de l'intelligence artificielle (IA) suscite des inquiétudes quant à son impact environnemental. Ainsi, Sundberg (2023) soutient que l'IA présente « *une empreinte carbone en croissance rapide* », liée à la consommation d'énergie due à son exploitation mais aussi à la fabrication du matériel. Ces alertes proviennent aussi de deux entités pourtant souvent antagonistes : les associations écologistes et les *bigtechs*. D'une part, dans un rapport publié en mars 2024, le CAAD (2024) s'alarmait d'un doublement possible du nombre de centres de données occasionnant « *une augmentation de 80 % des émissions globales de CO_2* » pour ces infrastructures. D'autre part, la production d'énergie décarbonée, notamment nucléaire (Jeans, 2025), ressort comme une préoccupation importante des *bigtechs* (IEA, 2025a). Côté science, les inquiétudes quant aux impacts environnementaux ont conduit au développement de recherches sur les IA durables dont l'objectif est « *de développer des outils d'IA plus frugaux* » (Vuarin et al., 2023). Sur quels constats chiffrés ces inquiétudes sont-elles basées ?

Concrètement, l'entraînement du grand modèle de langage (LLM, *Large Langage Model*) GPT-3 aurait nécessité 1,3 GWh, soit la consommation annuelle moyenne de 120 foyers étasuniens, occasionnant l'émission de 522 tonnes de CO_2 (Sundberg, 2023). Par ailleurs, une seule requête ChatGPT générerait « *100 fois plus de carbone qu'une recherche Google classique* » (Sundberg, 2023). Quant à BLOOM, l'énergie consommée pour son entraînement est estimée à 433 MWh (Luccioni et al., 2023). Les performances à un instant donné d'une technologie peuvent-elles valablement être projetées pour en analyser l'évolution future ?

Cette comparaison entre ChatGPT et Google pourrait en effet erronément laisser penser que cette proportion est immuable. En réalité, l'innovation technologique suit traditionnellement un cycle de vie industriel. Décrit par William J. Abernathy et James M. Utterback, il a été baptisé « *modèle A-U* ». Les auteurs distinguent trois phases (Roth, 2016 ; Trott, 2021). La première phase voit

l'innovation stimulée par les besoins. Elle porte dès lors essentiellement sur le produit jusqu'à ce qu'une classe de produit obtienne la reconnaissance du marché. Ce « *design dominant* » va accélérer l'innovation de processus. L'objectif devient alors moins d'améliorer les performances de la technologie que de réduire les coûts de production ou d'exploitation. La maturité s'accompagne progressivement de sa simplification et, dès lors, d'une réduction importante des prix. Le produit devient petit à petit une commodité tandis que l'innovation de produit peut se relancer sur de nouvelles niches de marché. Peut-on appliquer ce cycle aux applications d'intelligence artificielle ?

2 Gains d'efficacité en IA

Patterson et ses co-auteurs (2022) critiquent en tout cas la surestimation des émissions dans les prévisions du fait, d'une part, de la propagation de données erronées fournies dans des études fréquemment citées, d'autre part, l'ignorance des améliorations en continu des systèmes d'apprentissage logiciel. Ainsi, l'optimisation de l'exploitation des modèles conduirait à des gains constatés de l'ordre de 100 pour la consommation et de 1000 pour les émissions (du fait de la décarbonatation de l'approvisionnement en énergie chez les gestionnaires d'infrastructures). Sont dès lors distingués quatre points sur lesquels agir pour réduire les émissions de CO_2 : (1) le modèle en lui-même, (2) le matériel permettant son exécution, (3) le centre de données centralisant le matériel et (4) l'utilisation d'énergie décarbonée par ce dernier (Patterson et al., 2022). Les auteurs pointent ainsi que le PUE[1] moyen des centres de données industriels est de 1,58 contre 1,1 environ pour les fournisseurs d'infrastructures publiques de *cloud computing*. Au final, la consommation des centres de données apparaît globalement sous contrôle (croissance modérée), excepté dans certains pays comme l'Irlande (IEA, 2025b), et ce, malgré un contexte d'explosion du trafic (IEA, 2025a). Ces estimations optimistes s'appliquent-elles également aux intelligences artificielles génératives et, plus particulièrement, aux grands modèles de langage (LLM) ?

Plusieurs recherches tendent à montrer que les gains observés sur la durée en apprentissage profond s'observent également avec les grands modèles de langage. Patterson et ses co-auteurs (2022) fournissent ainsi l'exemple de GLaM, un modèle apparu 18 mois après GPT-3, dont la consommation énergétique a été réduite par un facteur 3. Ho et ses co-auteurs (2024) ont réalisé une évaluation sur plus de 400 LLM. Les chercheurs ont montré que les modèles voyaient leurs besoins en ressources de calcul réduits d'un facteur 2 tous les 8 mois environ. IEA (2025a) fournit la répartition de la consommation d'énergie par les intelligences artificielles. Jusqu'à 10 % concerne le développement du modèle (expérimentation), entre 20 et 40 %, son entraînement ainsi que de 60 à 70 % pour l'inférence c'est-à-dire l'utilisation en production du modèle. L'IAG apparaît donc ici comme une industrie propice aux économies d'échelle. La consommation liée à l'entraînement reste inchangée quelle que soit la base d'utilisateurs tandis que la centralisation des infrastructures permet un abaissement du coût unitaire associé à chaque requête. Des optimisations permettent donc, d'une part, une réduction des coûts d'entraînement, d'autre part, une réduction des coûts d'inférence. Cependant, ne cachent-elles pas une évolution de la consommation globale des infrastructures artificielles génératives compte tenu d'un possible « *effet rebond* » ?

3 Question de l'effet rebond

L'effet rebond correspond au fait que « *l'accroissement des consommations de matières et d'énergie induit par l'utilisation généralisée des TIC efface largement les réductions de l'empreinte écologique obtenues par unité de produit* » (Flipo & Gossart, 2009). Dit autrement le développement des usages feraient plus que compenser les gains dus à l'optimisation.

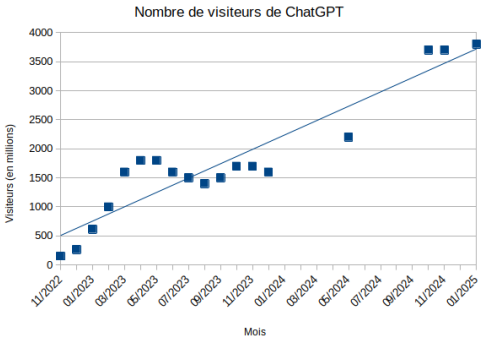


Figure 1. Nombre de visiteurs de ChatGPT (source : Similarweb).

Deux éléments sont susceptibles de contribuer à cet effet rebond : la croissance de la base d'utilisateurs et celle de la complexité des tâches déléguées à l'IA. Premièrement, la diffusion de la technologie d'intelligence artificielle générative conduit à un accroissement de la base d'utilisateurs. Cependant, force est de constater qu'après une croissance rapide en novembre 2022, pour atteindre 100 millions d'utilisateurs en deux mois (Hu, 2023), l'augmentation du nombre d'utilisateurs chez ChatGPT s'est ensuite déroulée de manière sensiblement plus lente (cf. Figure 1). Ce ralentissement[2] pourrait s'expliquer par plusieurs facteurs. D'une part, les agents conversationnels demeurent des outils pour adopteurs précoces (cf. Trott,

1 Le PUE, ou *Power Usage Effectiveness*, est une mesure d'efficacité définie comme le ratio entre l'énergie consommée par tout le centre de données et celle consommée par les seuls équipements informatiques (Patterson et al., 2022).

2 Ces statistiques, d'une part, ne couvrent pas l'usage des LLM au travers des API, d'autre part, négligent la croissance éventuellement captée, début 2025, par des *chatbots* concurrents (p. ex. Mistral et DeepSeek).

2021). En effet, ils sont majoritairement utilisés par les plus jeunes (Bianchi & Angulo, 2024 ; Ma et al., 2024). De plus, et malgré l'apparente convivialité des outils, les usages avancés supposent l'acquisition d'une expertise (Ma et al., 2024), notamment en promptologie (Yao et al., 2024). La diffusion à la majorité des utilisateurs potentiels est dès lors plus lente. D'autre part, le développement commercial de ces outils a été émaillé d'indisponibilités dues à la lourdeur des premiers modèles. Le manque de capacités de calcul a ainsi conduit à la mise en pause des inscriptions à la version payante ChatGPT Plus en novembre 2023[3]. Deuxièmement, les usages s'orientent progressivement vers des tâches complexes nécessitant des modèles capables d'exécuter des raisonnements (p. ex. OpenAI o1) ou d'automatiser le traitement de davantage de documents (p. ex. fonctions de recherche approfondie *i.e.* « Deep Search »). Samsi et ses co-auteurs (2023) montrent ainsi que l'utilisation de modèles de plus grande taille occasionne des coûts énergétiques accrus lors de l'inférence. Les modèles les plus simples, moins consommateurs, se trouveraient dès lors progressivement remplacés par des modèles aux capacités accrues, plus consommateurs. Qu'en est-il réellement ? Quelles motivations, quelles pratiques sous-tendraient des gains d'efficacité dans le cas des LLM ?

4 Optimisation des LLM

Sur le plan des motivations, trois éléments poussent à l'amélioration de l'efficacité énergétique des LLM : les coûts de production (entraînement), les coûts d'exploitation (inférence) et les tensions d'approvisionnement des GPU (Oremus, 2023 ; Isaac & Griffith, 2024 ; Stokel-Walker, 2024). Les plans tarifaires, qu'il s'agisse de ChatGPT Plus, à 20 dollars par mois, ou ChatGPT Pro, à 200 dollars par mois, peinent d'ailleurs à être rentables (Quiroz-Gutierrez, 2025 ; Oremus, 2023). C'est ce qui explique, par exemple, que les premiers utilisateurs payants de GPT-4 « *pouvaient envoyer 25 requêtes seulement toutes les 3 heures car il était trop coûteux à exécuter* » (Oremus, 2023). Les producteurs d'IAG sont dès lors pris en tenaille entre, d'une part, l'importance des coûts et la confrontation à une limite physique (devoir faire avec un stock limité de puces dans un contexte de forte demande ; Bradshaw & Morris, 2024), et, d'autre part, la croissance limitée des revenus. Celle-ci est liée aux tarifs modestes des abonnements (*chatbots*) et du paiement à l'usage (API), maintenus sous pression par la concurrence (OpenAI, Gemini, Microsoft, Mistral, Claude, DeepSeek-AI…). De plus, la recherche de rentabilité encourage les innovations de processus. En quoi ces dernières consistent-elles ?

Sur le plan des pratiques, et pour analyser les opportunités d'optimisation des LLM, deux axes vont être retenus, d'une part, les quatre dimensions identifiées par Patterson et ses co-auteurs (2022), à savoir le modèle, le matériel, les centres de données et les sources d'énergie, et, d'autre part, les trois étapes identifiées par l'IEA, à savoir le développement, l'entraînement et l'inférence. Parmi les quatre dimensions précitées, nous allons négliger les centres de données et les sources d'énergie, car elles relèvent de politiques d'optimisation plus générales au secteur du numérique.

Sur le plan du matériel, l'entraînement reste tributaire de la disponibilité (Smith, 2024), et des progrès, des GPU (*Graphics Processing Unit*). Par contre, l'inférence s'accommode de puces spécifiques telles que, chez la société Groq, les LPU (*Language Processing Unit*) . Ces derniers se distinguent par un meilleur temps de réponse et une efficacité accrue (Ward-Foxton, 2023). Par ailleurs, les gestionnaires de grands centres de données (*hyperscalers*) tels que Amazon, Google, META et Microsoft travaillent sur leurs propres processeurs dédiés à l'IA (Smith, 2024).

Sur le plan des modèles, les optimisations sont à la fois organisationnelles et techniques. Premièrement, la collaboration permet le partage des coûts de développement puis, surtout, d'entraînement. Dans le premier cas, les jeux de données et les modèles peuvent être mis en commun au sein d'un consortium ou d'une communauté. Dans le second cas, le modèle des fondations, propres aux logiciels libres, pourrait servir d'exemple (Viseur, 2024). Deuxièmement, les méthodes d'entraînement peuvent être améliorées (p. ex. *early stopping*, *sparse training* et *gradient accumulation*). Troisièmement, les modèles en eux-mêmes peuvent être optimisés. Cela passe par la réduction du poids des nœuds au sein du réseau de neurones, le principe du MoE (*Mixture of Experts*) et la réduction de la taille des modèles (Gent, 2023). Eldan et Li (2023) ont ainsi montré qu'il était possible de développer un SLM (*Small Langage Model*), à l'image de TinyStories, capable de rivaliser, sur le plan des capacités de génération de texte, avec un LLM (GPT-2), grâce à un effort sur la conception du jeu de données. Des modèles moins lourds (poids réduits), exécutés partiellement (MoE), plus petits (SLM) nécessitent moins de ressources de calcul et voient donc leur empreinte environnementale réduite. Quelles sont les incitations économiques à mettre en œuvre ces différentes méthodes ?

5 Rationalité économique

Afin d'illustrer notre réflexion, nous traitons le cas suivant. En prenant des informations publiées sur le coût d'entraînement et le coût d'inférence par mille *tokens* de GPT-4[4], ainsi que d'autres hypothèses incluant un nombre mensuel de requêtes par utilisateur de 100[5], une taille

3 Cf. https://x.com/sama/status/1724626002595471740.

4 Cf. https://patmcguinness.substack.com/p/gpt-4-details-revealed.

5 Ce nombre de 100 *prompts* par utilisateur en moyenne est estimé sur base des chiffres évoqués fin 2024 par Sam Altman, et donnant le nombre quotidien de messages et le nombre hebdomadaire d'utilisateurs

moyenne de requête de 250 *tokens*[6] et un taux de conversion de 1 % au modèle *freemium*[7], nous calculons, pour un nombre croissant de *prompts*, le nombre d'utilisateurs, une estimation du prix de revient au *prompt* et le surplus (chiffre d'affaires moins coûts d'entraînement moins coûts d'inférence). Cette évaluation très simplifiée (cf. Tableau 1) permet de dresser quelques conclusions intéressantes quant à la rationalité de certains choix.

Coût total d'entraînement :	$63.000.000
Durée de vie d'un modèle (mois) :	12
Coût d'inférence par 1000 *tokens* :	$0,002
Taille moyenne d'une requête (*tokens*) :	250
Coût d'inférence par *prompt* :	$0,0010
Prompts mensuels par utilisateur :	100
Taux de conversion (Plus) :	5 %
Revenus par utilisateur (Plus) :	$20

Prompts	**Utilisateurs**	**Prix de revient**	**Surplus**
1.000.000	10.000	$5,2510	-$5.241.000
100.000.000	1.000.000	$0,0535	-$4.350.000
583.333.333	5.833.333	$0,1090	$0
5.000.000.000	50.000.000	$0,0021	$39.750.000
50.000.000.000	500.000.000	$0,0011	$444.750.000
500.000.000.000	5.000.000.000	$0,0010	$4.494.750.000

Tableau 1. Simulation de prix de revient.

Premièrement, nous constatons que le coût total d'inférence devient rapidement, pour les prestataires internationaux devant satisfaire plusieurs centaines de millions d'utilisateurs, d'un ordre de grandeur comparable au coût d'entraînement (ramené à un coût fixe mensuel compte tenu de la durée de vie moyenne d'un modèle). Le modèle des fondations propre au logiciel libre peut donc se révéler pertinent pour le partage des coûts d'entraînement, en particulier pour les organisations de taille moyenne, devant gérer mensuellement quelques millions de *prompts* au maximum. Deuxièmement, la rentabilité des agents conversationnels n'est possible que pour des prestataires de grande taille. De fait, un surplus n'est dégagé qu'au-delà des 6 millions d'utilisateurs mensuels. Troisièmement, nous constatons qu'un prestataire dominant tel qu'OpenAI présente un problème plus global de modèle d'affaires. En effet, ce surplus reste éloigné d'un bénéfice d'exploitation. D'une part, il n'intègre pas d'autres coûts incluant par exemple les salaires. D'autre part, il est calculé en négligeant les coûts supplémentaires, liés aux usages avancés (p. ex. chaînes de raisonnement, recherche approfondie, multimodalité et agents) ainsi que pour l'entraînement et l'inférence d'autres modèles inclus dans l'abonnement. Cela concerne notamment DALL-E et Sora, respectivement pour les images et les vidéos. Luccioni et ses co-auteurs (2024) montrent ainsi que le coût d'inférence pour une génération d'images est en moyenne 61 fois plus important que pour du texte. Notre simulation permet donc de tirer des conclusions utiles à l'atteinte de la rentabilité. Tout d'abord, compte tenu de l'importance des coûts fixes (p. ex. coûts d'entraînement), des bases installées importantes sont nécessaires. Au-delà des 50 millions d'utilisateurs, le coût d'entraînement (mensualisé) devient en effet inférieur au coût d'inférence. De plus, les opportunités d'économies d'échelle poussent à la consolidation du secteur. Ensuite, le modèle *freemium* montre ses limites. La concurrence empêche de réduire les prestations de la version gratuite (ce qui met les taux de conversion sous pression) et d'augmenter les coûts d'abonnement. Augmenter les revenus se révèle difficile sauf à faire évoluer le modèle d'affaires. C'est notamment ce que fait Microsoft, dont l'IA est désormais financée au travers des abonnements aux logiciels de productivité (Crider, 2025). Sur ce plan, OpenAI pourrait par exemple développer une offre de liens sponsorisés adaptée à l'usage de ChatGPT comme moteur de recherche. Enfin, l'abaissement des coûts unitaires d'inférence impacte directement les coûts compte tenu du volume de requêtes exécuté sur les agents conversationnels internationaux. Quant à celui des coûts d'entraînement (fixes), il conduit à une réduction mécanique du prix de revient d'une requête. L'optimisation, tant de l'entraînement que de l'inférence, se révèle donc une nécessité, voire un impératif de survie, pouvant s'appuyer, d'une part, sur des innovations matérielles (p. ex. nouvelles puces), d'autre part, sur des innovations logicielles (p. ex. nouvelles stratégies d'entraînement et réduction du poids des modèles utilisés).

6 Cas de bonnes pratiques

Deux modèles se sont distingués début 2025 pour leurs pratiques propices à davantage de frugalité : LUCIE et DeepSeek-R1.

Le LLM LUCIE, développé par OpenLLM France et l'ESN Linagora, met en œuvre certaines des bonnes pratiques organisationnelles identifiées. Premièrement, le projet s'appuie sur une infrastructure partagée, le supercalculateur Jean Zay. Lucie-7B a ainsi été entraîné sur 512 GPU NVIDIA H100 pour un total d'environ 550.000 heures de calcul[8]. Cette infrastructure permet

respectivement à 1 milliard et à 300 millions ; cf. https://www.linkedin.com/posts/rowancheung_sam-altman-just-dropped-some-new-chatgpt-activity-7270172267223904257-lAfL/.

6 Cette valeur inclut les *tokens* donnés en entrée (*prompt*) et ceux inclus dans la réponse.

7 Cette valeur de 5 % a été retenue sur base du nombre de clients (soit environ 5 millions), extrapolables du chiffre d'affaires (2023) en rythme annuel (1,3 milliards), rapporté au nombre d'utilisateurs hebdomadaires (environ 100 millions), pris comme base d'utilisateurs actifs.

8 Cf. https://huggingface.co/OpenLLM-France/Lucie-7B.

d'atteindre une taille critique minimale, d'en accroître l'efficacité et d'en augmenter le taux d'utilisation. De plus, LUCIE-7B adopte une stratégie open-source complète. En effet sont publiés les données d'entraînement (cf. Lucie-Training-Dataset), les scripts d'entraînement (cf. Lucie-Training, un *fork* de Megatron-DeepSpeed) et les modèles eux-mêmes (cf. LUCIE-7B sur Hugging Face). Cette approche permet une optimisation des jeux de données, par ailleurs réutilisables sans nécessiter une collecte massive, et des méthodes d'entraînement, ainsi qu'une réutilisation plus aisée, sous sa forme originale ou après spécialisation.

Le LLM DeepSeek-R1, développé par la société chinoise DeepSeek-AI, a été publié dans le contexte de tensions et de concurrence croissante entre les États-Unis et la Chine. Le modèle se distingue par un entraînement partiel sur des GPU domestiques, sans que ne soit ici claire la frontière entre gains réels et propagande visant à démontrer l'inefficacité des restrictions étasuniennes à l'exportation des puces les plus puissantes. Par ailleurs, le modèle se distingue par, d'une part, l'adoption d'une stratégie open-source (publication du modèle et de la méthode d'entraînement), d'autre part, une architecture MoE permettant un abaissement des coûts d'inférence[9]. DeepSeek-R1 revendique des performances comparables aux meilleurs modèles étasuniens en matière de raisonnement (Guo et al., 2025). Malgré ses performances, DeepSeek-R1 relève davantage de l'innovation de processus du fait des efforts entrepris pour abaisser les coûts d'entraînement et d'inférence.

7 Conclusion

Le cycle décrit par le modèle A-U se retrouve dans le développement actuel de l'IAG. Premièrement, des modèles permettant des tâches simples ont été mis sur le marché (p. ex. GPT-3.5). Ces modèles ont démontré leur utilité, ainsi que le plébiscite pour le design de l'agent conversationnel, mais aussi leurs limites pour des tâches plus complexes impliquant par exemple des chaînes de raisonnement. Deux pressions à l'innovation en ont découlé. D'une part, les limitations ont encouragé la mise sur le marché de modèles plus performants et plus lourds (p. ex. OpenAI o1). D'autre part, l'existence d'une base d'utilisateurs pour des modèles simples a encouragé leur optimisation. La continuation de l'innovation de produit, sur de nouveaux modèles, s'est dès lors complétée d'une innovation de processus visant à réduire les coûts de production (nouvelles stratégies d'entraînement, publication de modèles open-sources…) et d'exploitation (optimisation des centres de données ; agents, puces spécialisées...), à performance inchangée. La *commoditisation* s'observe également avec la publication de modèles sensiblement plus petits, plus frugaux, mais néanmoins adaptés à la réalisation de tâches spécifiques (p. ex. RAG). Si des signaux de croissance de la consommation de ressources énergétiques et matérielles existent, ils s'accompagnent d'une amélioration continue des infrastructures et des modèles, propice à une perpétuation du découplage actuellement constaté entre les usages et les ressources consommées. L'IEA prévoit cependant une croissance régulière de la consommation électrique des centres de données, autour de 5 % par an sur 10 ans (ce qui aboutirait à un triplement de la consommation), avec de fortes disparités locales ou régionales (IEA, 2025b). Par ailleurs, dresser un bilan complet nécessiterait une mise en balance avec les économies éventuellement permises par l'automatisation à l'aide de modèles génératifs (Tomlinson et al., 2024). Enfin, cette recherche identifie un ensemble de thématiques sur lesquelles des activités de recherche ciblées permettraient de contribuer à la réduction du poids environnemental de ces systèmes techniques.

9 Cf. https://api-docs.deepseek.com/news/news250120.

Références

[1] Bianchi, T. & Angulo, F. (2024), Online search after ChatGPT: the impact of generative AI. *Semrush & Statista*. https://static.semrush.com/file/docs/evolution-of-online-after-ai/Online_Search_After_ChatGPT.pdf.

[2] Bradshaw, T. & Morris, S. (2024). Microsoft acquires twice as many Nvidia AI chips as tech rivals. *Financial Times*, 17 décembre 2024. https://www.ft.com/content/e85e43d1-5ce4-4531-94f1-9e9c1c5b4ff1.

[3] CAAD (2024). Artificial Intelligence Threats to Climate Change. *Climate Action Against Disinformation*, 7 mars 2024. https://foe.org/wp-content/uploads/2024/03/AI_Climate_Disinfo_v6_031224.pdf.

[4] Crider, M. (2025). En intégrant Copilot, Microsoft365 voit ses tarifs augmenter pour les particuliers. *Le Monde Informatique*, 21 janvier 2025. https://www.lemondeinformatique.fr/actualites/lire-ia-et-cybersecurite-priorites-des-ssii-et-editeurs-francais-en-2024-96067.html.

[5] Eldan, R., & Li, Y. (2023). TinyStories: How Small Can Language Models Be and Still Speak Coherent English?. *arXiv preprint*. https://doi.org/10.48550/arXiv.2305.07759.

[6] Flipo, F., & Gossart, C. (2009). Infrastructure numérique et environnement. L'impossible domestication de l'effet rebond. *Terminal. Technologie de l'information, culture & société*, (103-104). https://doi.org/10.4000/terminal.3093.

[7] Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30, 681-694. https://www.doi.org/10.1007/s11023-020-09548-1.

[8] Gent, E. (2023). When AI's Large Language Models Shrink. *IEEE Spectrum*, 31 mars 2023. https://spectrum.ieee.org/large-language-models-size.

[9] Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu,

R., ... & He, Y. (2025). Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint*. https://doi.org/10.48550/arXiv.2501.12948.

[10] Ho, A., Besiroglu, T., Erdil, E., Owen, D., Rahman, R., Guo, Z. C., ... & Sevilla, J. (2024). Algorithmic progress in language models. *arXiv preprint*. https://doi.org/10.48550/arXiv.2403.05812.

[11] Hu, K. (2023). ChatGPT sets record for fastest-growing user base - analyst note. *Reuters*, 2 février 2023. https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/.

[12] IEA (2025b), Energy and AI, *International Energy Agency*. https://www.iea.org/reports/energy-and-ai.

[13] IEA (2025a). Data Centres and Data Transmission Networks, *International Energy Agency*. https://www.iea.org/energy-system/buildings/data-centres-and-data-transmission-networks.

[14] Isaac, M., & Griffith, E. (2024). OpenAI Is Growing Fast and Burning Through Piles of Money. *The New York Times*, 27 septembre 2024. https://www.nytimes.com/2024/09/27/technology/openai-chatgpt-investors-funding.html.

[15] Jeans, D. (2025). Sam Altman's Fusion Power Startup Is Eyeing Trump's $500 Billion AI Play. *Forbes*, 5 février 2025. https://www.forbes.com/sites/davidjeans/2025/02/05/stargate-sam-altman-fusion-helion/.

[16] Luccioni, S., Jernite, Y., & Strubell, E. (2024). Power hungry processing: Watts driving the cost of ai deployment? In *Proceedings of the 2024 ACM conference on fairness, accountability, and transparency* (pp. 85-99). https://doi.org/10.1145/3630106.3658542.

[17] Luccioni, A. S., Viguier, S., & Ligozat, A. L. (2023). Estimating the carbon footprint of bloom, a 176b parameter language model. *Journal of Machine Learning Research*, 24(253), 1-15. https://www.jmlr.org/papers/volume24/23-0069/23-0069.pdf.

[18] Ma, L., Xu, X., He, Y., & Tan, Y. (2024). Learning to Adopt Generative AI. *arXiv preprint*. https://doi.org/10.48550/arXiv.2410.19806.

[19] Oremus, W. (2023). AI chatbots lose money every time you use them. That is a problem. *The Washington Post*, 5 juin 2023. https://www.washingtonpost.com/technology/2023/06/05/chatgpt-hidden-cost-gpu-compute/.

[20] Patterson, D., Gonzalez, J., Hölzle, U., Le, Q., Liang, C., Munguia, L. M., ... & Dean, J. (2022). The carbon footprint of machine learning training will plateau, then shrink. *Computer*, 55(7), 18-28. https://doi.org/10.1109/MC.2022.3148714.

[21] Quiroz-Gutierrez, M. (2025). Sam Altman says he's losing money on OpenAI's $200-per-month subscriptions: 'People use it much more than we expected'. *Fortune*, 7 janvier 2025. https://fortune.com/2025/01/07/sam-altman-openai-chatgpt-pro-subscription-losing-money-tech/.

[22] Roth, F. (2016). V. William J. Abernathy et James M. Utterback. *Le cycle des innovations technologiques*. In Les Grands Auteurs en Management de l'innovation et de la créativité (pp. 103-120). EMS Editions. ISBN : 978-2-84769-812-1.

[23] Samsi, S., Zhao, D., ... & Gadepally, V. (2023). From words to watts: Benchmarking the energy costs of large language model inference. In *2023 IEEE High Performance Extreme Computing Conference (HPEC)* (pp. 1-9). IEEE. https://doi.org/10.1109/HPEC58863.2023.10363447.

[24] Smith, M. S. (2024). Challengers are Coming for NVidia's Crown: In AI's Game of Thrones, Don't Count Out the Upstarts. *IEEE Spectrum*, 61(10), 40-44. https://doi.org/10.1109/MSPEC.2024.10705376.

[25] Stokel-Walker, C. (2024). AI chatbots are improving at an even faster rate than computer chips. *New Scientists*, 27 mars 2024. https://www.newscientist.com/article/2424179-ai-chatbots-are-improving-at-an-even-faster-rate-than-computer-chips/.

[26] Sundberg, N. (2023). Tackling AI's Climate Change Problem. *MIT Sloan Management Review*, 65(2), 38-41. https://sloanreview.mit.edu/article/tackling-ais-climate-change-problem/.

[27] Tomlinson, B., Black, R. W., Patterson, D. J., & Torrance, A. W. (2024). The carbon emissions of writing and illustrating are lower for AI than for humans. *Scientific Reports (Sci Rep)*, 14(1), 3732. https://doi.org/10.17605/OSF.IO/YHTMQ.

[28] Trott, P. (2021). *Innovation Management and New Product Development – Seventh Edition*. Pearson. ISBN : 978-1-2922-5152-3.

[29] van Wynsberghe, A. (2021). Sustainable AI: AI for sustainability and the sustainability of AI. *AI and Ethics*, 1(3), 213-218. https://www.doi.org/10.1007/s43681-021-00043-6.

[30] Viseur, R. (2024). Stratégies open-sources : opportunités et limitations dans le domaine des Large Language Models (LLM). *Inforsid*, Nancy (France), 31 mai 2024. https://hdl.handle.net/20.500.12907/50980.

[31] Vuarin, L., Lopes, P. G., & Massé, D. (2023). L'intelligence artificielle peut-elle être une innovation responsable? *Innovations*, 72(3), 103-147. https://doi.org/10.3917/inno.pr2.0153.

[32] Ward-Foxton, Sally (2023). Groq Demonstrates Fast LLMs on 4-Year-Old Silicon. *EETimes*, 12 septembre 2023. https://www.eetimes.com/groq-demos-fast-llms-on-4-year-old-silicon/.

[33] Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., & Narasimhan, K. (2024). Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36. https://proceedings.neurips.cc/paper_files/paper/2023/file/271db9922b8d1f4dd7aaef84ed5ac703-Paper-Conference.pdf.

Session 3 : Interprétabilité, explicabilité

L'insertion au service de l'intelligence artificielle : Modèle d'apprentissage pour l'annotation d'images satellite pour le consortium AI4GEO

Fabien Amarger[1], Nathalie Noblecourt[1], Jehanne Portefaix[1], Pierre-Marie Brunet[2]

[1] Digitanie
Cour Guillaut 09700 Saverdun
prénom.nom@digitanie.org
[2] CNES
Centre spatial de Toulouse
18 avenue Edouard Belin 31401 Toulouse
prénom.nom@cnes.fr

Résumé

Le consortium AI4GEO avait pour volonté de regrouper différents acteurs, publics ou privés, du milieu du traitement de données géographiques pour mutualiser les solutions et les moyens. Un des besoins communs à plusieurs membres du consortium a été de pouvoir générer une représentation en 3D à partir d'une image satellite. Pour cela, il est nécessaire d'avoir deux types de données : l'altitude des éléments et l'emprise des éléments. La hauteur est calculée par photogrammétrie à partir de plusieurs prises de vues du même point. L'emprise est plus complexe à déterminer puisqu'il faut un modèle permettant d'analyser l'image satellite, par exemple les contours d'un immeuble ou le tracé d'une route. Un modèle de segmentation sémantique d'images satellites nécessite une grande quantité de données d'apprentissage pour avoir un résultat suffisamment précis, de même qu'un enjeu fort de généralisation (conserver la performance de prédiction des zones géographiques variées). Pour obtenir ces données, la seule méthode viable est d'effectuer cette annotation manuellement pour optimiser la fiabilité. Ce travail est long et redondant, et nécessite une organisation spécifique de la production. L'entreprise d'insertion Digitanie s'est spécialisée dans la production industrielle de ce type de données. La collaboration entre le CNES (membre du consortium AI4GEO) et Digitanie a permis de produire une grande quantité de données précises et fiables. Nous proposons, dans cet article, de présenter le consortium AI4GEO, l'implication du CNES dans ce consortium et les objectifs visés. Ensuite, nous présenterons la collaboration avec l'entreprise Digitanie et l'organisation de la production qui a permis la création d'autant de données fiables. Enfin, nous conclurons sur les résultats obtenus, aussi bien en terme technique et scientifique que l'apport pour l'insertion.

Mots-clés

Apprentissage, Segmentation sémantique, Analyse d'image, Image satellite, AI4GEO, Insertion par l'activité économique

Abstract

The aim of the AI4GEO consortium was to bring together different actors, both public and private, in the field of geographic data processing, in order to share solutions and resources. One of the needs shared by several consortium members was to be able to generate a 3D representation from a satellite image. To achieve this, two types of data are required : the entity height and the entity ground surface. The height is calculated photogrammetrically from several shots taken from the same point. The ground surface, on the other hand, is more complex to determine, since it requires a model that can be used to analyze the satellite image, such as the contours of a building or the layout of a road. A semantic segmentation model for satellite imagery requires a large amount of training data to obtain a sufficiently accurate result, as well as a high level of generalizability (maintaining prediction performance for varied geographical areas). To obtain this data, the only viable method is to carry out this annotation manually, to optimize reliability. This work is time-consuming and redundant, and requires a specific production organization. The insertion company Digitanie specializes in the industrial production of this type of data. Collaboration between CNES (a member of the AI4GEO consortium) and Digitanie has produced a large quantity of accurate, reliable data. In this article, we present the AI4GEO consortium, CNES's involvement and its objectives. We will then present the collaboration with the Digitanie company and the production organization that enabled the creation of so much reliable data. Finally, we'll conclude with a look at the results obtained, both in technical and scientific terms, and in terms of their contribution to the integration process.

Keywords

Machine Learning, Semantic segmentation, Image analysis, Satelite image, AI4GEO, Integration company

1 Introduction

AI4GEO[1] est un programme français scientifique et industriel composé de plusieurs instituts publics (CNES, IGN, ONERA) et de groupes industriels (CS Group, AIRBUS Defence and SPace, CLS, GEOSAT, QUANTCUBE) couvrant toute la chaîne de valeur de l'information géospatiale. Il vise à produire automatiquement des cartes sémantiques en 3D à très haute résolution et à l'échelle mondiale. Pour ce faire, une partie essentielle consiste à produire des cartes d'occupation du sol de haute qualité.

Ces cartes sémantiques ont plusieurs objectifs applicatifs et peuvent apporter un intérêt majeur pour différents domaines comme les véhicules autonomes, l'urbanisme, l'intelligence économique ou encore l'agriculture.

Pour cette tâche, les méthodes d'apprentissage profond (Deep Learning) constituent aujourd'hui l'état de l'art. En particulier, les réseaux entièrement convolutifs (Fully Convolutional Networks - FCN) [4] sont désormais courants pour effectuer une telle tâche de segmentation sémantique. Ces modèles ont réussi à surpasser les méthodes traditionnelles grâce à leur capacité à extraire et à apprendre automatiquement les caractéristiques pertinentes pour la tâche considérée et en prenant en compte le contexte spatial [3, 5]. Ce dernier point est nécessaire pour améliorer les performances du traitement des images satellites de très haute définition en vue de produire des cartes précises de l'occupation des sols.

Dans cet article, nous allons présenter la méthode mise en place pour l'entraînement d'un modèle pour l'analyse d'images satellites en vue d'obtenir ces cartes précises de l'occupation des sols. Nous présenterons ensuite comment nous avons pu obtenir un grand nombre de données de labellisation, des images de grande qualité pour permettre cet entrainement grâce à l'aide de l'activité par l'insertion économique. Enfin, nous conclurons sur l'intérêt, pour l'avenir de l'intelligence artificielle, d'obtenir des données d'apprentissage de qualité, produites en France et permettant à des personnes éloignées de l'emploi de contribuer à ces sujets.

2 Vectorisation sémantique multi-classe d'images satellites très haute définition[6]

La base de données AI4GEO est constituée d'images satellites Pléiades [1] stockées sous forme de géotiffs optimisés pour le cloud (Cloud Optimized Geotiffs - COG) de 16 bits et d'une vérité de terrain annotée manuellement qui a été produite tout au long du projet. Les images utilisées pour l'entraînement et l'inférence sont des images ortho rectifiées d'une résolution de 50 cm avec une correction du sommet de l'atmosphère (Top Of Atmosphere - TOA). Il s'agit d'une opération de prétraitement standard qui génère des images de réflectance (corrigées radiométriquement à partir de l'étalonnage du capteur et des effets atmosphériques systématiques de l'atmosphère). Chaque image est composée de quatre bandes spectrales : Rouge, Vert, Bleu et Proche-Infrarouge. Quatorze villes présentant un intérêt industriel ont été choisies : Arcachon, Biarritz, Brisbane, Lagos, Maros, Montpellier, Munich, Nantes, Paris, Port-Elisabeth, Shanghai, Strasbourg, Tianjin, Toulouse. Pour toutes ces villes, les responsables du projet au CNES ont sélectionné 10 tuiles de 2048x2048 pixels représentant la diversité du paysage sémantique. Ces tuiles ont été segmentées manuellement, puis chaque polygone a été classé par des opérateurs en photo interprétation. Nous avons mis en place un processus incrémental tout au long de la production de la vérité terrain afin d'assurer un dialogue efficace entre les opérateurs et les experts en Machine Learning. Nous reviendrons plus en détails sur ce processus dans une section dédiée dans cet article. Nous obtenons une vérité terrain de haute qualité jusqu'à 137 classes différentes. Pour la phase d'apprentissage du modèle, uniquement 10 classes ont été utilisées. Pour ce faire, nous avons forcé le classement de chaque pixel parmi les classes suivantes : bâtiment, route, chemin de fer, végétation haute, végétation basse, eau et arrière-plan (autres). Ce choix a été fait après une discussion entre les différents partenaires pour décider quelles classes étaient les plus utiles pour les différents cas d'utilisation. Les classes sont déséquilibrées dans cet ensemble de données, mais cela représente une difficulté structurelle de la tâche. La figure 1 donne des exemples de vérités terrain obtenues par labellisation manuelle. OpenStreetMap (OSM) [2] est une base de données géographique ouverte et collaborative qui fournit des caractéristiques géographiques dans le monde entier. L'utilisation de ce jeu de

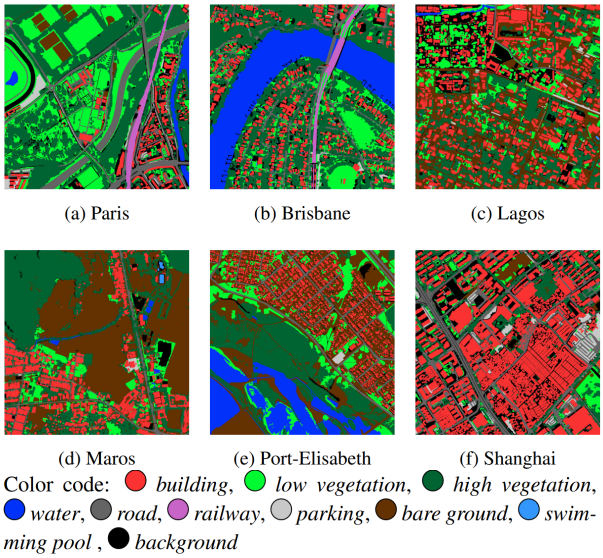


FIGURE 1 – Exemples de résultats de la vectorisation sémantique

1. https://dinamis.data-terra.org/en/pleiades-products/

données permet d'exploiter des images Pléiades complètes (environ 40000x40000 pixels) et de ne pas se limiter aux tuiles présélectionnées. Cependant, la qualité de la vérité terrain extraite de ce jeu de données peut varier en fonction de la zone considérée. Plus précisément, la vérité terrain de certains éléments peut être légèrement mal placée ou même complètement manquante. En outre, la date d'acquisition de l'image Pléiades peut impliquer des différences de mise à jour avec OSM. Les données OSM ne peuvent donc pas suffire comme base d'entrainement pour l'entraînement du modèle de vectorisation sémantique, mais ces données peuvent être une aide, précieuse, pour la vectorisation manuelle. C'est donc dans ce contexte que l'entreprise Digitanie a été sollicitée pour annoter manuellement des images Pléiades par photo-interprétation avec l'aide des données OSM.

3 Processus de production

Dans le cadre d'un partenariat entre le CNES (membre du consortium AI4GEO) et Digitanie, un processus de production a été mis en place pour répondre au besoin de données géographiques de très haute qualité pour l'apprentissage d'un modèle de segmentation sémantique. À l'instar d'une usine de production mécanique de précision, un processus de production a été mis en place pour labelliser des images satellites par photo-interprétation.
L'objectif est d'annoter un sous-ensemble d'images, représentatives, des villes à labelliser. Ces images sont fournies par le CNES, Digitanie réalise la vectorisation manuelle. Ce travail nécessite un gros volume horaire pour réussir à analyser et vectoriser l'intégralité des images, c'est pour cela que nous parlons d'usine du numérique pour de la production de données géographiques.

3.1 Organisation de production

Dans un objectif de proposer une labellisation de qualité et à moindre cout, il a été nécessaire de penser l'organisation de la production dans son ensemble pour optimiser l'intégralité du processus de production.
Pour effectuer la labellisation, nous utilisons l'outil QGIS [2] qui est un outil libre de traitement de données géographiques.

Échantillon, méthodologie et métriques

La première étape consiste à établir une méthodologie à partir d'un échantillon suffisamment représentatif de données ; À partir de cet échantillon, les référents techniques définissent des indicateurs ou métriques, qui vont déterminer un temps de production unitaire pour une surface d'image traitée (découpée en tuiles). La taille de l'image doit être suffisamment importante pour permettre une production sur un temps donné par un opérateur, mais ne doit pas être trop grande pour permettre une livraison et un contrôle de façon continue. Les référents techniques définissent une méthodologie précise, partagée et validée par le CNES, qui permet de déterminer les classes à labelliser et les interprétations à effectuer lors d'ambiguïtés. De plus, ils se rendent disponibles régulièrement pour aider les opérateurs. Cette proximité et la clarté de la méthodologie sont d'autant plus importants, et nous le verrons plus en détail plus tard, qu'elles permettent à des personnes n'ayant pas de connaissance en traitement de données géographiques de pouvoir contribuer aisément au projet.

Formation

Une fois cette étape de méthodologie établie à partir de l'échantillon, Les salariés doivent être formés. Effectivement, cette phase de formation permet aux opérateurs de s'approprier la méthodologie, de l'éprouver, voir de l'améliorer dans une démarche d'amélioration continue. Cette phase de montée en compétences, est essentielle pour permettre un rendement maximum dès la mise en production des opérateurs et ne doit pas être négligée.

Production

Une fois la méthodologie définie, les métriques établies, les opérateurs formés, il est maintenant possible de passer à la phase de production. Elle est organisée par vagues pour permettre d'observer les métriques à la fin de chaque vague et d'ajuster la méthodologie itérativement. Chaque vague correspond à un ensemble de tuiles à labelliser provenant d'une ville en particulier. Un travail préparatoire, effectué par les référents techniques, permet aux opérateurs de se concentrer sur la production. Une répartition des tâches de production est faite en début de semaine, pour que chaque opérateur n'ait ni trop, ni pas assez de travail pour la semaine et pour paralléliser la production au maximum. Les opérateurs remplissent, quotidiennement, les indicateurs pour pouvoir suivre l'état de la production et pouvoir planifier le travail pour permettre la continuité de la production. Digitanie et le CNES ont mis un point d'honneur, tout au long du projet, à s'assurer que cette répartition du travail était en adéquation avec des conditions respectueuses des opérateurs. Cette répartition prend en compte l'ancienneté, mais aussi la diversité des personnes. Ce respect des personnes permet de s'assurer que le travail n'est pas trop fastidieux et de ne pas démotiver les opérateurs. De plus, un espace de communication dédié au projet a été mis en place pour faciliter l'entraide et la collaboration entre les opérateurs.

Contrôle

Un contrôle croisé est effectué par les opérateurs, ce qui permet de sensibiliser à la qualité du travail attendu. Une fois les images produites et contrôlées, elles sont à nouveau contrôlées (de façon plus générale) par un référent technique avant d'être livrées pour les clients.

Retours d'expériences

Afin de garantir un niveau de qualité optimum de la production des données, il est nécessaire d'organiser des retours d'expériences régulièrement. Nous avons mis en place trois catégories de retours d'expériences. La première catégorie est un retour d'expériences hebdomadaire entre les référents techniques et les clients afin de s'assurer de la satisfaction de ces derniers quant à la qualité de la production, mais aussi de l'efficacité de celle-ci. Ensuite, nous organisons des retours d'expériences entre les référents techniques et

2. https://qgis.org/

les opérateurs pour clarifier certains points de la méthodologie qui auraient pu évoluer, mais aussi pour s'assurer qu'elle est soutenable par les opérateurs. Enfin, le dernier type de retours d'expériences est entre les opérateurs et les clients. Ces retours sont moins fréquents, mais sont très importants et ont plusieurs objectifs : permettre aux clients de se rendre compte du travail réalisé par les opérateurs et permettre aux opérateurs de comprendre réellement les besoins et les objectifs du projet pour donner du sens à leur travail quotidien. Tout au long du projet, un suivi des métriques est effectué par les référents techniques, en toute transparence avec les clients. Cette transparence est un pilier de tout le processus de production, autant d'un point de vue suivi économique, de qualité du résultat attendu mais aussi d'un point de vue scientifique. Effectivement, la transparence permet un ajustement permanent de la méthodologie et du processus de production. Par exemple, certaines difficultés relevées par les opérateurs ont influencé les choix des futures tuiles à labelliser.

3.2 L'intérêt de la production numérique pour l'insertion

Comme évoqué précédemment, nous recrutons des opérateurs à tous niveaux techniques concernant le traitement de données géographiques, mais aussi concernant l'utilisation de l'outil informatique de manière générale. Digitanie est une entreprise d'insertion et de ce fait, elle accompagne des personnes éloignées de l'emploi à s'insérer par l'activité économique (IAE) [3] et qui plus est dans un milieu rural et donc assez éloigné des grandes zones technologiques que peuvent connaitre les grandes villes. Nous avons mis en place tout un parcours de suivi de compétence de ces personnes pour les aider à acquérir le niveau attendu pour pouvoir contribuer aux projets de traitement de données géographiques et notamment à la labellisation d'images satellites pour le CNES dans le cadre du consortium AI4GEO. Nous avons observé une moyenne d'un mois de formation nécessaire pour qu'une personne, quel que soit son niveau initial, puisse commencer à contribuer à labelliser ces images. Il est particulièrement gratifiantpour des personnes qui n'avaient jamais travaillé sur un ordinateur auparavant, de pouvoir contribuer rapidement à un sujet de pointe autour de l'intelligence artificielle et l'observation de la Terre. Cette activité crée aussi des vocations puisqu'un certain nombre de personnes se réorientent vers la géomatique après leur passage à Digitanie.

4 Conclusion

Nous avons tout d'abord décrit les objectifs du consortium AI4GEO en termes de création de cartes géospatiales 3D de très hautes qualités. Nous avons détaillé les outils techniques et technologiques mis en œuvre pour atteindre ces objectifs avec un besoin en termes de labellisation manuelle d'images satellites. Nous avons ensuite précisé le processus de production manuelle mis en place pour permettre cette labellisation avec un très haut niveau de qualité. Cette production permet d'améliorer la qualité des modèles obtenus et donc d'améliorer la labellisation automatique des images satellites d'observation de la Terre. Enfin, nous avons expliqué en quoi l'organisation d'une entreprise d'insertion pour de la production numérique permet de répondre à ces besoins. Les modèles d'intelligence artificielle sont toujours plus performants en absorbant toujours plus de données. Mais la qualité des données utilisées pour l'apprentissage de ces modèles impact directement la qualité des modèles eux-mêmes. Pouvoir disposer des données de très haute qualité pour l'apprentissage permet d'optimiser cette phase d'apprentissage. De plus, ce travail qui est long et fastidieux, peut être une chance pour des personnes éloignées de l'emploi de retrouver une stabilité économique grâce à l'IAE. C'est cette complémentarité entre l'IA et l'IAE que nous avons voulu mettre en avant dans cet article et que nous souhaitons continuer à faire évoluer sur d'autres sujets de traitement de données (géographiques ou non) et plus généralement sur cette notion de production industrielle du numérique tout en continuant à accompagner et à aider des personnes éloignées de l'emploi, notamment en valorisant leurs compétences sur des sujets de pointes.

3. https://travail-emploi.gouv.fr/insertion-par-lactivite-economique-retour-sur-le-congres-des-5-et-6-decembre-2024

Références

[1] Pierre-Marie Brunet, Pierre Lassalle, Simon Baillarin, Bruno Vallet, Arnaud Le Bris, Gaëlle Romeyer, Guy Le Besnerais, Flora Weissgerber, Gilles Foulon, Vincent Gaudissart, et al. Ai4geo : A data intelligence platform for 3d geospatial mapping. In *24th ISPRS Congress Commission II : Imaging Today, Foreseeing Tomorrow*, volume 43, pages 817–823, 2021.

[2] OpenStreetMap Contributors. Planet dump retrieved from https://planet.osm.org. 2022.

[3] Alberto Garcia-Garcia, Sergio Orts-Escolano, Sergiu Oprea, Victor Villena-Martinez, and Jose Garcia-Rodriguez. A review on deep learning techniques applied to semantic segmentation. *arXiv preprint arXiv :1704.06857*, 2017.

[4] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[5] Bipul Neupane, Teerayut Horanont, and Jagannath Aryal. Deep learning-based semantic segmentation of urban features in satellite images : A review and meta-analysis. *Remote Sensing*, 13(4) :808, 2021.

[6] Axel Rochel, Clément Deschesne, Adrien Chan-Hon-Tong, and Pierre-Marie Brunet. Multiclass semantic segmentation with very high-resolution satellite images. In *Proceedings of the 2023 conference on Big Data from Space*, pages 21–24. Publications Office of the European Union, 2023.

Explications de diagnostic à base de modèle

Alban Grastien
Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France

alban.grastien@cea.fr

Résumé

Diagnostiquer consiste à détecter et identifier des défauts ou pannes dans un système complexe — réseau électrique ou de communication, chaîne de production, etc. Le diagnostiqueur est parfois incorrect en raison d'approximations inéluctables dans le modèle du système. Dans ce travail, nous proposons d'adjoindre au diagnostic une justification du résultat sous la forme d'une synthèse des observations reçues. Ainsi un opérateur peut vérifier la validité du diagnostic. Cette justification doit rester simple pour être compréhensible et nous faisons le lien avec les « chroniques », développées dans les années 90 pour le diagnostic.

Mots-clés

Diagnostic à base de modèle. IA explicative.

1 Introduction

Le diagnostic est la branche de l'IA qui s'intéresse à détecter, isoler et identifier l'occurrence de fautes dans les systèmes partiellement observables. Nous nous intéressons ici au diagnostic à base de modèle dont le raisonnement se base sur une comparaison entre les observations espérées et celles réellement obtenues.

Nous nous plaçons principalement dans un contexte industriel (chaîne de production, réseau électrique ou de communication), même si les techniques de diagnostic peuvent s'appliquer à une large panoplie de systèmes. Un frein important au développement du diagnostic en industrie est le caractère « boîte noire » de l'approche et le risque d'erreurs. Ce risque est particulièrement important car le modèle nécessite souvent des approximations voire inclut des erreurs dû à la difficulté de collecter des informations pour les systèmes complexes.

Étant donné ce constat, nous voyons le diagnostiqueur comme un outil d'aide à la décision plutôt qu'un décideur final que l'utilisateur humain doit suivre aveuglément. Pour ce faire, nous souhaitons étendre le diagnostiqueur pour le faire renvoyer, en plus du diagnostic, une *explication* qui aide l'utilisateur à suivre le raisonnement suivi par celui-ci. Ainsi, en plus d'aider à la compréhension finale du diagnostic et de ses répercutions, l'explication doit permettre à l'utilisateur de déterminer si le diagnostiqueur a commis une erreur. Ce second objectif est très important pour nous : nous pensons que le diagnostiqueur a de grandes chances de commettre des erreurs, et que ces erreurs pourraient être désastreuses si elles conduisent à une mauvaise réponse de la part de l'utilisateur. Il est donc important de s'assurer que ces explications sont « simples ». En ce sens, un diagnostic qui ne serait pas accompagné d'une explication simple serait inutile.

Dans cet article, nous définissons une explication de diagnostic comme un objet mathématique (par exemple, un prédicat) qui permet à l'utilisateur de produire le diagnostic. Si la validité du prédicat et l'inférence du diagnostic à partir de ce prédicat sont toutes deux « simples », alors l'explication est compréhensible pour un utilisateur humain. S'il existe une explication simple à tous les comportements fautifs, on dit que le système est explicable.

Dans un second temps, nous proposons une implémentation d'explication sous forme de « chronique ». Une chronique est un motif d'événements observables. L'usage des chroniques a été proposé dans les années 90 [7] pour le diagnostic : une ou plusieurs chroniques sont construites pour chaque faute et la tâche de diagnostic se résume à identifier les chroniques dans le flot d'observation. L'avantage des chroniques est qu'elles peuvent être utilisées sans modèle car les chroniques peuvent être construites à partir de connaissance experte. D'autre part, des algorithmes efficaces ont été développés [6].

Nous représentons donc les explications comme des chroniques. Ceci permet de réhabiliter les chroniques en montrant qu'une qualité de celles-ci, qui n'était pas mentionnée dans la litérature scientifique, est qu'elles sont interprétables par les utilisateurs.

Un autre problème des systèmes de diagnostic à base de chronique est qu'il est parfois nécessaire de construire de nombreuses chroniques qui plus est de grande taille. Il est parfois même impossible de construire un ensemble de chroniques complet quand bien même chaque faute est diagnosticable. Nous montrons cependant que si le système est explicable — ce que nous considérons comme une condition indispensable pour un système à diagnostiquer — alors le nombre de chroniques est nécessairement faible et celle-ci sont petites (puisque chaque comportement a une explication simple). Nous en concluons qu'une approche à base de chronique est viable.

Le reste de cet article est divisé comme suit. La prochaine section présente les définitions classiques de diagnostic. La notion d'explication est décrite dans la section 3 et la question d'explicabilité dans la section 4. Finalement, nous discutons de travaux similaires avant de conclure.

2 Contexte

2.1 Diagnostic de SÉD

Un SÉD (système à événements discrets) est un modèle de système dynamique. Nous reprenons une définition proche de celle proposée initialement par Sampath et co-auteurs [15]. Un *SÉD* est un tuple $M = \langle Q, \Sigma, T, I, \Sigma_{\mathrm{o}}, \Sigma_{\mathrm{F}} \rangle$ tel que Q est un ensemble d'*états* dont $I \subseteq Q$ le sous-ensemble des *états initiaux*, Σ est l'ensemble des *événements* dont Σ_{o} le sous-ensemble d'*événements observables* et $\Sigma_{\mathrm{F}} \subseteq \Sigma$ le sous-ensemble d'*événements de faute* et $T \subseteq Q \times \Sigma \times Q$ est l'ensemble des *transitions*. Pour simplifier la discussion, nous considérons qu'il n'y a qu'un seul événement de faute ($\Sigma_{\mathrm{F}} = \{e_{\mathrm{F}}\}$), mais ces travaux se généralisent facilement à un non singleton.

La sémantique du SÉD repose sur la notion de chemin. Un *chemin* $\rho \in \mathit{Chs}(M)$ de longueur k est une double séquence $q_0 \xrightarrow{e_1} q_1 \xrightarrow{e_2} \dots \xrightarrow{e_k} q_k$ d'états et de événements telle que $q_0 \in I$ est un état initial et pour chaque $i \in \{1, \dots, k\}$, $\langle q_{i-1}, e_i, q_i \rangle \in T$ est une transition. La restriction de la séquence d'événements $[e_1, \dots, e_k]$ aux événements observables est appelé une *trace*. Un chemin comprenant un événement de faute ($\exists i \in \{1, \dots, k\}.\ e_i = e_{\mathrm{F}}$, noté $e_{\mathrm{F}} \in \rho$) est qualifié de *fautif* ; sinon, le chemin est *normal*. S'il y a au moins d événements après la faute ($i \leq k - d$), on parle de chemin d-fautif.

Un *problème de diagnostic* est une paire $P = \langle M, O \rangle$. Le diagnostic est le problème consistant à déterminer si la faute e_{F} a eu lieu. Formellement, le *diagnostic* de O est défini par

$$\Delta_M(O) = \begin{cases} \text{N} & \text{si } \exists \rho \in \mathit{Chs}(M). \\ & \quad \mathit{obs}(\rho) = O \wedge e_{\mathrm{F}} \notin \rho, \\ \text{F} & \text{sinon.} \end{cases}$$

Les symboles N et F signifient respectivement « normal » et « fautif ».

On notera que la définition de diagnostic est « optimiste » en ce sens que le diagnostiqueur se contente de trouver une trace nominale en accord avec les observations pour renvoyer un diagnostic normal. Cette définition est acceptable dans la mesure où nous faisons une seconde hypothèse, à savoir que le système est diagnosticable, c'est-à-dire que nous supposons qu'il existe un délai borné à l'issue duquel toute faute sera détectée. Formellement, la *diagnosticabilité* pour un délai $d \in \mathbf{N}$ est la propriété

$$\forall \rho \in \mathit{Chs}(M).\ \rho \text{ est } d\text{-fautif} \Rightarrow \Delta_M(\mathit{obs}(\rho)) = \text{F}.$$

Exemple *Considérons l'exemple de la figure 1 (gauche). Le chemin* $\rho = 0 \xrightarrow{a} 1 \xrightarrow{a} 2 \xrightarrow{b} 1' \xrightarrow{c} 1 \xrightarrow{a} 2 \xrightarrow{e_{\mathrm{F}}} 3 \xrightarrow{a} 4$ *produit la trace* $O = \mathit{obs}(\rho) = aabcaa$. *Le diagnostic de cette trace est* $\Delta_M(O) = \text{F}$.

2.2 Chroniques

Les chroniques sont un formalisme de système expert développé dans les années 90 pour diagnostiquer les SÉD de manière efficace et sans qu'il soit nécessaire de construire un modèle. Une chronique décrit un motif d'événements symptomatique d'une faute.

Un *graphe orienté* est une paire $G = \langle V, E \rangle$ telle que V est un ensemble de *nœuds* et $E \subseteq V \times V$ un ensemble d'*arêtes*. Un *cycle* dans G est une séquence de $k > 1$ nœuds $v_1, \dots, v_k$ qui satisfait les deux propriétés $\forall i \in \{1, \dots, k-1\}.\ \langle v_i, v_{i+1} \rangle \in E$ et $v_1 = v_k$. Un *graphe orienté acyclique* (GOA) est un graphe orienté qui ne comporte aucun cycle ; en particulier, il n'admet aucune boucle ($\forall v \in V.\ \langle v, v \rangle \notin E$).

Étant donné l'ensemble d'événements observables Σ_{o}, une chronique (atemporelle) [7] est un triplet $\theta = \langle G, L_{\mathrm{V}}, L_{\mathrm{E}} \rangle$ tel que $G = \langle V, E \rangle$ est un GOA, $L_{\mathrm{V}} : V \to \Sigma_{\mathrm{o}}$ et $L_{\mathrm{E}} : E \to 2^{\Sigma_{\mathrm{o}}}$ sont deux fonctions qui associent, respectivement, l'ensemble des nœuds et l'ensemble des arêtes à, respectivement, un événement observable et un ensemble d'événements observables. [1]

On appelle les événements mentionnés par L_{V} des *observations nécessaires* et les événements de L_{E} des *observations interdites*. Une chronique est reconnue si les observations nécessaires apparaissent dans la trace tandis que les événements interdits n'y apparaissent pas, ceci dans l'ordre précisé par le graphe. Les événements interdits représentent le fait que l'absence de certains événements est inattendue ; ainsi, si un événement et son annulation (par exemple, ouverture/fermeture ou entrée/sortie) sont observables, on s'attendra à voir une annulation entre deux occurrences du premier événement : ouverture, suivi de fermeture, suivi d'ouverte. C'est plus ou moins ce que les événements a et b représentent dans l'exemple de la figure 1 (gauche).

Formellement, étant données une trace comportant n événements observables $O = o_1, \dots, o_n$ et une chronique $\theta = \langle G, L_{\mathrm{V}}, L_{\mathrm{E}} \rangle$, une *(fonction de) correspondance* entre O et θ (indiquant quand, dans la trace O, les événements nécessaires de θ ont lieu) est une fonction $t : V \to \{1, \dots, n\}$ qui satisfait les propriétés suivantes :

1. $\forall v_1, v_2 \in V.\ v_1 \neq v_2 \Rightarrow t(v_1) \neq t(v_2)$,
2. $\forall \langle v_1, v_2 \rangle \in E.\ t(v_1) < t(v_2)$,
3. $\forall v \in E.\ o_{t(v)} = L_{\mathrm{V}}(v)$ et
4. $\forall \langle v_1, v_2 \rangle \in E.\ \forall i \in \{t(v_1) + 1, \dots, t(v_2) - 1\}.\ o_i \notin L_{\mathrm{E}}(\langle v_1, v_2 \rangle)$.

S'il existe une correspondance entre O et θ, on dit que θ est une chronique de O et que O suit θ. On utilise aussi la notation $O \in [[\theta]]$.

Lien avec le diagnostic La chronique θ est *typique de la faute* e_{F} (ou, simplement, *typique*) si toute trace O cohérente avec le modèle M et qui suit θ provient d'une chemin fautif :

$$\forall \rho \in \mathit{Chs}(M).\ \mathit{obs}(\rho) \in [[\theta]] \Rightarrow e_{\mathrm{F}} \in \rho.$$

En conséquence de quoi, toute trace qui suit θ peut être diagnostiquée comme fautive :

Lemme 1 *Si la chronique* θ *est typique, alors le résultat suivant est correct*

$$O \in [[\theta]] \Rightarrow \Delta_M(O) = \text{F}.$$

1. Les définitions classiques ajoutent des informations temporelles aux arêtes du graphe ; étendre les résultats pour ce type de définition est un des axes de recherche futurs de nos travaux.

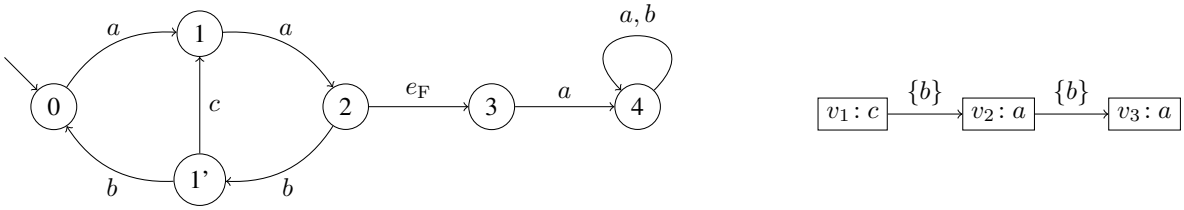


FIGURE 1 – Gauche : Exemple de SÉD diagnosticable si l'ensemble Σ_o comporte a et b. Il est explicable (cf. Section 4) pour une taille d'explication de $S = 3$ si Σ_o vaut $\{a, b, c\}$; sinon, il n'est pas explicable. Droite : exemple de chronique typique si a, b et c sont observables.

Exemple *Reprenons l'exemple de la figure 1 et considérons la chronique sur la droite de la figure. Cette chronique comprend trois nœuds associés aux événements nécessaires c, a et a respectivement. L'arête entre v_2 et v_3 comporte également un événement interdit b. On peut facilement démontrer que cette chronique est typique puisqu'elle trahit la séquence de transitions* $1' \xrightarrow{c} 1 \xrightarrow{a} 2 \xrightarrow{e_F} 3 \xrightarrow{a} 4$. *Ainsi, une trace qui suit cette chronique peut facilement être diagnostiquée quelque soit sa longueur.*

Un système de reconnaissance de chroniques (SRC ou *chronicle recognition system* CRS en anglais) est un diagnostiqueur qui utilise un ensemble Θ de chroniques typiques. Le SRC lit la trace fournie par le système diagnostiqué et vérifie si celle-ci suit une des chroniques. Tant que ceci n'arrive pas, le SRC renvoie N ; mais dès qu'une des chroniques est suivie par la trace, alors le diagnostiqueur renvoie F. On notera qu'un tel diagnostiqueur n'est pas toujours exact mais ses approximations peuvent être acceptables sous certaines conditions. Par exemple, un tel diagnostiqueur introduit parfois des délais entre la détection d'une faute et la date à laquelle on disposait de suffisamment d'informations pour la détection. Si ce délai est borné et si cette borne est suffisamment faible, alors ce délai n'est pas problématique en pratique (on se rappellera qu'il y a déjà un délai entre l'occurrence de la faute et la date à laquelle celle-ci peut-être détectée).

Un diagnostiqueur à base de chronique est surtout très facile à implémenter et très efficace. Il est donc intéressant de pouvoir remplacer un diagnostiqueur complet tel que [15, 13, 19] pour un simple SRC. S'il existe un ensemble fini et « petit » Θ de chroniques tel que toute occurrence de faute sera détectée par un SRC muni de Θ, on dit que Θ couvre les comportements de faute du système. On dit alors que le diagnostic à base de chronique est une méthode *viable* pour le SÉD M.

3 Explications

Nous sommes intéressés par le calcul d'explications pour le diagnostic. De manière informelle, une explication est une information permettant à un utilisateur humain de comprendre le raisonnement du diagnostiqueur et de se convaincre de la correctitude du diagnostic. Le cas échéant, une explication pourrait être utilisée pour détecter une erreur de la part du diagnostiqueur. En effet, même si le diagnostiqueur est correctement implémenté, la théorie du diagnostic repose sur des hypothèses dont on peut légitimement contester la validité. Ainsi, le diagnostic à base de modèle présume l'existence d'un modèle (le SÉD) correct. Or, pour de nombreuses raisons, cette présomption peut être malvenue :

— Le modèle peut contenir des erreurs. Par exemple, nous avons pu constater lors de nos collaborations avec certaines compagnies de distribution d'électricité que celles-ci ne disposaient pas de base de données centralisées de leur réseau électrique, ce qui rendait toute modélisation du réseau approximative *a minima*.

— Le modèle peut ignorer certaines situations exceptionelles requérant une modélisation plus fine. Pour reprendre l'exemple du réseau de distribution d'électricité, un événement sportif ou culturel inhabituel pourra modifier les paramètres habituels du réseau et tromper le diagnostiqueur. Une explication du diagnostic devrait permettre à l'opérateur en charge du réseau de comprendre que la panne prédite est possiblement liée à ces conditions inhabituelles et l'encourager à chercher d'autres symptômes avant d'agir prématurément.

3.1 Qu'est-ce qu'une explication

Nous considérons donc qu'une explication est une information fournie en parallèle du diagnostic qui permet — d'après l'agent en charge de générer l'explication — d'arriver à la même conclusion que le diagnostiqueur mais avec un effort de raisonnement plus faible. Par exemple, une explication peut pointer les observations importantes dans la trace fournie par le système.

Nous proposons de définir une explication comme suit :

Définition 1 *Étant donné un problème de diagnostic $P = \langle M, O\rangle$ où M est un SÉD et O est une trace telle que $\Delta_M(O) =$ F, une* explication *est un object mathématique x tel que les assertions suivantes sont vraies :*

1. *x est une propriété de O, ce qu'on notera $O \models x$;*
2. *x est une preuve du diagnostic, à savoir que*

$$\forall \rho \in \mathit{Chs}(M).\ (\mathit{obs}(\rho) \models x) \Rightarrow \Delta_M(\mathit{obs}(\rho)) = \text{F},$$

ce que l'on notera $x \models$ F.

Une explication est simple *s'il est simple de prouver $O \models x$ et $x \models$ F.*

En pratique, on n'est intéressé que par des explications simples.
Il convient de faire plusieurs remarques sur cette définition. Tout d'abord, elle ne donne pas de définition exacte d'explication. C'est intentionel parce que le but est justement de fournir une définition générale qui permette à différents chercheurs de proposer différentes définitions. Nous donnons ci-après une définition basée sur les chroniques mais d'autres définitions seraient possibles. Par ailleurs, la définition est ici donnée pour des problèmes de diagnostic de SÉD mais elle s'applique à n'importe quel type de formalisme. Ensuite, la définition reste informelle sur ce qu'on entend par « simple ». C'est notamment lié au fait que la définition d'explication reste ouverte. C'est aussi parce que les critères précis et spécifiques déterminant ce qui rend une explication « simple » devront être définis au cas par cas. Est-ce la taille de l'explication qui compte ? La taille du modèle qu'il faut considérer ? Le type de raisonnement à appliquer ? etc. Enfin, on notera que la définition présume que le diagnostic est fautif ($\Delta_M(O) = \text{F}$). La raison est qu'un comportement nominal est généralement plus difficile à prouver qu'un comportement fautif parce qu'il implique l'absence d'observations anormales ; la raison pour laquelle un comportement est nominal est que *toutes les observations* restent dans le périmètre décrit par le modèle. [2]
On peut donc voir une explication comme un prédicat satisfait par la trace tel que, d'une part, la vérification que la trace satisfait ce prédicat est facile et, d'autre part, l'occurrence de faute se déduit facilement depuis le prédicat. Ainsi par exemple, un prédicat qui capturerait l'information « J'ai tourné la clef mais le moteur n'a pas démarré » indique, *a priori*, une panne. Dans cet exemple, l'explication ressemble à une chronique, ce qui n'est pas anodin : nous proposons en effet d'utiliser des chroniques pour représenter des explications. La définition est cependant ouverte. Ainsi, si le SÉD est représenté de manière symbolique (à savoir que les états et transitions ne sont pas listées explicitement mais de manière implicite via des variables d'états), l'explication peut inclure des faits (prédicats sur les variables d'états) qui peuvent être dérivés à partir des observations.

3.2 Explications et chroniques

Nous proposons d'instantier la définition 1 avec une chronique typique. L'idée est que la difficulté du diagnostic pour un humain est le plus souvent le volume des observations (taille de la trace). Le rôle de l'explication est d'extraire les observations (ou leur absence) importantes de la trace pour les mettre en avant. Vérifier qu'une trace suit une chronique est facile (d'autant plus que nous demandons que la correspondance soit incluse dans l'explication), et nous partons de l'hypothèse que prouver l'occurrence de la faute à partir d'une chronique typique (si celle-ci est suffisamment petite) est également facile, au moins pour un utilisateur expert.

2. Pour un système diagnosticable, on ne peut généralement prouver la nominalité du comportement que jusqu'à la date actuelle moins d, le délai de diagnosticabilité ; on pourrait donc également définir une explication pour la nominalité du comportement jusqu'à cette date.

Définition 2 *Étant donné un problème de diagnostic $P = \langle M, O\rangle$ où M est un SÉD et O est une trace telle que $\Delta_M(O) = \text{F}$, une* explication à base de chronique *est une paire $x = \langle \theta, t\rangle$ telle que θ est une chronique typique de O et t est une correspondance entre O et θ.*

L'explication comprend donc une chronique typique qui est la condition suffisante pour diagnostiquer la faute ($x \models \text{F}$). De plus, l'explication contient une correspondance entre O et θ qui prouve — et facilite la preuve — que O suit θ ($O \models x$). On dira parfois que θ explique la faute, en ce sens qu'il existe une correspondance t telle que $\langle \theta, t\rangle$ est une explication.

Définition 3 *La* taille *d'une chronique θ, notée $|\theta|$, est le nombre de nœuds dans le graphe. Un* seuil de complexité *est un entier $S \in \mathbf{N}$. Étant donné un seuil de complexité $S \in \mathbf{N}$, la chronique est* simple *si sa taille est au plus S : $|\theta| \leq S$.*

Dans la définition 3, on utilise la taille de la chronique, c'est-à-dire son nombre de nœuds, comme proxy pour la complexité. On pourrait choisir d'autres mesures telles que le nombre de composants qu'elle mentionne. Cette définition reste donc ouverte.
On fait l'hypothèse suivante :

Hypothèse 1 *Une explication à base de chronique $x = \langle \theta, t\rangle$ est simple si θ est simple.*

On note en particulier que, étant donnée la correspondance t entre O et θ, vérifier $O \models x$ est trivial puisqu'il suffit de vérifier que les observations nécessaires sont présentes aux dates indiquées tandis que les observations interdites sont, quant à elles, absentes. La seconde condition, que la preuve $\theta \models \text{F}$ est simple, est loin d'être évidente ; c'est pour cela que nous parlons ici d'« hypothèse ». Cependant, nous considérons qu'une chronique simple devrait naturellement capturer les informations *évidemment* suffisantes pour conclure à une faute. Par ailleurs, en pratique, on pourra imaginer une procédure incrémentale et interactive dans laquelle un utilisateur pourra demander une explication différente ou plus complète. Ceci est une piste de recherche future.
Les explications à base de chronique peuvent être vues comme un cas spécial d'observations critiques [5]. Ces dernières sont définies comme une abstraction de la trace qui permettent de produire le même (minimal) diagnostic, c'est-à-dire que toutes les informations redondantes ou non pertinentes de la trace ont été supprimées. Nous avons choisi ici de privilégier le lien avec les chroniques à cause de l'existence des SRC qui offrent l'explication de manière immédiate sans nécessiter de traitement post-diagnostic.
Exemple *Appliquons les définitions précédentes au problème de diagnostic appliqué au SÉD de la figure 1 (gauche) pour la trace $O = aabcaabab$. Le diagnostic est $\Delta_M(O) = \text{F}$. Une explication pour ce diagnostic est $\langle \theta, f\rangle$ où θ est la chronique de la figure 1 (droite) et $t = \{c \mapsto 4, a \mapsto 5, a \mapsto 6\}$ est la correspondance entre θ et O.*

4 Explicabilité

Nous cherchons maintenant à déterminer si un SÉD est explicable, c'est-à-dire s'il est construit de telle manière que non seulement toute faute sera diagnostiquée mais également que ce diagnostic pourra être expliqué. Si le système n'est pas explicable, on pourra alors ajouter des capteurs qui donneront plus d'options pour générer une explication simple. On pourra aussi modifier le comportement du système (par exemple via ses contrôleurs) de telle manière que celui-ci renvoie des indices permettant d'expliquer le diagnostic.

4.1 SÉD non explicables

Puisque la simplicité d'une explication repose sur sa taille, on définit la notion de complexité d'explicabilité d'une trace comme étant la taille de sa plus petite explication. On ne cherchera pas forcément à trouver cette explication (ou une de ces explications), mais cette notion est utile pour décider l'existence d'une explication simple.

Définition 4 *Étant donnés un SÉD M et une trace O telle que $\Delta_M(O)$ est* F*, la* complexité d'explicabilité *de O est la taille de sa plus petite explication :*

$$comp(O) = \min_{O \models \langle \theta, t \rangle \models \mathrm{F}} |\theta|.$$

Nous illustrons la question d'explicabilité sur un exemple avant de fournir une définition formelle.

Exemple *Nous reprenons l'exemple de la figure 1 mais nous considérons que c n'est plus observable ($\Sigma_\mathrm{o} = \{a, b\}$). La trace $O = aabcaaba$ que nous considérions jusqu'alors est maintenant $O' = aabaaba$. Le diagnostic de cette trace est toujours le même : $\Delta_{M'}(O') =$* F*. La plus simple chronique qu'on puisse construire et qui sert d'explication pour O' est illustrée sur la figure 2. Si le seuil de complexité S est inférieur à 5, alors cette chronique n'est pas simple et il n'y a pas de simple explication pour la trace O'.*

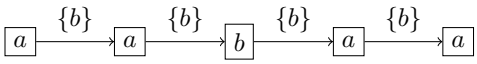


FIGURE 2 – Une chronique non simple pour le système de la figure 1.

Cet exemple nous conduit à la définition suivante :

Définition 5 *Étant donné un seuil de complexité S et un SÉD M diagnosticable avec un délai d, M est (d, S)-explicable si pour tout chemin d-fautif ρ, la complexité d'explicabilité de sa trace est S ou moins.*

Exemple *Selon la définition 5, l'exemple de la figure 1 n'est pas explicable lorsque c n'est pas observable. En réalité, toute trace fautive de M peut être expliquée (et même, étant donné notre définition d'explication, il est toujours possible de créer une explication pour n'importe quel système diagnosticable parce qu'on peut créer une chronique qui corresponde précisément à la trace), mais ces explications peuvent être de longueur rédhibitoire.*

En revanche, dans le même exemple pour $\Sigma_\mathrm{o} = \{a, b, c\}$, le système est explicable. En effet, deux chroniques sont suffisantes pour couvrir tous les cas de fautes : la première est celle de la figure 1 (droite), la seconde est identique si ce n'est que l'événement observable c du premier nœud est remplacé par l'événement a.

Il est possible de prouver que la notion d'explicabilité satisfait des propriétés de monotonicité telles que celle-ci :

Lemme 2 *Soient $M = \langle Q, \Sigma, T, I, \Sigma_\mathrm{o}, \Sigma_\mathrm{F} \rangle$ et $M' = \langle Q, \Sigma, T, I, \Sigma'_\mathrm{o}, \Sigma_\mathrm{F} \rangle$ deux SÉD tels que $\Sigma_\mathrm{o} \subseteq \Sigma'_\mathrm{o}$. Si M est (d, S)-explicable, alors pour tout $d' \geq d$, M' est (d', S)-explicable.*

Le lemme 2 peut se démontrer facilement par le fait que toutes les explications simples trouvées pour M sont également des explications simples pour M'. Augmenter la taille de d ou le nombre d'événements observables ne peut qu'augmenter l'espace des explications.

Notons cependant que le lemma 2 repose sur un certain nombre d'hypothèses implicites. Tout d'abord, il considère que le système est d-diagnosticable. En effet, si le système n'était pas diagnosticable, alors on pourrait faire face à des scénarios bizarres. Ainsi, il serait possible que tous les comportements de M qui sont diagnosticables soient explicables, tandis que certains comportements de M' (qui sont diagnosticables pour M' mais pas pour M) seraient non explicables. Autrement dit, le diagnostiqueur deviendrait plus puissant (capable de diagnostiquer plus de comportements fautifs) mais ne serait pas capable d'expliquer le diagnostic de ces comportements.

Par ailleurs, on notera l'importance du fait que la simplicité de l'explication est basée uniquement sur la taille de la chronique et non pas sur la fonction de correspondance. Et en effet, imaginons une autre définition de la taille de l'explication. Dans cette définition, le but est de pouvoir afficher sur un écran tous les événements observés pendant la chronique. On peut afficher autant d'événements observés que de lignes. La taille d'une explication devient la différence entre l'index du dernier événement mentionné par la fonction de correspondance et celui du premier. On voit alors qu'ajouter des événements à l'ensemble Σ_o *pollue* l'écran et rend l'explication non simple.

Il convient donc de faire attention à la définition d'explication et de simplicité.

Pour finir, nous montrons qu'il existe des exemples pour lesquels il n'y a aucun seuil de complexité permettant de garantir l'explicabilité.

Définition 6 *Pour un SÉD d-diagnosticable, une famille de chemins* divergente *est une séquence infinie de chemin $\rho_1, \rho_2, \ldots$ telle que chaque chemin ρ_i est d-fautif et la complexité d'explicabilité augmente de manière strictement monotone :*

$$\forall i \in \mathbf{N}.\ comp(obs(\rho_i)) < comp(obs(\rho_{i+1})).$$

Exemple *Nous revenons à l'exemple de la figure 1 où $\Sigma_o = \{a, b\}$ et considérons la trace fautive $\rho_n = a(ab)^n aa$. La plus petite chronique pour cette trace est similaire à la figure 2 et comporte $3 + 2n$ nœuds, ce qui signifie que $\rho_1, \rho_2, \dots$ forme une famille de chemins divergente. On conclue donc que, bien que diagnosticable, il n'existe aucun seuil qui rende ce système explicable.*
Puisque la complexité est un entier, une famille de chemins divergente dans un SÉD démontre qu'il n'y a pas de borne maximale pour la plus petite explication d'une trace de ce SÉD.

Lemme 3 *Soient m, ℓ, et d trois entiers. Il n'existe pas de fonction ub tel que pour tout SÉD d-diagnosticable comportant m états et ℓ événements et tout chemin ρ d-fautif, il existe une explication de taille au plus $ub(m, \ell, d)$.*

Ce lemme est prouvé par l'exemple précédent. En revanche, on peut prouver le lemme suivant.

Lemme 4 *Étant donné un SÉD d-diagnosticable comportant m états, si il existe un chemin d-fautif dont l'explication la plus courte a une taille supérieure à 2^m, alors le SÉD a une famille de chemins divergente.*

La preuve du lemme 4 repose sur l'idée suivante. Considérons un chemin ρ et sa trace $obs(\rho)$, et considérons que sa plus courte explication utilise la chronique θ de taille supérieure à 2^m. On peut d'ailleurs supposer que les arêtes de la chronique définissent un ordre total sur les nœuds du graphe parce que les chroniques basées sur des ordres strictement partiels peuvent être interprétées comme des unions de chroniques totalement ordonnées. De manière générale, il est possible d'associer à chaque nœud v_j d'une chronique un ensemble de croyance $\mathcal{Q}$ qui représente l'ensemble des états dans lequel le système pourrait se trouver étant donné le préfixe de ce nœud. Si le même ensemble de croyance $\mathcal{Q}$ est associé à de nœuds v_j, v_k, alors le raisonnement a effectué une boucle. Il est possible de modifier ρ pour dupliquer cette boucle de telle manière que la chronique θ n'est plus appropriée mais qu'il faille la remplacer par une autre chronique θ' plus longue (dans laquelle la boucle a été dupliquée le même nombre de fois). Il est possible qu'une autre chronique de longueur identique existe qui explique à la fois ρ et ρ', mais si on applique le même raisonnement et si on duplique la boucle dans ρ' pour chacune des chroniques, on obtient invariablement après un nombre fini de modifications un nouveau chemin ρ^k dont la complexité d'explicabilité est supérieure à celle de ρ. En appliquant cette approche récursivement, on génère une famille de chemins divergente.
Ce lemme démontre donc qu'on peut vérifier si le SÉD contient une famille de chemins divergente en vérifiant si l'ensemble des chroniques typiques de taille 2^m couvre tous les chemins d-fautifs, même si la taille des chroniques rend impossible cette vérification en pratique.

4.2 Explicabilité et viabilité du SRC

On peut à présent définir le problème d'explicabilité.

Définition 7 *Étant donnés une valeur d, un SÉD M qui est d-diagnosticable et un seuil de complexité S, le problème d'explicabilité consiste à déterminer si M est (d, S)-explicable.*

Notre définition d'explicabilité repose sur les chroniques. À supposer que l'on considère que le seuil S est bas, alors le nombre de chroniques qu'on peut construire est aussi faible. Essentiellement, il est de l'ordre de $\ell^S \times (2^\ell)^S$ où le premier terme indique le nombre possible de combinaisons d'événements nécessaires tandis que le second terme fait référence aux événements interdits. En pratique cependant, nous pensons que le nombre de chroniques dont un SRC aura besoin pour diagnostiquer correctement le système sera beaucoup plus faible. (Ou alors, on pourra ajuster la définition de chronique. Par exemple, l'assertion « l'événement observable e est suivi de k événements observables différents de e' » ne s'exprime pas facilement avec les chroniques parce que un nœud est associé avec un seul événement ; à la place, on pourrait associer les nœuds avec un ensemble d'événements pour éviter ce cas de figure.)
Ceci nous autorise à produire le résultat suivant :

Lemme 5 *En première approximation, si un SÉD est explicable pour un seuil de complexité faible, alors le diagnostic à base de chronique est une approche viable pour le diagnostic de ce SÉD.*

Nous considérons que ce résultat est très intéressant parce qu'il révèle une propriété intéressante du diagnostic par chronique : à savoir que non seulement il produit de manière naturelle des diagnostics explicables, mais que pour tout problème explicable, l'approche par chronique est applicable. Ainsi, si l'explicabilité est une contrainte utilisateur forte, il est possible d'utiliser un SRC pour implémenter le diagnostic. Étant donné la simplicité des SRC, c'est un résultat très encourageant.

4.3 Vérifier l'explicabilité

Nous souhaitons maintenant vérifier si un système est explicable et, par la même occasion, générer un ensemble de chroniques typiques qui couvrent tous les cas fautifs prédits par le modèle.
Tout d'abord, déterminer si une chronique est typique est relativement simple à implémenter. En effet, une chronique représente un langage, le langage de toutes les séquences d'événements observables qui suivent cette chronique. Ce langage est régulier et facile à représenter. Vérifier si la chronique est typique consiste à effectuer un produit synchrone de ce langage avec le modèle pour vérifier s'il existe un chemin autorisé par le modèle qui suit la chronique mais qui ne comporte pas de faute.
De même, étant donnés un modèle et un ensemble de chroniques, on peut, avec des opérations classiques sur les automates, chercher un chemin qui ne suit aucune de ces chroniques et qui est d-fautive.
En combinant ces deux routines, on peut vérifier l'explainabilité du SÉD en suivant la procédure 1. Elle consiste à chercher des chemins d-fautifs et vérifier s'ils sont explicables

simplement. À chaque étape de l'algorithme, le chemin d-fautif généré doit ne pas être explicable par les chroniques déjà calculées. L'algorithme s'arrête quand il n'arrive pas à expliquer le chemin d-fautif actuel (le SÉD n'est pas explicable) ou s'il ne trouve plus de chemin d-fautif (le SÉD est explicable et on a trouvé un ensemble couvrant de chronique).
Par ailleurs, puisque le nombre de chroniques simples est borné, l'algorithme termine en un temps fini.

Procédure 1 Vérifier l'explainabilité d'un modèle

entrées : SÉD $M = \langle Q, \Sigma, T, I, \Sigma_o, F\rangle$, seuil de complexité S, délai d
sorties : `Explicable` et une liste de chroniques qui couvrent les chemins d-fautifs ou `Non explicable` et un chemin d-fautif non explicable
$\Theta = \emptyset$
boucle
  $\rho :=$ `chercher_chemin`(M, Θ, d)
  si $\rho = \bot$ **alors renvoyer** `Explicable`, Θ
  fin de si
  $\langle \theta, t\rangle :=$ `expliquer`(M, ρ, S)
  si $\theta = \bot$ **alors renvoyer** `Non explicable`, ρ
  fin de si
  $\Theta := \Theta \cup \{\theta\}$
fin de boucle

4.4 Calculer une chronique expliquant une trace

Étant donné une trace, calculer une chronique expliquant celle-ci est une tâche complexe. Nous avons déjà expliqué que toute trace peut être transformée en une chronique, mais celle-ci ne sera probablement pas simple.
À partir d'une chronique, on peut définir un espace de recherche consistant en un ensemble partiellement ordonné de chroniques. Deux chroniques $\theta_1 \preceq \theta_2$ ordonnées dans cet espace de recherche sont telles que θ_1 est une « abstraction » de θ_2, c'est-à-dire que toute séquence d'événements observables qui suit θ_1 suit également θ_2. Plus une chronique est abstraite, plus elle est simple (selon notre définition). D'autre part, on peut facilement prouver des résultats de monotonicité : si θ_2 n'est pas typique, alors θ_1 ne l'est pas non plus.
Dès lors, on peut explorer cet espace de recherche pour trouver une chronique typique simple. C'est une tâche coûteuse du point de vue calculatoire ; nos implémentations actuelles ne fonctionnent que pour des problèmes ridiculement petits. Nous proposons donc quelques pistes pour améliorer les temps de calcul :
D'une part, on souhaitera généralement partir d'une trace aussi petite que possible car l'espace de recherche sera d'autant plus petit.
Ensuite, il faudra vraisemblablement tâcher de développer des algorithmes incrémentaux : comme nous l'avons indiqué plus haut, vérifier qu'une chronique est typique requiert un produit synchrone avec le modèle. Étant donné un tel produit synchrone pour une chronique, le produit pour une autre chronique similaire comportera pour la majeure partie les mêmes états. Une procédure intelligente devrait pouvoir identifier quelles branches du produit ne nécessitent pas d'être réexplorées.
Par ailleurs, une modélisation décentralisée du système devrait permettre une étude de l'impact des observations sur le diagnostic de la faute. Les diagnostiqueurs dits dynamiques par exemple, [18, 4], identifient quels capteurs peuvent être éteints puis doivent être rallumés durant la surveillance du système. De même, si l'on est capable de déterminer qu'une observation est sans importance dans un contexte donné, on peut immédiatement l'ignorer.
D'autre part, nous prédisons que des techniques d'apprentissage automatique devraient permettre de prédire les régions de l'espace de recherche les plus intéressantes. Par exemple, on pourra chercher à générer un large nombre traces fautives et non fautives et effectuer des études statistiques pour déterminer quelles observations sont les plus fréquentes aux alentours de l'occurrence de la faute. Ces variables sont vraissemblablement impliquées dans la détection des fautes. Toute expertise extérieure pourra aider à déduire l'espace de recherche.
Enfin, nous notons que cette étape peut échouer. Ainsi, si notre algorithme est incapable de calculer une explication simple pour une trace donnée alors que celle-ci existe, on peut juste renvoyer un échec et conseiller à l'utilisateur de modifier le système pour le rendre plus explicable quand bien même ce n'est pas strictement indispensable.

5 Travaux similaires

À notre connaissance, nous sommes les premiers à proposer une définition formelle d'explication et d'explicabilité dans le cadre du diagnostic à base de modèle, même si Christopher et Grastien [5] avaient posé des jalons avec la théorie des observations critiques qui se propose d'extraire les observations importantes dans un problème de diagnostic. En comparaison, nous proposons une définition plus générique d'explication dans laquelle d'autres informations pourraient être ajoutées telle que certaines inférences intermédiaires.
Bertoglio et coauteurs [2] proposent de renvoyer une explication qui consiste à renvoyer l'ordre dans lequel les fautes sont censés avoir eu lieu, y compris si une faute a eu lieu plusieurs fois.
Des techniques de diagnostic ont été développées en IA connexionniste. Elles permettent de mettre en lumière les observations qui ont contribué le plus au diagnostic. Ainsi, Belikov et coauteurs ainsi que Brito et coauteurs [1, 3] utilisent Grad-CAM [16] tandis que Jang et coauteurs [9] se basent sur SHAP [10]. Comparées à notre travail, ces techniques manquent de garanties formelles. On peut aussi se demander comment les comportements négatifs (absence de certaines observations) peuvent être capturés par ces méthodes.
Notre approche se situe dans le cadre formel de type explication abductive [11]. Les explications y sont comprises

comme une fraction de l'entrée au processus de décision / classificateur qui garantit (quelques soient les autres entrées) la sortie effectivement observée. Comparées aux approches existantes, les entrées d'un SÉD contiennent des informations implicites sous la forme d'absence d'observation qu'il convient donc de gérer de manière spécifique.

Il existe des travaux pour l'apprentissage automatique de chroniques [12, 8, 17]. Nous pensons que les résultats présentés ici pourraient relancer cette ligne de travail.

6 Conclusion

Dans cet article, nous avons proposé la première définition d'explication de diagnostic. Une explication est un objet mathématique devant permettre à un utilisateur d'inférer le même résultat que le diagnostiqueur. Nous montrons qu'il est possible d'utiliser les chroniques pour représenter cette explication. Nous montrons également que tout système explicable peut, en première approximation, être diagnostiqué par un système de reconnaissance de chronique.

Nous considérons que ces travaux sont fondateurs pour les explications en diagnostic. Parmi les travaux futurs, il sera d'abord nécessaire de développer des techniques plus efficaces pour le calcul d'explications / chroniques. Par exemple, nous souhaitons être capable d'analyser le système pour rapidement trouver des candidats d'explication. Des techniques similaires ont été développées pour prouver la diagnosticabilité de grands systèmes [14].

Nous souhaitons également développer plus de types d'explications. Quelles informations seraient utiles pour un utilisateur ? Lorsque le diagnostiqueur mentionne telle ou telle observation, il y a peut-être une raison concrète qu'il serait utile d'expliciter. Par exemple, l'explication pourrait avoir la forme : « la lumière est allumée, ce qui prouve que la ligne est électrifiée ». Dans cet exemple, la seconde clause n'est pas informative au sens de la théorie de l'information (c'est une conséquence de la première clause), mais mentionner ce fait permet de comprendre que toute autre information qu'on pourrait déduire de la première clause peut être ignorée.

Enfin, il serait nécessaire d'étudier les critères précis qui rendent une explication simple ou qui permettent à un utilisateur de détecter une erreur de la part du diagnostiqueur. Cette question devra être résolue en conjonction avec les sciences sociales.

Références

[1] J. Belikov, M. Meas, R. Machlev, A. Kose, A. Teplja-kov, L. Loo, E. Petlenkov, and Y. Levron. Explainable AI based fault detection and diagnosis system for air handling units. pages 271–279, 2022.

[2] N. Bertoglio, G. Lamperti, M. Zanella, and X. Zhao. Explanatory monitoring of discrete-event systems. In *Intelligent Decision Technologies : Proceedings of the 12th KES International Conference on Intelligent Decision Technologies (KES-IDT 2020)*, pages 63–77. Springer, 2020.

[3] L. C. Brito, G. A. Susto, J. N. Brito, and M. A. V. Duarte. Fault diagnosis using explainable AI : A transfer learning-based approach for rotating machinery exploiting augmented synthetic data. *Expert Systems with Applications*, 232 :120860, 2023.

[4] F. Cassez and S. Tripakis. Fault diagnosis with static or dynamic diagnosers. *Fundamenta Informaticae (FI)*, 88(4) :497–540, 2008.

[5] C. J. Christopher and A. Grastien. Critical observations in model-based diagnosis. *Artificial Intelligence*, 331 :104116, 2024.

[6] C. Dousson and P. Le Maigat. Chronicle recognition improvement using temporal focusing and hierarchization. In *20th International Joint Conference on Artificial Intelligence (IJCAI-07)*, pages 324–329, 2007.

[7] Christophe Dousson. *Suivi d'évolutions et reconnaissance de chroniques*. PhD thesis, Toulouse, 1994.

[8] B. Guerraz and C. Dousson. Chronicles construction starting from the fault model of the system to diagnose. In *Fifteenth International Workshop on Principles of Diagnosis (DX-04)*, pages 51–56, 2004.

[9] K. Jang, Karl K. Pilario, N. Lee, I. Moon, and J. Na. Explainable artificial intelligence for fault diagnosis of industrial processes. *IEEE Transactions on Industrial Informatics*, 232 :1–8, 2023.

[10] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems (NeurIPS-17)*, 30, 2017.

[11] J. Marques-Silva and A. Ignatiev. Delivering trustworthy AI through formal XAI. In *36th Conference on Artificial Intelligence (AAAI-22)*, 2022.

[12] E. Mayer. Inductive learning of chronicles. In *Thirteenth European Conference on Artificial Intelligence (ECAI-98)*, pages 471–472, 1998.

[13] Y. Pencolé and M.-O. Cordier. A formal framework for the decentralised diagnosis of large scale discrete event systems and its application to telecommunication networks. *Artificial Intelligence (AIJ)*, 164(1–2) :121–170, 2005.

[14] Y Pencolé. Diagnosability analysis of distributed discrete event systems. In *Fifteenth International Workshop on Principles of Diagnosis (DX-04)*, pages 173–178, 2004.

[15] M. Sampath, R. Sengupta, S. Lafortune, K. Sinnamohideen, and D. Teneketzis. Diagnosability of discrete-event systems. *IEEE Transactions on Automatic Control (TAC)*, 40(9) :1555–1575, 1995.

[16] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM : Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision (ICCV-17)*, pages 618–626, 2017.

[17] A. Subias, L. Travé-Massuyès, and E. Le Corronc. Learning chronicles signing multiple scenario instances. *IFAC Proceedings Volumes*, 47(3) :10397–10402, 2014. Nineteenth IFAC World Congress.

[18] D. Thorsley and D. Teneketzis. Active acquisition of information for diagnosis and supervisory control of discrete event systems. *Journal of Discrete Event Dynamical Systems (JDEDS)*, 17 :531–583, 2007.

[19] M. Zanella and G Lamperti. *Diagnosis of active systems*. Kluwer Academic Publishers, 2003.

Une IA hybride pour la surveillance de la santé du système de lubrification et de refroidissement de la BTP d'un hélicoptère

Ammar Mechouche, Matthis Houles, Jérôme Belmonte et Pierre-Loic Maisonneuve

Airbus Helicopters, 13700 Marignane, France

Ammar.Mechouche@Airbus.com

Résumé

Ce résumé présente une approche combinant l'apprentissage automatique et l'intelligence artificielle symbolique pour la surveillance du système de lubrification et de refroidissement de la boîte de transmission principale (BTP) d'un hélicoptère. Les modèles d'apprentissage automatique détectent les anomalies, tandis que l'IA symbolique intervient pour en diagnostiquer l'origine ou les invalider à l'aide de règles floues définies par des experts métier. Les résultats obtenus sont prometteurs et ouvrent la voie à un déploiement de la solution dans Flyscan, un service de maintenance prédictive proposé aux clients d'Airbus Helicopters.

Mots-clés

IA hybride, surveillance santé, hélicoptère.

Abstract

In this abstract we present an approach combining machine learning and symbolic artificial intelligence for monitoring the lubrication and cooling system of a helicopter's main gearbox (MGB). Machine learning models detect anomalies, while symbolic AI diagnoses their origin or invalidates them using fuzzy rules defined by domain experts. The results are promising and pave the way for deploying the solution in Flyscan, a predictive maintenance service offered to Airbus Helicopters customers.

Keywords

Hybrid AI, health monitoring, helicopter.

1 Introduction

La surveillance de l'état de santé des systèmes dynamiques des hélicoptères est un enjeu majeur pour améliorer la sécurité des vols, optimiser la disponibilité des appareils et réduire les coûts de maintenance. Traditionnellement, cette surveillance repose sur l'analyse des signaux vibratoires captés au plus près des composants mécaniques. Bien que ces techniques de traitement du signal soient éprouvées et offrent des résultats fiables, elles ne permettent pas de couvrir l'ensemble des systèmes critiques. Pour pallier cette limitation, d'autres sources de données, collectées en vol via le système HUMS (Health & Usage Monitoring System), peuvent être exploitées afin d'enrichir et d'améliorer la surveillance.

Dans ce résumé, nous proposons une approche combinant l'apprentissage automatique et l'intelligence artificielle symbolique pour surveiller le système de refroidissement et de lubrification de la BTP d'un hélicoptère. Cette solution, détaillée en section 3, vise à améliorer la détection précoce d'anomalies et à renforcer la maintenance prédictive.

Les résultats obtenus sont prometteurs et ouvrent la voie à un déploiement dans Flyscan, le service de maintenance prédictive d'Airbus Helicopters. Une version détaillée de ce système de surveillance a été présentée à l'European Rotorcraft Forum 2023 [1] et s'inscrit dans la continuité des travaux publiés dans [2]. L'originalité de cette approche réside dans l'association de modèles de normalité construits par apprentissage automatique, de règles de détection établies par les experts métiers et de la logique floue pour le suivi de l'état de santé d'un système complexe.

2 Description de la méthode

Le système de lubrification et de refroidissement joue un rôle essentiel dans le bon fonctionnement de la BTP (Figure 1). Il est composé de deux circuits pressurisés, principal et secondaire, permettant d'assurer la lubrification des pièces en contact pour la transmission de puissance, l'évacuation de la chaleur générée via le radiateur, ainsi que l'élimination des polluants au niveau du filtre à huile. La solution développée exploite les capteurs de pression et de température d'huile situés à différents endroits du système (éléments 1, 2, 3, 6 et 7 sur la Figure 1).

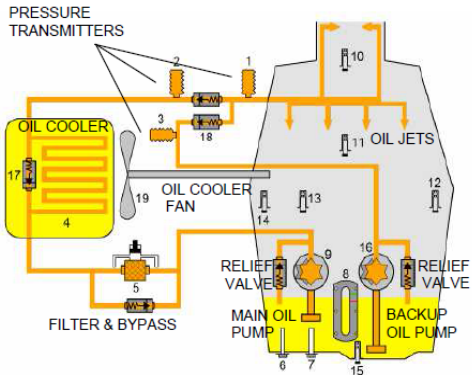


Figure 1 : Système de lubrification et de refroidissement de la BTP d'un hélicoptère.

La solution développée repose sur deux étapes principales.

2.1 Détection des anomalies

La première étape consiste à construire des modèles de normalité grâce à des algorithmes d'apprentissage automatique (Forêt d'arbres aléatoires) exploitant les vastes quantités de données de vol collectées (de l'ordre de dizaines de milliers d'heures de vol). Pour garantir la robustesse des modèles, les vols présentant des anomalies avérées ont été exclus de la base d'apprentissage. De plus, certaines règles métier ont été appliquées afin de limiter l'apprentissage aux phases de vol thermiquement stables, propices à l'estimation de valeurs de normalité.
Après chaque vol client – et uniquement durant ces phases spécifiques – les valeurs nominales des capteurs de pression et de température d'huile sont estimées à partir d'un ensemble de paramètres de vol collectés par le Flight Data Continuous Recorder (FDCR) (altitude, vitesse, puissance motrice, température extérieure, etc. - rectangles bleus en haut de la Figure 2). Ces valeurs inférées sont ensuite comparées aux mesures réelles des capteurs physiques. Trois situations peuvent alors être identifiées : les mesures sont conformes aux prédictions ; les mesures sont significativement supérieures ou inférieures aux valeurs attendues : les paramètres sont jugés anormalement élevés ou anormalement bas, respectivement. Afin de gérer les incertitudes, un formalisme de logique floue est employé. Chaque paramètre de pression et de température se voit ainsi attribuer un score de normalité compris entre 0 et 1, reflétant son état (normal, anormalement élevé ou anormalement bas). Cette première étape permet ainsi de détecter les anomalies dans le système.

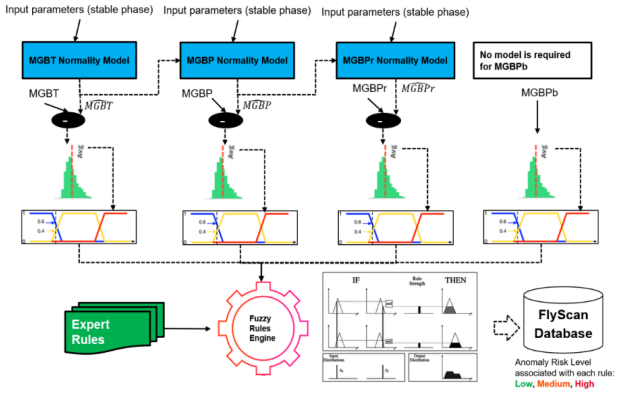


Figure 2 : Vue globale de la solution développée pour la surveillance du système.

2.2 Diagnostic des anomalies

La seconde étape vise à déterminer l'origine de l'anomalie détectée ou, le cas échéant, à invalider une alerte erronée issue des modèles d'apprentissage automatique. Pour ce faire, un ensemble de règles de diagnostic, définies par les experts métier, est appliqué. Ces règles permettent d'identifier avec précision le mode de panne en question ainsi que le sous-système concerné, en s'appuyant sur l'état de normalité de chaque paramètre évalué lors de l'étape précédente. Par exemple, l'une de ces règles s'énonce comme suit : **SI** la pression d'huile principale (capteur 2) est anormalement élevée **ET** la pression de secours (capteur 3) est normale **ET** la pression à la rampe (capteur 1) est anormalement élevée **ET** la température d'huile au fond de la boîte (capteur 6) est normale, **ALORS** l'anomalie est due au colmatage des gicleurs d'huile. Une douzaine de modes de panne du système sont ainsi surveillés grâce à ces règles, qui sont exécutées par un moteur d'inférence flou. Ce dernier attribue à chaque diagnostic un niveau de criticité (bas, moyen ou élevé). Le support HUMS vérifie ces diagnostics et décide d'alerter ou non le client.

3 Résultats

Le système décrit est implémenté et testé sur des données réelles. La Table 1 décrit les résultats d'évaluation, plutôt satisfaisants, des modèles de normalité de la pression principale (MGBP), température principale (MGBT) et la pression à la rampe (MGBPr). La normalité de la pression de secours étant simple à gérer sans modèle de normalité.

Model / Metrics	R2 (%)	RMSE
MGBP	74.4	0.27 (bar)
MGBT	85.1	3.89 (°C)
MGBPr	96.0	0.10 (bar)

Table 1 : Résultats de l'évaluation du système.

La solution dans la globalité, incluant les règles de l'expert, a été testée sur des données réservées pour le test et comprenant un historique de plusieurs milliers de vols d'hélicoptères pour lesquels, par ailleurs, les données de maintenance, d'incidents et de révisions étaient disponibles. Ainsi, pour chaque anomalie détectée par le système, il est vérifié si, aux alentours de la date d'occurrence de l'anomalie, une action de maintenance a été effectuée par le client sur le système, ou si un incident incriminant le système a été enregistré ou si une dégradation a été constatée sur le système lors de sa révision. Sur **14** hélicoptères et **62000** heures de vol, le système a pu détecter l'anomalie dans **7** cas sur **9**, un cas ambigu et seulement un cas non détecté.

4 Suite des travaux

La suite des travaux se focalise d'une part sur l'amélioration des modèles de normalité par l'utilisation de nouveaux algorithmes d'IA. D'autre part, ces travaux abordent les aspects de confiance dans les modèles d'IA auprès des professionnels du support HUMS.

5 Références

[1] Mechouche A., Houles M., Belmonte J., Maisonneuve P.-L., "Monitoring of MGB Lubrication and Cooling System based on Big Data Normality Models and Fuzzy Expert Rules", European Rotorcraft Forum, 2023.

[2] Daouayry N., Mechouche A., Maisonneuve P.L., Scuturici V.M., Petit J.M., "Data-centric helicopter failure anticipation: The MGB oil pressure virtual sensor case", IEEE International Conference on Big Data, 2019.

Session 4 : Grands modèles de langage

Génération d'une base de courriers électroniques synthétiques par des grands modèles de langue dans le domaine de la relation client

Fatma-Zohra Hannou[1], Isabelle Renault[1], Florent Mely[2], Anne-Laure Guénet[3], Guillaume Dubuisson Duplessis[3], Sabrina Campano [1]

[1] EDF Lab Paris Saclay, SEQUOIA
[2] AI&Data
[3] EDF Commerce, Direction des Systèmes d'Information et du Numérique (DSIN)

fatma-zohra.hannou@edf.fr, isabelle.renault@edf.fr, anne-laure.guenet@edf.fr, guillaume.dubuisson-duplessis@edf.fr, sabrina.campano@edf.fr

Résumé

Dans le domaine de la relation client, exploiter les données textuelles offre un levier essentiel pour développer des systèmes d'IA performants, mais présente d'importants défis de confidentialité et de conformité réglementaire, et nécessite souvent l'annotation manuelle et coûteuse de données. Cet article décrit une approche d'utilisation des grands modèles de langue pour générer des e-mails synthétiques, sans intégrer de données client réelles. L'objectif est de permettre d'entrainer des systèmes d'IA sur ces données synthétiques, en offrant de meilleures garanties de protection de la vie privée, et en permettant de minimiser le volume de données manuellement annotées. Nous décrivons la chaîne de traitement et son implémentation notamment la phase de création de prompts, qui capture la diversité des sujets, des styles et les types de données personnelles. Au-delà de la génération, l'insertion d'entités fictives dans le texte permet de reformer un email automatiquement annoté similaire à un email réel. Les résultats d'évaluation sur un jeu de données de 1600 emails indiquent une piste prometteuse pour l'entraînement de systèmes d'IA tout en offrant de meilleures garanties du point de vue du respect de la vie privée et de la conformité réglementaire.

Mots-clés

Grands modèles de langue, génération de données synthétiques, ingénierie du prompt, RGPD, protection des données à caractère personnel, reconnaissance d'entités nommées.

Abstract

Training AI systems for customer relations tasks is hindered by the high costs of manual data annotation and stringent privacy regulations. This article describes an approach using large language models to generate synthetic emails without incorporating real customer data. The aim is to train AI systems on these synthetic data, providing better privacy guarantees and minimizing the volume of manually annotated data. We describe the processing pipeline and its implementation, particularly the prompt creation phase, which captures the diversity of topics, styles, and types of personal data. Beyond generation, inserting fictitious entities into the text allows for the automatic annotation of an email similar to a real one. Evaluation results on a dataset of 1,600 emails indicate a promising approach for training AI systems while offering better guarantees in terms of privacy and regulatory compliance.

Keywords

Large language models, synthetic data generation, prompt engineering, GDPR, protection of personal data named-entity recognition.

1 Introduction

La relation client d'un grand Groupe comme EDF produit un volume important de données textuelles. Ces données, issues d'e-mails, de commentaires de satisfaction ou encore de conversations téléphoniques retranscrites se caractérisent par leur diversité et leur variété. Elles peuvent être courtes et spontanées ou, au contraire, longues et structurées, présentant une hétérogénéité en termes de style, de ton et de niveau de langue, soulevant ainsi de véritables défis pour les explorer efficacement [7]. En particulier, les e-mails, souvent rédigés en français, sont utilisés pour répondre au mieux aux attentes de nos clients en suivant le cadre réglementaire du « règlement général sur la protection des données » (RGPD) [6]. L'analyse de ces données textuelles par des techniques d'intelligence artificielle (IA) permet d'améliorer la qualité du service client, de personnaliser les interactions et d'optimiser les processus. Par exemple, l'IA peut être utilisée pour automatiser le traitement des e-mails (par exemple le routage vers le bon service [8]) ou encore détecter les irritants et les motifs de satisfaction [25].

Cependant, l'utilisation de données réelles pour entraîner ces systèmes d'IA soulève des défis importants en matière de protection de la vie privée, dans le contexte du RGPD

et dans le cadre d'une démarche *privacy-by-design*, visant à intégrer la protection des données dès la conception des systèmes d'IA. Une approche classique pour répondre à ces enjeux consiste à désidentifier les données réelles en supprimant ou en masquant les informations identifiantes, notamment à l'aide de techniques de reconnaissance d'entités nommées (NER) [6]. Or, l'entraînement de ces modèles de désidentification nécessite lui-même l'utilisation de données réelles pour être performant et robuste.

Dans ce contexte, l'essor des grands modèles de langue (LLM) [1, 9] a ouvert de nouvelles perspectives pour la génération de données synthétiques textuelles, offrant une alternative à des situations où la disponibilité des données est un frein. En effet, les LLM peuvent générer des données textuelles réalistes et variées [3, 12], permettant ainsi d'éviter l'utilisation de données personnelles des clients présentes dans les systèmes d'information privés des entreprises.

Cependant, cette approche s'accompagne de plusieurs défis. Au-delà du risque d'hallucinations (génération de contenu incorrect ou impertinent), il est généralement difficile de contraindre les LLM à produire des données variées conformes à des caractéristiques observées dans les données réelles telles que les contextes relatifs à des sujets de mails, le style de rédaction, ou l'orthographe [18].

Cet article explore le potentiel des LLM pour la génération de données synthétiques anonymes dans le domaine de la relation client. La chaine de traitement présentée permet de créer une base d'e-mails synthétiques en utilisant les LLM, et sans accéder à des données client du Groupe EDF. Elle repose notamment sur une phase de conception du prompt (*prompt design*) qui guide la génération au moyen de caractéristiques descriptives des données réelles. La base générée se décline en deux versions : la première contient des textes de mails désidentifiés qui font uniquement référence à des types d'entité (comme nom, prénom, adresse), et la deuxième version instancie dans chaque e-mail les types d'entités qu'il mentionne par des données de clients fictifs générées automatiquement. Cette deuxième version de la base fournit un corpus dont les entités sont annotées par le LLM, ce qui augmente son utilité pour les tâches d'entraînement. Nous analysons ensuite la qualité des données générées (1600 mails), en évaluant leur coût, leur diversité et leur utilité pour des applications concrètes. Conscients qu'il est difficile de qualifier d'anonyme l'algorithme de génération d'une part, les données synthétiques d'autre part [22, 14], nous analysons ce processus de génération au travers du prisme des risques liés à l'utilité, à la réidentification, et à l'exactitude. Enfin, nous discutons du caractère anonyme ou non des jeux de données générées, en tenant compte des spécificités du domaine de la relation client et des exigences du RGPD.

Cet article s'articule autour de quatre sections principales. La Section 2 présente un état de l'art sur les principaux travaux d'anonymisation et l'usage des LLM pour la génération de données synthétiques. La Section 3 détaille la chaîne de traitement implémentée pour la génération des e-mails synthétiques. La Section 4 expose les résultats obtenus et présente l'évaluation de la qualité des données synthétiques et les coûts liés à leur génération. Ces résultats sont discutés en Section 5, puis la Section 6 dresse les principales conclusions et quelques perspectives prometteuses.

2 Travaux antérieurs

2.1 Techniques de désidentification

La désidentification utilisée pour pseudonymiser voire anonymiser les données personnelles est une question centrale. La pseudonymisation est définie comme « ...un traitement de données personnelles réalisé de manière à ce qu'on ne puisse plus attribuer les données relatives à une personne physique sans information supplémentaire. », l'anonymisation comme « ...un traitement qui consiste à utiliser un ensemble de techniques de manière à rendre impossible, en pratique, toute identification de la personne par quelque moyen que ce soit et de manière irréversible. » [4]. Différentes techniques de désidentification des données existent. Majeed et al. [15] distinguent les modèles de désidentification de texte basés sur le traitement du texte de ceux associés à la publication de données préservant la vie privée (PPDP). Les premiers détectent puis suppriment les entités sensibles à l'aide de règles, d'apprentissage automatique ou encore de réseaux neuronaux Ils nécessitent de disposer de données annotées dont le coût de production reste élevé. Les seconds introduisent du bruit pour masquer les données identifiantes et réduire le risque de divulgation. Zhao et Chen [27] abordent le compromis confidentialité - utilité des données des modèles PPDP, et soulignent la nécessité d'évaluer les risques de réidentification.

2.2 Génération de données synthétiques

Les LLM ouvrent de nouvelles perspectives avec la génération de données synthétiques à l'aide d'instructions (prompts). Le contrôle de la conformité des sorties des LLM à ces instructions fait l'objet de nombreux travaux de recherche. Sahoo et al. [20] proposent un état de l'art des techniques d'ingénierie de prompts en fonction des domaines d'application (réalisation de nouvelles tâches sans nouvel apprentissage, réduction des hallucinations...). Sahoo et al. [5] se focalisent sur l'apprentissage en contexte (ICL). Long et al. [12] proposent une vision unifiée des travaux menés sur la génération de textes, la curation et l'évaluation des textes générés. Ils soulignent que les principaux défis à relever restent de garantir la fidélité des contenus des textes générés et de s'assurer de leur diversité.

L'utilisation de données synthétiques fait l'objet de nombreux travaux, notamment dans le domaine médical. Différentes études montrent des gains à finetuner des modèles basés sur une architecture *transformer* comme BERT [1], RoBERTa [2], BioBERT [3] ou encore ClinicalBERT [4] pour des

1. Hugging Face - Modèle BERT : https://huggingface.co/docs/transformers/model_doc/bert
2. Hugging Face - Modèle RoBERTa : https://huggingface.co/docs/transformers/model_doc/roberta
3. BioBERT : https://github.com/dmis-lab/biobert
4. ClinicalBERT : medicalai/ClinicalBERTÃůHuggingFace

tâches aval telles que l'extraction d'information et la détection de symptômes à l'aide de jeux de données incluant des données générées par des LLM. Li et al. [10] explorent deux approches d'augmentation de données de dossiers médicaux électroniques à l'aide de LLM : l'annotation de jeux de données publics et la génération de dossiers médicaux fictifs. Tang et al. [23] utilisent des données synthétiques étiquetées générées par un LLM à l'aide de prompts sélectionnés lors d'un processus itératif intégrant une évaluation humaine. Chung et al. [3] adoptent une démarche alliant stratégie de paramétrage des LLM et interactions humaines pour la correction d'étiquettes afin d'accroître la diversité des textes générés. Ils montrent que cette correction augmente de 14,4% la précision absolue de modèles BERT entrainés avec leurs jeux de données diversifiés.

Plusieurs stratégies ont également été conçues pour contrôler les sorties des LLMs. Väth et al. [24] proposent une appproche de génération de données synthétiques par des LLMs basée sur le parcours d'un arbre de dialogue et constatent l'apport de cette structure pour l'entraînement d'agents d'apprentissage par renforcement. [26] proposent *KnowGPT* un système permettant de créer des prompts enrichis avec des informations issues de graphes de connaissances pour une meilleure contrôlabilité des LLMs.

L'utilisation de données créer des données textuelles synthétiques n'est cependant pas synonyme d'anonymisation. Ainsi, Staab et al. [21] montrent des risques de réidentification, par les LLMs, de données personnelles incluses dans leurs données d'entraînement y compris lors d'inférences réalisées à partir de textes anonymisés. Par ailleurs, Stadler et al. [22] montrent empiriquement que des données tabulaires synthétiques ("différentiellement privées") produites par des modèles génératifs n'offrent pas un meilleur compromis entre confidentialité et utilité que celles issues de techniques d'anonymisation traditionnelles.

L'utilisation des LLMs pour synthétiser des données, notamment dans le cadre de l'entraînement ou de l'amélioration d'autres modèles d'IA, pose plusieurs défis. Il est crucial de vérifier si la licence du LLM permet la réutilisation de ses sorties et de s'assurer que la propriété intellectuelle est respectée. Cela implique de garantir que le LLM n'a pas été entraîné sur des données soumises à des restrictions de propriété intellectuelle. La traçabilité des données utilisées pour l'entraînement est essentielle pour assurer la conformité, mais les documentations actuelles manquent souvent de transparence [11] (le site web "opening up ChatGPT" [5] étudie l'ouverture des LLMs). Conscients de ces manques, Longpre et al. [13] ont réalisé avec l'aide d'experts juridiques et d'experts en apprentissage automatique, un audit sur plus de 1800 jeux de données textuelles. Ils montrent des omissions de licence de plus de 70% et des taux d'erreurs de plus de 50% sur les sites d'hébergement de données largement utilisés, et proposent une interface interactive [6] pour permettre de retracer la provenance des jeux de données de finetuning open-source les plus populaires.

5. https://opening-up-chatgpt.github.io/
6. Data Provenance Explorer : https://www.dataprovenance.org/

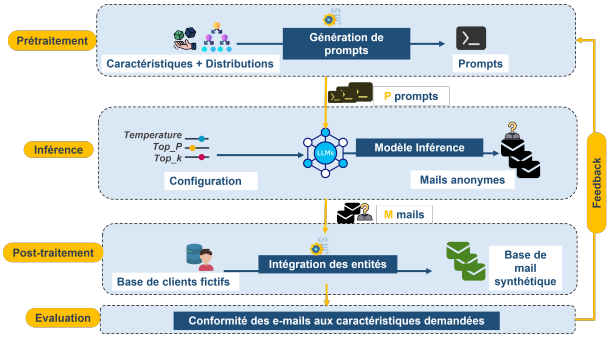


FIGURE 1 – Chaîne de traitement implémentée pour la création des e-mails synthétiques. Le processus comporte 4 phases : le **prétraitement** pour la conception de prompts diversifiés capturant les caractéristiques souhaitées, **l'inférence** utilisant des LLMs pour la génération de texte avec des types d'entités non instanciés, le **post-traitement** qui intègre des instances d'entités fictives issues de données ouvertes. L'**évaluation** qui mesure la qualité et l'utilité des données créées, et ces résultats permettent d'ajuster le prétraitement via une boucle de feedback.

Dans ces travaux, nous prenons le parti de générer des données synthétiques sans utiliser des données presonnelles des clients et nous étudions leur intérêt dans une approche « privacy by design ». L'utilisation des LLM pour la génération de données synthétiques semble d'autant plus pertinente dans le domaine de la relation client que ces outils sont de plus en plus utilisés par les clients eux-mêmes pour leurs communications écrites avec les entreprises.

3 Méthodologie

La méthodologie de génération des e-mails synthétiques repose sur une chaîne de traitement constituée de quatre étapes principales, illustrées dans la Figure 1.

Une première étape de *prétraitement* permet de créer des prompts diversifiés, représentatifs des caractéristiques des corpus client réels. Ces prompts fournis en entrée d'un LLM permettent de contrôler la génération des e-mails. Le paramétrage du LLM via ses hyperparamètres est un moyen de gérer les aspects de créativité, cohérence, longueur du texte en sortie... Les e-mails bruts générés par le LLM sont des e-mails incluant des libellés de type d'entités tel que *nom* ou *prénom*, et qui n'incluent pas d'exemple de ces types d'entités (tels que "*Dupont*" ou "*Marie*"). Cette spécificité a été explicitement demandée dans les prompts, afin de faciliter le processus d'annotation (détaillé dans la Section 4). Par la suite, une phase de *post-traitement* traite les textes des e-mails pour réintégrer des entités fictives issues de données ouvertes, augmentant ainsi le réalisme et l'utilité du corpus créé. La dernière phase consiste à évaluer la conformité du corpus généré aux caractéristiques encodées dans les prompts lors du prétraitement, et permet ainsi de réajuster le processus de génération en fonction des résultats obtenus. La Section 4 détaille l'application de cette méthodologie pour notre cas d'usage.

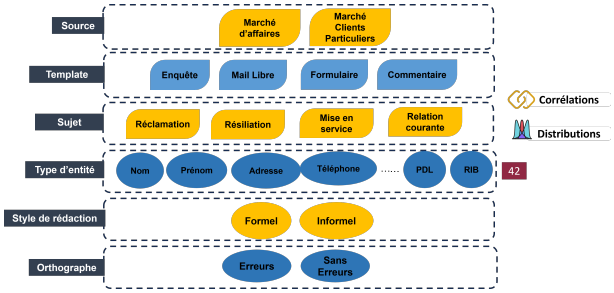


FIGURE 2 – Liste des caractéristiques considérées pour la création des prompts. Les 6 dimensions capturent la diversité observée dans un corpus de données clients réel : la source métier, le gabarit (template), le sujet, les entités à intégrer, le style de rédaction et la qualité orthographique.

3.1 Données d'entrée

Un ensemble de descriptions statistiques des e-mails a été collecté pour permettre d'enrichir le processus de génération, et de diversifier les résultats. Aucun extrait provenant d'e-mails réels n'a été utilisé dans la chaîne de traitement. Cette stratégie vise à garantir la conformité de la base de données créée aux réglementations européennes qui régissent la qualification des jeux de données anonymes [2]. Cela permet notamment d'éviter les risques d'individualisation, de corrélation ou d'inférence, entre autres, pouvant mener à la ré-identification.
Ces descriptions statistiques incluent :

1. La liste des types d'entités directement ou indirectement identifiantes couramment présentes dans les e-mails clients (nom, adresse, point de livraison,...).
2. Des métadonnées décrivant les caractéristiques des e-mails (sujets traités, types des canaux sources).
3. Des distributions probabilistes modélisant la fréquence des types d'entités, directement ou indirectement identifiantes, et des thématiques abordées.

3.2 Stratégies de création des prompts

La création des prompts représente une étape clé de la chaîne de traitement, car elle permet de guider le LLM vers caractéristiques attendues et de capturer la diversité des e-mails souhaitée.
La Figure 2 montre les caractéristiques considérées pour la création des prompts : la source des e-mails (marché d'affaire ou marché des clients particuliers), leur gabarit ou *template* (ex : enquêtes, e-mails libres), leur sujet (thématique, ex : réclamation, mise en service, résiliation), les types d'entités possibles pour les e-mails ainsi que le style (formel ou informel) et enfin le niveau la qualité orthographique de ces derniers.
Chaque prompt est constitué de deux parties :

— Un prompt système générique <syst> définissant le rôle du LLM pour cette tâche de génération des textes synthétiques.
— Un prompt d'instruction structuré </INST> pour garantir une variété linguistique et contextuelle, définissant notamment le sujet de l'e-mail, le style de rédaction et la qualité orthographique attendue.

Exemple 1. *Extrait de prompt pour générer un e-mail client avec pour thème :* réclamation.

```
<s> Tu es un générateur de courriers é
lectroniques. Le courrier électronique est
envoyé par un client à un fournisseur d'énergie.
Le courrier électronique doit être rédigé
uniquement en français.
<INST> Génère un courrier électronique de
réclamation. Le courrier électronique doit être
rédigé par un client qui décrit l'objet de sa
réclamation, en fournissant les informations
nécessaires pour expliquer le contexte.
Utilisez un style de rédaction décontracté et
simple.</INST></s>
```

Nous avons envisagé plusieurs approches pour l'intégration des types d'entités à l'instruction du prompt : approche basée sur la distribution des types d'entités dans le corpus réel, approche basée sur la distribution du nombre d'entités par e-mail, ... Pour ces travaux, la première approche est retenue.
L'objectif de cette stratégie est de contrôler la fréquence d'apparition de chaque type d'entité dans le corpus final, afin d'imiter un corpus réel. En effet, un corpus d'entraînement synthétique avec des entités plus fréquentes ou rares que dans le corpus réel pourrait affecter les performances du système d'IA. Cette approche prend en compte les corrélations entre les entités pour augmenter le réalisme des e-mails générés. Par exemple, lors d'une première mise en service dans une nouvelle maison, un client ne mentionnera probablement pas le point de livraison (référence du point de soutirage ou d'injection de l'électricité du client). La corrélation entre le thème *"mise en service"* et l'entité *"point de livraison"* est faible voire inexistante.

3.3 Génération

La phase de génération utilise un LLM pour la création des textes des e-mails. Plusieurs paramètres contrôlant la part d'aléatoire, tels que la *température, le top-p et le top-k*, ont été ajustés au cours des différents tests afin de déterminer la configuration optimale, permettant un meilleur compromis entre la diversité/créativité d'un côté et le respect des prompts sur les aspects de formalité et styles de rédaction (voir Section 4 pour les résultats et évaluation et la Section 5 pour la discussion associée). Plusieurs valeurs de $maxnumtokens$, contrôlant la longueur maximum de la séquence générée par le LLM, peuvent être spécifiées avec un potentiel impact sur les performances du modèle (temps de génération).
Par ailleurs, une stratégie d'optimisation par lot (*« batch strategy optimization »*) permet de réduire le temps de génération, en exécutant, à partir du même prompt, plusieurs inférences en parallèle sur différents GPU réduisant ainsi le coût total de génération.

Exemple 2. *Le prompt défini dans l'Exemple 1, permet la génération de l'e-mail synthétique suivant* [7] *:*

Madame, Monsieur,
Je me permets de vous écrire ce mail car j'ai rencontré un problème avec mon approvisionnement en électricité chez [Adresse], sous le point de comptage et d'estimation [pce].
En effet, ma maison est actuellement privée d'électricité. Je suis ... confort minimum.
Malgré mes tentatives de vous contacter par téléphone, je n'ai pu obtenir aucune information sur la durée de cette coupure et sur les raisons de celle-ci.
Par conséquent, je vous serais reconnaissant de bien vouloir m'informer ...
Dans l'attente de votre réponse, je vous prie d'agréer, Madame, Monsieur, mes salutations distinguées.
Cordialement,
[nom]
[email]"

3.4 Post-traitement

Les e-mails générés par le LLM sont des e-mails incluant des entités génériques (ex. *"[nom de famille]"*) et non pas des entités instanciées (ex. *"Durand"*) offrant ainsi deux avantages : minimiser le risque d'intégration d'entités instanciées utilisées lors de la phase de pré-entraînement du LLM et permettre l'annotation automatique du corpus d'e-mails généré, évitant une annotation humaine coûteuse.
L'instanciation de chaque e-mail comprend quatre étapes principales. La première étape consiste à détecter les entités génériques contenues dans l'e-mail sur la base d'une expression régulière matchant tous les patrons de chaînes de caractères délimitées par des crochets ouvrant et fermant (ex. : *"[nom de famille]"*). Le prompt demande explicitement à ce que ces types d'entités soient délimités par des crochets. L'identification du type de chacune de ces entités est ensuite effectuée en parcourant un dictionnaire associant à chaque type d'entité une expression régulière couvrant les différents patrons (les différentes formes prises par les entités génériques) de reconnaissance de cette entité (ex. *"nom"* : "nom", "nom de famille"). Puis, un dictionnaire d'entités instanciées est généré sur la base des types d'entités détectés par l'outil : les clés représentent les types d'entités et les valeurs les entités instanciées (ex. "nom" : *" Durand"*). Cette instanciation est faite, suivant les types d'entités, soit par tirage aléatoire dans des échantillons issus de bases de données ouvertes , soit sur la base d'expressions régulières conformes aux formats des types d'entités concernées, ou alors via le module python *Faker* [8]. La dernière étape consiste à remplacer chaque entité générique présente dans l'e-mail par sa valeur correspondante dans le dictionnaire des entités instanciées généré. Pour les entités citées plusieurs fois dans le même e-mail (co-référence), un traitement *naïf* a été effectué, en attribuant la même valeur pour ces entités du même type.

Exemple 3. *Le post-traitement de l'e-mail de l'Exemple 2 produit l'e-mail suivant :*

Madame, Monsieur,
Je me permets de vous écrire ce mail car j'ai rencontré un problème avec mon approvisionnement en électricité chez 13 rue Verdier Bagneux, sous le point de comptage et d'estimation 09992424547933.
En effet, ma maison est actuellement privée d'électricité. Je suis ... confort minimum.
Malgré mes tentatives de vous contacter par téléphone, je n'ai pu obtenir aucune information sur la durée de cette coupure et sur les raisons de celle-ci.
Par conséquent, je vous serais reconnaissant de bien vouloir m'informer ...
Dans l'attente de votre réponse, je vous prie d'agréer, Madame, Monsieur, mes salutations distinguées.
Cordialement,
DURAND
jean.durand@gmail.com"

Lors du processus de post-traitement, toutes les informations requises lors d'une annotation sont mises à jour de façon incrémentale (type d'entité et position de chaque entité) garantissant ainsi l'utilisation du corpus instancié et annoté pour les tâches d'entraînement futures de systèmes IA.

4 Résultats & Evaluation

Cette section présente l'application de la méthodologie décrite dans la précédente section pour la création d'une base de données synthétique d'e-mails clients. Le périmètre de la première base synthétique se focalise sur des e-mails de clients particuliers sous forme de texte libre.

4.1 Implémentation

L'implémentation de la chaîne de traitement décrite dans la Section 3 s'effectue sur un DGX A-100 [9], en utilisant 5 GPUs, de 40GB en mémoire chacun. Au-dessous de 4 GPUs, le chargement du modèle *Mixtral 8x7b* ne peut s'effectuer (voir la discussion dans la Section 5). Tous les tests ont été effectués avec deux LLMs : *Mistral-7B-Instruct-v0.2* [10] et *Mixtral-8x7B-Instruct-v0.1* [11]. Après plusieurs séries d'expérimentations, seuls les résultats issus de *Mixtral-8x7B* ont été jugés suffisamment qualitatifs et diversifiés pour être intégrés à la base finale. Avant l'usage de ces modèles, une vérification de licence a été effectuée. Certaines licences peuvent, en effet, restreindre l'utilisation des données générées, interdisant ainsi leur intégration dans des ensembles de données d'entraînement pour d'autres modèles.

7. Email raccourci pour des raisons d'espace.
8. https://pypi.org/project/Faker/
9. https://resources.nvidia.com/en-us-dgx-systems/dgxa100-system?xs=489761
10. https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2
11. https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1

Les modèles utilisés (Mistral 7b et Mixtral 8x7b) sont sous licence Apache 2.0 permissive pour ce type d'utilisation.
Un ensemble de jeux de données ouverts a été employé pour réintégrer des entités réalistes dans les e-mails : les noms [12], prénoms [13], et les adresses [14]. Tous ces jeux de données sont publiés sous licence ouverte qui autorise leur usage.
Afin d'exploiter le potentiel créatif des LLMs, la création de la base de 1600 e-mails se fait en plusieurs générations (15) chacune contenant 100 ou 200 instances, et correspondant à un paramétrage spécifique (par exemple des valeurs de température différentes). Le tableau 1 résume les caractéristiques intégrées comme instructions des prompts et leurs distributions souhaitées. Le style de rédaction varie entre formel et informel, l'orthographe peut être sans erreurs gramaticales ou avec certaines erreurs. Quatre sujets d'e-mails ont été considérés : coupure d'électricité, mise en service, résiliation, ou réclamation. Douze types d'entités ont été retenus pour la première base d'e-mails. La prise en compte des distributions permet de contrôler la diversité de la génération des e-mails afin d'augmenter l'utilité de la base synthétique comparée à un corpus réel. A noter que les probabilités dans la Table 1 ne sont données qu'à titre indicatif, et ne reflètent pas la description d'un corpus réel, pour des raisons de confidentialité.

Caractéristique	Valeurs (Pourcentage %)
Style	Formel (70%) Informel (30%)
Sujet	Coupure d'électricité (30%) Mise en service (25%) Résiliation (25%) Réclamation (20%)
Orthographe	Correcte (90%) Avec fautes (10%)
Entités	Nom (90%), Prénom (90%), Adresse (60%), Téléphone (30%), Contrat (50%) Point de Livraison (50%), Point de comptage et d'estimation (50%), Numéro client (60%) RIB (30%), IBAN (20%), BIC (20%)

TABLE 1 – Distribution des caractéristiques des prompts utilisés pour la génération de la base d'e-mails synthétiques. Les probabilités capturent la diversité des e-mails en termes de style de rédaction (formel /informel), qualité orthographique (avec ou sans erreurs), les sujets des e-mails, et la fréquence d'occurence des types d'entités.

4.2 Performance de génération

L'évaluation des performances permet de mesurer le temps de création de la base d'e-mails, avec un focus sur le temps d'inférence. Nous analysons les paramètres qui impactent le temps d'inférence, ainsi que la répartition des différentes étapes de la chaîne de traitement dans le temps global.

Analyse du temps d'inférence La Figure 3 illustre le temps d'inférence individuel d'un e-mail, l'analyse est réalisée sur un corpus de 100 e-mails générés. On remarque une variabilité des temps d'inférence qui peut s'expliquer par les longueurs variables des prompts (car incluent plus d'entités par exemple), ou leur complexité comme des prompts relatifs à des contextes de thématiques ayant plus de détails, donc une longueur de texte générée supérieure.
La plupart des e-mails sont générés avec des temps d'inférence entre 10 et 60 secondes, mais certaines rares valeurs atteignent les 80 secondes. La moyenne des temps d'inférence observés est de $\tilde{4}6$ secondes et est équivalente à la moyenne générale ($\tilde{4}4$ secondes) sur tous les jeux de données constituant la base synthétique.

Analyse de l'impact du nombre de tokens sur le temps d'inférence Au regard de la variabilité des temps d'inférence observée, une mesure de l'impact de la longueur des prompts a été effectuée. Les résultats obtenus permettent d'abord de souligner une forte diversité de longueur de prompts (entre 200 et 400 tokens) qui peut s'expliquer par la phase de création de prompts suivant une approche d'intégration de caractéristiques détaillées (expliquée dans la Section 3). En dépit de ces longueurs différentes, les temps d'inférence mesurés ne sont pas corrélés avec les longueurs de prompts utilisés dans la générations des e-mails.

Nous effectuons donc une deuxième analyse qui porte cette fois-ci sur la longueur du texte généré, dont les résultats confirment une corrélation linéaire.

Nous observons à travers les résultats une corrélation linéaire entre la longueur du texte généré et le temps d'inférence. La complexité de certains prompts conduit à des textes plus longs avec plus de détails relatifs au contexte. Des e-mails au sujet de réclamations pour coupure d'électricié peuvent énumérer les dommages causés au client, quand parfois une simple mention du problème est incluse.

Analyse de la décomposition du temps de génération global Pour illustrer le temps nécessaire à la création d'un jeu de données, la Figure 4 détaille les temps requis pour chaque phase de la chaîne de traitement. La phase d'inférence requiert 98 % du temps global du création du jeu de données, quand la phase de post-traitement (instanciation des types d'entités par des données ouvertes) s'effectue en 20 secondes, fournissant ainsi un corpus annoté. Le temps du chargement du modèle Mixstral_8x7b est de 67 secondes. Le chargement du modèle s'effectue une seule fois quel que soit le nombre d'e-mails générés et peut être mutualisé pour créer la base en un seul chargement.

12. https://www.data.gouv.fr/fr/datasets/liste-de-prenoms-et-patronymes/
13. https://www.data.gouv.fr/fr/datasets/fichier-des-prenoms-depuis-1900/
14. https://adresse.data.gouv.fr/donnees-nationales

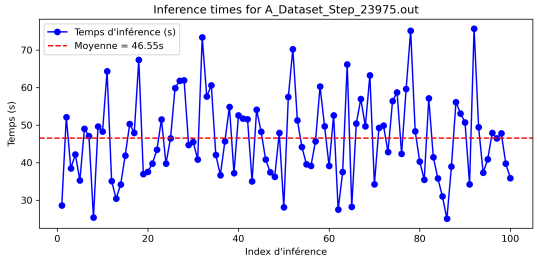


FIGURE 3 – Répartition du temps d'inférence des e-mails individuels. La variabilité des temps correspond à la diversité des prompts, leur longueur et la longueur des e-mails générés. Ce jeu de données contient 100 e-mails générés avec *Mixtral-8x7b*, température= 1.2, max_tokens_length= 4000, top_p = 0.95, top_k = 40

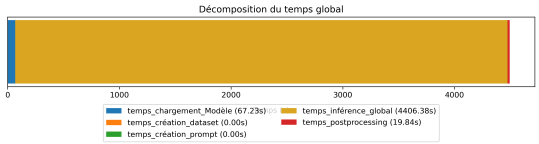


FIGURE 4 – Décomposition du temps de la création d'un jeu de données de 100 e-mails sur les différentes étapes de la chaîne de traitement. Les temps de création du jeu de prompts et son paramétrage étant insignifiants ils apparaissent comme nuls. La majeure partie du temps est consommée par la phase d'inférence (98 %).

4.3 Conformité qualitative

La conformité des e-mails générés à la langue, au sujet, au style de rédaction et à la qualité orthographique requis dans les prompts est évaluée de manière qualitative. Un échantillon de 410 e-mails a été constitué à partir des 1600 e-mails générés en se restreignant aux e-mails générés avec des températures égales à 0.9, 1.2 et 1.4. Une vérification est faite sur les distributions des caractéristiques croisées "Style" et "Qualité orthographique" de l'échantillon qui restent proches de celles du corpus.

Pour chaque caractéristique de chaque e-mail généré , la conformité aux consignes de prompt est évaluée en comparant la modalité de la caractéristique présente dans l'e-mail généré à celle du prompt de génération de l'e-mail. Une non-conformité est déclarée lorsqu'un écart entre les modalités requise et constatée, écart auquel peut être associé un seuil de tolérance, est observé. Le protocole d'annotation est le suivant pour les caractéristiques considérées :

1. Langue : la présence de plus d'un mot non français dans l'e-mail généré implique une non-conformité, tous les e-mails devant être en français.
2. Sujet : lorsque le sujet de l'e-mail généré diffère de la consigne donnée dans le prompt, une non-conformité est déclarée (ex. le prompt requiert une demande de mise en service alors que l'e-mail généré porte sur une réclamation suite à une coupure).
3. Qualité orthographique : la présence d'une faute d'orthographe ou de grammaire implique que l'e-mail généré est "avec erreurs", dans le cas contraire l'e-mail généré est "sans erreurs".
4. Style : les caractéristiques du style formel suivantes ont été considérées : 1- l'utilisation des formules de politesse, de salutations, de titres. 2- emploi du registre soutenu, 3-absence d'expressions verbales, 4-structure claire et organisée. Par exemple, lorsque l'e-mail comporte des formulations non formelles (ex. "Bonjour, J'espère que vous allez bien."), son style est considéré comme "informel", dans le cas contraire comme "formel".

La campagne d'annotation a été réalisée par deux annotateurs différents sur le même échantillon avec les mêmes règles d'annotation hormis pour le Style. En effet, cette caractéristique reste subjective, et d'autant plus difficile à appréhender qu'elle porte sur l'entièreté de chaque e-mail qui peut inclure plusieurs phrases dont certaines peuvent être formelles et d'autres informelles [19].

Les résultats de l'annotation montrent que la conformité des e-mails à la langue et au sujet est quasi toujours respectée, à la différence du style et de la qualité orthographique pour lesquels une variabilité est observée, en fonction de la température considérée. La température de 1.4 permet de gagner en conformité des e-mails générés aux prescriptions des prompts pour toutes les modalités. Une analyse plus approfondie révèle des disparités en fonction des modalités des caractéristiques de style et de qualité orthographique : la modalité "style formel" (resp. "sans erreurs") atteint un taux de conformité plus élevé que celui associé à la modalité "style informel" (resp. "avec erreurs", hors temp 1.4). Un écart est observé entre les deux annotations, plus marqué pour le style informel, où la qualification d'un texte entier (plusieurs lignes) diffère. Certains mails comportent à la fois des formules formelles et des tournures informelles, ce qui peut expliquer cet écart d'annotation. Cependant, les discussions restent quasi similaires. Le LLM a beaucoup plus de difficultés à créer des e-mails formels avec erreurs, mais ceci s'améliore avec des températures plus élevées.

La génération d'e-mails avec un style informel ou une qualité orthographique dégradée reste plus difficile à atteindre pour les températures les plus faibles (cf. Table 2 avec les résultats de l'annotation croisée des caractéristiques de style et de qualité orthographique).

4.4 Conformité de l'intégration des entités

Une analyse des e-mails générés a été effectuée pour observer la fréquence d'intégration des entités, comparée à la distribution de ces entités demandée dans le prompt de génération. Un fichier de configuration a été utilisé pour générer des prompts capturant cette variabilité (voir Section 3). La Figure 5 illustre la comparaison entre la distribution demandée, et la distribution observée dans le corpus des e-mails synthétiques. Globalement, les fréquences demandées dans les prompts ont été respectées par le LLM durant la génération, à l'exception des informations bancaires, où

Style	Orthographe	Emails conformes aux attendus		
		Temp. 0.9 A 1 / A 2	Temp 1.2 A 1 / A 2	Temp. 1.4 A 1 / A 2
Formel	Sans erreurs	97,1% / 96,4 %	98,5% / 97.2 %	98,4% / 100 %
Formel	Avec erreurs	25,0% / 40 %	52,6% / 50 %	100,0% / 100 %
Informel	Sans erreurs	86,2% / 85,7 %	36,2% / 94,7 %	88,9% / 81,81 %
Informel	Avec erreurs	50,0% / 83,3 %	100,0% / 100 %	100,0% / 100 %

TABLE 2 – Conformité des e-mails générés par modalité croisée Style × Qualité orthographique en fonction de la température. Chaque valeur représente le pourcentage de conformité pour l'annotateur A1 ou l'annotateur A2.

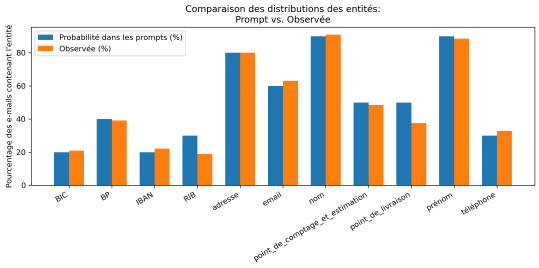


FIGURE 5 – Comparaison des distributions des entités dans les prompts et dans la base des e-mails synthétiques générée. La majorité des distributions sont respectées à part un léger déséquilibre sur les informations bancaires.

on observe moins d'entités *RIB* que demandées mais compensées par plus d'*IBAN* et de BIC.

5 Discussion

5.1 Contrôle de la génération des LLM

L'un des principaux défis observés durant la création de la base de données synthétiques était de contraindre les modèles LLM à se conformer aux instructions de prompts. La qualité des e-mails générés a donc été évaluée sur le respect de ces instructions et non sur l'adéquation du corpus généré pour des tâches aval comme la reconnaissance d'entités nommées ou autres. Lors des tous premiers tests effectués sur le modèle *Mistral 7b*, les résultats contenaient encore beaucoup de parties de textes en anglais, malgré les instructions spécifiques de se limiter à la langue française. Sur la conformité de la langue, l'usage de *Mixtral 8x7b* a nettement amélioré l'adhérence à la langue. Cela a également été le cas en ce qui concerne la formalité et la qualité orthographique. Les LLM sont souvent entraînés pour produire un contenu *propre, correctement écrit*. La génération d'e-mails qui reproduisent les comportements clients observés dans un corpus réel implique la présence de fautes d'orthographe et un usage fréquent du style informel. Pour forcer le LLM à mieux respecter ces dernières consignes, différentes températures ont été testées pour la génération, avec un impact sur la qualité des e-mails générés. Des hallucinations (peu fréquentes) ont été observées : insertion de code Python, reprise de parties d'instructions dans les corps d'e-mails (malgré les consignes de prompts), succession d'e-mails dans un même corps d'e-mail, répétition en fin d'e-mail de phrases du corps de l'e-mail sous forme de "Note : une phrase de l'e-mail". Pour pallier une partie de ces hallucinations, une étape de curation des e-mails générés a été intégrée au post-traitement. La génération de plusieurs types d'entités en lieu et place d'une unique entité générée (*[Nom Prénom]* au lieu de *[Nom]*, *[Prénom]*) par le LLM a aussi été constatée. Enfin, bien que peu fréquente, la réinjection d'entités instanciées issues du pré-entraînement du LLM dans les e-mails générés a aussi été observée.

L'analyse de la conformité met en exergue le compromis à faire entre créativité du LLM et respect des instructions de prompts par le LLM. Des améliorations de la chaîne de traitement proposée restent à faire et devront porter sur un affinement conjoint des prompts et des post-traitements (expressions régulières, curation des données). Par exemple, pour le style et l'orthographe, l'ajout d'une instruction de conformité à des standards tel que *Common European Framework of Reference for Languages* ou encore la mise en place d'une génération de texte multi-étapes avec, dans un premier temps, une demande faite au LLM de caractériser un e-mail formel vs un e-mail informel, puis la génération d'e-mails sur cette base. La diversité des e-mails générés reste à évaluer en termes de vocabulaire utilisé.

5.2 Performance et Coûts

L'utilisation d'un LLM tel que *Mixtral-8x7B* pour la génération d'e-mails synthétiques implique des coûts associés à l'infrastructure GPU. Dans cette étude, les expérimentations ont été réalisées sur une infrastructure serveur DGX-A100 exploitant 5 GPUs de 40GB chacun, et un modèle chargé en local. Aucun coût supplémentaire n'a été engendré. Pour estimer le coût d'utilisation d'une telle infrastructure, on peut comparer les coûts directs d'infrastructure locale à ceux d'une solution cloud. Sur une plateforme cloud, le coût horaire d'utilisation d'un GPU s'élève en moyenne à 3,7$ [15]. Ainsi, l'utilisation de 5 GPUs simultanément pour un cycle de génération de 15 heures entraîne un coût approximatif de 278$. A noter que ces résultats peuvent s'optimiser via *vLLM* (vs. pipeline Hugging Face [16]).

A titre de comparaison, l'annotation de 1600 e-mails réels nécessiterait une équipe d'annotateurs et un temps d'annotation cumulé de 2 à 3 minutes par e-mail, donc pouvant atteindre 80 h pour l'entièreté du cycle d'annotation (incluant les annotations croisées et leur validation). A ce titre, la génération d'e-mails s'avère compétitive (facteur × 4).

L'utilisation d'un LLM permet une économie de coût tout en offrant un corpus annoté scalable, avec une qualité homogène, et qui capture la diversité d'un corpus réel (distributions, cas particuliers,..). De plus, l'usage d'une base d'e-mails synthétiques présente un avantage de conformité réglementaire. Contrairement à l'annotation des données réelles, qui implique la manipulation d'entités identifiantes et des risques de divulgation, la génération de données synthétiques ne requiert aucun accès à des données réelles.

15. https://aws.amazon.com/fr/ec2/pricing/on-demand/

5.3 Utilité des données synthétiques

L'utilisation de données synthétiques soulève la question de leur utilité, i.e. de savoir si ces données synthétiques sont une bonne approximation de la réalité pour construire des modèles d'IA sur des données texte de la relation client. Des tests préliminaires sur la reconnaissance d'entités nommées dans des e-mails client à partir d'un modèle d'IA entraîné exclusivement sur les données synthétiques générées dans ces travaux indiquent une performance honorable se situant à 15 points en deça des modèles opérationnels les plus performants. Ces résultats encourageants nécessitent des travaux supplémentaires. Ils indiquent la probable nécessité d'un mix de données réelles et de données de synthèse pour l'entraînement de modèles opérationnels performants, permettant ainsi la minimisation de l'utilisation de données réelles. En outre, ces travaux devront veiller à ce que l'usage de données de synthèse ne renforce pas les biais existants ou ne crée pas de nouvelles discriminations [14].

5.4 Pseudonymisation ou anonymisation

Le caractère pseudonyme des données synthétiques créées dans ces travaux semble incontestable car elles sont synthétisées de manière désidentifiées avec insertion d'entités factices. Notamment, ces données synthétiques ne peuvent pas être considérées comme exactes et étant associées à des individus. En outre, le processus décrit ne repose pas sur l'utilisation de données réelles présentes dans les systèmes d'information de notre entreprise. Néanmoins, il reste des incertitudes sur le caractère anonyme des données synthétiques [14]. En effet, le LLM lui-même est pré-entraîné sur des données réelles (e.g., les données extraites de l'internet public) et, dans certaines conditions, peut générer des données identiques aux données de pré-entraînement du LLM [17]. Même si cela peut sembler improbable, il n'est pas impossible que des données suffisamment proches des données de pré-entraînement du LLM soient générées permettant une réidentification d'un individu. Ce dernier point nécessite des approfondissements pour quantifier la probabilité de réidentification dans ce type de situation.

6 Conclusion et perspectives

Nous avons présenté dans cet article une méthode permettant de produire des e-mails synthétiques de relation client répondant à deux enjeux : (i) offrir la meilleure garantie possible de la protection de la vie privée, et (ii) minimiser le temps d'annotation de données requis pour l'entraînement de systèmes d'IA. Pour cela, nous avons proposé une chaîne de traitement en plusieurs étapes générant un corpus d'e-mails annoté en entités et sans aucune référence directe à des e-mails réels. Les e-mails synthétiques sont générés au moyen de LLMs par des instructions reposant sur des caractéristiques descriptives générales des e-mails réels (ex : thèmes, distributions des types d'entités), et contiennent des entités fictives issues de bases ouvertes, ou générées artificiellement par des règles. Nos expérimentations ont mis en lumière des défis propres à la génération de données synthétiques avec des LLMs, en particulier le contrôle de l'adhérence du texte généré avec les instructions fournies dans le prompt. Nous avons montré qu'il est possible d'obtenir un corpus d'e-mails synthétiques respectant le format et la distribution des types d'entités souhaitées, la langue demandée, ou le thème. Il subsiste des limites dans les capacités du modèle à adopter un style d'expression informel et à faire des fautes d'orthographe, qui peuvent être compensées en augmentant la température. L'augmentation de la température produit une diversité qui peut engendrer un post-traitement plus coûteux.

En perspective, nous prévoyons d'apporter des compléments aux méthodes proposées, pour atteindre un niveau de performance permettant d'utiliser ce corpus de données synthétiques afin d'entrainer un système d'IA interne EDF de détection d'entités nommées. Des tests préliminaires à partir d'un modèle d'IA entraîné exclusivement sur les données synthétiques générées dans ces travaux ont montré une performance se situant à 15 points en deça des modèles opérationnels les plus performants Pour affiner davantage le contenu généré et aller plus loin dans le degré d'adhérence au prompt, nous prévoyons d'étendre la méthodologie par l'exploration arborescente pour la génération guidée de prompts et l'usage des graphes [26]. Si les résultats sont satisfaisants, ces travaux seront étendus à d'autres périmètres (segments de marché différents, et formats différents comme des formulaires ou des fils de discussion client conseiller). Un traitement plus précis des co-références est également envisagé, pour augmenter l'utilité des données pour d'autres tâches. Des guides fournis aux conseillers, donnant par exemple des informations sur le "ton de voix" (*tone of voice*) à adopter avec un client, pourraient être intégrés au prompt au moyen d'une architecture RAG, pour permettre de traiter les textes rédigés par un conseiller.

Remerciements

Nous remercions chaleureusement toutes les personnes qui sont intervenues de près ou de loin sur ce projet : Sofiane Kerroua, Anne Gayet, Laetitia Leroux, Sonia Audheon, Dominique Manzoni-Quantin, François Raynaud.

Références

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv :2303.08774*, 2023.

[2] Article 29 Data Protection Working Party. Opinion 05/2014 on Anonymisation Techniques (WP216). Technical Report WP216, European Commission, April 2014.

[3] John Joon Young Chung, Ece Kamar, and Saleema Amershi. Increasing diversity while maintaining accuracy : Text data generation with large language models and human interventions. *arXiv preprint arXiv :2306.04140*, 2023.

[4] CNIL. L'anonymisation de données personnelles, May 2020. Accessed : 2025-02-25.

[5] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, et al. A survey on in-context learning. *arXiv preprint arXiv :2301.00234*, 2022.

[6] Guillaume Dubuisson Duplessis, Elliot Bartholme, Sofiane Kerroua, Mathilde Poulain, Ahès Roulier, and Anne-Laure Guénet. Désidentification de données texte produites dans un cadre de relation client. In *Conférence Traitement Automatique des Langues Naturelles (TALN) – démonstrations*, pages 10–13, 2020.

[7] Guillaume Dubuisson Duplessis, François Bullier, and Anne-Laure Guénet. Démonstration : exploration sémantique de données texte de la relation client. In *Conférence Nationale sur les Applications Pratiques de l'Intelligence Artificielle*, pages 103–106, 2023.

[8] Guillaume Dubuisson Duplessis, Sofiane Kerroua, Ludivine Kuznik, and Anne-Laure Guénet. Cameli@ : analyses automatiques d'e-mails pour améliorer la relation client. In *Conférence sur le Traitement Automatique des Langues Naturelles (TALN). Volume IV : Démonstrations*, pages 623–626, 2019.

[9] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv :2401.04088*, 2024.

[10] Rumeng Li, Xun Wang, and Hong Yu. Two directions for clinical data generation with large language models : data-to-label and label-to-data. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2023, page 7129, 2023.

[11] Andreas Liesenfeld, Alianda Lopez, and Mark Dingemanse. Opening up chatgpt : Tracking openness, transparency, and accountability in instruction-tuned text generators. In *Proceedings of the 5th International Conference on Conversational User Interfaces*, CUI '23, New York, NY, USA, 2023. Association for Computing Machinery.

[12] Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. On llms-driven synthetic data generation, curation, and evaluation : A survey. *arXiv :2406.15126*, 2024.

[13] Shayne Longpre, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, William Brannon, Niklas Muennighoff, Nathan Khazam, Jad Kabbara, Kartik Perisetla, et al. A large-scale audit of dataset licensing and attribution in ai. *Nature Machine Intelligence*, 6(8) :975–987, 2024.

[14] Alexis Léautier. [Données synthétiques] - Et l'Homme créa les données à son image 2/2. *Laboratoire d'Innovation Numérique de la CNIL (LINC)*, août 2022. Consulté le 24 janvier 2025.

[15] Abdul Majeed and Sungchang Lee. Anonymization techniques for privacy preserving data publishing : A comprehensive survey. *IEEE access*, 9 :8512–8545, 2020.

[16] Matias Martinez. The impact of hyperparameters on large language model inference performance : An evaluation of vllm and huggingface pipelines. *arXiv preprint arXiv :2408.01050*, 2024.

[17] Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. Scalable extraction of training data from (production) language models, 2023.

[18] Rajvardhan Patil and Venkat Gudivada. A review of current trends, techniques, and challenges in large language models. *Applied Sciences*, 14(5) :2074, 2024.

[19] Ellie Pavlick and Joel Tetreault. An empirical analysis of formality in online communication. *Transactions of the association for computational linguistics*, 4 :61–74, 2016.

[20] Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. A systematic survey of prompt engineering in large language models : Techniques and applications. *arXiv preprint arXiv :2402.07927*, 2024.

[21] Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. Beyond memorization : Violating privacy via inference with large language models. *arXiv preprint arXiv :2310.07298*, 2023.

[22] Theresa Stadler, Bristena Oprisanu, and Carmela Troncoso. Synthetic data–anonymisation groundhog day. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 1451–1468, 2022.

[23] Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. Does synthetic data generation of llms help clinical text mining ? *arXiv preprint arXiv :2303.04360*, 2023.

[24] Dirk Väth, Lindsey Vanderlyn, and Ngoc Thang Vu. Towards a zero-data, controllable, adaptive dialog system. *arXiv preprint arXiv :2403.17582*, 2024.

[25] Nicolas Vautier, Marc Héry, Mourad Miled, Irène Truche, François Bullier, Anne-Laure Guénet, Guillaume Dubuisson Duplessis, Sabrina Campano, and Philippe Suignard. Utilisation de llms pour la classification d'avis client et comparaison avec une approche classique basée sur camembert. In *Conférence Nationale sur les Applications Pratiques de l'Intelligence Artificielle*, 2024.

[26] Qinggang Zhang, Junnan Dong, Hao Chen, Daochen Zha, Zailiang Yu, and Xiao Huang. Knowgpt : Knowledge graph based prompting for large language models. *Advances in Neural Information Processing Systems*, 37 :6052–6080, 2025.

[27] Ying Zhao and Jinjun Chen. A survey on differential privacy for unstructured data content. *ACM Computing Surveys (CSUR)*, 54(10s) :1–28, 2022.

Une architecture multi-agents pour la génération automatique de tickets en environnement industriel : focus sur l'agent de classification

Ying Zhang[1], Sébastien Bonnet[2], Matthieu Petit Guillaume[1],
Muriel Hug[1], Aurélien Krauth[1], Rémi Uhartegaray[2]

[1] Leviatan, 725 Bd Robert Barrier, 73100 Aix-les-Bains
[2] AntemetA, 19 Av. Georges Pompidou, Bâtiment Le Danica, 69003 Lyon

{y.zhang, matthieu, m.hug, aurelien}@leviatan.fr
{sebastien.bonnet, remi.uhartegaray}@antemeta.fr

Résumé

La génération automatique de tickets à partir d'e-mails constitue un enjeu majeur pour le support client en environnement industriel, où confidentialité et ressources limitées sont critiques. Nous proposons une architecture multi-agents composée d'un agent de classification pour identifier la catégorie de l'e-mail et d'agents d'extraction spécialisés pour collecter les informations requises. Dans cet article, nous présentons la conception, l'entraînement et l'évaluation de l'agent de classification, basé sur un grand modèle de langage open source quantifié en 4 bits et affiné via PEFT/LoRA. Les résultats montrent qu'avec environ 13 milliards de paramètres, il est possible d'atteindre plus de 88% d'exactitude tout en répondant aux contraintes industrielles. Cette approche modulaire ouvre la voie à une automatisation complète du processus, chaque agent étant dédié à une tâche spécifique du workflow.

Mots-clés

Classification d'e-mails, génération automatique de tickets, grands modèles de langage, fine-tuning, multi-agents, quantification 4 bits, industrie.

Abstract

Automatically generating tickets from e-mails is a key challenge for industrial customer support, especially under constraints of data confidentiality and limited resources. We propose a multi-agent architecture with a single classification agent that determines the ticket category and specialized extraction agents that gather the required fields. In this paper, we focus on the design, training, and evaluation of the classification agent, built upon an open-source large language model quantized to 4 bits and fine-tuned via PEFT/LoRA. Our results show that a mid-sized model (approximately 13 billion parameters) can achieve over 88% accuracy while complying with industrial constraints. This modular approach paves the way for fully automated workflows, with each agent dedicated to a specific stage of the ticket-generation pipeline.

Keywords

E-mail classification, automatic ticket generation, large language models, fine-tuning, multi-agent, 4-bit quantization, industry.

1 Introduction

Dans le secteur du support client, le nombre croissant d'e-mails reçus chaque jour pose d'importants défis aux équipes de maintenance et d'assistance. À partir de ces e-mails, des tickets doivent être créés. Ce processus consiste à trier, classer et affecter ces tickets aux équipes responsables en fonction de leur nature et de leur urgence, puis, selon les différentes catégories de tickets, à générer un fichier JSON dont la structure (attributs) varie, afin de l'intégrer directement dans notre outil de gestion.

Les processus manuels de lecture et de tri entraînent souvent des délais de traitement élevés, des erreurs de classification et une surcharge de travail. Face à ces contraintes, l'automatisation de la création de tickets à partir d'e-mails constitue donc un levier stratégique pour accroître la réactivité et la qualité du service.

Dans le cadre de notre projet, nous adoptons une approche multi-agents pour couvrir les différentes étapes de l'analyse et du traitement des tickets. Bien que ce système comporte plusieurs « agents », nous concentrerons ici notre attention sur l'agent de classification. L'objectif est de démontrer qu'un modèle de langage, correctement entraîné et adapté aux contraintes industrielles (sécurité des données et ressources matérielles limitées), peut catégoriser automatiquement les e-mails clients tout en maintenant un haut niveau de précision.

Après avoir présenté le contexte, la problématique et l'état de l'art, nous décrirons l'architecture multi-agents dans ses grandes lignes, puis nous focaliserons sur le fonctionnement, l'entraînement et l'évaluation de l'agent de classification. Enfin, nous discuterons des résultats obtenus et évoquerons les perspectives d'évolution.

2 Contexte et problématique

Dans un environnement où le support client doit traiter chaque jour un volume considérable d'e-mails, la gestion manuelle de ces messages devient rapidement un goulot d'étranglement. Les opérateurs doivent lire et comprendre chaque demande, puis la classer dans une catégorie de ticket appropriée (incident, demande de service, question d'ordre technique, etc.). Cette opération, très coûteuse en temps, reste par ailleurs sujette aux erreurs humaines, pouvant entraîner une mauvaise affectation des demandes et retarder la prise en charge de problèmes critiques.

L'émergence des grands modèles de langage (Large Language Models, LLM) ouvre la voie à une automatisation plus poussée de ce processus. Toutefois, dans un cadre industriel, l'adoption de ces modèles soulève des défis spécifiques :

- **Sécurité des données** : Les contenus échangés par e-mail peuvent être sensibles et ne doivent pas être exposés à des services publics.
- **Contraintes matérielles** : Déployer un LLM de très grande taille requiert des ressources (GPU, VRAM) considérables, souvent incompatibles avec l'infrastructure standard d'une PME (Petite et Moyenne Entreprise).
- **Fiabilité de la classification** : Le modèle doit offrir une précision suffisante pour que l'automatisation ne génère pas de nouvelles sources d'erreurs.

Ainsi, la problématique principale consiste à concilier les avantages offerts par les LLM (compréhension fine du langage, capacité à traiter rapidement de gros volumes de données) avec les contraintes d'un déploiement industriel (infrastructures matérielles limitées et données confidentielles), tout en préservant une qualité de classification à la hauteur des exigences d'un service client professionnel.

3 État de l'art

Ces dernières années, les avancées en traitement automatique du langage ont été largement portées par l'émergence des modèles de langage de grande taille (LLM). Ces modèles, tels que LLaMA [8], GPT [10], Qwen [12], et d'autres architectures émergentes, ont permis des améliorations notables dans des domaines comme la compréhension du langage, la génération de texte, et la traduction automatique.

Cependant, lorsqu'il s'agit d'implémenter ces modèles dans des environnements industriels, plusieurs défis apparaissent. Comme indiqué précédemment, la confidentialité des données et l'importante consommation de ressources matérielles constituent deux freins majeurs à l'adoption des LLM. Même un modèle puissant doit être adapté aux données métier (domain adaptation), ce qui implique un micro-ajustement (fine-tuning) potentiellement coûteux.

Pour faire face à ces contraintes, différentes stratégies d'optimisation sont proposées, comme la quantification (passage en int8, int4, etc.) [2, 6], ou les approches de Parameter-Efficient Fine-Tuning (PEFT) [7], telles que LoRA [5, 2], qui réduisent la taille des poids à ajuster et permettent de se rapprocher des performances d'un modèle complet. Parallèlement, l'approche multi-agents [13] en IA gagne en popularité, chaque agent se voyant confier une tâche spécialisée (classification, extraction d'information, etc.), ce qui facilite la modularité et l'évolutivité du système. Dans le cadre du présent travail, nous nous inscrivons dans cette continuité, en centrant l'analyse sur la partie « classification » pour répondre aux besoins de tri automatisé des e-mails.

4 Conception en multi-agents et focalisation sur l'agent de classification

Un large éventail de modèles open source a vu le jour, chacun offrant plusieurs versions de différentes tailles en nombre de paramètres. Cependant, il existe une différence de qualité notable entre les versions à faible et à fort nombre de paramètres : les modèles plus légers (quelques milliards de paramètres) sont plus rapides et moins gourmands en ressources, mais leurs performances en compréhension et en génération de texte sont souvent inférieures, en particulier pour les tâches complexes.

Dans notre environnement industriel, nous privilégions le déploiement de LLM d'environ 13 milliards de paramètres. Néanmoins, ce type de modèle reste moins performant qu'un modèle de 70 milliards [3], sans même parler de ceux qui comptent plusieurs centaines de milliards de paramètres. Cette différence exige des optimisations algorithmiques, une adaptation des données ainsi qu'un entraînement spécifique à chaque tâche pour compenser le manque de puissance brute.

Dans notre cas d'usage, le nombre de catégories de tickets peut dépasser les 300. Toutefois, nous avons choisi, dans un premier temps, de cibler uniquement certaines catégories principales. Chacune d'elles requiert l'extraction de champs d'informations spécifiques. Pour répondre à ces besoins variés, il serait peu judicieux de définir un unique « agent » chargé à la fois de la classification et de l'extraction de tous les champs, car le prompt deviendrait trop complexe et potentiellement ambigu. Pour un modèle de 13 milliards de paramètres, cela s'avérerait quasiment impossible.

Afin de mieux structurer notre système, nous avons opté pour une architecture multi-agents [13]. Concrètement :

- Un premier agent se consacre exclusivement à la classification des e-mails afin d'identifier la catégorie.
- Un agent d'extraction spécifique à chaque catégorie prend ensuite le relais pour extraire les champs requis en fonction de la catégorie identifiée.

Un orchestrateur central coordonne ces agents : dès qu'un e-mail est reçu, il est classé, puis redirigé vers l'agent d'extraction adéquat, si besoin. Ainsi, nous obtenons un total de $N+1$ agents, où N représente le nombre de catégories. Cette approche offre :

- **Modularité accrue** : chaque catégorie peut évoluer indépendamment (ajout de nouveaux champs à extraire, ajustement du prompt, etc.).

— **Lisibilité renforcée** : chaque agent d'extraction est axé sur un ensemble restreint d'informations à récupérer, ce qui améliore la qualité.
— **Maintenance facilitée** : en cas de nouvelle catégorie, il suffit d'ajouter un nouvel agent d'extraction et de réajuster l'agent de classification, sans perturber les autres agents.

Toutefois, au stade actuel du projet, seul l'agent dédié à la classification a atteint la maturité requise pour un déploiement et une évaluation rigoureuse. Les autres agents sont encore en phase d'expérimentation et feront l'objet de travaux futurs.

Dans ce contexte, le présent article se concentre exclusivement sur l'agent de classification, depuis la sélection et l'adaptation du modèle de langage (LLM) jusqu'à l'évaluation de ses performances. Ce choix nous permet de proposer une étude détaillée d'une solution déjà opérationnelle, tout en soulignant l'intérêt et la faisabilité de la démarche multi-agents dans un environnement industriel.

5 Jeu de données

L'ensemble de données utilisé dans cette étude provient de la base de gestion des tickets de support client. Son objectif principal est de permettre l'entraînement et l'évaluation d'un agent de classification des e-mails en fonction des catégories de tickets.

5.1 Structure et origine des données

Les données initiales comprenaient 175 813 entrées issues du système de gestion des tickets, chaque entrée étant caractérisée par les attributs suivants :

— **Numéro de ticket** : Identifiant unique pour chaque ticket.
— **Description** : Contenu des échanges de mails entre le client et le support technique.
— **Catégorie du ticket** : Classification du type de ticket (ex. : incident technique, facturation, demande de service, etc.).
— **Message** : Réponses du support envoyées aux clients, comprenant mises à jour et modifications du ticket.

L'objectif de cette classification repose sur l'analyse de la première communication client pour chaque ticket, afin d'identifier et de caractériser la nature initiale des demandes. Toutefois, l'étude des données brutes révèle des contraintes méthodologiques liées à l'absence d'un ensemble clairement défini de « premiers e-mails » envoyés par les clients, ce qui rend difficile l'extraction rigoureuse de la requête initiale. Par ailleurs, de nombreux tickets présentent déjà plusieurs échanges par e-mail avant même leur enregistrement, compromettant l'isolement de la première interaction et compliquant l'application de procédures de classification fiables.

5.2 Nettoyage des données

Pour assurer la qualité et la pertinence du jeu de données destiné à la classification, plusieurs étapes de préparation et de nettoyage ont été mises en œuvre :

1. Filtrage initial des colonnes
Les enregistrements ne respectant pas la structure à quatre colonnes prédéfinies ont été éliminés, réduisant l'ensemble à 154 414 lignes conformes.
2. Regroupement par Numéro de ticket
Les données ont été agrégées par Numéro de ticket afin de ne conserver que la première ligne associée à chaque identifiant, tandis que les lignes subséquentes pour un même ticket ont été supprimées. En effet, chaque enregistrement correspond à un état de mise à jour d'un ticket (par exemple : création, modification, fermeture, etc.), pouvant ainsi générer plusieurs enregistrements pour un même identifiant. Cette opération a permis de ramener l'ensemble de données à 29 376 tickets uniques.
3. Extraction des premiers e-mails
Étant donné que la classification repose sur l'analyse de la première communication reçue du client, un traitement supplémentaire a été appliqué à la colonne « Description » (où se trouvent souvent plusieurs échanges). Ce traitement visait à isoler précisément les informations correspondant au premier e-mail.
Une règle fondée sur la détection de motifs textuels récurrents (par exemple : « To… », « Sujet… », etc.) a été mise en place pour segmenter les différentes parties du fil de discussion et extraire la première communication du client.
Cette approche de segmentation reste rudimentaire et pourrait être optimisée dans de futurs travaux. Néanmoins, elle permet d'isoler de manière satisfaisante le texte relatif à la première prise de contact.
4. Analyse des catégories
Enfin, les fréquences des différentes catégories de tickets ont été étudiées afin de déceler d'éventuels déséquilibres de classe, susceptibles d'influencer la performance des algorithmes de classification.

5.3 Sélection des catégories

L'analyse des 29 376 tickets a mis en évidence 365 catégories distinctes. Une répartition inégale des occurrences a été constatée : la catégorie la plus fréquente regroupe 10 963 tickets, tandis que plus de 200 catégories ne comptent chacune que moins de cinq occurrences.

Afin de garantir la pertinence et la qualité des expérimentations, les catégories ont été soumises à un filtrage rigoureux selon deux critères :

1. Sélection de 24 catégories stratégiques
Ces catégories ont été jugées essentielles pour l'entreprise en raison de leur impact opérationnel et de leur récurrence.
2. Seuil minimal de représentativité de 2‰
Les catégories dont la proportion dans l'ensemble de données est inférieure à 2‰ ont été considérées comme sous-représentées et donc écartées de l'étude.

À l'issue de ce processus, neuf catégories ont été retenues pour la phase de recherche et développement (R&D) :

— Demande de service/Backup BCS/Autre
— Demande de service/Backup BCS/Demande de renseignement
— Demande de service/Backup BCS/Restauration qualifiée

— Demande de service/Backup BCS/Stratégie de sauvegarde/Création
— Demande de service/Backup BCS/Stratégie de sauvegarde/Modification
— Demande de service/Backup BCS/Stratégie de sauvegarde/Suppression
— Demande de service/Cyber Sécurité CS2/Bastion/Création-Modification d'entrées
— Incidents/Backup BCS/Sauvegarde
— Incidents/Supervision

Cette sélection vise à concentrer les efforts de modélisation sur les catégories les plus pertinentes, tant sur le plan stratégique qu'en termes de volume de données.

5.4 Constitution de l'ensemble de données final

L'ensemble final est composé de 4500 tickets, répartis entre un ensemble d'entraînement (80 %) et un ensemble de test (20 %), selon la distribution suivante :

Catégorie	Train Set	Test Set
Incidents/Supervision	2397	565
Demande de service/Backup #BCS/Restauration qualifiée	361	83
Incidents/Backup #BCS/Sauvegarde	346	95
Demande de service/Backup #BCS/Autre	200	52
Demande de service/Backup #BCS/Stratégie de sauvegarde/Création	78	21
Demande de service/Backup #BCS/Stratégie de sauvegarde/Suppression	69	25
Demande de service/Backup #BCS/Demande de renseignement	58	17
Demande de service/Backup #BCS/Stratégie de sauvegarde/Modification	50	14
Demande de service/Cyber Sécurité #CS2/Bastion/Création-Modification d'entrées	49	20

TABLE 1 – Répartition des tickets

L'ensemble de données ainsi préparé garantit :

— Une couverture des principales catégories rencontrées en support client.
— Un équilibre entre les différentes classes, tout en évitant les catégories sous-représentées.
— Une qualité de données optimisée, grâce à la sélection des premières interactions client et l'élimination des doublons et bruits textuels.

Ce jeu de données constitue une base fiable pour entraîner et évaluer un modèle de classification automatique des e-mails en environnement industriel. L'ensemble de nos entraînements a été réalisé sur ce jeu de données d'entraînement (train set), tandis que toutes les évaluations de benchmark ont été effectuées sur ce jeu de test (test set).

6 Configuration matérielle

Les expérimentations de cette tâche ont été réalisées sur une machine dotée des caractéristiques suivantes :

— GPU : 1 × NVIDIA A10 avec 24 Go de VRAM.
— CPU : 30 cœurs.
— RAM : 214,7 Go.
— Stockage : 1,5 To SSD.

Cette configuration permet de répondre aux exigences en termes de calcul intensif pour le benchmarking des modèles tout en offrant une capacité de stockage suffisante pour gérer les données expérimentales.

7 Benchmark basé sur les modèles initiaux

7.1 Ingénierie des prompts et des plateformes

Pour évaluer les performances initiales des modèles dans la tâche de classification, nous avons mis en place deux approches distinctes de benchmark, reposant sur des plateformes et des types de prompts différents.

La première approche utilise la plateforme Ollama [9], avec un prompt conçu pour générer une sortie au format JSON standardisé. Celui-ci est structuré de manière à inclure une explication de la tâche, une présentation des catégories, des règles spécifiques de classification, ainsi qu'un exemple annoté et une définition explicite du format JSON attendu. Cette méthode vise à assurer une sortie normalisée et directement exploitable. Un exemple détaillé du prompt est disponible en Annexe A.1.

La seconde approche repose sur la plateforme Unsloth [1], où le prompt est formulé en langage naturel, sans contrainte de structure JSON dans les résultats. Bien que sa conception s'appuie sur des principes similaires à ceux du prompt JSON, il privilégie une approche plus souple pour évaluer la capacité du modèle à comprendre et exécuter la classification de manière plus libre. Un exemple détaillé de ce prompt est disponible en Annexe A.2.

Ces deux méthodologies permettent d'analyser les performances des modèles sous différents angles, notamment en termes de précision, de cohérence des résultats et de conformité aux exigences du projet, en particulier pour l'intégration des résultats dans le système de gestion des tickets.

7.2 Expérimentations choix des modèles et combinaisons testés

Après avoir comparé plusieurs modèles de 13 milliards de paramètres, nous avons retenu les combinaisons suivantes de plateformes, de prompts et de modèles pour nos expérimentations.

Sur la plateforme Ollama, avec un prompt au format JSON, nous avons testé

— Llama3.2-11B-Vision, Qwen2.5-14B et Mistral-Nemo-12B.

En parallèle, sur la plateforme Unsloth, avec un prompt en langage naturel, nous avons évalué

— Llama3.2-11B-Vision, Qwen2.5-14B et Pixtral-12B.

Deux modèles, Llama3.2-11B-Vision et Qwen2.5-14B, ont été déployés sur les deux plateformes, permettant ainsi d'analyser leurs performances dans des contextes distincts et d'évaluer leur capacité d'adaptation à différentes structures de prompts.

7.3 Vérification et normalisation des sorties générées en langage naturel et au format JSON

Dans le cadre de la génération de sorties en langage naturel ou au format JSON, il est fondamental d'évaluer si les prédictions correspondent aux catégories attendues. Cette vérification repose sur un ensemble de critères normalisés permettant d'assurer la cohérence et la fiabilité des résultats.

7.3.1 Définition d'une correspondance valide

Une prédiction est considérée comme *matched* si elle peut être associée de manière univoque à l'une des catégories définies dans notre référentiel. Cette correspondance est établie en fonction de critères linguistiques et statistiques.

7.3.2 Normalisation des prédictions

Afin de réduire les variations introduites par des erreurs de formatage ou de syntaxe, un processus de normalisation est appliqué aux prédictions. Celui-ci comprend :

— Suppression des caractères spéciaux, afin d'éliminer les éléments non significatifs pour la comparaison.

— Conversion en minuscules, permettant d'uniformiser les entrées et de minimiser les biais liés à la casse.

7.3.3 Méthodes d'évaluation des correspondances

L'association entre une prédiction et une catégorie attendue repose sur une combinaison de mesures textuelles permettant d'évaluer la similarité sémantique et syntaxique :

— *Distance de Levenshtein* [11] : Quantification du nombre minimal d'opérations (insertion, suppression, substitution) nécessaires pour transformer une chaîne de caractères en une autre.

— Ratio de similarité de séquences : Application de l'algorithme *SequenceMatcher* [4] pour mesurer le degré de similitude entre les chaînes textuelles.

— Ratio de correspondance des mots (*word_match_ratio*) : Comparaison basée sur la proportion de mots communs entre la prédiction et la catégorie de référence.

7.3.4 Seuils d'acceptation des correspondances

Pour qu'une prédiction soit considérée comme valide, elle doit satisfaire les critères suivants :

— $DistanceLevenshtein \leq 3$, garantissant une proximité lexicale acceptable.

— $RatioSimilarité \geq 85\%$, reflétant un degré élevé de correspondance textuelle.

— $RatioCorrespondanceMots \geq 80\%$, assurant une cohérence terminologique suffisante.

Cette approche permet d'optimiser la robustesse de la catégorisation automatique et d'améliorer la fiabilité des prédictions en minimisant les erreurs de classification.

7.4 Résultats des expérimentations

Les résultats obtenus sont les suivants :

Mod.	Plat.	Mat. Pred.	Unmat. Pred.	Match Rate	Acc.
Llama	Olla.	891	1	0,9988	0,8049
Qwen	Olla.	850	42	0,9529	**0,8520**
Mistral-Nemo	Olla.	659	233	0,7645	0,6446
Pixtral	Unsl.	465	427	0,5213	0,4159
Qwen	Unsl.	523	369	0,5863	0,4776
Llama	Unsl.	746	146	0,8363	0,4540

TABLE 2 – Benchmark des modèles initiaux (Mod. : Modèle, Plat. : Plateforme, Mat. Pred. : Matched Predictions, Unmat. Pred. : Unmatched Predictions, Match Rate : Matching Rate, Acc. : Accuracy (exactitude), Olla. : Ollama (prompt JSON), Unsl. : Unsloth (prompt en langage naturel)

Voici une explication détaillée des concepts :

— Matched Predictions : Cela correspond aux prédictions du LLM qui génèrent un nom de catégorie valide (parmi les catégories attendues).

— Unmatched Predictions : Ces prédictions correspondent à des résultats où le modèle génère une réponse qui n'appartient pas à l'ensemble des catégories pré-définies.

— Matching Rate :

$$\frac{MatchedPredictions}{MatchedPredictions + UnmatchedPredictions}$$

— Accuracy (exactitude) : l'exactitude mesure le pourcentage de prédictions correctes.

7.5 Analyse des résultats

Les modèles utilisant Ollama et le prompt JSON présentent des taux de correspondance (matching rate) et des exactitudes nettement supérieurs, en particulier pour Llama3.2-11B-Vision et Qwen2.5-14B.

Les performances des modèles sur la plateforme Unsloth, en particulier avec le prompt en langage naturel, sont globalement plus faibles, ce qui peut être attribué à l'absence de structure formelle dans les résultats générés. Pour une analyse plus approfondie des performances, nous avons généré des matrices de confusion pour les modèles Ollama-Llama3.2-11B-Vision et Ollama-Qwen2.5-14B.

Ollama-Llama3.2-11B-Vision :

— Le modèle excelle dans les catégories majoritaires comme *incidents/supervision* (526 prédictions correctes).

FIGURE 1 – Matrice de confusion pour Ollama-Llama3.2-Vision

— Cependant, des confusions subsistent dans les catégories moins fréquentes telles que *demande de service/backup bcs/stratégie de sauvegarde/suppression*, avec une faible exactitude.
— La catégorie *UNMAPPED* correspond à « impossible à classer ». Il y a deux cas distincts. Premièrement, conformément aux consignes du prompt, le modèle retourne *UNKNOWN* en cas d'incertitude sur la classification, ce résultat étant affiché comme *UNMAPPED* dans la figure. Deuxièmement, le LLM génère des catégories totalement impossibles à identifier, même après vérification et normalisation, comme mentionné précédemment.

FIGURE 2 – Matrice de confusion pour Ollama-qwen2.5

Ollama-Qwen2.5-14B :

— Avec une exactitude globale de 85,20 %, ce modèle se distingue par sa gestion efficace de catégories comme demande de *service/backup bcs/stratégie de sauvegarde/suppression*, où il surpasse Llama3.2 en termes de classifications correctes (23 contre 14).
— Bien que ses performances globales soient élevées, le taux d'erreurs dans *UNMAPPED* est légèrement supérieur à celui de Llama3.2.

Ces expérimentations mettent en lumière l'importance d'une ingénierie des prompts adaptée et d'une plateforme robuste pour maximiser les performances des modèles. Les résultats obtenus serviront de base pour les étapes ultérieures d'optimisation et de fine-tuning.

8 Procédure d'entraînement (Fine-Tuning)

Dans cette section, nous décrivons le processus de fine-tuning appliqué aux modèles Llama 3.2 Vision 11B et Qwen 2.5 14B, en mettant en avant les motivations, les configurations techniques, et les résultats préliminaires.

8.1 Contexte et sélection des modèles

Sur la base des résultats des expérimentations initiales, deux modèles ont été retenus pour le fine-tuning :
Llama 3.2 Vision 11B : Modèle multimodal performant, ayant démontré une exactitude initiale élevée dans les tâches de classification.
Qwen 2.5 14B : Modèle purement linguistique, reconnu pour ses capacités à gérer les tâches de compréhension et d'extraction dans des contextes complexes.
Le choix de ces modèles repose principalement sur deux critères essentiels. Tout d'abord, leur performance initiale s'est révélée nettement supérieure lors des évaluations comparatives, attestant de leur efficacité et de leur pertinence pour les tâches ciblées.
Ensuite, la question de la compatibilité avec les plateformes a également joué un rôle déterminant. En effet, bien que le modèle Llama 3.2 Vision 11B présente des capacités multimodales avancées, les contraintes actuelles de la plateforme Ollama en matière d'intégration de modèles multimodaux privés compliquent son déploiement [1] [2]. Cette limitation souligne l'importance d'optimiser Qwen 2.5 14B, un modèle exclusivement textuel, afin d'assurer une intégration plus fluide et efficiente au sein de l'environnement ciblé.

8.2 Configurations techniques du fine-tuning

Dans notre démarche d'optimisation du fine-tuning, nous avons adopté la plateforme Unsloth, privilégiant une approche de quantification du modèle afin d'améliorer l'efficacité en termes de consommation mémoire et de rapidité de convergence. L'utilisation de modèles pris en charge par Unsloth permet ainsi d'exploiter pleinement les ressources matérielles disponibles, tout en garantissant des performances optimales.
Tous nos fine-tuning sont réalisés à partir de la version quantifiée (4 bits) des modèles d'origine, ce qui réduit significativement l'empreinte mémoire et permet l'entraînement sur des GPU disposant d'une VRAM limitée.
Dans notre cas spécifique, la tâche est exclusivement linguistique et ne nécessite pas de traitement d'images. Toute-

1. https ://github.com/unslothai/unsloth/issues/1504
2. https ://github.com/ollama/ollama/issues/7912

fois, Llama 3.2 Vision 11B étant un modèle multimodal, nous avons mis en place des mesures visant à préserver ses capacités en vision dans l'éventualité d'un usage futur, comme l'intégration du traitement d'images jointes aux e-mails. Pour ce faire, nous figeons les poids de la partie « image » (*image tower*) et ciblons le fine-tuning uniquement sur les couches linéaires du modèle de langage, en excluant notamment la *lm_head* afin de préserver la stabilité du modèle [5, 2]. Concrètement, nous nous concentrons sur l'ajustement

- des couches d'attention (*q_proj*, *k_proj*, *v_proj*, *o_proj*)
- ainsi que des couches de MLP (*gate_proj*, *up_proj*, *down_proj*).

Pour le modèle Qwen 2.5 14B, nous appliquons un entraînement sur l'ensemble des couches linéaires, à l'exception de la *lm_head*.

Au cours de nos expériences, nous avons testé plusieurs configurations de paramètres afin d'évaluer leur impact sur les performances du modèle. Dans cet article, nous présentons les résultats d'entraînement obtenus à partir d'un ensemble de paramètres optimisés. Ce choix vise à équilibrer l'utilisation des ressources computationnelles et la performance du modèle, garantissant ainsi un entraînement stable et efficace, même dans un environnement matériel contraint.

Pour garantir un benchmark équitable, nous avons utilisé les mêmes configurations pour le fine-tuning des deux modèles : Llama 3.2 Vision 11B et Qwen 2.5 14B. Ci-dessous, la configuration spécifique des modèles fine-tunés :

- Paramètres PEFT (Parameter Efficient Fine-Tuning)
  - *r=16* : Rank élevé pour garantir une précision optimale.
  - *lora_alpha=16* : Aligné avec la valeur de r.
  - *lora_dropout=0* : Aucun dropout appliqué.
  - *bias="none"* : Pas de biais utilisé.
  - *use_rslora=True* : Stabilise et optimise l'entraînement.
  - *loftq_config=None* : LoftQ non activé.
- Paramètres d'entraînement :
  - Batch size par appareil : 2.
  - Gradient accumulation steps : 4.
  - Étapes de warmup : 5.
  - Taux d'apprentissage : 2e-4.
  - Précision : bf16.
  - Optimiseur : *adamw_8bit*.
  - Décay : 0,01.
  - Longueur maximale des séquences : 2048.
  - Quantification : 4 bits.

8.3 Expérimentations des fine-tunings

Pour chaque modèle, deux versions ont été entraînées : une avec 30 steps de mise à jour des poids, et une autre avec 1 époque complète (correspondant à environ 440 steps). Ces versions permettent d'évaluer l'impact du nombre d'itérations d'entraînement sur les performances des modèles.

Pour analyser les performances des modèles, nous avons examiné l'évolution de la perte (loss) au cours de l'entraînement. L'évaluation s'est concentrée sur la stabilité et la convergence de la perte dans le cadre d'une époque complète (1ep).

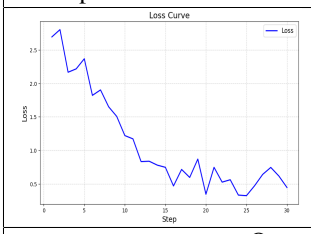
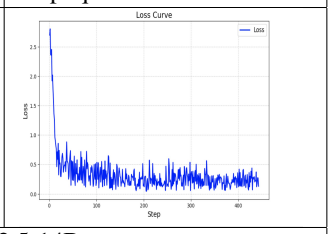
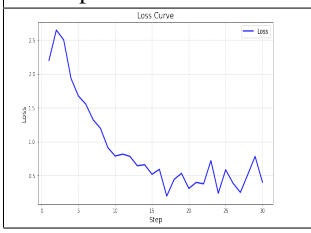
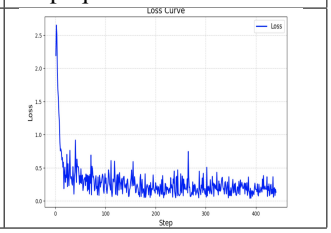
Llama-3.2-11B-Vision	
30 steps	1 époque
	
Qwen-2.5-14B	
30 steps	1 époque
	

TABLE 3 – Courbes de perte (Loss curves) des fine-tunings

Modèle Qwen-2.5-14B : Ce modèle a montré une perte moyenne de 0,8505 avec un écart-type de 0,6461 sur les 30 dernières étapes. Ces résultats suggèrent une bonne capacité de convergence avec une perte relativement faible et des variations limitées, témoignant d'une grande stabilité en fin d'entraînement. De plus, la valeur minimale de perte atteinte, 0,1962, illustre son potentiel à fournir des prédictions précises.

Modèle LLaMA-3.2-11B-Vision : En comparaison, LLaMA a affiché une perte moyenne légèrement plus élevée de 1,0489, accompagnée d'un écart-type de 0,6963. Les variations de la perte, bien que modérées, sont légèrement supérieures à celles de Qwen. La valeur minimale de perte atteinte, 0,3238, bien qu'encourageante, reste au-dessus de celle observée pour Qwen.

En conclusion, sur le critère de la perte, le modèle Qwen surpasse LLaMA, non seulement par une perte moyenne inférieure, mais également par une meilleure stabilité en fin d'entraînement. Ces résultats confirment que Qwen est mieux adapté à converger rapidement tout en minimisant les erreurs.

9 Benchmark des modèles fine-tunés

Cette section présente les performances des modèles fine-tunés sur deux plateformes (Ollama et Unsloth) et selon deux formats de sortie (JSON imposé et texte libre simulant du JSON). Au total, six variantes ont été évaluées :

- *Qwen-Ollama-JSON* : 1 époque et 30 steps
- *Qwen-Unsloth–"naturel"* : 1 époque et 30 steps
- *Llama-Unsloth–"naturel"* : 1 époque et 30 steps

Sur Ollama, la sortie est systématiquement contrainte au format JSON. En revanche, le modèle Llama fine-tuné,

étant multimodal, présente des problèmes de compatibilité avec cette plateforme (voir Section 8.1).
Concernant Unsloth, bien que les sorties soient qualifiées de « naturelles », le prompt oriente tout de même le modèle vers une structure JSON (ex. *Output : {'categorisation' : ...}*). Un post-traitement similaire à celui décrit en Section 7.3 est ensuite appliqué pour harmoniser les résultats.

9.1 Résultats des expérimentations des modèles fine-tunés

Mod.	Plat.	Mat. Pred.	Unmat. Pred.	Match Rate	Acc.
Qwen 1ep.	Olla.	891	1	0,9989	0,8285
Qwen 30steps	Olla.	873	19	0,9787	0,8509
Qwen 1ep.	Unsl.	**892**	**0**	**1,0**	**0,8812**
Qwen 30steps	Unsl.	886	6	0,9933	0,8565
Llama 1ep.	Unsl.	889	3	0,9966	0,8610
Llama 30steps	Unsl.	870	22	0,9753	0,8105

TABLE 4 – Benchmark des modèles fine-tunés (Mod. : Modèle, Plat. : Plateforme, Mat. Pred. : Matched Predictions, Unmat. Pred. : Unmatched Predictions, Match Rate : Matching Rate, Acc. : Accuracy (exactitude), Olla. : Ollama (prompt JSON), Unsl. : Unsloth (prompt en langage naturel)

9.2 Analyse des performances des modèles fine-tunés

Les modèles Qwen et LLaMA montrent une amélioration significative de leurs performances après fine-tuning. Par exemple, LLaMA passe d'une exactitude initiale de 0,4540 (avant micro-ajustement) à plus de 0,8610, surpassant ainsi le score de 0,8049 obtenu avec la version non ajustée sous Ollama.
Le meilleur score global (0,8812) est obtenu avec Qwen (1 époque) sur Unsloth, en cohérence avec la tendance observée dans l'évolution de la fonction de perte (loss).

9.3 Analyse des matrices de confusion

Une analyse des matrices de confusion des modèles fine-tunés (1 époque) sur Unsloth révèle les observations suivantes :

Qwen - 1 époque - Unsloth (sortie naturelle)

- La majorité des classes sont correctement identifiées, en particulier les catégories majoritaires telles que *Incidents/Supervision* ou *Incidents/Backup BCS/Sauvegarde*.
- Les catégories moins fréquentes (ex. *Demande de service/Backup #BCS/Stratégie de sauvegarde/Suppression*) sont également bien prédites,

FIGURE 3 – Matrice de confusion pour Unsloth-qwen2.5 affiné en 1 époque

avec un faible taux de confusion.

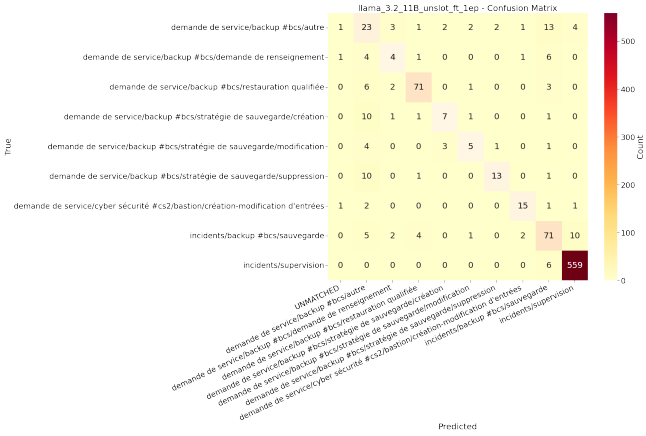


FIGURE 4 – Matrice de confusion pour Unsloth-Llama3.2-Vision affiné en 1 époque

LLaMA - 1 époque - Unsloth (sortie naturelle)

- Une amélioration significative est observée par rapport à la version non ajustée, y compris pour les classes minoritaires, grâce au fine-tuning.
- Les confusions résiduelles concernent principalement des catégories proches, mais restent limitées compte tenu de la taille du jeu de données.

Dans l'ensemble, les fines distinctions entre catégories proches sont mieux gérées lorsque le modèle est spécifiquement entraîné sur le jeu de données annoté. Le fine-tuning permet ainsi d'affiner la capacité de classification, de réduire le taux de prédictions *UNMAPPED* et d'augmenter l'exactitude globale.

9.4 Discussion sur l'écart de performance sous Ollama

Les légers reculs observés sur Ollama peuvent être attribués à plusieurs facteurs. Tout d'abord, la conversion et la quan-

tification du modèle vers le format *gguf* introduisent des approximations supplémentaires liées à la re-quantification et à la compression, ce qui peut altérer légèrement la précision. De plus, des différences dans les pipelines d'entraînement et d'exécution pourraient également expliquer cette baisse de performance. Le fine-tuning ayant été réalisé via Unsloth, le modèle a ensuite été porté sous Ollama, un processus impliquant plusieurs transformations susceptibles d'affecter la cohérence des poids finaux.

10 Conclusion et perspectives

10.1 Conclusion

Dans le cadre de ce travail, nous avons évalué plusieurs modèles de langage d'environ 13 milliards de paramètres dans un environnement industriel, démontrant qu'il est possible d'atteindre un niveau de performance élevé (une exactitude de 88,12%) tout en respectant des contraintes matérielles et de confidentialité.

L'approche multi-agents retenue, qui sépare la classification et l'extraction d'informations en différents « agents », a été mise en œuvre dans le cadre de ce projet. Nous pensons qu'elle offre une modularité propice à de futures évolutions et un gain de maintenance, mais une étude plus approfondie reste à mener pour valider formellement son efficacité à grande échelle.

De plus, nos expérimentations soulignent l'importance de la quantification (en 4 bits) et du *fine-tuning* ciblé (via PEFT/LoRA), solutions qui permettent de concilier efficacité opérationnelle et qualité des prédictions dans un contexte de ressources limitées.

L'ensemble des travaux présentés dans cet article, incluant le fine-tuning et les évaluations comparatives, est mis à disposition en accès libre sur le dépôt GitHub[3], à l'exception des jeux de données soumis à des contraintes de confidentialité.

10.2 Perspectives

À court terme, les perspectives s'organisent principalement autour de deux axes :

1. Améliorations de l'agent de classification.

Bien que la première version de l'agent de classification ait démontré une performance prometteuse, des enrichissements restent prévus : l'équipe projette d'élargir et d'actualiser le jeu de données d'entraînement afin d'affiner à nouveau le modèle Qwen2.5. L'objectif est de mieux couvrir l'éventail des catégories de tickets et de consolider la robustesse du système dans des scénarios réels encore plus variés.

2. Mise en place des agents d'extraction.

Le développement d'agents spécialisés dans l'extraction d'informations à partir des e-mails clients représente la prochaine étape clé du projet. Une première preuve de concept a été réalisée sur la catégorie « Restauration qualifiée », nécessitant l'identification de cinq champs : la date de restauration, le chemin source, le type, le chemin de destination, le nom d'hôte. Nous avons déjà fait les benchmarks des modèles initiaux pour cette tâche. Les résultats indiquent que Qwen 2.5 14B et Llama 3.2 13B Vision restent deux candidats de choix pour cette tâche, mais la constitution d'un corpus annoté demeure un défi. L'équipe prévoit donc d'exploiter une stratégie d'auto-annotation [14], puis de valider un échantillon des données par des experts humains avant de procéder au fine-tuning final. Cette approche devrait accélérer le processus d'étiquetage, tout en garantissant la qualité des annotations indispensables à l'entraînement d'un modèle d'extraction fiable.

À long terme, notre ambition est d'intégrer l'ensemble de ces agents (classification et extraction) dans une architecture unifiée, afin de faciliter la génération automatique de tickets et leur suivi complet. Enrichir et diversifier les données, affiner les prompts et perfectionner les algorithmes de fine-tuning feront partie des travaux futurs. Ainsi, nous espérons rendre ces technologies d'intelligence artificielle de plus en plus accessibles et performantes pour les besoins industriels de traitement d'e-mails et de support client.

3. https://github.com/LeviatanAI/email-ticket-classifier

Références

[1] Michael Han Daniel Han and Unsloth team. Unsloth, 2023.

[2] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora : Efficient finetuning of quantized llms, 2023.

[3] Hugging Face. Open llm leaderboard. Accessed : 2025-02-21.

[4] Python Software Foundation. *difflib — Helpers for computing deltas*, 2025. Accessed : 2025-02-21.

[5] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora : Low-rank adaptation of large language models, 2021.

[6] Shiyao Li, Xuefei Ning, Luning Wang, Tengxuan Liu, Xiangsheng Shi, Shengen Yan, Guohao Dai, Huazhong Yang, and Yu Wang. Evaluating quantized large language models, 2024.

[7] Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. Peft : State-of-the-art parameter-efficient fine-tuning methods, 2022.

[8] Meta. Llama 3.2 : Revolutionizing edge ai and vision with open, customizable models, September 2024. Accessed : 2025-02-21.

[9] Ollama. Ollama. Accessed : 2025-02-21.

[10] OpenAI. Openai platform models. Accessed : 2025-02-21.

[11] Wikipedia contributors. Distance de levenshtein — wikipedia, the free encyclopedia, 2025. Accessed : 2025-02-21.

[12] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian

Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2024.

[13] Yingxuan Yang, Qiuying Peng, Jun Wang, Ying Wen, and Weinan Zhang. Llm-based multi-agent systems : Techniques and business perspectives, 2024.

[14] Ying Zhang, Matthieu Petit Guillaume, Aurélien Krauth, and Manel Labidi. Cryptogpt : a 7b model rivaling gpt-4 in the task of analyzing and classifying real-time financial news, 2024.

A Annexe

A.1 Prompt avec sortie en format JSON sous Ollama

```
Tu es un assistant spécialisé dans la
classification de tickets à partir de
leur contenu textuel. Ton objectif est
d'analyser la description fournie et
de l'associer à l'une des catégories
ci-dessous:
```
- Demande de service/Backup #BCS/Autre
- Demande de service/Backup
#BCS/Demande de renseignement
- Demande de service/Backup
#BCS/Restauration qualifiée
- Demande de service/Backup
#BCS/Stratégie de sauvegarde/Création
- Demande de service/Backup
#BCS/Stratégie de sauvegarde/Modification
- Demande de service/Backup
#BCS/Stratégie de sauvegarde/Suppression
- Demande de service/Cyber Sécurité
#CS2/Bastion/Création-Modification
d'entrées
- Incidents/Backup #BCS/Sauvegarde
- Incidents/Supervision
```
### Règles :
1. Réponds uniquement par la catégorie
exacte, sans texte supplémentaire.
2. La catégorie associée doit être
unique.
3. Si aucune catégorie ne correspond,
la valeur de "categorisation" doit être
"unknown".
### Exemple :
La description:
Bonjour¤ Merci de relancer les
sauvegardes FULL¤ si ils ne sont pas
repassées en automatique. Ghislain
Envoyé de mon iPhone Début du message
transféré
### Sortie :
{{
"categorisation": "Incidents/Backup
#BCS/Sauvegarde "
}}
Analyse uniquement la description
suivante :
{content}
```

A.2 Prompt avec sortie en langage naturel sous Unsloth

```
Tu es un assistant spécialisé dans la
classification de tickets à partir de
leur contenu textuel. Ton objectif est
d'analyser la description fournie et
de l'associer à l'une des catégories
ci-dessous:
```
- Demande de service/Backup #BCS/Autre
- Demande de service/Backup
#BCS/Demande de renseignement
- Demande de service/Backup
#BCS/Restauration qualifiée
- Demande de service/Backup
#BCS/Stratégie de sauvegarde/Création
- Demande de service/Backup
#BCS/Stratégie de sauvegarde/Modification
Demande de service/Backup
#BCS/Stratégie de sauvegarde/Suppression
- Demande de service/Cyber Sécurité
#CS2/Bastion/Création-Modification
d'entrées
- Incidents/Backup #BCS/Sauvegarde
- Incidents/Supervision
"'
Règles :
1. Réponds uniquement par la catégorie
exacte, sans texte supplémentaire, la
réponse doit être encapsuler par deux
*.
2. Une seule catégorie maximum sera
retenue.
3. Si aucune catégorie ne correspond,
réponds simplement par *unknown*.
Exemple :
La description:
Bonjour¤ Merci de relancer les
sauvegardes FULL¤ si ils ne sont pas
repassées en automatique. Ghislain
Envoyé de mon iPhone Début du message
transféré
Sortie :
Incidents/Backup #BCS/Sauvegarde
Analyse uniquement la description
suivante :
{content}
```
```

Vers une optimisation de RAG en français : conception d'un reranker open source, fine-tuning et évaluation

Ying Zhang[1], Matthieu Petit Guillaume[1], Aurélien Krauth[1]

[1] Leviatan, 725 Bd Robert Barrier, 73100 Aix-les-Bains

{y.zhang, matthieu, aurelien}@leviatan.fr

Résumé

Dans les systèmes RAG, la qualité des documents extraits conditionne fortement la pertinence des réponses générées. Nous proposons un reranker français open source, basé sur un modèle de type Cross-Encoder, afin de réordonner les documents candidats et ainsi améliorer la cohérence du texte généré. Nous détaillons la construction d'un jeu de données combinant un corpus de similarité sémantique et des données de questions-réponses en français, puis comparons plusieurs stratégies de fine-tuning (ajustement fin). Les résultats montrent qu'un modèle open source optimisé peut rivaliser avec des solutions commerciales, à condition de bien choisir le modèle pré-entraîné et d'apporter une attention particulière à la génération d'exemples négatifs. Nous proposons enfin des pistes d'amélioration pour renforcer davantage la performance globale des systèmes RAG.

Mots-clés

RAG (Retrieval-Augmented Generation), Reranker, Retriever (module de recherche), Cross-Encoder, Fine-tuning, Questions-Réponses, Open Source, Traitement automatique du français

Abstract

In Retrieval-Augmented Generation (RAG) systems, the quality of retrieved documents is crucial for producing accurate and coherent answers. We introduce an open-source French reranker based on a Cross-Encoder architecture, designed to reorder candidate documents and enhance the coherence of the generated text. We describe the creation of a dataset that integrates semantic similarity corpora with French question-answering data and compare several fine-tuning strategies. Our results demonstrate that a carefully optimized open-source model can rival commercial rerankers, provided there is thoughtful selection of the pretrained model and careful design of negative examples. We conclude by suggesting improvements to further enhance overall RAG performance.

Keywords

RAG (Retrieval-Augmented Generation), Reranker, Retriever, Cross-Encoder, Fine-tuning (ajustement fin), Question Answering, Open Source, French NLP

1 Introduction

Dans les systèmes de génération augmentée par la recherche (Retrieval-Augmented Generation, RAG), la qualité des documents extraits par le module de recherche (Retriever) est un facteur déterminant pour la fiabilité et la pertinence des réponses produites. Les techniques de recherche traditionnelles, bien qu'efficaces pour un large éventail d'applications, peuvent toutefois introduire du bruit informationnel en sélectionnant des documents moins pertinents [3, 13]. Cette limitation se répercute sur la qualité du texte généré, qui s'appuie sur ces documents comme source d'information.

Pour pallier ce problème, l'intégration d'un reranker se révèle cruciale. Ce composant permet de réordonner et de prioriser les documents déjà extraits afin de mettre en avant les plus pertinents et les plus fiables. De cette façon, le générateur de texte dispose d'un contexte documenté mieux ciblé et plus pertinent, ce qui améliore la cohérence et la qualité globale de la réponse finale.

Dans cet article, nous mettons en évidence le rôle central joué par le reranker dans un système RAG. Nous montrons comment ce module peut être affiné (micro-ajusté) à l'aide de données privées pour renforcer davantage la pertinence du réordonnancement. Nous décrivons la préparation du jeu de données ainsi que l'expérimentation de micro-ajustement menée dans le cadre de plusieurs configurations expérimentales. Nous comparons ensuite nos résultats à ceux de solutions commerciales et open source dans un benchmark, mettant en évidence des performances contrastées : certaines configurations s'avérant très efficaces, d'autres moins. Nous analysons ces résultats et proposons des pistes d'optimisation.

Enfin, il convient de souligner que les modèles, les données d'entraînement, et les procédures de fine-tuning décrits dans cette étude sont entièrement accessibles en open source[1], afin de promouvoir la transparence et de faciliter l'adoption de ces travaux par la communauté.

1. Le projet est disponible à l'adresse suivante : https://github.com/LeviatanAI/reranker-cross-encoder, et les modèles ainsi que les jeux de données sont accessibles à l'adresse : https://huggingface.co/LeviatanAIResearch.

2 Contexte et problématique

La technique de RAG combine un module de recherche (Retriever) et un générateur de texte (par exemple GPT [18] ou Llama [16]) afin d'améliorer la pertinence et l'exactitude des réponses produites. Contrairement aux modèles purement pré-entraînés qui se reposent exclusivement sur leurs connaissances internes, un système RAG s'appuie également sur une base de connaissances ou un ensemble de documents externes (base documentaire, corpus textuel, etc.), permettant ainsi d'offrir des réponses plus complètes et régulièrement mises à jour.

Dans sa forme la plus élémentaire, un système RAG comporte deux composantes principales :

1. Le Retriever (module de recherche) : il interroge une base de connaissances ou une base de données pour identifier les documents ou passages les plus proches de la requête de l'utilisateur. Parmi les méthodes de recherche les plus courantes, on retrouve la recherche sémantique par vecteurs (particulièrement adaptée aux textes longs ou non structurés) [11, 22], la recherche par mots-clés (souvent efficace pour des textes courts ou structurés) [23, 24] ou encore la recherche hybride (qui combine mots-clés et représentations vectorielles) [28].

2. Le Générateur : il s'agit généralement d'un grand modèle de langage (par exemple GPT [18] ou Llama [16]) qui, pour construire sa réponse, tient compte à la fois des documents extraits et de ses propres connaissances linguistiques.

L'un des avantages majeurs de RAG réside dans la réduction du phénomène d'« hallucination » — lorsque le modèle invente ou déforme des informations. En consultant des documents externes, le générateur est en mesure de vérifier le contenu avant de l'incorporer dans sa réponse, ce qui augmente la fiabilité du texte produit.

Par ailleurs, RAG offre d'autres avantages importants :

- Flexibilité et mise à jour des connaissances : en séparant les données du modèle lui-même, RAG permet d'adapter facilement les informations accessibles sans nécessiter un nouvel entraînement. Cela facilite l'intégration de nouvelles connaissances et permet l'exploitation de données privées dans un environnement sécurisé, sans les exposer ni les inclure dans l'apprentissage du modèle.
- Meilleure traçabilité : le lien explicite entre un document et la réponse générée renforce la confiance des utilisateurs, qui peuvent vérifier l'origine des informations fournies.

Dans la majorité des scénarios industriels de RAG, les données manipulées sont principalement non structurées (textes libres, documents PDF, pages web, etc.). Une méthode répandue pour en extraire des passages pertinents repose sur la recherche sémantique vectorielle, où la requête de l'utilisateur est convertie en une représentation vectorielle pour identifier les passages les plus proches, transmis ensuite au Générateur pour produire la réponse.

L'idée sous-jacente est de projeter à la fois la requête et les documents (ou passages) dans un espace vectoriel de dimension fixe, puis de comparer leurs représentations pour en mesurer la proximité sémantique. La méthode du bi-encoder [22] est couramment utilisée dans ce cadre :

- Un encodeur (souvent basé sur BERT [6], RoBERTa [14] ou des modèles similaires) transforme chaque document ou passage en un vecteur d'embedding. Ce calcul peut être effectué hors ligne et le résultat stocké dans une base de données vectorielle (comme FAISS [7], Milvus [30] ou Elasticsearch [8]).
- Lors de la requête, le même encodeur (ou un encodeur symétrique) génère un vecteur d'embedding représentant la question de l'utilisateur.
- Enfin, une mesure de similarité (produit scalaire ou distance cosinus) est calculée entre le vecteur de requête et les vecteurs stockés. Les documents présentant le score de similarité le plus élevé sont considérés comme les plus pertinents.

Si cette approche offre l'avantage d'un traitement rapide grâce à l'indexation optimisée et d'une bonne couverture sémantique, elle présente toutefois certaines limites :

- Sensibilité aux biais : le modèle, entraîné pour produire des embeddings généraux, peut négliger certaines nuances contextuelles ou linguistiques propres à la requête.
- Scores de pertinence approximatifs : pour des requêtes complexes (plusieurs sous-questions, contraintes de style, etc.), l'utilisation exclusive d'une similarité vectorielle peut demeurer insuffisante pour hiérarchiser correctement les documents.

Pour contourner ces limites et maximiser les chances de fournir au Générateur les documents les plus adéquats, une pratique courante consiste à extraire un grand nombre de candidats au stade du Retriever (bi-encoder). Cependant, l'envoi d'un volume important de documents au Générateur pose deux problèmes majeurs : d'une part, la consommation de ressources de calcul explose ; d'autre part, la qualité du texte généré peut se dégrader.

C'est pour répondre à ces défis qu'intervient le Reranker [17, 22] : en réordonnant les documents sélectionnés par le bi-encoder, il permet d'augmenter la précision globale du système RAG tout en réduisant la charge computationnelle du Générateur, comme illustré à la figure 1. Le module Reranker se place ainsi comme un maillon essentiel dans la chaîne de traitement, garantissant une sélection plus ciblée et plus pertinente des informations destinées à la phase de génération.

3 État de l'art

3.1 Principales approches de reranking

Plusieurs approches de reranking coexistent, chacune présentant des avantages et des inconvénients. Parmi les méthodes les plus courantes, on distingue notamment :

1. Le Cross-Encoder [17] basé sur des Transformers [29]
Le Cross-Encoder constitue aujourd'hui l'une des approches de reranking les plus populaires. Contrairement aux bi-encoders, qui génèrent un embedding séparé pour la requête et pour chaque document, son principe repose sur

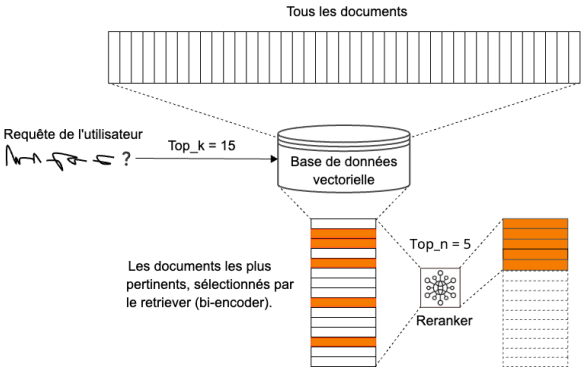


FIGURE 1 – Système de recherche en deux étapes [1]

l'encodage simultané de la requête et du document dans une même séquence en entrée du modèle (par exemple, [CLS] QUESTION [SEP] DOCUMENT [SEP]) afin de produire un score de pertinence. Cette méthode permet de capturer de manière fine les interactions entre le texte de la requête et celui du document, ce qui se traduit souvent par des performances remarquables en termes de classement. Cependant, cette granularité accrue implique un coût computationnel élevé : en effet, chaque paire (REQUÊTE, DOCUMENT) doit être traitée conjointement par le réseau, ce qui peut devenir onéreux lorsque le nombre de candidats à réordonner est important.

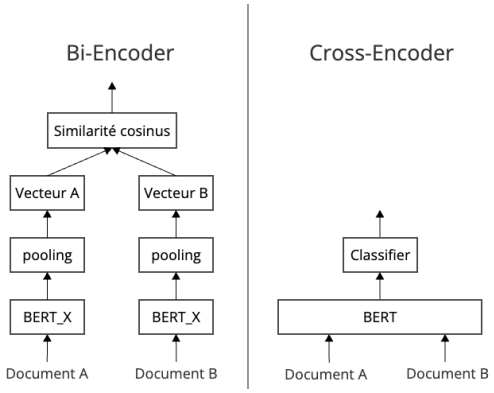


FIGURE 2 – Structure de bi-encoder et de cross-encoder [22]

2. Le Reranker génératif
Dans cette catégorie, on utilise un modèle génératif (par exemple T5 [20], GPT [18] ou d'autres grands modèles de langage) pour évaluer la pertinence ou produire directement un score de classement. Les approches génératives peuvent se subdiviser en deux sous-catégories :

— Reranker basé sur T5 [9]

T5, entraîné à réaliser de multiples tâches de traitement automatique du langage naturel, peut être adapté pour « générer » un score de pertinence ou un label de pertinence. Les formulations dites « prompt-based » permettent d'exploiter la flexibilité du modèle pour réaliser l'inférence des scores de classement. L'avantage de cette approche réside dans sa polyvalence et sa capacité à généraliser sur divers jeux de données. Toutefois, son efficacité dépend de la qualité du prompt et du fine-tuning réalisé, et son déploiement peut s'avérer coûteux en ressources si le modèle est de grande taille.

— Reranker basé sur les LLM [21]

À l'instar de GPT ou d'autres grands modèles de langage, cette variante suit une logique similaire, mais tire parti de la puissance (et parfois des connaissances internalisées) des modèles très volumineux. D'un côté, cela peut conduire à de meilleures performances, notamment pour des questions ambiguës ou nécessitant des connaissances larges. De l'autre, l'important coût de calcul et les contraintes d'accès (modèles propriétaires ou services payants) peuvent limiter l'adoption à grande échelle.

3.2 Comparaison et enjeux

Les Cross-Encoders se distinguent par leur capacité à capturer les interactions requête-document avec un haut niveau de précision, au prix toutefois d'un coût de calcul plus conséquent. Les méthodes génératives, quant à elles, offrent une plus grande flexibilité et peuvent intégrer des signaux sémantiques complexes, mais restent encore très dépendantes de la taille du modèle et du prompt utilisé.
Dans cette étude, nous nous concentrons sur l'approche la plus largement adoptée en pratique : le Cross-Encoder. Notre objectif est de mettre au point un reranker multilingue, facile à déployer, peu coûteux en ressources et disponible en open source. Dans un premier temps, les expériences présentées porteront sur un ensemble de données en langue française, afin d'illustrer la faisabilité et l'efficacité de l'approche proposée dans un contexte concret.

4 Expérimentation

4.1 Composition du jeu de données

4.1.1 Source des jeux de données

Dans le cadre de nos travaux, nous nous sommes appuyés sur deux grandes catégories de données pour entraîner et évaluer notre reranker :

1. Données dédiées à la similarité sémantique

Nous avons utilisé le STS Benchmark multilangue (stsb) [2], un corpus de référence pour l'évaluation de la similarité sémantique entre paires de phrases. Sa structure, composée de paires de textes et de scores de similarité, se prête particulièrement bien à l'apprentissage de modules de réordonnancement (ou reranking).

2. Données de questions-réponses en langue française

Afin de disposer d'un jeu de données riche et varié, nous avons rassemblé quatre sources ouvertes de questions-réponses en français :

— PIAF [12]
— FQuAD [15]
— SQuAD-French [10]
— pandora-s/neural-bridge-rag-dataset-12000-google-translated [19], filtrée pour la langue française (pandora-rag-fr)

L'utilisation conjointe de ces corpus vise à couvrir un large éventail de thématiques et de formulations linguistiques, permettant ainsi de tester la robustesse du reranker à différents styles de questions et à divers types de documents.

4.1.2 Processus général de construction des jeux de données

Pour le jeu de données STS Benchmark multilingue, nous avons d'abord extrait la portion en français, puis normalisé les scores de similarité afin de les contraindre à l'intervalle [0, 1].
Dans le but de garantir une cohérence optimale dans l'entraînement du reranker, un protocole d'agrégation et de transformation a été établi pour le deuxième type de jeu de données (données de questions-réponses), comprenant notamment les étapes suivantes :

1. Chargement et filtrage linguistique

Chaque corpus (PIAF, FQuAD, SQuAD-French, pandora-rag-fr) est d'abord chargé individuellement, afin de repérer et d'extraire les échantillons pertinents (en l'occurrence, uniquement ceux rédigés en français lorsque le corpus présente une composante multilingue).

2. Génération des paires positives

Les entrées constituées d'un contexte et d'une ou plusieurs questions associées forment les exemples « positifs » (LABEL = 1). Chaque couple (CONTEXTE, QUESTION) ainsi validé illustre une correspondance avérée entre la question et l'extrait textuel sélectionné.

3. Création d'exemples négatifs

Pour renforcer la capacité du modèle à discriminer le pertinent de l'impertinent, des paires artificiellement « négatives » (LABEL = 0) sont générées. Concrètement, celles-ci associent une question aléatoire issue d'un autre document ou d'une autre entrée, de sorte qu'elle ne corresponde pas au contexte considéré.

4. Fusion et structuration finale

Les jeux de données issus de chaque corpus sont ensuite fusionnés en un ensemble unique, organisé sous forme de tableaux structurés (champs contexte, question, label et source). Cette étape permet de regrouper de manière homogène les différents types d'exemples (positifs et négatifs).

5. Préparation pour l'entraînement

Afin de faciliter l'intégration dans les pipelines de traitement du langage, nous avons intégré et mis en ligne le jeu de données sur Hugging Face. La compilation des exemples finaux couvre un large spectre de cas où la question est (ou non) liée au contexte fourni. Concernant la répartition des données, nous respectons la division train/test d'origine. Autrement dit, le jeu de données initialement dédié à l'entraînement demeure utilisé pour le train set, et celui destiné aux tests reste inchangé. Seul le jeu de données « PIAF » n'ayant pas de découpage préexistant, nous avons procédé à une répartition séquentielle en affectant les 70 % des données initiales au train set et les 30 % restants au test set.

Grâce à ce processus, un volume conséquent de données étiquetées, composé de paires cohérentes et incohérentes, est mis à la disposition du modèle de reranking. L'objectif est d'entraîner celui-ci à attribuer un score de pertinence élevé aux paires contextes-questions justifiées, et un score faible dans les cas où la question ne correspond pas au contenu textuel.

4.2 Fine-tuning (ajustement fin)

4.2.1 Combinaison des jeux de données et des modèles

Pour l'entraînement de notre reranker, nous nous sommes appuyés sur la méthodologie proposée par l'équipe Sentence Transformers. Celle-ci fournit une classe dédiée CROSSENCODER qui s'appuie en interne sur la classe AUTOMODELFORSEQUENCECLASSIFICATION de Hugging Face, tout en offrant des fonctionnalités simplifiées pour l'entraînement et la prédiction des scores de similarité ou de pertinence [26].
Afin de procéder au fine-tuning, nous avons dû sélectionner un modèle pré-entraîné issu de la famille Transformer compatible avec AUTOMODEL. Pour évaluer l'impact du volume et de la diversité des données sur les performances du reranker, dans la première phase de nos tests, nous avons expérimenté 4 combinaisons de jeux de données et de modèles de base :

- **EXPÉRIENCE 1.** BERT de Google [6] + STS Benchmark en français (stsb-fr)
- **EXPÉRIENCE 2.** DistilRoBERTa-base [25] + (PIAF, FQuAD, SQuAD-french)
- **EXPÉRIENCE 3.** DistilRoBERTa-base + (PIAF, FQuAD, SQuAD-french, pandora-rag-fr)
- **EXPÉRIENCE 4.** DistilRoBERTa-base + (PIAF, FQuAD, SQuAD-french, pandora-rag-fr, stsb-fr)

Cette comparaison vise à mettre en évidence l'influence conjointe de la nature des données (questions-contextes versus paires de phrases corrélées sémantiquement) et du modèle de base (BERT vs. DistilRoBERTa) sur la qualité du réordonnancement.

4.2.2 Configuration du fine-tuning

Tous nos processus de fine-tuning ont été réalisés sur une carte GPU T4 et reposent exclusivement sur l'ensemble d'entraînement. Il est toutefois important de souligner que, contrairement à ces processus, les étapes d'inférence et de benchmark, détaillées par la suite, ont été effectuées sur CPU.
Afin de garantir une comparaison équitable entre les différentes configurations de fine-tuning, nous avons, après une phase de tests préliminaires, adopté une configuration unique pour l'ensemble des entraînements. Nous avons fixé les paramètres d'entraînement suivants, tandis que les autres valeurs ont été maintenues par défaut :

- EPOCH : 4
- TRAIN_BATCH_SIZE : 16
- SAVE_BEST_MODEL : True
- WARMUP_STEPS : 10 % des données d'entraînement

Concernant l'évaluation durant le processus d'entraînement, une distinction importante a dû être faite en fonction de la nature de nos jeux de données. En effet, notre jeu de données

comprend deux types de données :

1. Des scores continus (dans le cas du jeu de données STS Benchmark, normalisé pour produire des valeurs dans l'intervalle [0, 1])
2. Des scores binaires (issus des jeux de données de questions-réponses)

Par conséquent, le choix de l'évaluateur a varié selon les différentes expériences. Pour les expériences 2 et 3, où les jeux de données de fine-tuning ne contenaient pas le STS Benchmark et comportaient exclusivement des étiquettes binaires, l'évaluation a été réalisée en utilisant la précision moyenne et le meilleur score F1 possible. Dans ce contexte, nous avons employé le CEBinaryClassificationEvaluator [27].
En revanche, pour les expériences 1 et 4, qui intégraient le jeu de données STS Benchmark, nous avons opté pour le CECorrelationEvaluator [27], plus adapté à l'évaluation de la corrélation entre les scores prédits et les scores de référence continus.

4.3 Résultats préliminaires des fine-tunings

4.3.1 Comparaison des résultats entre les expériences

Dans cette section, nous proposons une analyse comparative des quatre expériences de fine-tuning réalisées. Pour chaque expérience, les modèles ont été évalués sur plusieurs ensembles de données, à savoir :

1. Les ensembles de développement et de test du STS Benchmark en français (stsb-fr Dev et stsb-fr Test). Voir la Table 1.

Expérience	stsb-fr Dev Set		stsb-fr Test Set	
	Pearson	**Spearman**	**Pearson**	**Spearman**
Exp. 1	0,8722	0,8692	**0,8362**	**0,8245**
Exp. 2	0,4710	0,6119	0,3492	0,4673
Exp. 3	0,5258	0,6990	0,4225	0,5908
Exp. 4	**0,9219**	**0,9187**	0,7565	0,7460

TABLE 1 – Résultats des différentes configurations sur le jeu de données STS Benchmark

2. Trois jeux de données internes élaborés à partir de données prétraitées provenant de PIAF, FQuAD, SQuAD-French, pandora-rag-fr et stsb-fr :
   (a) Un premier jeu intégrant l'ensemble des sources, évalué via la corrélation (Table 2).
   (b) Un second jeu constitué de PIAF, FQuAD et SQuAD-French, évalué à l'aide d'indicateurs de classification binaire (Table 3).
   (c) Un troisième jeu rassemblant PIAF, FQuAD, SQuAD-French et pandora-rag-fr, également évalué à l'aide d'indicateurs de classification binaire (Table 3).

Expérience	Pearson	Spearman
Exp. 1	0,7741	0,7608
Exp. 2	0,8892	0,8372
Exp. 3	0,8985	0,8594
Exp. 4	**0,9522**	**0,8966**

TABLE 2 – Résultats sur l'ensemble combiné des données PIAF, FQuAD, SQuAD-French, pandora-rag-fr et stsb-fr

Exp.	PIAF, FQuAD et SQuAD-Fr				
	Acc.	**F1**	**Prec.**	**Rec.**	**Avg. Prec.**
Exp. 1	0,9527	0,9529	0,9455	0,9603	0,9889
Exp. 2	**0,9765**	**0,9765**	**0,9785**	0,9745	0,9931
Exp. 3	0,9763	0,9762	0,9769	0,9756	**0,9955**
Exp. 4	0,9753	0,9754	0,9720	**0,9788**	0,9954
Exp.	**PIAF, FQuAD, SQuAD-Fr et pandora-rag-fr**				
	Acc.	**F1**	**Prec.**	**Rec.**	**Avg. Prec.**
Exp. 1	0,9468	0,9472	0,9410	0,9534	0,9858
Exp. 2	0,9747	0,9747	0,9778	0,9716	0,9935
Exp. 3	**0,9771**	**0,9770**	**0,9803**	0,9736	**0,9951**
Exp. 4	0,9767	0,9767	0,9791	**0,9743**	0,9952

TABLE 3 – Résultats sur les jeux de données de questions-réponses (Exp. : Expérience, Acc. : Accuracy (exactitude), F1 : F1 score, Prec : Precision, Rec. : Recall, Avg. Prec. : Average Precision)

4.4 Analyse comparative

Les résultats obtenus révèlent plusieurs points intéressants. D'une part, l'impact du modèle de base et des données utilisées est significatif. L'expérience 1, qui utilise le modèle BERT de Google avec uniquement le STS Benchmark, montre de bonnes performances en termes de corrélation sur stsb-fr. Toutefois, ses résultats sur les jeux de données internes restent inférieurs à ceux obtenus avec les modèles basés sur DistilRoBERTa-base. En effet, les expériences 2 à 4, qui se fondent sur DistilRoBERTa-base, affichent une amélioration notable sur les jeux de données internes, notamment en ce qui concerne les métriques de classification (accuracy, F1, etc.). Cependant, l'inclusion ou non de certains jeux de données (notamment pandora-rag-fr et stsb-fr) influence significativement les performances sur stsb-fr.
D'autre part, des disparités apparaissent entre les évaluations sur stsb-fr et les jeux de données internes. Pour l'expérience 2, par exemple, les scores de corrélation sur le stsb-fr (Pearson de 0,4710 en dev et de 0,3492 en test) sont nettement inférieurs à ceux obtenus sur le jeu de données combiné interne (Pearson 0,8892). Cette disparité suggère que le fine-tuning sur des données spécifiques à la tâche (questions-contextes) peut favoriser les performances sur ces derniers, au détriment d'un benchmark généraliste comme stsb-fr. En outre, l'expérience 4 présente les meilleures performances sur le stsb-fr Dev Set (Pearson 0,9219, Spearman 0,9187) et sur le jeu de données combiné interne (Pearson 0,9522, Spearman 0,8966), bien que la performance sur le stsb-fr Test Set soit

moins élevée (Pearson 0,7565, Spearman 0,7460). Ce résultat pourrait être lié à une suradaptation aux jeux de données internes ou à la nature hétérogène du benchmark stsb-fr.
Enfin, l'effet de l'ajout progressif des corpus se révèle pertinent. L'intégration de pandora-rag-fr, comme observé dans l'expérience 3, améliore globalement les performances sur les jeux de données internes par rapport à l'expérience 2, tant en termes de corrélation qu'en classification. Cela indique que la diversité des exemples contribue à la robustesse du reranker. De plus, l'intégration complète – c'est-à-dire l'ajout de stsb-fr en plus de pandora-rag-fr dans l'expérience 4 – permet d'atteindre des scores de corrélation maximaux sur le stsb-fr Dev Set et de maintenir des performances stables en classification, démontrant ainsi l'intérêt de combiner des jeux de données hétérogènes pour couvrir une grande variété de scénarios.

5 Benchmarks de nos modèles, du reranker commercial et du reranker open source

Dans cette section, nous comparons les performances de nos quatre modèles issus des expériences précédentes avec deux rerankers de référence : le reranker commercial COHERE RERANK-MULTILINGUAL-V2.0 [4] et le reranker open source DANGVANTUAN/CROSSENCODER-CAMEMBERT-LARGE [5].
L'évaluation a été réalisée en utilisant deux ensembles de test distincts, issus respectivement des jeux de données FQuAD et PIAF. Dans chaque cas, le système RAG utilise l'ensemble de test comme base documentaire.

5.1 Configuration de l'évaluation

Pour ces deux benchmarks, nous avons employé le modèle d'embedding INTFLOAT/MULTILINGUAL-E5-LARGE de Microsoft [31] en tant que bi-encoder. Le premier filtrage, réalisé par ce bi-encoder, repose sur une mesure de similarité cosinus qui permet de sélectionner les 30 candidats les plus pertinents.
Le benchmark basé sur le jeu de données FQuAD Test comprend un total de 3188 questions, tandis que celui reposant sur le jeu de données PIAF Test en inclut 1151. Dans le cas de PIAF, en l'absence d'une partition initiale, 70 % des données ont été utilisées pour l'entraînement et les 30 % restants pour le test, conformément à la méthode décrite en section 4.1.2. Il est à noter qu'aucune donnée de l'ensemble de test n'a été intégrée durant l'entraînement.

5.2 Résultats obtenus

Les performances ont été évaluées en mesurant le taux de rappel selon la métrique Recall@k, considérée pour trois niveaux de sélection : top 5, top 7 et top 10. Les résultats obtenus pour chacun des deux ensembles de test sont présentés dans la Table 4 et la Table 5.

Modèle	Top 5	Top 7	Top 10
Cohere Reranker	**92,50%**	**93,48%**	**94,26%**
D.V. CrossEncoder	52,23%	62,33%	71,46%
Expérience 1	**84,54%**	**87,92%**	**90,90%**
Expérience 2	30,14%	38,33%	49,53%
Expérience 3	39,12%	47,33%	56,81%
Expérience 4	**72,49%**	**78,67%**	**84,07%**

TABLE 4 – Benchmark réalisé sur l'ensemble de test de FQuAD (Cohere Reranker : COHERE RERANK-MULTILINGUAL-V2.0, D.V. CrossEncoder : DANGVANTUAN/CROSSENCODER-CAMEMBERT-LARGE)

Modèle	Top 5	Top 7	Top 10
Cohere Reranker	**95,57%**	**96,35%**	**97,22%**
D.V. CrossEncoder	61,77%	69,24%	78,63%
Expérience 1	**94,87%**	**96,00%**	**96,96%**
Expérience 2	38,66%	48,05%	58,99%
Expérience 3	56,73%	64,90%	73,59%
Expérience 4	**90,70%**	**93,83%**	**95,83%**

TABLE 5 – Benchmark réalisé sur l'ensemble de test de PIAF (Cohere Reranker : COHERE RERANK-MULTILINGUAL-V2.0, D.V. CrossEncoder : DANGVANTUAN/CROSSENCODER-CAMEMBERT-LARGE)

5.3 Analyse des résultats

Les résultats obtenus permettent de dégager plusieurs observations pertinentes :

1. Performances comparées

Le reranker commercial COHERE RERANK-MULTILINGUAL-V2.0 affiche des scores très élevés sur les deux ensembles de test, atteignant près de 97 % en top 10 sur le jeu de données PIAF. En comparaison, le reranker open source DANGVANTUAN/CROSSENCODER-CAMEMBERT-LARGE obtient des performances nettement inférieures, avec un écart particulièrement prononcé sur le FQuAD test set (71,46 % en top 10 contre 94,26 % pour cohere).

2. Performances des expériences

Parmi nos expériences, l'expérience 1 présente des résultats compétitifs, se rapprochant des performances du modèle commercial sur le jeu de données PIAF et obtenant de bonnes performances sur le jeu de données FQuAD. En revanche, les expériences 2 et 3 montrent des scores faibles, suggérant que la configuration et les données utilisées dans ces tests n'ont pas permis d'exploiter pleinement le potentiel du reranker. L'expérience 4, quant à elle, s'appuie sur la configuration la plus riche en données parmi nos modèles développés, lui permettant d'atteindre des performances proches de celles du modèle commercial, notamment sur le jeu de données PIAF.

5.4 Analyse des erreurs

Les performances relativement faibles observées dans nos expériences s'expliquent principalement par la méthode de construction des exemples négatifs. Dans les données d'origine issues des questions-réponses, un même contexte est associé à une ou plusieurs questions ; chaque couple (QUESTION, CONTEXTE) constitue un exemple positif puisque le contexte contient la réponse. Or, pour créer des exemples négatifs, nous avons associé aléatoirement d'autres questions

au contexte, ce qui ne reflète pas la réalité de la tâche, car ce choix aléatoire produit des couples (QUESTION, CONTEXTE) qui abordent deux sujets complètement différents. En effet, au lieu de déterminer si la réponse se trouve dans le contexte, le modèle apprend à détecter si la question et le contexte traitent du même sujet.

Prenons l'exemple de la question « Quelle dynastie accélère la croissance de Babylone ? » issue du jeu de données FQuAD. Dans l'expérience 3, les top 10 documents ont tous des scores supérieurs à 0,98, car ils proviennent du même article sur Babylone, ne faisant que varier de passage en passage. Cette homogénéité entraîne une similarité thématique généralisée, rendant difficile la distinction du document contenant réellement la réponse.

Ces observations suggèrent qu'il serait pertinent de repenser la stratégie de génération des exemples négatifs, en incluant des cas présentant des similarités thématiques plus fines et des écarts de score plus discriminants, afin de mieux simuler la véritable complexité de la tâche.

6 Conclusion et perspectives

Dans cette étude, nous avons examiné le rôle central du reranker dans les systèmes de génération augmentée par la recherche (RAG). Nos expériences ont permis de comparer différentes configurations de fine-tuning, mettant en exergue l'impact du choix des données et du modèle de base sur la qualité du réordonnancement. Ce benchmark met en évidence la compétitivité des solutions commerciales par rapport aux approches open source et nos propres expériences. Tandis que le reranker commercial cohere offre des performances de très haut niveau, nos expériences, notamment l'expérience 1 et l'expérience 4, démontrent que des configurations optimisées peuvent se rapprocher de ces performances. Ces résultats ouvrent la voie à des travaux futurs visant à affiner davantage nos modèles, notamment en explorant des stratégies de fine-tuning plus robustes et en intégrant des jeux de données complémentaires.

Sur le plan des perspectives, nous envisageons d'ajuster la construction de la partie négative du jeu de données. Pour une question donnée, nous ne nous contenterons pas de sélectionner aléatoirement des documents sur des thématiques différentes, mais nous intégrerons également les autres segments de l'article ne contenant pas la réponse afin de constituer des exemples négatifs plus représentatifs. Une fois ce nouveau jeu de données finalisé, nous réitérerons l'ensemble des analyses et élargirons l'évaluation à d'autres modèles de type BERT base.

À long terme, nous souhaitons également construire un reranker multilingue open source, capable de fonctionner sur CPU, et dont les performances seraient comparables à celles des solutions commerciales.

Références

[1] James Briggs. Rerankers and two-stage retrieval. https://www.pinecone.io/learn/series/rag/rerankers/. In : *Retrieval Augmented Generation*, Accessed : 2025-03-05.

[2] Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. SemEval-2017 task 1 : Semantic textual similarity multilingual and crosslingual focused evaluation. In Steven Bethard, Marine Carpuat, Marianna Apidianaki, Saif M. Mohammad, Daniel Cer, and David Jurgens, editors, *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada, August 2017. Association for Computational Linguistics. https://aclanthology.org/S17-2001/.

[3] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading Wikipedia to answer open-domain questions. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 1870–1879, Vancouver, Canada, July 2017. Association for Computational Linguistics. https://aclanthology.org/P17-1171/.

[4] Cohere. Improve search performance with a single line of code. https://cohere.com/rerank, 2025. Accessed : 2025-03-05.

[5] Van Tuan DANG. dangvantuan/crossencoder-camembert-large. https://huggingface.co/dangvantuan/CrossEncoder-camembert-large, 2022. Accessed : 2025-03-05.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT : pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. http://arxiv.org/abs/1810.04805.

[7] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. 2024. https://arxiv.org/abs/2401.08281.

[8] Elastic. Elasticsearch : The official distributed search & analytics engine for all types of data. https://www.elastic.co/elasticsearch/, 2021. Accessed : 2025-03-05.

[9] Kai Hui, Tao Chen, Zhen Qin, Honglei Zhuang, Fernando Diaz, Mike Bendersky, and Don Metzler. Retrieval augmentation for t5 re-ranker using external sources. https://arxiv.org/abs/2210.05145, 2022.

[10] Ali Kabbadj. French-squad : French machine reading for question answering. https://github.com/Alikabbadj/French-SQuAD, 2019. Accessed : 2025-03-05.

[11] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. *CoRR*, abs/2004.04906, 2020. https://arxiv.org/abs/2004.04906.

[12] Rachel Keraron, Guillaume Lancrenon, Mathilde Bras, Frédéric Allary, Gilles Moyse, Thomas Scialom, Edmundo-Pavel Soriano-Morales, and Jacopo

Staiano. Project piaf : Building a native french question-answering dataset. https://arxiv.org/abs/2007.00968, 2020.

[13] Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. *CoRR*, abs/2005.11401, 2020. https://arxiv.org/abs/2005.11401.

[14] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta : A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. http://arxiv.org/abs/1907.11692.

[15] d'Hoffschmidt Martin, Vidal Maxime, Belblidia Wacim, and Brendlé Tom. FQuAD : French Question Answering Dataset. *arXiv e-prints*, page arXiv :2002.06071, Feb 2020. https://arxiv.org/abs/2002.06071.

[16] Meta. Llama 3.2 : Revolutionizing edge ai and vision with open, customizable models. https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/, September 2024. Accessed : 2025-02-21.

[17] Rodrigo Nogueira and Kyunghyun Cho. Passage reranking with BERT. *CoRR*, abs/1901.04085, 2019. http://arxiv.org/abs/1901.04085.

[18] OpenAI. Openai platform models. https://platform.openai.com/docs/models. Accessed : 2025-02-21.

[19] pandora s. pandora-s / neural-bridge-rag-dataset-12000-google-translated. https://huggingface.co/datasets/pandora-s/neural-bridge-rag-dataset-12000-google-translated, 2024. Accessed : 2025-03-05.

[20] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140) :1–67, 2020. http://jmlr.org/papers/v21/20-074.html.

[21] Mandeep Rathee, Sean MacAvaney, and Avishek Anand. Guiding retrieval using llm-based listwise rankers. https://arxiv.org/abs/2501.09186, 2025.

[22] Nils Reimers and Iryna Gurevych. Sentence-bert : Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. https://arxiv.org/abs/1908.10084.

[23] S. E. Robertson and K. Sparck Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27(3) :129–146, 1976. https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.4630270302.

[24] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing Management*, 24(5) :513–523, 1988. https://www.sciencedirect.com/science/article/pii/0306457388900210.

[25] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert : smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108, 2019. https://arxiv.org/abs/1910.01108.

[26] SBERT.net. Cross encoder - training overview. https://sbert.net/docs/cross_encoder/training_overview.html, 2025. Accessed : 2025-03-05.

[27] SBERT.net. Evaluation. https://sbert.net/docs/package_reference/cross_encoder/evaluation.html, 2025. Accessed : 2025-03-05.

[28] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR : A heterogenous benchmark for zero-shot evaluation of information retrieval models. *CoRR*, abs/2104.08663, 2021. https://arxiv.org/abs/2104.08663.

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. http://arxiv.org/abs/1706.03762.

[30] Jianguo Wang, Xiaomeng Yi, Rentong Guo, Hai Jin, Peng Xu, Shengjun Li, Xiangyu Wang, Xiangzhou Guo, Chengming Li, Xiaohai Xu, et al. Milvus : A purpose-built vector data management system. In *Proceedings of the 2021 International Conference on Management of Data*, pages 2614–2627, 2021. https://dl.acm.org/doi/pdf/10.1145/3448016.3457550.

[31] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Multilingual e5 text embeddings : A technical report. https://arxiv.org/abs/2402.05672, 2024.

Session 5 : APIA-RJCIA – IA dans les systèmes embarqués

LeYOLO, nouvelle architecture embarquée pour la détection d'objets

Lilian Hollard[1], Lucas Mohimont[1], Nathalie Gaveau[2], Luiz Angelo Steffenel[1]

[1] Université de Reims Champagne-Ardenne,CEA, LRC DIGIT, LICIIS, Reims, France
[2] Université de Reims Champagne-Ardenne, INRAE, RIBP USC 1488, Reims, France

Résumé

La réduction du coût de calcul des réseaux neuronaux profonds est essentielle pour la détection d'objets en temps réel. Pourtant, les récents progrès reposent surtout sur le matériel plutôt que sur l'optimisation des modèles. Cela se remarque notamment dans les dernières architectures YOLO, où la vitesse prime sur la légèreté.

Pour répondre à ce défi, nous introduisons deux contributions majeures. D'abord, LeNeck, un cadre de détection rapide et précis, réduisant le nombre de paramètres. Ensuite, LeYOLO, un modèle optimisé pour YOLO, combinant compacité et haute précision. Ces solutions sont idéales pour les dispositifs à faible consommation, y compris les microcontrôleurs.

Mots-clés

Vision par ordinateur, Optimisation des réseaux de neurones, Architecture de réseaux de neurones, Microcontrôleur.

Abstract

Reducing the computational cost of deep neural networks is essential for real-time object detection. However, recent progress has been based mainly on hardware rather than model optimization. This is particularly noticeable in the latest YOLO architectures, where speed precedes lightness.

To address this challenge, we are introducing two significant contributions. Firstly, LeNeck is a fast and accurate detection framework that reduces the number of parameters. Secondly, LeYOLO is a model optimized for YOLO that combines compactness and high precision. These solutions are ideal for low-power devices, including microcontrollers.

Keywords

Computer vision, Neural Network Optimization, Neural Network Architecture, Microcontrollers.

1 Introduction

Un calcul efficace, un traitement en temps réel et une exécution à faible latence sont essentiels pour les dispositifs edge alimentés par l'IA, y compris les drones autonomes, les systèmes de surveillance, l'agriculture intelligente et les caméras intelligentes. Bien que le cloud computing offre une alternative pour exécuter des modèles puissants, il présente des inconvénients tels que la latence, les contraintes de bande passante et les risques de sécurité [1, 43, 39]. Dans les applications pratiques de détection d'objets, les avancées de l'apprentissage profond se sont principalement concentrées sur l'optimisation de la vitesse pour les GPU à haute performance, souvent au détriment de l'efficacité sur le matériel à faible puissance.

Initialement introduit par Joseph Redmon et al. [25], les modèles YOLO sont connus pour leur vitesse d'inférence en détection d'objets. Ces modèles ont connu des améliorations architecturales significatives au fil des ans, tirant parti des puissances de calcul modernes.

Malgré leur vitesse inhérente, il y a eu un mouvement notable dans le développement des modèles YOLO ces dernières années. Avec les rapides avancées des capacités des GPU et les nouvelles innovations matérielles, l'accent s'est déplacé des modèles légers à ceux privilégiant la vitesse d'inférence [15, 17, 37, 14, 38]. En conséquence, les modèles YOLO sont devenus nettement plus rapides malgré l'augmentation des paramètres et des FLOP [1].

Notre travail met en évidence le fait que, malgré leur vitesse impressionnante sur les GPU, les modèles YOLO ont du mal avec le matériel sans accélération pour l'IA tel que les microcontrôleurs et les micro-ordinateurs embarqués. Par exemple, sur les microcontrôleurs STMicroelectronics - largement utilisés dans la robotique et les applications IoT - les modèles YOLO modernes prennent plus d'une seconde par inférence sur les puces les plus puissantes (Section 4, Tableau 4), les rendant inadaptés aux applications en temps réel. Sur les microcontrôleurs moins puissants, des améliorations supplémentaires sont nécessaires pour réduire le temps d'inférence, un défi que nous abordons dans cette étude. Ces contraintes posent un défi critique pour les industries dépendant de l'IA à faible puissance, où l'efficacité énergétique, la petite taille du modèle et l'utilisation optimisée des ressources sont essentielles.

Dans les tâches de classification, les recherches sur l'optimisation des comptes de paramètres et des coûts computationnels ont produit des modèles notables comme Mobile-

1. Nous décrivons les opérations en virgule flottante comme FLOP, définissant toutes les opérations arithmétiques que le réseau de neurones nécessite pour effectuer une inférence. Dans notre article, 1 FLOP est environ 2 MADD ou 2 MACC. Ainsi, la variation des benchmarks tels que MobileNet diffère de leur article original.

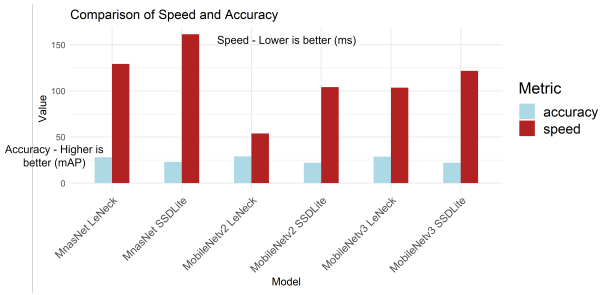


FIGURE 1 – Différence de vitesse (ms) et de précision (mAP) entre SSDLite et LeNeck sur STM32N6570-DK.

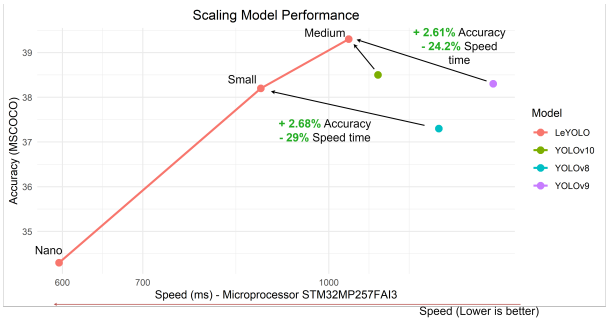


FIGURE 2 – Comparaison entre LeYOLO et les YOLO modernes, démontrant une meilleure précision pour moins de temps d'exécution sur STM32MP257FAI3.

Nets [12, 27, 11] et EfficientNets [30, 31]. Bien que ces modèles soient remarquables, ils sont principalement reconnus pour leurs capacités de classification exceptionnelles plutôt que pour la détection d'objets. Les recherches se sont principalement concentrées sur les classificateurs légers, souvent associés à un ajout de détection d'objets comme SSDLite [20, 27]. Bien que les classificateurs à faible nombre de paramètres combinés avec SSDLite offrent une meilleure vitesse sur les microcontrôleurs, leur précision est inférieure à celle de YOLO.

Nos recherches ont identifié un écart crucial : il y a peu d'effort sur l'optimisation des architectures de détection d'objets qui équilibrent l'efficacité des paramètres et le coût computationnel tout en maintenant une précision au niveau des YOLO modernes. Cet écart oblige les développeurs à choisir entre des modèles YOLO haute performance mais coûteux en calcul et des alternatives à faible puissance comme SSDLite, qui sacrifient la précision pour la vitesse. Notre travail vise à combler cette partie manquante de recherche en introduisant des modèles de détection d'objets plus efficaces adaptés aux applications d'IA edge.

Cet article introduit deux contributions principales.

1. La première est une alternative à SSDLite appelée LeNeck qui comble l'écart entre les classificateurs à faible nombre de paramètres et les petits modèles YOLO. En utilisant LeNeck au lieu de SSDLite, nous maintenons une vitesse d'inférence similaire tout en obtenant une bien meilleure précision (Figure 1)).

2. La deuxième contribution est LeYOLO - une nouvelle famille de modèles YOLO légers et efficaces. LeYOLO correspond à la précision des échelles YOLO plus petites tout en améliorant considérablement la vitesse d'inférence sur les microcontrôleurs (Figure 2).

Nos résultats montrent que cette approche rivalise avec les modèles YOLO à des échelles comparables. Nous démontrons qu'il est possible d'optimiser l'architecture des réseaux de neurones pour la détection d'objets grâce à une nouvelle méthode d'échelle entre les classificateurs légers et les modèles YOLO.

2 Etat de l'art

Notre travail se concentre sur le développement d'une architecture optimale pour la détection d'objets en combinant deux approches clés : les détecteurs d'objets optimisés pour la vitesse et les classificateurs à faible coût conçus pour minimiser le nombre de paramètres en utilisant des techniques bien établies. LeYOLO et LeNeck intègrent des éléments connus pour leur efficacité dans la réduction de paramètres. Plus précisément, nous utilisons des Inverted Bottleneck, initialement introduits dans MobileNetV2 [27] et ensuite affinés par EfficientNet [30, 31] et GhostNet [9, 33]. Les convolutions pointwise [18] et depthwise sont des composants cruciaux dans l'optimisation de l'architecture, contribuant de manière significative à des modèles comme MNASNet [29].

L'essor des classificateurs à faible coût a conduit à SSDLite, une variante optimisée de SSD intégrant des convolutions groupées basées sur MobileNets. Initialement conçu pour réduire les coûts de détection en utilisant VGG [28], SSDLite partage des similitudes avec les premiers modèles YOLO [26]. Depuis lors, aucune méthode n'a significativement surpassé SSDLite, bien que SSDLiteX [16] ait tenté d'améliorer ses performances.

Du côté de YOLO, les recherches ont exploré la réduction des paramètres dans les architectures principales. Les efforts de tinier-yolo, efficient yolo, mobile densenet et autres [4, 41, 7, 40] ont intégré des éléments de classificateurs légers comme les convolutions depthwise et des techniques plus anciennes telles que les modules fire [13] pour minimiser l'utilisation des paramètres.

EfficentDet [40] partage la philosophie centrale de notre modèle : utiliser des couches à faible coût computationnel (concaténation et additions, convolutions depthwise et pointwise). Cependant, EfficientDet nécessite trop d'informations sémantiques et trop d'états de blocage (attente des couches précédentes, graphes complexes), ce qui le rend difficile à optimiser pour une vitesse d'exécution rapide.

Les auteurs de YOLOF [3] ont opté pour un modèle avec une seule entrée et une seule sortie dans le Neck[2]. Bien

2. Partie du modèle qui partage plusieurs niveaux d'informations sémantiques

que cette conception soit théoriquement plus rapide et plus efficace en calcul, l'article YOLOF révèle une baisse significative de la précision lors de la comparaison d'un Neck à sortie unique (Single-in, Single-out - SiSO) avec un Neck à sorties multiples (Single-in, Multiple-out - SiMO).

Plus récemment, YOLOX [5] et YOLOv9 [38] ont introduit des alternatives légères avec des paramètres réduits. YOLOX remplace les convolutions standard par des convolutions depthwise de tailles de noyau plus grandes et traite des entrées d'image plus petites. YOLOv9 contribue de manière substantielle à l'optimisation des paramètres mais se concentre sur l'échelle YOLO standard plutôt que sur les architectures adaptées aux mobiles.

Enfin, Tinyssimo YOLO [24], basé sur les premiers modèles YOLO [25], se concentre sur la réduction des coûts computationnels pour permettre la détection d'objets sur les microcontrôleurs fonctionnant dans la gamme de puissance des milliwatts. Cependant, il ne parvient pas à atteindre la précision et l'efficacité même des plus petites variantes de YOLO ou des classificateurs basés sur SSDLite.

3 Optimisation des détecteurs d'objets en temps réel pour les microcontrôleurs

Les détecteurs d'objets modernes reposent sur des blocs d'architecture qui exploitent pleinement le matériel moderne. Les convolutions standard et les structures parallèles ou multi-branches sont couramment utilisées. Cependant, ces conceptions sont trop gourmandes en ressources pour les microcontrôleurs. Conçu pour être hautement efficace, le bloc de construction principal de LeYOLO optimise à la fois les paramètres et le mAP (mesure de précision pour la détection d'objet). Il s'appuie sur une structure bien connue appelée Inverted Bottleneck, couramment utilisée dans les réseaux de neurones efficaces comme MobileNets [12, 27, 11] et EfficientNets [40, 31].

Au lieu d'utiliser de grands filtres coûteux pour traiter les images, LeYOLO décompose le processus en étapes plus petites et plus efficaces en utilisant trois couches de convolution principales. Notre bloc applique une convolution 1×1 qui projette les cartes de caractéristiques des canaux C de $x \in R^{B,C,H,W}$ en un tenseur de dimension d (où $d \geq C$). Ensuite, une convolution depthwise $k \times k$ traite efficacement les caractéristiques spatiales. Enfin, une autre convolution pointwise 1×1 ramène les canaux à leur taille d'origine. Bien que de nombreux articles utilisant des Inverted Bottleneck modifient la convolution pointwise finale pour produire un nombre de canaux différent de l'entrée, LeNeck et LeYOLO ne suivent pas cette approche. Au lieu de cela, nous nous appuyons uniquement sur des convolutions standard séparées lors de la transition entre les tailles de cartes de caractéristiques après le sous-échantillonnage. Ces convolutions ajustent intrinsèquement à la fois le nombre de canaux et la taille de la carte de caractéristiques, éliminant ainsi le besoin de transformations supplémentaires au sein du Inverted Bottleneck.

Astuce d'optimisation. Normalement, la première convolution 1×1 étend les canaux avant le traitement. Cependant, si le nombre de canaux n'a pas besoin de changer (si $C == d$), nous supprimons la première convolution pointwise. Ce petit changement réduit considérablement le nombre de calculs, en particulier dans les premières couches où les images sont grandes.

Impact sur la vitesse et la précision. L'élimination des calculs inutiles rend le réseau plus rapide et plus efficace tout en maintenant une haute précision (Section 3.3). Cette optimisation est particulièrement bénéfique pour l'exécution de modèles de détection d'objets sur des dispositifs à faible puissance et à ressources limitées. Pour comparaison, SSDLite ne commence à partager des informations sémantiques qu'au niveau P4 [3] tandis que les détecteurs d'objets classiques et modernes commencent au niveau P3, qui fournit des détails spatiaux plus riches mais à un coût computationnel plus élevé. En réduisant stratégiquement les calculs redondants dans les premières couches, LeNeck atteint la même vitesse que SSDLite tout en exploitant le niveau P3 plus informatif, résultant en une meilleure performance de détection sans surcoût computationnel. Le modèle utilise la fonction d'activation SiLU σ, comme dans les versions modernes de YOLO (YOLOv7, YOLOv9) pour une meilleure performance.

Nous définissons les dimensions d'entrée et de sortie comme C et la dimension étendue comme d. Pour les filtres $W_1 \in R^{1,1,C,d}$, $W_2 \in R^{k,k,1,d}$, et $W_3 \in R^{1,1,d,C}$, notre approche peut être représentée comme suit :

$$y = \begin{cases} W_3 \otimes \sigma[W_2 \otimes \sigma(W_1 \otimes x)] & \text{si } d \neq C \\ W_3 \otimes \sigma[W_2 \otimes \sigma(W_1 \otimes x)] & \text{si } d = C \text{ et } W_1 = \text{Vrai} \\ W_3 \otimes \sigma[W_2 \otimes (x)] & \text{si } d = C \text{ et } W_1 = \text{Faux} \end{cases} \quad (1)$$

3.1 LeNeck - Détecteur d'objets polyvalent

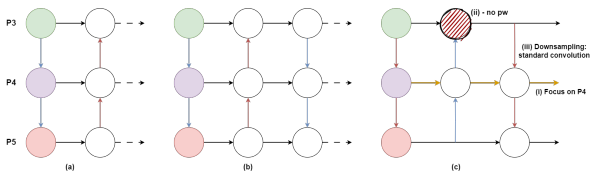


FIGURE 3 – Différence entre le Neck LeYOLO proposé et un agrégateur de caractéristiques sémantiques efficace. (a) Correspond à FPN [6]. (b) Représente PANnet [42]. Enfin, (c) est notre solution proposée.

Dans la détection d'objets, nous appelons le Neck la partie du modèle qui agrège plusieurs niveaux d'informations sémantiques, partageant les informations de couches distantes aux premières couches. Historiquement, les chercheurs ont utilisé un PANet [42] ou FPN [19] pour partager efficacement les cartes de caractéristiques, permettant plusieurs

3. P4 est le niveau sémantique de l'information correspondant à la taille de l'entrée divisée par 2^4.

niveaux de détection en reliant plusieurs informations sémantiques P_i au PANet et leurs sorties respectives comme illustré dans la Figure 3(a). Pour créer LeNeck, nous avons identifié un aspect très important dans la composition des réseaux de neurones profonds. Nous avons remarqué qu'il y a constamment une répétition significative des couches au niveau sémantique équivalent à P4. Nous avons trouvé cela dans tous les MobileNets [12, 27, 11], dans l'optimisation des Inverted Bottleneck dans EfficientNets [30, 31] et EfficientDet [40], ainsi que dans les architectures plus récentes avec des mécanismes de self-attention comme MobileViTs [22, 23, 35], EdgeNext [21], et FastViT [34], qui sont conçus pour la vitesse. Plus intéressant encore, les modèles conçus par Neural Architecture Search (NAS) [29, 11, 30] utilisent également ce schéma. Par conséquent, nous introduisons LeNeck, un agrégateur de caractéristiques sémantiques efficace qui utilise le niveau sémantique P4 comme principal conducteur pour fusionner les informations de P3 et P5 (Figure 3.(i)). Le calcul à P3 et P5 n'est effectué qu'une seule fois, garantissant l'efficacité (P3 utilise trop de taille spatiale, et P5 utilise un nombre très étendu de canaux).

Nous réduisons le calcul - en particulier au niveau P3 en raison de la grande taille spatiale - en supprimant la première convolution pointwise (Figure 3.(ii)). Après une étude comparative (Section 3.3) réalisée sur le backbone de LeYOLO à l'échelle nano, nous avons saisi l'opportunité de supprimer les convolutions pointwise coûteuses en temps puisque les canaux d'entrée de P3 concaténés avec les caractéristiques suréchantillonnées de P4 résultent en la dimension d requise par la convolution depthwise intermédiaire de notre Inverted Bottleneck optimisé présenté dans la section 3. Chaque nombre de canaux d'entrée, ainsi que le nombre de canaux étendus du Inverted Bottleneck, ne dépasse jamais 6. L'entrée de P3 est $32C$ tandis que la dernière couche cachée du Neck de LeYOLO étend les canaux d égale 192.

Comme les convolutions standard ne sont pas très efficace en nombre de paramètres et de calcul, nous nous limitons à l'utiliser deux fois. De P3 à P4, et de P4 à P5 pour effectuer un sous-echantillonage (Figure 3.iii)).

3.2 Backbone de LeYOLO

Notre mise en œuvre implique la minimisation de l'échange d'informations inter-couches sous la forme de $I(X;h_1) \geq I(X;h_2) \geq \ldots \geq I(X;h_n)$, avec n égal à la dernière couche cachée du backbone du réseau de neurones, en garantissant que le nombre de canaux d'entrée/sortie ne dépasse jamais une différence de ratio de 6 de la première couche cachée à la dernière. De plus, plutôt que d'augmenter la complexité computationnelle de notre modèle comme [37, 38, 10, 2], nous avons opté pour une mise à l'échelle plus efficace, intégrant la théorie du goulot d'étranglement inversé de Dangyoon Han et al. [8] qui stipulait que les convolutions pointwise ne devraient pas dépasser un ratio de 6 dans le Inverted Bottleneck.

TABLE 1 – Architecture du backbone de LeYOLO

Input	Operator	exp size	out size	NL	s
P0	conv2d, 3x3	-	16	SI	2
P1	conv2d, 1x1	16	16	SI	1
P1	bneck, 3x3, **pw=False**	16	16	SI	2
P2	bneck, 3x3	**96**	32	SI	2
P3	bneck, 3x3	96	32	SI	1
P3	bneck, 5x5	96	64	SI	2
P4	bneck, 5x5	192	64	SI	1
P4	bneck, 5x5	192	64	SI	1
P4	bneck, 5x5	192	64	SI	1
P4	bneck, 5x5	192	64	SI	1
P4	bneck, 5x5	**576**	96	SI	2
P5	bneck, 5x5	576	96	SI	1
P5	bneck, 5x5	576	96	SI	1
P5	bneck, 5x5	576	96	SI	1
P5	bneck, 5x5	576	96	SI	1

TABLE 2 – Amélioration de LeNeck et LeYOLO (best of).

Améliorations	mAP	GFLOP
base (LeYOLO nano)	34.3	2.64
+3x3	32.9	2.877
+5x5	34.9	3.946
+5x5 après P4	34.2	3.19
+Sous-echantillonage 3x3	34.6	3.011
+aucun pw backbone et neck	34.1	2.823
+Ratio d'expansion de 2 au lieu de 3 dans LeNeck	34.3	2.64

3.3 Etude comparative

Une étude comparative en apprentissage automatique est une méthode de recherche qui teste l'impact de couches, de caractéristiques ou de techniques spécifiques en les désactivant ou en les remplaçant. L'objectif est d'identifier quels paramètres sont cruciaux pour la performance du modèle, guidant le développement d'un détecteur d'objets entièrement optimisé. Nous utilisons LeYOLO dans son intégralité (backbone + LeNeck) pour l'étude comparative afin d'affiner les deux contributions (Tableau 2).

Nous avons d'abord exploré diverses configurations de taille de noyau. Bien que des noyaux plus grands améliorent généralement les performances, ils nécessitent également plus de ressources de calcul. Le choix optimal était une convolution 5×5 après le sous-échantillonnage P4.

En suivant les idées de ConvNeXt, nous avons utilisé des convolutions séparées pour le sous-échantillonnage. Cependant, l'utilisation d'un noyau 3×3 au lieu de 5×5 dans cette configuration a conduit à de meilleurs résultats.

Enfin, nous avons fait deux optimisations critiques : la réduction du ratio d'expansion dans le Inverted Bottleneck de 3 à 2 et l'élimination de la première convolution pointwise coûteuse dans les premières couches du backbone et au niveau P3 dans le Neck. Ces modifications ont considérablement réduit le coût computationnel du modèle tout en entraînant une perte de précision de -0,3 mAP, que nous avons jugée négligeable compte tenu des gains d'efficacité.

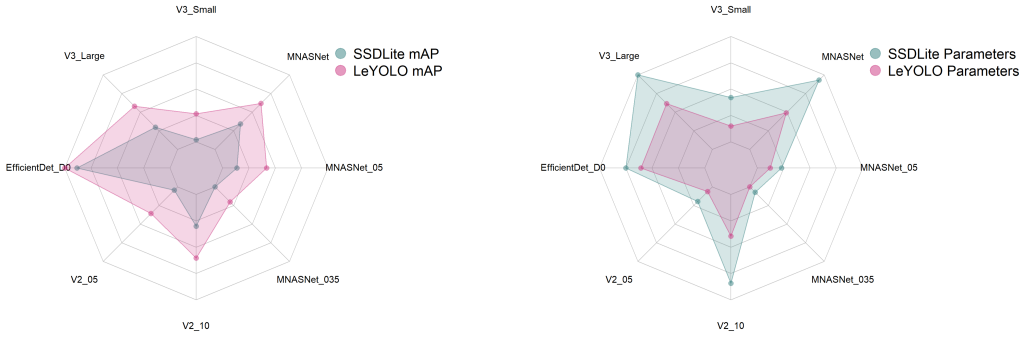


FIGURE 4 – LeYOLO comparé à SSDLite, avec un meilleur ratio paramètre-précision.

4 Résultats experimentaux

Nous entraînons chaque réseau de neurones avec les mêmes hyperparamètres et techniques d'augmentation de données, tels que SGD, avec un taux d'apprentissage de 0,01 et un momentum de 0,9. Nous nous appuyons principalement sur l'augmentation de données mosaïque ainsi que sur hsv de $\{0,015, 0,7, 0,4\}$ et une translation d'image de 0,1. En ce qui concerne les spécificités de l'entraînement, nous avons utilisé une taille de batch de 96 sur 4 GPU P100. La performance est évaluée sur l'ensemble de validation de MSCOCO en utilisant la précision moyenne (mAP). Pour LeYOLO, nous offrons une variété de modèles inspirés de la base architecturale présentée ci-dessus. Une approche classique implique la mise à l'échelle du nombre de canaux, de couches et de la taille d'entrée de l'image. Traditionnellement, la mise à l'échelle met l'accent sur les configurations de canaux et de couches, intégrant parfois divers schémas de mise à l'échelle (Tableau 3).

TABLE 3 – Différentes échelles pour l'entrainement de LeYOLO ainsi que leurs résultats.

Models	Nano	Small	Medium	Large
Input spatial size	640	640	640	768
Channels ratio	x1	x1.33	x1.33	x1.33
Layer ratio	x1	x1	x1.33	x1.33
mAP	34.3	38.2	39.3	41

LeYOLO évolue de la version Nano à la version Large avec une mise à l'échelle liée à ce qu'EfficientDet a apporté : répétition des canaux de 1.0 à 1.33, couches de 1.0 à 1.33, et taille spatiale pour les besoins d'entraînement de 640×640 à 768×768. Plusieurs tailles spatiales sont utilisées à des fins d'évaluation, allant de 320×320 à 768×768 Nous évaluons LeYOLO à des tailles spatiales réduites, tous les résultats étant présentés dans le Tableau 8.

Nous évaluons la vitesse du modèle sur deux microprocesseurs : STM32MP257FAI3 et STM32N6570-DK. Les deux utilisent des cœurs Arm Cortex, équilibrant faible consommation d'énergie et capacité de calcul efficace. Ces microcontrôleurs peuvent réaliser une inférence en temps réel à une résolution de 320×320 à 640×640 le calcul devient plus exigeant, mais LeYOLO traite toujours chaque inférence en moins d'une seconde, surpassant les modèles YOLO modernes (Tableau 4).

TABLE 4 – Vitesse d'inférence de LeYOLO (640×640) et sa précision sur appareil embarqué (Onnx - STM32MP257FAI3).

Models	mAP	Speed(ms)
LeYOLO Nano	**34.3**	**596**
LeYOLO Small	**38.2**	**877.9**
LeYOLO Medium	**39.3**	**1039**
YOLOv10 Nano [36]	38.5	1099
YOLOv8 Nano [14]	37.3	1235
YOLOv9 Tiny [38]	38.3	1371

4.0.1 Détection d'objets mobile

LeYOLO surpasse les détecteurs d'objets de type YOLO sur les dispositifs embarqués ou ceux avec une puissance de calcul limitée. Nous fournissons un tableau détaillé (Tableau 8) montrant le nombre de FLOPs, et nous observons une corrélation entre cette métrique et la vitesse d'exécution sur les dispositifs à faible ressource de calcul (Tableau 4).

TABLE 5 – Vitesse d'inférence de LeNeck (320x320) et sa précision sur appareil embarqué (Onnx - STM32MP257FAI3).

Models	SSDLite	**LeNeck**	SSDLite	**LeNeck**
	Speed(ms)		mAP.95	
V3-Small	**146.2**	165.9	16.0	**21.3**
V3-Large	**286.5**	292.3	22	**28.1**
V2-1.0	259.2	**256.3**	22.1	**28.6**
MNASNet 0.5	167.7	**155.3**	18.5	**24.6**
MNASNet	306.2	**262.3**	23	**28.9**
LeYOLO Nano		165.4		25.2

Nous intégrons l'état de l'art des backbones à faible nombre de paramètres avec LeNeck. Quel que soit le backbone utilisé, tous les nombres de canaux, P3,P4 et P5 spécificités de répétition restent les mêmes. À P3, la première convolution pointwise n'est jamais utilisée, comme dans le LeYOLO de base, ce qui entraîne le premier filtre étant la convolution

TABLE 6 – Vitesse d'inférence de LeNeck (320x320) et sa précision sur appareil embarqué (Onnx - STM32N6570-DK).

Models	SSDLite	**LeNeck**	SSDLite	**LeNeck**
	Speed(ms)		mAP.95	
V3-Small	**46.91**	50.19	16.0	**21.3**
V3-Large	**121.6**	129.2	22	**28.1**
V2-1.0	104	**103.5**	22.1	**28.6**
MNASNet 0.5	86.31	**36.03**	18.5	**24.6**
MNASNet	161.3	**53.78**	23	**28.9**

TABLE 7 – Performance de LeNeck comparée à d'autres modèles de l'état de l'art combinés avec SSDLite sur MSCOCO.

Models	SSDLite	**LeNeck**	SSDLite	**LeNeck**
	Parameters(M)		mAP.95	
V3-Small	2.49	**1.34**	16.0	**21.3**
V3-Large	4.97	**3.33**	22	**28.1**
EfficientDetD0	3.9	**3.29**	34.6	**37.1**
V2-0.5	1.54	**0.98**	16.6	**23.3**
V2-1.0	4.3	**2.39**	22.1	**28.6**
MNASNet 0.35	1.02	**0.7**	15.6	**20.0**
MNASNet 0.5	1.68	**1.22**	18.5	**24.6**
MNASNet	4.68	**2.8**	23	**28.9**

depthwise de la taille exacte du nombre de canaux d'entrée équivalent du backbone.

LeNeck, en tant que détecteur d'objets général pour les classificateurs légers, conserve le même nombre de canaux [4] et répétition des couches [5] que la version Nano de LeYOLO. À partir d'une variété de classificateurs légers avec un faible nombre de paramètres et de FLOP, LeNeck a surpassé SSDLite dans tous les aspects de ce que nous attendons d'un modèle à faible coût - meilleure échelle de paramètres, meilleure précision, et enfin, bonne vitesse d'inférence - avec les résultats de vitesse d'inférence décrits dans les tableaux 5 - 6 et l'efficacité paramètres-précision décrite dans le Tableau 7 et la Figure 4.

4.0.2 Microcontrôleurs bas de gamme

Au-delà du traitement en temps réel, nous avons également benchmarké LeYOLO par rapport aux modèles YOLO modernes sur divers microcontrôleurs bas de gamme, comme le montre le Tableau 9, en utilisant la variante LeYOLO-Small (YOLOv10 n'est pas compatible avec ces types de microcontrôleurs). Selon le Tableau 8, LeYOLO-Small correspond à YOLOv8 à YOLOv10 en précision. De plus, il s'avère plus efficace en inférence sur ces microcontrôleurs, étendant la capacité de YOLO à fonctionner efficacement sur des dispositifs bas de gamme.

4.1 Analyse plus approfondie

Pour analyser plus en détail les résultats de vitesse, nous mettons en avant l'objectif de notre Inverted Bottleneck avec une couche pointwise optionnelle (voir la Section 3 pour plus d'informations). Dans LeYOLO Small, la convolution pointwise à haute résolution spatiale dans le backbone et au niveau P3 dans LeNeck entraîne une amélioration minimale de la précision, comme le montre la Section 3.3. Par rapport à l'Inverted Bottleneck classique, notre solution économise 8.5% de la vitesse d'inférence sur tous les STM32 benchmarkés dans l'article (Tableau 10-pw). Contrairement aux architectures YOLO standard, qui reposent sur une répétition de couches profondes avec moins de canaux, LeYOLO obtient une meilleure efficacité en utilisant un ratio d'extension de 2 au lieu de 3 tout en maintenant une profondeur de répétition minimale de 3. Ce choix de conception améliore la vitesse d'inférence de 17% tout en préservant la précision, comme le confirme notre étude expérimentale (Tableau 9-exp x3).

Notre modèle est presque aussi rapide que SSDLite tout en obtenant une précision significativement meilleure. La légère différence de vitesse provient de la taille de la carte de caractéristiques - SSDLite commence à P4, tandis que nous commençons à P3. Cependant, LeNeck reste suffisamment léger pour rivaliser avec SSDLite en vitesse. Étant donné qu'il conserve des informations spatiales plus riches, LeNeck fonctionne également mieux pour détecter diverses tailles d'objets (Tableau 8).

5 Discussions

LeNeck : Compte tenu de l'efficacité coût-efficacité de LeNeck, il existe une opportunité significative pour l'expérimentation sur différents backbones de modèles de classification de pointe. LeYOLO émerge comme une alternative prometteuse à SSD et SSDLite. Les résultats prometteurs obtenus sur MSCoco avec notre solution suggèrent une applicabilité potentielle à d'autres modèles de classification.

Efficacité computationnelle : Nous avons mis en œuvre une nouvelle mise à l'échelle pour les modèles YOLO, prouvant qu'il est possible d'atteindre des niveaux de précision très élevés tout en utilisant très peu de ressources computationnelles (FLOP). LeYOLO fournit des résultats très rapides sur les dispositifs embarqués.

6 Conclusion

Tout au long de cet article, nous avons introduit plusieurs optimisations clés :

1. Amélioration des performances en aval du classificateur : Pour un budget de paramètres donné, LeNeck surpasse SSDLite en réduisant le nombre de paramètres tout en améliorant la précision sur MSCOCO. L'intégration de LeNeck avec les backbones existants à faible nombre de paramètres améliore la précision et l'efficacité à plusieurs échelles.
2. Une alternative viable aux modèles YOLO de petite taille : Le backbone optimisé de LeYOLO et LeNeck surpassent les variantes équivalentes des YOLO en détection d'objets. Les choix architecturaux derrière le backbone de LeYOLO entraînent une meilleure mise à l'échelle et un meilleur rapport précision-paramètres et FLOP.

4. de 32 couches à 96 avec une ratio d'expansion de 2
5. répétition $l = 3$

TABLE 8 – Etat de l'art des détecteur d'objets compatible avec les microcontrôleurs STM32.

Models	Input Size	mAP	mAP50	mAP75	S	M	L	FLOP(G)	Parameters (M)
MobileNetv3-S[12]	320	16.1	-	-	-	-	-	**0.32**	1.77
MobileNetv2-x0.5[27]	320	16.6	-	-	-	-	-	0.54	1.54
MnasNet-x0.5[29]	320	18.5	-	-	-	-	-	0.58	1.68
LeYOLO-Nano	320	**25.2**	**37.7**	**26.4**	**5.5**	**23.7**	**48.0**	0.66	**1.1**
MobileNetv3[11]	320	22	-	-	-	-	-	1.02	3.22
LeYOLO-Small	320	**29**	42.9	30.6	6.5	29.1	53.4	1.126	1.9
LeYOLO-Nano	480	**31.3**	**46**	**33.2**	**10.5**	**33.1**	**52.7**	1.47	1.1
MobileNetv2[27]	320	22.1	-	-	-	-	-	1.6	4.3
MnasNet[29]	320	23	-	-	-	-	-	1.68	4.8
LeYOLO-Small	480	**35.2**	50.5	37.5	13.3	38.1	55.7	**2.53**	**1.9**
MobileNetv1[12]	320	22.2	-	-	-	-	-	2.6	5.1
LeYOLO-Medium	480	**36.4**	**52.0**	**38.9**	**14.3**	**40.1**	**58.1**	3.27	2.4
LeYOLO-Small	640	**38.2**	**54.1**	**41.3**	**17.6**	**42.2**	**55.1**	**4.5**	**1.9**
YOLOv5-n[15]	640	28	45.7	-	-	-	-	4.5	1.9
EfficientDet-D0[32]	512	33.80	52.2	35.8	12	38.3	51.2	5	3.9
LeYOLO-Medium	640	**39.3**	**55.7**	**42.5**	**18.8**	**44.1**	**56.1**	**5.8**	**2.4**
YOLOv9-Tiny[38]	640	38.3	53.1	41.3	-	-	-	7.7	2
LeYOLO-Large	768	**41**	**57.9**	**44.3**	**21.9**	**46.1**	**56.8**	**8.4**	**2.4**

TABLE 9 – Vitesse d'inférence et précision de LeYOLO (640x640) sur des dispositifs embarqués.

Device	LeYOLO Small	YOLOv8	YOLOv9
	Speed (s)		
STM32H74l-DISCO	**12.3**	13.7	13.6
STM32F769l-DISCO	**19**	21.5	22.1
STM32F746G-DISCO	**20**	25	24.5
STM32F469I-DISCO	**54.6**	73.6	72.5

TABLE 10 – Amélioration de la vitesse d'inférence de LeYOLO Small (640x640)

Device	LeYOLO Small	pw	exp x3
	Speed (s)		
STM32H74l-DISCO	**12.37**	13.52	14.87
STM32F769l-DISCO	**19.04**	20.64	22.68
STM32F746G-DISCO	**20**	22.36	24.73
STM32F469I-DISCO	**54.6**	59.23	65.28

3. Vitesse d'inférence améliorée : LeYOLO et LeNeck obtiennent une meilleure vitesse d'inférence que les détecteurs d'objets à faible nombre de paramètres de pointe, grâce à leur architecture optimisée.

Nos contributions sont particulièrement efficaces sur les dispositifs mobiles, embarqués et à faible puissance, se rapprochant d'un équilibre idéal entre l'efficacité des paramètres et la performance de détection. La réduction de la taille du modèle tout en maintenant la précision permet la détection d'objets directement sur de petits dispositifs avec une surcharge computationnelle minimale. Ce raffinement étape par étape rapproche les modèles YOLO des applications pratiques de l'IA pour le edge.

Nous encourageons une expérimentation plus approfondie avec notre proposition, en explorant diverses variantes de jeux de données adaptées aux besoins spécifiques de l'industrie. Nous visons à fournir une gamme plus large de comparaisons pour LeYOLO dans des scénarios impliquant des dispositifs mobiles avec des ressources computationnelles très limitées.

Remerciements

Ce travail a été soutenu par Chips Joint Undertaking (Chips JU) dans le projet EdgeAI "Technologies Edge AI pour une performance embarquée optimisée" du projet, accord de subvention n° 101097300.

Références

[1] Rajkumar Buyya, Chee Shin Yeo, Srikumar Venugopal, James Broberg, and Ivona Brandic. Cloud computing and emerging IT platforms : Vision, hype, and reality for delivering computing as the 5th utility. *Future Generation Computer Systems*, 25(6) :599–616, June 2009.

[2] Yuxuan Cai, Yizhuang Zhou, Qi Han, Jianjian Sun, Xiangwen Kong, Jun Li, and Xiangyu Zhang. Reversible column networks. *arXiv preprint arXiv :2212.11696*, 2023.

[3] Qiang Chen, Yingming Wang, Tong Yang, Xiangyu Zhang, Jian Cheng, and Jian Sun. You only look one-level feature. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13039–13048, 2021.

[4] Wei Fang, Lin Wang, and Peiming Ren. Tinier-YOLO : A Real-Time Object Detection Method for Constrained Environments. *IEEE Access*, 8 :1935–1944, 2020.

[5] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox : Exceeding yolo series in 2021. *arXiv preprint arXiv :2107.08430*, August 2021.

[6] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le. Nas-fpn : Learning scalable feature pyramid architecture for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7036–7045, 2019.

[7] Mohammad Hajizadeh, Mohammad Sabokrou, and Adel Rahmani. MobileDenseNet : A new approach to object detection on mobile devices. *Expert Systems with Applications*, 215 :119348, April 2023.

[8] Dongyoon Han, Sangdoo Yun, Byeongho Heo, and YoungJoon Yoo. Rethinking channel dimensions for efficient model design. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 732–741, 2021.

[9] Kai Han, Yunhe Wang, Qi Tian, Jianyuan Guo, Chunjing Xu, and Chang Xu. Ghostnet : More features from cheap operations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1580–1589, 2020.

[10] Geoffrey Hinton. How to Represent Part-Whole Hierarchies in a Neural Network. *Neural Computation*, 35(3) :413–452, February 2023.

[11] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019.

[12] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets : Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv :1704.04861*, 2017.

[13] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet : Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size. *arXiv preprint arXiv :1602.07360*, 2016.

[14] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics YOLO, January 2023.

[15] Glenn Jocher, Ayush Chaurasia, Alex Stoken, Jirka Borovec, NanoCode012, Yonghye Kwon, Kalen Michael, TaoXie, Jiacong Fang, imyhxy, Lorna, (Zeng Yifu), Colin Wong, Abhiram V, Diego Montes, Zhiqiang Wang, Cristi Fati, Jebastin Nadar, Laughing, UnglvKitDe, Victor Sonck, tkianai, yxNONG, Piotr Skalski, Adam Hogan, Dhruv Nair, Max Strobel, and Mrinal Jain. ultralytics/yolov5 : v7.0 - YOLOv5 SOTA Realtime Instance Segmentation, November 2022.

[16] Hyeong-Ju Kang. Ssdlitex : Enhancing ssdlite for small object detection. *Applied Sciences*, 13(21), 2023.

[17] Chuyi Li, Lulu Li, Hongliang Jiang, Kaiheng Weng, Yifei Geng, Liang Li, Zaidan Ke, Qingyuan Li, Meng Cheng, Weiqiang Nie, et al. Yolov6 : A single-stage object detection framework for industrial applications. *arXiv preprint arXiv :2209.02976*, September 2022.

[18] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv preprint arXiv :1312.4400*, 2013.

[19] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[20] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. Ssd : Single shot multibox detector. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 21–37, Cham, 2016.

[21] Muhammad Maaz, Abdelrahman Shaker, Hisham Cholakkal, Salman Khan, Syed Waqas Zamir, Rao Muhammad Anwer, and Fahad Shahbaz Khan. EdgeNeXt : Efficiently Amalgamated CNN-Transformer Architecture for Mobile Vision Applications. In Leonid Karlinsky, Tomer Michaeli, and Ko Nishino, editors, *Computer Vision – ECCV 2022 Workshops*, Lecture Notes in Computer Science, pages 3–20, Cham, 2023.

[22] Sachin Mehta and Mohammad Rastegari. Mobilevit : Light-weight, general-purpose, and mobile-friendly vision transformer. In *International Conference on Learning Representations*, 2022.

[23] Sachin Mehta and Mohammad Rastegari. Separable Self-attention for Mobile Vision Transformers, June 2022.

[24] Julian Moosmann, Marco Giordano, Christian Vogt, and Michele Magno. Tinyissimoyolo : A quantized, low-memory footprint, tinyml object detection network for low power microcontrollers. In *2023 IEEE 5th International Conference on Artificial Intelligence Circuits and Systems (AICAS)*, pages 1–5, 2023.

[25] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once : Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, June 2016.

[26] Joseph Redmon and Ali Farhadi. Yolov3 : An incremental improvement. *arXiv preprint arXiv :1804.02767*, 2018.

[27] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2 : Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[28] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv :1409.1556*, 2014.

[29] Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V. Le. MnasNet : Platform-Aware Neural Architecture Search for Mobile. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2815–2823, Long Beach, CA, USA, June 2019.

[30] Mingxing Tan and Quoc Le. EfficientNet : Rethinking Model Scaling for Convolutional Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning*, pages 6105–6114. PMLR, May 2019.

[31] Mingxing Tan and Quoc Le. EfficientNetV2 : Smaller Models and Faster Training. In *Proceedings of the 38th International Conference on Machine Learning*, pages 10096–10106. PMLR, July 2021.

[32] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet : Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10781–10790, 2020.

[33] Yehui Tang, Kai Han, Jianyuan Guo, Chang Xu, Chao Xu, and Yunhe Wang. GhostNetV2 : Enhance Cheap Operation with Long-Range Attention. *Advances in Neural Information Processing Systems*, 35 :9969–9982, December 2022.

[34] Pavan Kumar Anasosalu Vasu, James Gabriel, Jeff Zhu, Oncel Tuzel, and Anurag Ranjan. Fastvit : A fast hybrid vision transformer using structural reparameterization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5785–5795, 2023.

[35] Shakti N Wadekar and Abhishek Chaurasia. Mobilevitv3 : Mobile-friendly vision transformer with simple and effective fusion of local, global and input features. *arXiv preprint arXiv :2209.15159*, October 2022.

[36] Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, and Guiguang Ding. Yolov10 : Real-time end-to-end object detection. *arXiv preprint arXiv :2405.14458*, 2024.

[37] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7 : Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7464–7475, June 2023.

[38] Chien-Yao Wang, I-Hau Yeh, and Hong-Yuan Mark Liao. Yolov9 : Learning what you want to learn using programmable gradient information. *arXiv preprint arXiv :2402.13616*, 2024.

[39] Fangxin Wang, Miao Zhang, Xiangxiang Wang, Xiaoqiang Ma, and Jiangchuan Liu. Deep Learning for Edge Computing Applications : A State-of-the-Art Survey. *IEEE Access*, 8 :58322–58336, 2020.

[40] Zixuan Wang, Jiacheng Zhang, Zhicheng Zhao, and Fei Su. Efficient Yolo : A Lightweight Model For Embedded Deep Learning Object Detection. In *2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6, July 2020.

[41] Wang Yang, Ding BO, and Li Su Tong. TS-YOLO :An efficient YOLO Network for Multi-scale Object Detection. In *2022 IEEE 6th Information Technology and Mechatronics Engineering Conference (ITOEC)*, volume 6, pages 656–660, March 2022.

[42] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

[43] Zhi Zhou, Xu Chen, En Li, Liekang Zeng, Ke Luo, and Junshan Zhang. Edge Intelligence : Paving the Last Mile of Artificial Intelligence With Edge Computing. *Proceedings of the IEEE*, 107(8) :1738–1762, August 2019.

Défauts ferroviaires : vers une détection visuelle embarquée

Sasa RADOSAVLJEVIC [1,2], Kevin HOARAU [1], Sergio RODRÍGUEZ FLÓREZ [1], Abdelhafid EL OUARDI [1], Alain RIVERO [2]

[1] Université Paris-Saclay, ENS Paris-Saclay, CNRS, SATIE, 91190, Gif-sur-Yvette, France
[2] SNCF Réseau

sasa.radosavljevic@ens-paris-saclay.fr

Résumé

Les infrastructures ferroviaires nécessitent une surveillance continue pour prévenir les défauts pouvant affecter la sécurité et l'efficacité du réseau. Cet article propose une évaluation expérimentale d'un système de détection de défauts visuels sur rails, basé sur le détecteur YOLOv8 et déployé sur des architectures CPU-GPU embarquées. L'étude analyse l'atteinte des objectifs de détection sous contraintes temps-réel et évalue la capacité du système embarqué à maintenir une précision de détection satisfaisante tout en respectant les exigences industrielles en matière de traitement embarqué. Les résultats obtenus avec des données issues d'acquisitions réelles montrent que l'approche retenue représente un compromis viable entre précision de détection, temps de traitement et contraintes matérielles.

Mots-clés

Détection de défauts visuels sur rails YOLOv8, Systèmes embarqués, Traitement temps-réel.

Abstract

Railway infrastructures require continuous monitoring to prevent defects that could impact the safety and efficiency of the network. This paper presents the detection of visual rail defects, an overview of the state of the art, and deployment results of a method based on the YoloV8 detector on embedded CPU-GPU architectures. Evaluations with real-world data show that, thanks to an algorithmic complexity study, the proposed system represents a compromise between defect detection accuracy and computation time.

Keywords

Railway visual defect detection, YOLOv8, Embedded systems, Real-time processing.

1 Introduction

Le secteur ferroviaire est confronté à des problèmes d'une grande complexité technique. L'expansion technologique de ce secteur (trains autonomes), la variété des métiers abordés, les contraintes environnementales et sociétales contribuent à accroître cette complexité. Face à ces enjeux, les entreprises ferroviaires se tournent de plus en plus vers des outils performants, frugaux en énergie et embarquables dans les trains commerciaux, avec pour objectif d'apporter au client final une meilleure qualité de service tout en optimisant les coûts. Ces dernières années, les systèmes développés ont cherché à répondre à cette complexité en combinant différentes approches, dont l'Intelligence Artificielle (IA) et l'information multimodale. La transformation des modes de transport, l'anticipation des besoins des clients, et la nécessité de prévenir les incidents, font que le secteur ferroviaire est naturellement concerné par les avancées de cette technologie. Cette nouvelle problématique de recherche rassemble de nombreuses équipes de chercheurs mutualisant leurs savoir-faire et générant des solutions alternatives, permettant de passer de la maintenance corrective à la maintenance prédictive et de sécuriser le trafic.
Les réseaux de neurones artificiels (RNA) et en particulier les réseaux convolutifs profonds (CNN) ont trouvé de nombreuses applications dans la maintenance des infrastructures et du matériel roulant. Dans ces domaines techniques, bien que les progrès réalisés par l'IA soient exponentiels, nous sommes encore loin d'exploiter son plein potentiel [1]. Son succès et sa diffusion dans notre domaine dépendent de nombreux paramètres : l'acceptation de ces techniques par les principaux acteurs métiers, et la mise à disposition de données annotées et de leurs qualité [2]. Cependant, et malgré de multiples avantages, l'utilisation des RNA ou des CNN souffre de nombreux défauts, notamment de la nécessité d'une base de données d'apprentissage cohérente et de la capacité à les annoter [3]. Cet article explore une approche de détection embarquée et temps réel des défauts ferroviaires à l'aide de réseaux de neurones optimisés. La section 2 donne un bref aperçu de l'état de l'art et de l'avancement dans la détection visuelle des défauts ferroviaires. La section 3 décrit le système embarqué proposé, tandis que la section 4 détaille la méthodologie d'évaluation de la détection de défauts par vision. La section 5 présente les résultats expérimentaux, analysant la précision de détection et les performances de temps de traitement en inférence. Enfin, la section 6 présente une discussion sur les perspectives d'amélioration.

2 Aperçu de l'état de l'art

Les recherches dans le traitement d'image ont été bousculées par l'émergence du machine learning et des avan-

cées dans les réseaux de neurones. La vision par ordinateur a significativement amélioré la détection automatique des défauts ferroviaires. Les caméras de haute résolution associées à des algorithmes d'intelligence artificielle (IA) permettent la détection d'une diversité de défauts tels que les fissures, les ondulations et l'usure. Les travaux de Kou et al. (2022) ont encouragé l'utilisation de réseaux de neurones convolutifs (CNN) pour la détection de défauts de surface des rails pour améliorer la détectabilité et les temps de détection [4]. De la même manière, les contributions de Lei et Jia (2019) ont démontré la possibilité d'utiliser des machines à vecteurs de support (SVM) et l'analyse en composantes principales (PCA) pour le traitement des ensembles de données visuelles pour la caractérisation des défauts [5].

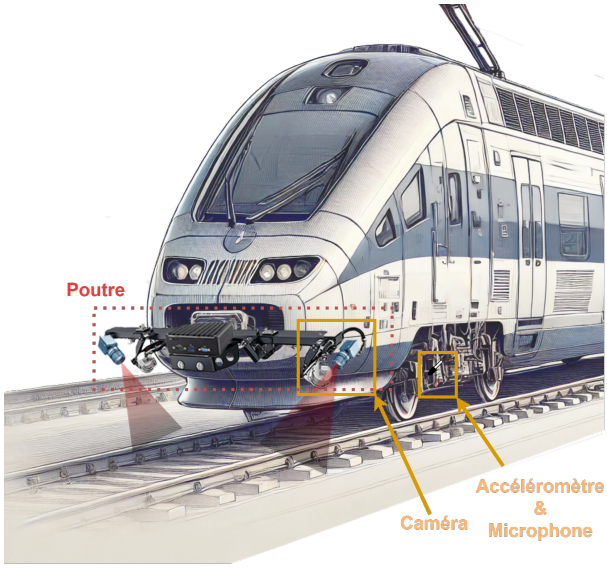


FIGURE 1 – Croquis du système embarqué train

Des études plus récentes, telles que la revue de Kumar et Harsha (2024) [6], ont approfondi l'utilisation des réseaux de neurones profonds, mettant en avant leur capacité à traiter des défauts complexes avec une grande précision. Par ailleurs, l'utilisation de systèmes d'imagerie embarqués sur drones a été explorée par Xiong et al. (2023) [7], soulignant leur potentiel pour accéder à des zones difficiles d'accès et garantir une qualité de données constante. Malgré ces avancées, les techniques de vision par ordinateur restent fortement influencées par des conditions externes telles que l'éclairage et les conditions météorologiques [8, 2]. Plus récemment encore, les architectures basées sur les transformeurs ont également été étudiées pour la détection des défauts ferroviaires. RailFormer, un système innovant basé sur les réseaux du type Transformer, a démontré des performances supérieures pour la détection des défauts en surface des rails en combinant des mécanismes d'attention Criss-Cross pour une extraction plus efficace des features [9]. De même, RailTrack-DaViT, une approche basée sur les transformeurs visuels, a surpassé les réseaux CNN traditionnels comme ResNet et EfficientNet sur plusieurs jeux d'entraînement, atteignant une précision inégalée pour l'inspection automatique des infrastructures ferroviaires [10].

Les techniques traditionnelles de contrôle non destructif (CND), bien que très précises, manquent de scalabilité. Les systèmes de vision par ordinateur, quant à eux, bien qu'efficaces, sont sensibles aux variations de l'environnement et aux contraintes liées à l'efficacité énergétique, au stockage des données et à l'encombrement des équipements. Ces limitations soulignent la nécessité d'une approche plus globale et adaptable pour la détection des défauts ferroviaires. Dans ce contexte, l'objectif de cet article est de démontrer la faisabilité du déploiement de YOLOv8n sur des architectures embarquées en conditions réelles. Pour cela, nous posons deux hypothèses de travail : (i) l'émulation d'un flux d'images à une vitesse équivalente à 160 km/h permet de simuler fidèlement un environnement embarqué ; (ii) la plateforme Jetson Orin Nano est capable de soutenir une fréquence de traitement suffisante ($\geq$ 30 FPS) tout en maintenant des performances de détection satisfaisantes. Bien que l'application visée puisse être assimilée à de la classification binaire (présence de défaut ou non), l'utilisation d'un modèle de détection d'objets se justifie par la présence de plusieurs éléments sur l'image d'une part, et la connaissance de la position de l'objet pour la corrélation à d'autres modalités d'autre part.

3 Modèle du système proposé

Afin de soulager les contraintes d'inspections sans ajouter de nouveaux engins dédiés à la surveillance de l'infrastructure, nous proposons l'intégration d'un système multimodal sur une poutre fixée à l'avant ou à l'arrière des trains (Fig. 1). De cette manière, l'analyse à la volée doit être réalisée avec la contrainte de la vitesse maximale du train.

3.1 Capteurs embarqués

Le système repose sur une combinaison de caméras linéaires, d'accéléromètres et de microphones, permettant une surveillance multimodale adaptée aux contraintes ferroviaires. Les caméras linéaires exploitent le déplacement rectiligne du train pour capturer des images haute vitesse avec une grande précision. L'acquisition est synchronisée avec l'odométrie du train, garantissant que chaque déclenchement d'acquisition correspond précisément à une avance mesurée du train. Les caméras sont placées de manière à observer l'intégralité des rails, que ce soit les éléments de l'âme du rail ou de son champignon (voir Fig. 2).

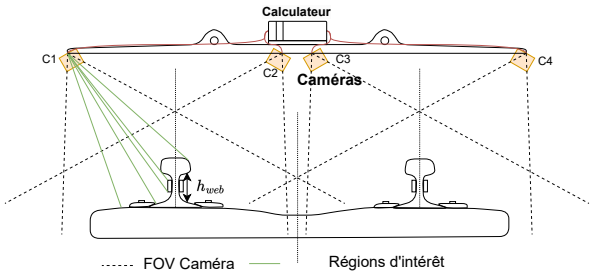


FIGURE 2 – Placement des caméras linéaires.

L'ajout des capteurs d'accélération et sonores permet d'élargir la détection aux défauts non visuels, en complétant l'analyse des caméras par des informations physiques supplémentaires sur l'interaction roue-rail. En effet, celle-ci ne peut avoir de sensation quant à ce qu'il se passe au niveau de l'interaction train/rail. Un accéléromètre est placé sur l'axe de la roue afin de mesurer la réponse dynamique du contact rail-roue lors du passage sur des défauts de surface, des anomalies structurelles ou des jonctions de rails. Un microphone complète l'analyse en apportant une redondance avec les mesures de l'accéléromètre et des caméras. Il permet également la détection de nouveaux défauts, tels que le grippage de frein ou des anomalies acoustiques au niveau des bogies. Toutefois, la combinaison de ces capteurs présentée dans cette section et les nouveaux défauts dont elle permet la détection ne sont pas explorés dans cet article, qui se limitera à l'exploitation des données visuelles.

3.2 Description des défauts

L'inspection automatisée des rails vise à identifier différents types de défauts (Fig. 3) pouvant impacter la sécurité et la maintenance de l'infrastructure du réseau ferroviaire. On peut les classer en plusieurs catégories selon leur zone d'apparition et leur moyen de détection. Outre les ruptures, la Table 1 regroupe les types de défauts et leurs descriptions selon deux catégories principales, les défauts de surface représentant les défauts d'usure du rail et les défauts structurels représentant les défauts des éléments constituant les rails. L'ajout des capteurs d'accélérométrie et de son permet d'identifier deux catégories supplémentaires difficilement détectables par les caméras, à savoir l'instabilité du ballast et les anomalies autour des bogies. A cela s'ajoute la détection d'éléments contextuels pouvant apporter des informations supplémentaires quant à la présence d'un défaut. Par exemple, la présence de cés de serrage enlève la nécessité de présence de tous les boulons. La présence d'une éclisse affirme un joint de rail à la place d'une rupture (si le modèle s'est trompé).

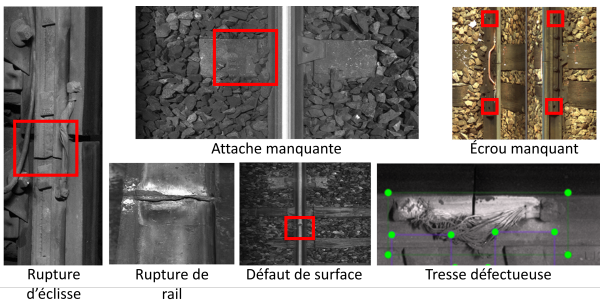


FIGURE 3 – Exemples de défauts.

3.3 Description de la chaîne de traitement

Le prototypage du système repose sur une architecture ROS2 (Robot Operating System 2) [13], permettant une communication efficace entre les différentes composantes logicielles et une gestion optimisée des flux de données en temps réel. Son paradigme de communication publish/subscribe (Fig. 4) assure une gestion efficace des flux de données provenant de capteurs hétérogènes, en maintenant une faible latence. Les données issues de la caméra sont simulées à l'aide d'un rosbag permettant de cadencer un flux de données.

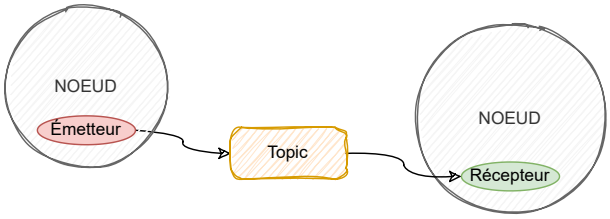


FIGURE 4 – Communication entre noeuds ROS2.

Dans notre cas, il n'y aura qu'un type de capteur (noeud caméra), un algorithme de traitement (noeud Yolov8n) et un élément de stockage des résultats (noeud bag). Les données sont alors transmises sous forme de topic aux subscribers concernés. La chaîne monomodale composée d'une caméra est présentée dans Fig. 6

L'architecture de réseau de neurones convolutifs (CNN) utilisée pour la détection des défauts est YOLO (You Only Look Once [14]). Ce modèle est conçu pour détecter les objets en une seule passe en générant des zones d'intérêt autour des éléments identifiés. Il associe simultanément à chaque région une classe prédite, permettant une classification efficace en temps réel.

YOLOv5 [15] a été l'un des premiers détecteurs d'objets à intégrer la prédiction des boîtes englobantes et des étiquettes de classe dans un réseau différentiable de bout en bout. Dans cette architecture, la partie convolutive du réseau est prédéfinie et figée, tandis qu'une partie des couches entièrement connectées reste ajustable pour adapter le modèle à une résolution spécifique.

Dans cette étude, nous utilisons YOLOv8 [16], qui apporte plusieurs améliorations significatives par rapport à YOLOv5. Les versions plus récentes ne sont pas étudiées car leurs améliorations n'impactent que peu les datasets à faible nombre de classes. Tout d'abord, l'optimisation des performances se traduit par une meilleure gestion de la quantification des poids (FP16, INT8), ce qui permet une exécution plus rapide sur des plateformes embarquées comme les Jetson Orin Nano/Xavier, tout en réduisant la consommation des ressources matérielles. Ensuite, l'architecture a été repensée avec un backbone amélioré, basé sur CSPDarknet, et l'intégration de mécanismes d'attention qui affinent la détection en concentrant les ressources de calcul sur les zones les plus pertinentes de l'image.

4 Méthodologie d'évaluation

Afin de simuler les contraintes réelles sur le système de calcul, un montage Hardware in the Loop (HiL, fig .5) a été proposé pour à la fois émuler l'acquisition des images et leur transfert vers l'élément de calcul, et avoir un environnement communicant permettant d'évaluer les temps de

TABLE 1 – Classification des défauts ferroviaires et leurs caractéristiques selon documents normatifs [11, 12]

Catégorie	Type de défaut	Description	Taille (mm)
Défauts de surface	Fissures	Microfissures ou fissures sur le rail	2 - 50
	Usure excessive	Érosion anormale du champignon du rail	> 10
	Corrugation	Ondulations périodiques sur le rail	10 - 1000
Défauts structurels	Rupture de rail	Fracture complète ou partielle du rail	50 - 12 000
	Boulons manquants	Fixations absentes	-
	Rupture d'éclisse	Défaillance aux joints de rails	50 - 150
Défauts contextuels*	Objets sur la voie	Corps étrangers pouvant perturber le trafic	Variable
	Niveau de ballast	Déformation sous les traverses	-
Interaction rail-roue*	Défauts de contact	Anomalies dans l'interface rail-roue	-
	Vibrations excessives	Dégradations mécaniques révélées par l'accéléromètre	-
	Bruits anormaux	Indicateur d'un problème de freinage ou bogie	-

*Non présenté dans l'article ou pas encore implémenté.

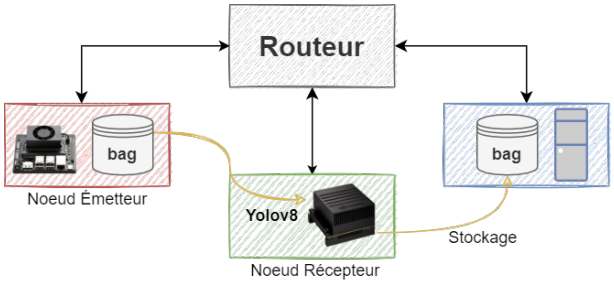


FIGURE 5 – Schéma du montage.

transfert et le comportement entre les différents éléments.

4.1 Entraînement du modèle

Dans une première étape, le modèle a été entraîné à partir d'un jeu de données issu d'images provenant de l'engin de surveillance SIM (Switch Inspection and Measurement) opéré par Eurailscoot. Ces données ont été manuellement annotées par un expert ferroviaire afin d'assurer une classification précise des défauts détectés. Cependant, la subjectivité de l'expert peut entraîner des variations dans la caractérisation des défauts, notamment pour ceux liés aux irrégularités de surface. Les données sont constituées de deux types de vues distinctes : une vue de dessus (RC - Rail centric), composée d'images en niveaux de gris de 1504×1500 pixels, et une vue latérale (SV - Side View), composée d'images en couleur RGB de 768×1508 pixels.

L'ensemble de données contextuelles [C] (Table 3) contient 1327 images RC et 9585 images SV pour l'entraînement, 167 images RC et 1296 images SV pour la validation, ainsi que 173 images RC et 1304 images SV pour le test. De son côté, l'ensemble des données des défauts [D] (Table 2) contient 3456 images RC et 5779 images SV pour l'entraînement, 282 images TV et 414 images SV pour la validation, ainsi que 263 images RC et 409 images SV pour le test.

Les jeux de données présentent un déséquilibre des données qui est dû à la fréquence d'apparition des éléments recherchés sur le réseau SNCF. Ces images étant annotées à

TABLE 2 – Jeu de données des défauts [**D**]

Classe	Composition					
	Train	***%***	***Val***	***%***	***Test***	***%***
Attache	8019	64.3	593	67.3	590	68.1
Surface	2949	23.6	195	22.1	183	21.1
Boulon	1502	12.1	93	10.6	93	10.8
Total	12470	87.7	881	6.2	866	6.1

Train. images : 9235, Val. images : 696, Test images : 672

TABLE 3 – Jeu de données des éléments de contexte [**C**]

Classe	Composition					
	Train	***%***	***Val***	***%***	***Test***	***%***
Tresse	3786	15.9	522	16.5	524	15.7
Cés de serrage	578	2.4	56	1.8	60	1.8
Éclisse	6045	25.5	795	25.2	872	26.2
Joint	6661	28.1	868	27.6	935	28.1
Soudure	2818	11.9	374	11.9	373	11.2
Marque	3828	16.2	533	17	563	17
Total	23716	78.5	3148	10.5	3327	11

Train. images : 10912, Val. images : 1463, Test images : 1477

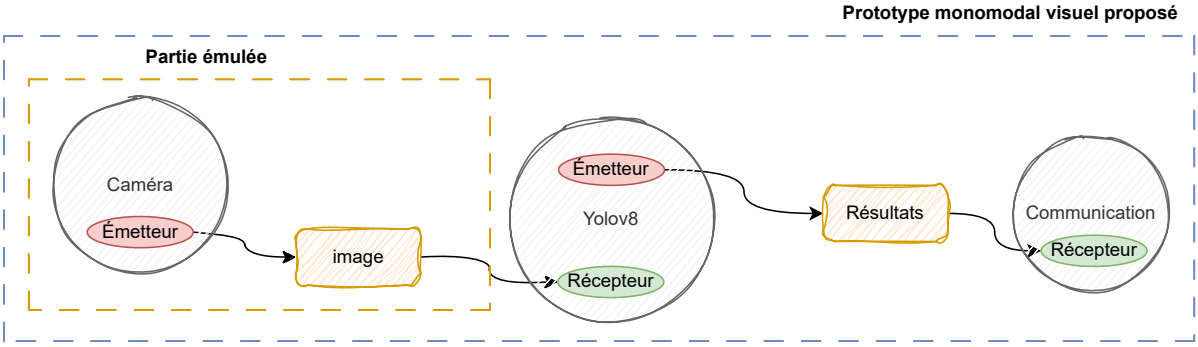


FIGURE 6 – Chaîne de traitement de l'architecture logicielle du prototype monomodal.

partir de plusieurs tournées, les objets ont une quantité proportionnelle aux défauts présents sur les voies, expliquant l'absence de ruptures dans les jeux de données. Afin de quantifier les performances du modèle de détection, nous définissons plusieurs métriques basées sur les notions de Vrai/Faux Positif/Négatif (Table 4).

TABLE 4 – Notions utiles pour la mesure d'efficacité d'un modèle

	Appartient à la classe	**N'appartient pas à la classe**
Attribué à la classe	Vrai Positif (VP)	Faux Positif (FP)
Non attribué à la classe	Faux Négatif (FN)	Vrai Négatif (VN)

Avec ces éléments, nous pouvons définir deux grandeurs qui seront par la suite analysées pour déterminer le point de fonctionnement d'un modèle pour une classe donnée.

$$\text{Précision} = \frac{VP}{VP + FP} \tag{1}$$

La Précision représente la capacité du modèle à minimiser les erreurs d'attribution d'un objet à une classe donnée.

$$\text{Rappel} = \frac{VP}{VP + FN} \tag{2}$$

Le Rappel évalue la capacité du modèle à ne pas omettre des objets appartenant à une classe donnée.

$$AP = \int_0^1 P(r)dr, \tag{3}$$

$$mAP = \frac{1}{N} \sum_{i=1}^{N} AP_i \tag{4}$$

La précision moyenne (4) résume la courbe Précision-Rappel (PR Curve) à un recouvrement de boîtes de détections donné (IoU, Intersection over Union = 0,5 veut dire que l'aire de l'intersection de la détection vaut la moitié de l'union de celle-ci et la vérité terrain). Elle permet d'indiquer la précision du modèle selon la rigueur appliquée à la localisation des objets détectés. Une valeur élevée indique que le modèle repère la majorité des objets en faisant peu d'erreurs de classification. Dans notre cas, un IoU de 0,5 est suffisant pour avoir une bonne idée de l'emplacement de l'objet sans détériorer le taux de détection.

4.2 Évaluation de la chaîne sur cible embarquée

L'évaluation du système en mode HiL a été réalisée sur une plateforme GPU embarquée de type Nvidia Jetson, la Jetson Xavier (Version 32 Go, GPU Nvidia Volta à 512 cœurs dont 64 cœurs Tensor, cadencé à 1377 MHz) en mode MAXN permettant d'utiliser le plein potentiel de la carte. Afin de simuler un environnement de production et d'évaluer les performances en conditions réelles, nous avons utilisé ROS2 et son outil "rosbag", permettant d'enregistrer et de rejouer des flux d'images à une cadence définie, imitant ainsi le déplacement du train.
La Jetson Xavier transmet les images à une fréquence définie (correspondant à la vitesse du train simulé). La Jetson Orin Nano exécute le noeud qui traite les images par réseaux de neurones. Les images sont publiées sur le Topic « /topic_image ». Les résultats des traitements sont ensuite enregistrés dans un rosbag, permettant d'analyser la latence et l'efficacité du pipeline sur les Topics « /Results_topic » et « /CnnTime_topic », respectivement le résultat de l'inférence et le tableau contenant les temps de calcul des différentes étapes de traitement.

4.3 Performances recherchées

Afin d'améliorer la réactivité de la détection des défauts et de limiter les contraintes de stockage, le passage d'un traitement "hors ligne" vers un traitement "en ligne" (à la volée) devient un objectif naturel et technologique. Pour cela, le système doit être temps-réel et répondre à une contrainte temporelle de traitement des données capteurs. L'objectif étant d'atteindre une vitesse de diagnostic confondue avec la vitesse de déplacement du train pour éviter l'accumulation de données dans le système. La vitesse de train visée est celle d'un TER pouvant circuler jusqu'à 160 km/h. La Table 1 indique une taille minimale d'un défaut visible de 2 mm imposant une définition minimale de la caméra de

TABLE 5 – Performances de détection des défauts en FP32 et FP16

Classe	FP32		FP16	
	mAP@0.5	**mAP@0.5-0.95**	**mAP@0.5**	**mAP@0.5-0.95**
Attache défectueuse	0.975	0.667	0.972	0.661
Défaut de surface	0.715	0.4	0.723	0.341
Boulon manquant	0.971	0.449	0.952	0.439
Toutes classes	0.887	0.487	0.882	0.480

TABLE 6 – Performances de détection du contexte en FP32 et FP16

Classe	FP32		FP16	
	mAP@0.5	**mAP@0.5-0.95**	**mAP@0.5**	**mAP@0.5-0.95**
Tresse	0.962	0.557	0.962	0.552
Cés de serrage	0.991	0.570	0.994	0.554
Éclisse	0.992	0.676	0.994	0.673
Joint	0.866	0.360	0.864	0.352
Soudure	0.953	0.501	0.950	0.499
Marque	0.937	0.528	0.938	0.523
Toutes classes	0.950	0.532	0.950	0.526

1 millimètre par pixel. Ainsi, l'objectif recherché est un système qui répond à un traitement temps-réel (30 FPS) avec des images de résolutions 1504x1504 pixels et une vitesse maximale de 160 km/h (éq. 44 445 pixels par seconde). Dans le même objectif, nous cherchons à atteindre une précision de détection de l'ordre de 95% avant compromis. Afin de réduire la charge de calcul tout en maintenant une précision acceptable, une sous-résolution des images à 2 mm/pixel peut être envisagée. Ce compromis permet de réduire les besoins en puissance de calcul sans altérer significativement la qualité de la détection. La pertinence de cette sous-résolution sera justifiée dans la section 5, où nous analyserons son impact sur les performances de classification et de reconnaissance des défauts.

5 Résultats et analyse

L'unité de calcul a été configurée de manière à exploiter au maximum la fréquence d'horloge et les capacités de calcul du système. En complément des temps de traitement du réseau de neurones, nous mesurons également le temps nécessaire à la désérialisation de l'image lors de la réception du message encodé par le noeud d'inférence. Cette approche permet ainsi d'évaluer le temps de traitement global de l'ensemble de la chaîne, en prenant en compte l'intégralité des étapes, de l'acquisition à l'inférence.

5.1 Performances de détection

La perte en précision relative au gain apporté en temps d'inférence pour une sous-résolution de 2,35 (Table 7) est très négligeable dans le cas des objets de contexte et est un compromis acceptable pour les éléments de défauts, notamment car la perte vient des défauts de surface que l'on vise à améliorer par l'ajout de modalités au système. De plus, on constate que le rapport de résolution et le rapport de temps d'inférence n'est pas quadratique comme on pourrait l'imaginer. Ce qui nous laisse imaginer pouvoir gagner encore un

TABLE 7 – Résultats de mAP pour différentes résolutions de pixel

Résolution (px)	**Contexte**		**Défauts**	
	640	**1504**	**640**	**1508**
mAP@0.5	0.950	0.953	0.887	0.904
mAP@0.5-0.95	0.523	0.538	0.487	0.497
Temps d'inférence (rapport)	x1	x1,8	x1	x1,7

ratio en faisant l'inférence sur plusieurs images côte à côte à condition d'avoir les capacités mémoire intégrées dans le système.

La première évaluation porte sur la capacité du modèle à détecter les objets recherchés. Un premier rosbag contenant la partie test (évaluation) du modèle est présenté au nœud d'inférence. Ce bag est décrit dans la colonne test des Tables 2 et 3. Les résultats de détection sont présentés Table 5 et 6. L'étude de la quantization des modèles permet de montrer que les pertes apportées par le passage de flottants sur 32 bits (FP32) vers 16 bits (FP16) sont très faibles [17], notamment pour les éléments de contexte. Un écart plus important est présent entre les défauts de surface et les boulons manquants, qui, combinés, ne provoquent pas une perte importante.

5.2 Résultats en temps d'inférence

Finalement, pour mesurer les capacités des architectures embarquées à traiter les images en temps réel, un rosbag contenant une tournée de l'engin SIM sur une voie parisienne composé de 4416 images avec ou sans objets à détecter est envoyé sur le noeud de traitement pour obtenir une mesure statistique des différents temps de traitement. L'analyse a été portée, d'une part, sur la mesure des temps de traitement autour de l'inférence tel que la désérialisation des messages ROS, le pré-traitement ainsi que

TABLE 8 – Résultats pour le temps de traitement sur la Jetson Xavier avec coefficients sur 32 bits et 16 bits.

Format	Désérialisation (ms)	Prétraitement (ms)	Inférence (ms)	Post-traitement (ms)	Temps total (ms)	Fréquence (FPS)
FP32						
Moyenne	1.25	4.07	9.06	3.12	17.50	57.1
Écart-type	0.23	0.75	0.06	0.10	1.14	3.56
Médiane	1.16	3.80	9.06	3.12	17.14	58.34
Minimum	1.07	3.64	9.01	1.73	15.45	64.72
Maximum	2.05	7.65	9.16	3.59	22.45	44.54
FP16						
Moyenne	1.22	4.01	5.36	3.10	13.69	73.0
Écart-type	0.21	0.05	0.76	0.10	1.12	5.48
Médiane	1.15	3.75	5.36	3.09	13.35	74.9
Minimum	1.05	3.58	5.31	2.85	12.79	78.18
Maximum	2.01	7.50	5.46	3.79	18.76	53.3

le post-traitement, et d'autre part sur l'inférence du réseau sur le GPU de la Jetson AGX Xavier. Les 1% des temps d'inférence les plus bas et les 1% les plus élevés sont exclus des statistiques afin d'éliminer les valeurs aberrantes. L'inférence est réalisée à partir d'un modèle TensorRT (engine) exporté depuis les poids du modèle PyTorch (pt). Ce dernier ayant en moyenne 50% en plus en temps d'inférence sur l'architecture Xavier, ce qui n'est pas envisageable pour un déploiement sur cible embarquée. Afin d'optimiser les performances d'inférence du modèle sur GPU embarqué, une étude de quantification a été réalisée. Cette opération consiste à convertir les poids du modèle initial, stockés sur 32 bits flottants (FP32), vers un format réduit à 16 bits flottants (FP16). L'objectif est de diminuer les besoins en mémoire et d'accélérer les calculs en bénéficiant du support matériel des GPUs modernes pour la précision FP16. Les mesures réalisées comparent les temps de traitement moyens des étapes de la chaîne d'inférence (désérialisation, prétraitement, inférence, post-traitement) entre les deux formats. Les résultats (voir Table 8) montrent que la quantification FP16 permet de réduire significativement le temps total de traitement par image, passant de 17,5 ms à 13,69 ms en moyenne, soit un gain d'environ 21% pour l'ensemble de la chaîne et de 40% pour l'inférence. Cela se traduit par une augmentation de la fréquence d'images traitées de 57 FPS à 73 FPS, ouvrant la possibilité de traiter deux flux caméra simultanément avec une seule unité de calcul. Ce gain est obtenu sans dégradation notable de la qualité de détection.

6 Perspectives et conclusion

Les résultats des expériences réalisées montrent la possibilité de l'utilisation du modèle YOLOv8 nano pour la détection de défauts ferroviaires en temps réel sur des architectures embarquées. Les compromis et les optimisations réalisés permettent d'obtenir des résultats en termes de détection et de temps d'inférence viables pour le déploiement de tel modèle sur un système embarqué.

L'un des premiers axes d'amélioration concerne la robustesse du modèle face aux conditions réelles d'exploitation. Bien que les performances soient évaluées en laboratoire avec des données issues de relevés réels, l'impact des variations environnementales, telles que les changements de luminosité, les conditions météorologiques et les vibrations du train, reste à explorer en profondeur. Il est essentiel de valider l'efficacité du modèle dans ces conditions réelles pour garantir une détection fiable sur l'ensemble du réseau ferroviaire.

Par ailleurs, l'optimisation des modèles pour l'embarqué reste un enjeu crucial. L'étude a montré que la quantification en FP16 permet une exécution plus rapide sans perte significative de précision, mais d'autres méthodes d'optimisation, comme l'usage de modèles quantifiés en INT8, le pruning de réseaux ou des accélérateurs spécialisés dans les inférences d'IA pourraient encore améliorer les performances, notamment à exécuter plusieurs flux vidéo en parallèle et amener à un système encore plus frugal et compact. Enfin, pour garantir une adoption généralisée, il serait pertinent d'évaluer la généralisation du modèle sur d'autres réseaux ferroviaires à l'international. Chaque réseau possède des spécificités techniques et environnementales qui peuvent influencer la performance des algorithmes de détection. Un réapprentissage spécifique à chaque réseau, basé sur des données issues de leurs infrastructures ferroviaires, permettrait d'adapter le modèle à diverses configurations et d'améliorer sa robustesse face à des scénarios variés.

En conclusion, cette étude a permis de démontrer l'intérêt des méthodes de détection basées sur l'IA pour la maintenance ferroviaire en temps réel. L'intégration du modèle YOLOv8 sur des architectures embarquées a prouvé sa capacité à détecter efficacement les défauts tout en respectant les contraintes de calcul inhérentes aux systèmes embarqués. Toutefois, plusieurs axes de recherche restent à explorer pour améliorer la robustesse, l'évolutivité et l'intégration du système dans une architecture globale de maintenance prédictive des infrastructures ferroviaires. Ces perspectives ouvrent la voie à une nouvelle génération de solutions intelligentes pour la surveillance et l'entretien des voies ferrées, contribuant ainsi à améliorer la sécurité et l'efficacité du réseau ferroviaire.

Références

[1] Wassamon Phusakulkajorn, Alfredo Núñez, Hongrui Wang, Ali Jamshidi, Arjen Zoeteman, Burchard Ripke, Rolf Dollevoet, Bart De Schutter, and Zili Li. Artificial intelligence in railway infrastructure : current research, challenges, and future opportunities. *Intelligent Transportation Infrastructure*, 2023. https://doi.org/10.1093/iti/liad016.

[2] Guoqing Jing, Xuanyang Qin, Haoyu Wang, and Chengcheng Deng. Developments, challenges, and perspectives of railway inspection robots. *Automation in Construction*, 2022. https://doi.org/10.1016/j.autcon.2022.104242.

[3] Andrei Popescu-Belis. Managing multimodal data, metadata and annotations : Challenges and solutions. *Multimodal signal processing : Theory and applications for human-computer interaction*, page 207, 2009. https://doi.org/10.1016/B978-0-12-374825-6.00013-7.

[4] Lei Kou. A Review of Research on Detection and Evaluation of the Rail Surface Defects. *Acta Polytechnica Hungarica*, 2022. https://doi.org/10.12700/APH.19.3.2022.3.14.

[5] Lei Jia, Ming Zhu, Ryan Sherman, Jeewoong Park, Yingtao Jiang, and Hualiang Teng. Rail defect detection technology : A review of the current methods and an acoustic based method proposed for high-speed-rail. 11 2019. https://doi.org/10.1177/0361198105191600110.

[6] Ankit Kumar and S. P. Harsha. A systematic literature review of defect detection in railways using machine vision-based inspection methods. *International Journal of Transportation Science and Technology*, 2024. https://doi.org/10.1016/j.ijtst.2024.06.006.

[7] Longhui Xiong, Guoqing Jing, Jingru Wang, Xiubo Liu, and Yuhua Zhang. Detection of Rail Defects Using NDT Methods. *Sensors*, 2023. https://doi.org/10.3390/s23104627.

[8] Milica Mićić, Ljiljana Brajović, Luka Lazarević, and Zdenka Popović. Inspection of RCF rail defects – Review of NDT methods. *Mechanical Systems and Signal Processing*, 2023. https://doi.org/10.1016/j.ymssp.2022.109568.

[9] Feng Guo, Jian Liu, Yu Qian, and Quanyi Xie. Rail surface defect detection using a transformer-based network. *Journal of Industrial Information Integration*, 2024. https://doi.org/10.1016/j.jii.2024.100584.

[10] Aniwat Phaphuangwittayakul, Napat Harnpornchai, Fangli Ying, and Jinming Zhang. RailTrack-DaViT : A Vision Transformer-Based Approach for Automated Railway Track Defect Detection. *Journal of Imaging*, 2024. https://doi.org/10.3390/jimaging10080192.

[11] Federal Railroad Administration. Manual of rail defects. *FRA Technical Documents*, 2015.

[12] International Union of Railways. Rail defect manual - uic7-2500-1. 2007.

[13] Steve Macenski, Tully Foote, Brian Gerkey, Chris Lalancette, and William Woodall. Robot Operating System 2 : Design, Architecture, and Uses In The Wild. *Science Robotics*, 2022. https://doi.org/10.48550/arXiv.2211.07752.

[14] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once : Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016. https://doi.org/10.48550/arXiv.1506.02640.

[15] Glenn Jocher. Ultralytics yolov5, 2020. https://github.com/ultralytics/yolov5.

[16] Q. Wang, W. Feng, H. Zhao, B. Liu, and S. Lyu. Yolov8 : A novel object detection algorithm with enhanced performance and robustness. In *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, 2024. https://doi.org/10.1109/adics58448.2024.10533619.

[17] Tailin Liang, John Glossner, Lei Wang, Shaobo Shi, and Xiaotong Zhang. Pruning and Quantization for Deep Neural Network Acceleration : A Survey, June 2021. https://doi.org/10.48550/arXiv.2101.09671.

Session 6 : APIA-RJCIA – IA pour l'analyse de graphes, de textes et d'images

Détection de communautés dans les graphes de connaissance d'activités

Marthe Désirée Olivia HABACK[1], Serge SONFACK SOUNCHIO[2], Orlane SONKENG TSAFACK[1], Halguièta TRAWINA[3], Ho Tuong Vinh[1]

[1] École internationale, Université nationale du Vietnam, Hanoï, Vietnam
[2] Robert Bosch Sp. z o. o. Jutrzenki 105, POLAND
[3] Laboratoire de Mathématique et d'Informatique; Université Joseph KI-ZERBO, Burkina Fasso

5 mai 2025

Résumé

La gestion des connaissances produites au cours des activités au sein des organisations joue un rôle important dans leur développement et leur succès. La formalisation de ces connaissances sous forme de graphe de connaissance d'activités (Activity Knowledge Graph : AKG) permet de représenter et de réutiliser ces connaissances pour résoudre des problèmes et prendre des décisions tactiques ou stratégiques. Cependant, les graphes de connaissance d'activités sont difficiles à interpréter par des cadres, car ils disposent d'une structure complexe due aux interconnexions des activités, ressources partagées. De plus, les AKGs sont spécifiques aux domaines d'activités des organisations.
Pour faciliter l'interprétation des graphes de connaissance d'activités, cette étude propose une approche basée sur la détection de communautés. Cette approche repose sur trois étapes : (1) un pré-traitement pour adapter l'AKG à l'algorithme de Louvain, (2) l'application de l'algorithme de Louvain pour la détection de communautés et (3) un post-traitement pour intégrer les communautés détectées à l'AKG.
Un cas d'étude dans le domaine de l'enseignement illustre l'applicabilité de notre approche, et les résultats démontrent la simplicité obtenue dans la compréhension des connaissances représentées par le graphique de connaissance d'activités.

Mots-clés

Graphe de connaissance, graphe de connaissance d'activité, détection de communauté, algorithme de Louvain.

Abstract

Managing the knowledge generated by organizational activities is important in their development and success. Formalizing this knowledge in the form of an Activity Knowledge Graph (AKG) enables it to be represented and reused to solve problems and make tactical or strategic decisions. However, activity knowledge graphs are difficult for managers to interpret, as they have a complex structure due to the interconnections between activities and shared resources. In addition, AKGs are specific to an organization's field of activity. This study proposes an approach based on community detection to facilitate the interpretation of activity knowledge graphs. This approach relies on three steps : (1) pre-processing to adapt the AKG to the Louvain algorithm, (2) application of the Louvain algorithm for community detection, and (3) post-processing to integrate the detected communities into the AKG. A case study in the educational domain illustrates the applicability of the approach, and the results demonstrate the simplicity achieved in understanding the knowledge represented by the activity knowledge graph.

Keywords

Knowledge graph, activity knowledge graph, community detection, Louvain alogorithm.

1 Introduction

Un graphe de connaissance est une structure permettant de représenter et d'interconnecter des informations, facilitant ainsi la représentation de la connaissance en se basant sur la logique relationnelle qui existe entre entités d'un domaine. Il constitue un outil pour les entreprises, car peut-être utilisé les connaissances de leurs activités et leur exploitation peut servir à prendre des décisions pour rester compétitives. De plus, la gestion des connaissances au sein d'une entreprise s'avère essentielle pour préserver et transmettre les savoirs accumulés, en particulier ceux des employés expérimentés proches de la retraite ou ayant eu une longue carrière au sein de l'entreprise. Une telle gestion favorise le partage d'informations, la transmission des compétences, et améliore l'efficacité organisationnelle.

Le graphe de connaissance d'activités (Activity Knowledge Graph : AKG) est une représentation de la connaissance liée aux activités d'une entreprise qui permet de conserver leur savoir-faire de l'entreprise. Cette connaissance capture le contexte des activités comme la temporalité, les règles, les objectifs, les sujets et les bénéficiaires, les outils et les lieux [11]. L'AKG d'une entreprise repose

sur l'ontologie de haut niveau (Core Activity Ontology), qui permet de créer les ontologies dans divers domaines [9].

Cependant, la forte interconnexion entre les éléments de contexte des activités et les activités elles-mêmes et la spécificité des AKG aux domaines des entreprises les rendent difficiles pour les managers appelés à prendre des décisions en entreprise sans toutefois avoir les compétences techniques du domaine de l'entreprise. En outre, l'utilisation des requêtes permet de répondre aux questions techniques, mais pas d'abstraire les connaissances pour faciliter la compréhension aux niveaux non-techniques.

Dans cette étude, nous allons proposer une approche pour la détection de communautés dans un graphe de connaissance d'activités (Activity Knowledge Graph : AKG), qui permettra de simplifier la compréhension de la connaissance des activités d'une entreprise et par conséquent améliorer la prise de décision.
Pour la suite, nous proposons d'organiser cet article en trois sections. La prochaine section (Section 2) présentera les travaux connexes relatifs à la détection de communautés, la section qui suivra (Section 3) détaillera notre méthodologie et les résultats de notre expérimentation seront illustrés à la Section 4, et nous proposerons une conclusion à la dernière section.

2 État de l'art

Dans notre travail, nous étudions les graphes de connaissances d'activités au sein des entreprises, en nous concentrant particulièrement sur le domaine éducatif. L'objectif est d'extraire de ces graphes une nouvelle forme de connaissance, simple à exploiter, qui reflète le graphe initial et apporte des informations utiles pour la prise de décision. Pour répondre à ce besoin, nous nous sommes appuyés sur plusieurs travaux récents relatifs aux graphes de connaissances et à la détection de communautés.

Une approche a proposé une structuration des grands graphes réels en communautés, basée sur une technique de clustering qui organise les données de manière hiérarchique en utilisant l'algorithme de Louvain [4]. Cependant, cette approche révèle des limitations significatives dans la gestion des communautés évolutives. En effet, malgré la combinaison de l'algorithme de Louvain avec des techniques de cartographie pour l'analyse des clusters, l'auteur n'est pas parvenu à développer une méthode efficace pour la construction et le suivi des communautés évolutives. Cette contrainte, associée aux difficultés d'utilisabilité des méthodes de détection de communauté, met en évidence les défis persistants dans l'analyse dynamique des réseaux sociaux.

D'autres efforts ont exploité les algorithmes de classification et de clustering pour les graphes de connaissances, comme le montrent [12] à travers leur approche basée sur les techniques d'apprentissage automatique pour la détection des communautés et l'analyse des relations entre entités. Les résultats montrent que ces techniques améliorent la précision des modèles d'analyse de graphiques et facilitent la prise de décision dans divers contextes organisationnels.

Dans leur étude **Organisations' Interpersonal Activity Knowledge Representation**, [11] explorent les connaissances explicites et implicites, essentielles au fonctionnement organisationnel et leurs limites. Leur approche, basée sur l'ontologie, met en évidence les défis de l'extraction de connaissances spécifiques pour une utilisation pratique, et pose les bases d'une détection de communautés pour optimiser les décisions organisationnelles.

Enfin, [7] nous introduit une approche basée sur **A motif - based probabilistic approach for community detection in complex networks**, en utilisant des algorithmes comme la descente de coordonnées de bloc. Bien que des difficultés aient été rencontrées lors de la mise en œuvre, cette approche ouvre la voie à des solutions futures utilisant des modèles génératifs.

Pour réaliser la détection de communautés, il est nécessaire de suivre un canevas bien défini, impliquant un processus structuré pour atteindre les résultats attendus. Ainsi, le choix de l'algorithme à utiliser est une étape clé dans ce processus tel que décrit par [17] et [8].

- **Méthode de Louvain** : L'algorithme de Louvain est populaire pour la détection de communautés dans les réseaux statiques. Et il peut également être adapté aux réseaux dynamiques. Il optimise la modularité du réseau en fusionnant et en divisant les communautés de manière itérative. Dans le contexte dynamique, il peut être étendu pour tenir compte des changements au fil du temps.
- **Infomap** : Cet algorithme est basé sur la théorie de l'information. Il cherche à minimiser la longueur moyenne de la description d'un parcours aléatoire sur le réseau et peut être utilisé pour détecter des communautés dynamiques en analysant les transitions entre les états du réseau.
- **LD-ABCD (Local Dissimilarities - Agent Based Cluster Discovering)** : Cet algorithme est basé sur la détection de clusters dans un graphe pondéré, ce qui correspond bien à notre contexte d'étude des interactions entre entités ou activités. Il fournit des groupes ou des clusters qui sont étroitement liés les uns aux autres en fonction de leurs interactions.
- **Clustering** : L'algorithme de Clustering est défini comme un processus automatique qui regroupe des textes non étiquetés (comme des tweets ou des descriptions de produits) en clusters ou en groupes de textes similaires. L'objectif est de détecter des motifs de similarité entre les données pour les organiser de manière sémantique [1].

Après avoir étudié plusieurs méthodes, nous avons opté pour l'algorithme de **Louvain** en raison de son efficacité dans la détection de communautés, mais également du fait qu'il permet l'optimisation de la modularité du graphe et qu'il permet de regrouper des activités similaires en communautés, ce qui facilite ainsi l'analyse et l'exploitation des graphes de connaissances.

3 Approche proposée

3.1 Notion de graphe de connaissance d'activités (AKG)

Un graphe de connaissance d'entreprise ou réseau sémantique d'entreprise, est considéré comme étant un ensemble de données, d'informations et d'objet d'une entreprise représentée sous forme de graphe structuré contenant les entités, les attributs et les relations entre ces entités et présenté sur un format compréhensible par un système [15] [6].
Un graphe de connaissance d'activités (Activity Knowledge Graph : AKG), comme illustré dans la Figure 1 est un graphe de connaissances sémantique des activités d'une entreprise et reposant sur une extension de l'ontologie de haut niveau, Core Activity Ontology (CAO) [1] [14]. Dans un AKG, des nœuds représentent les activités d'une entreprise, chacune associée à des ressources contextuelles comme la temporalité, les contraintes, les sujets, les bénéficiaires, les outils, la localisation et la motivation [13] [12].

3.2 Communauté dans l'AKG

Une communauté en général est un ensemble d'entités qui entretiennent des liens privilégiés parce qu'elles ont des affinités particulières, ou présentent des caractéristiques similaires, ou encore partagent des centres d'intérêt, ainsi que d'autres éléments [5]. Dans le sens des graphes, une communauté du point de vue structurel est un sous-ensemble de nœuds fortement connectés, c'est-à-dire qu'il existe entre eux un nombre important de connexions, davantage que celles que ces nœuds entretiennent avec les autres nœuds du même graphe [10].

Une communauté dans le cas d'un graphe de connaissance d'activités (AKG) est un ensemble d'activité fortement liée par des éléments de contexte partagés comme le temps, la motivation, les sujets (employés de la structure réalisant les activités), les bénéficiaires ou même les lieux des activités.

La détection de communautés dans un graphe sans tenir compte de la sémantique fournie par l'ontologie est simple et directe, car l'utilisation d'un algorithme ne nécessite pas des modifications de celui-ci. Néanmoins, pour les graphes sémantiques, il devient nécessaire de soit adapter les algorithmes de détection de communautés, soit d'adapter les graphes sémantiques (avec ontologie) aux algorithmes de détection de communautés. C'est cette deuxième approche que nous avons adoptée pour la détection des communautés dans les graphes de connaissances d'activités.

Pour détecter les communautés dans les graphes de connaissance d'activités, l'approche proposée repose sur trois étapes comme illustrées par la Figure 2 :

1. Pré-traitement : cette étape transforme un graphe de connaissance d'activités en un graphe plus simple dans lequel les seuls nœuds sont les activités et les relations entre les activités correspondent au fait que les deux activités ont une ressource en commun.
2. La deuxième étape est l'application de l'algorithme de Louvain pour la détection de communautés.
3. Post-traitement : cette étape permet de reconstituer le graphe de connaissance d'activités initial en tenant compte des différentes communautés identifiées à l'étape précédente.

3.2.1 Pré-traitement

L'étape de pré-traitement est celle qui permet de transformer un graphe de connaissance d'activités en un graphe simple pouvant être utilisé par l'algorithme de Louvain sans apporter des modifications particulière à l'algorithme. Cette étape allège le graphe de connaissance d'activités en convertissant la sémantique de l'ontologie du graphe à des nouvelles interactions entre les nœuds activité du graphe.

L'algorithme de transformation 1 est le suivant : toutes les activités du graphe de connaissance d'activités sont des nœuds du graphe simplifié. Par la suite, on ajoute une relation entre deux nœuds du graphe simplifié si les deux activités en question partagent un élément de contexte en commun comme la temporalité, le lieu, les ressources.

Algorithm 1 Pré-traitement

Require: Graphe de connaissance d'activités (AKG)
Ensure: AKG : G
Ensure: Graphe : $G^{'} = (V, E)$
V$\leftarrow \emptyset$, E$\leftarrow \emptyset$
for Activité :A $\in$ G **do**
  V$\leftarrow A$
end for
for Activité :A $\in$ V **do**
  for Activité :A$^{'}\in$ V **do**
    if {Nœud, A} $\in E(G)$ **and** {Nœud, A$^{'}$}$\in E(G)$ **then**
      E$\leftarrow \{A, A^{'}\}$
    end if
  end for
end for
Résult : Graphe : G$^{'}$ ▷ Graphe d'activités simplifié

1. github.com/sonfack/ActivityOntology/blob/main/core_activity_ontology.ttl Cette transformation implique évidemment une perte

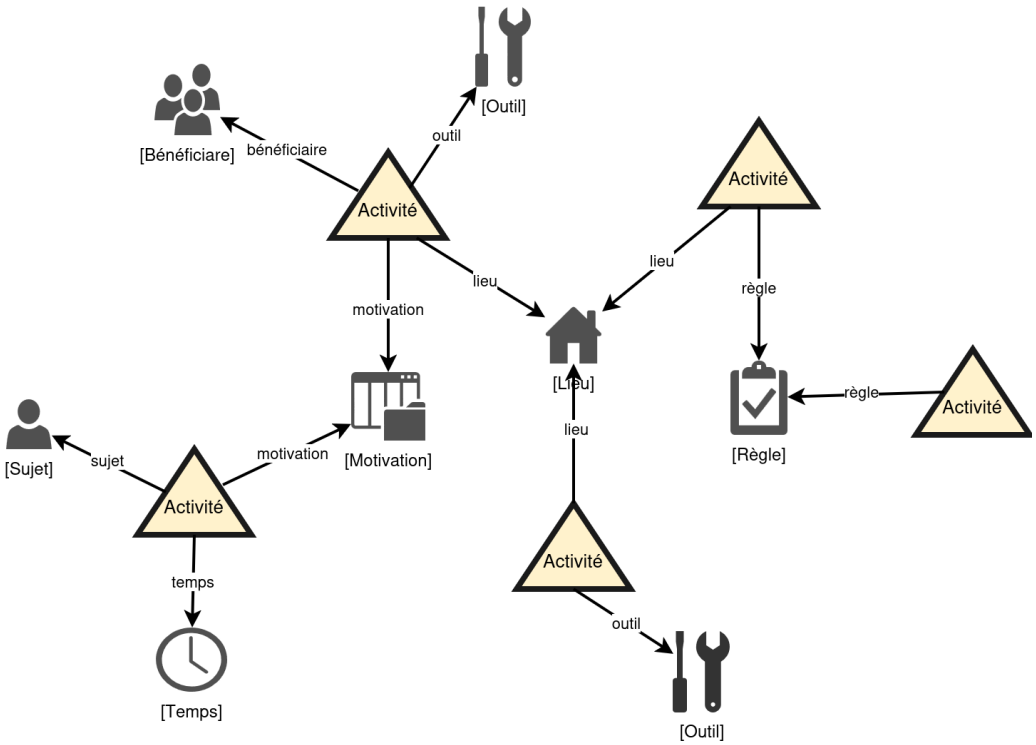


FIGURE 1 – Illustration d'un graphe de connaissance d'activité montrant les activités en triangles jaunes ainsi que les relations qui existent entre elles par l'intermédiaire des ressources ou éléments de contexte qu'elles partagent.

d'information sur le graphe résultant. Mais, elle nous permet d'avoir une sémantique claire pour la détection de communautés tout en simplifiant la structure du graphe pour l'algorithme de Louvain. En outre, l'étape du post-traitement nous permettra de remédier la perte de sémantique causée par la simplification du graphe initial.

3.2.2 Algorithme de Louvain

Pour notre travail, nous avons opté pour l'algorithme de Louvain. Cet algorithme a été sélectionné en raison de son efficacité et de sa capacité à gérer des graphes dynamiques. L'algorithme est hiérarchique et repose sur l'optimisation de la modularité, un critère mesurant la densité des liens à l'intérieur des communautés par rapport aux liens externes.

Cet algorithme optimise la modularité, qui est un indicateur qui permet de mesurer la densité des connexions à l'intérieur des communautés comparées à l'extérieur. Cet algorithme est divisé en deux phases [2, 3] :

— Phase 1
Chaque nœud du réseau constitue initialement une communauté à lui seul : à ce stade, il y a donc autant de communautés que de nœuds. Pour chaque nœud i, on examine ses voisins directs j et on évalue le gain de modularité potentiel si le nœud i est déplacé dans la communauté de j. Le nœud i est alors transféré dans la communauté offrant le gain de modularité positif le plus élevé. Si aucun gain positif n'est obtenu, i reste dans sa communauté d'origine. Cette procédure est répétée jusqu'à ce qu'aucune amélioration supplémentaire ne soit possible.

— Phase 2
Reconstruire le nouveau réseau dont les nœuds sont les communautés identifiées dans la phase 1 précédente. A cet effet, le poids de la relation entre deux nœuds du nouveau réseau est la somme des poids des relations entre les nœuds des communautés intervenant dans la nouvelle relation. A la fin de cette phase 2 on recommence avec la phase 1 ainsi de suite jusqu'à ce qu'on observe plus de changements et qu'on ait une modularité maximale. En effet, à chaque itération, on observera une réduction du nombre de nœuds et du temps de calcul.

Ce processus continu nous permet de créer une hiérarchie de communauté qui facilitera l'analyse des interactions entre activités [16].
La modularité Q est définie d'apres cet approche :

$$Q = \frac{1}{2m} \sum_{i,j} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j)$$

Où :

— A_{ij} est l'élément de la matrice d'adjacence, valant 1 si les nœuds i et j sont connectés, et 0 sinon.
— k_i et k_j sont les degrés des nœuds i et j, c'est-à-dire le nombre de connexions de chaque nœud.
— m est le nombre total d'arêtes dans le graphe.

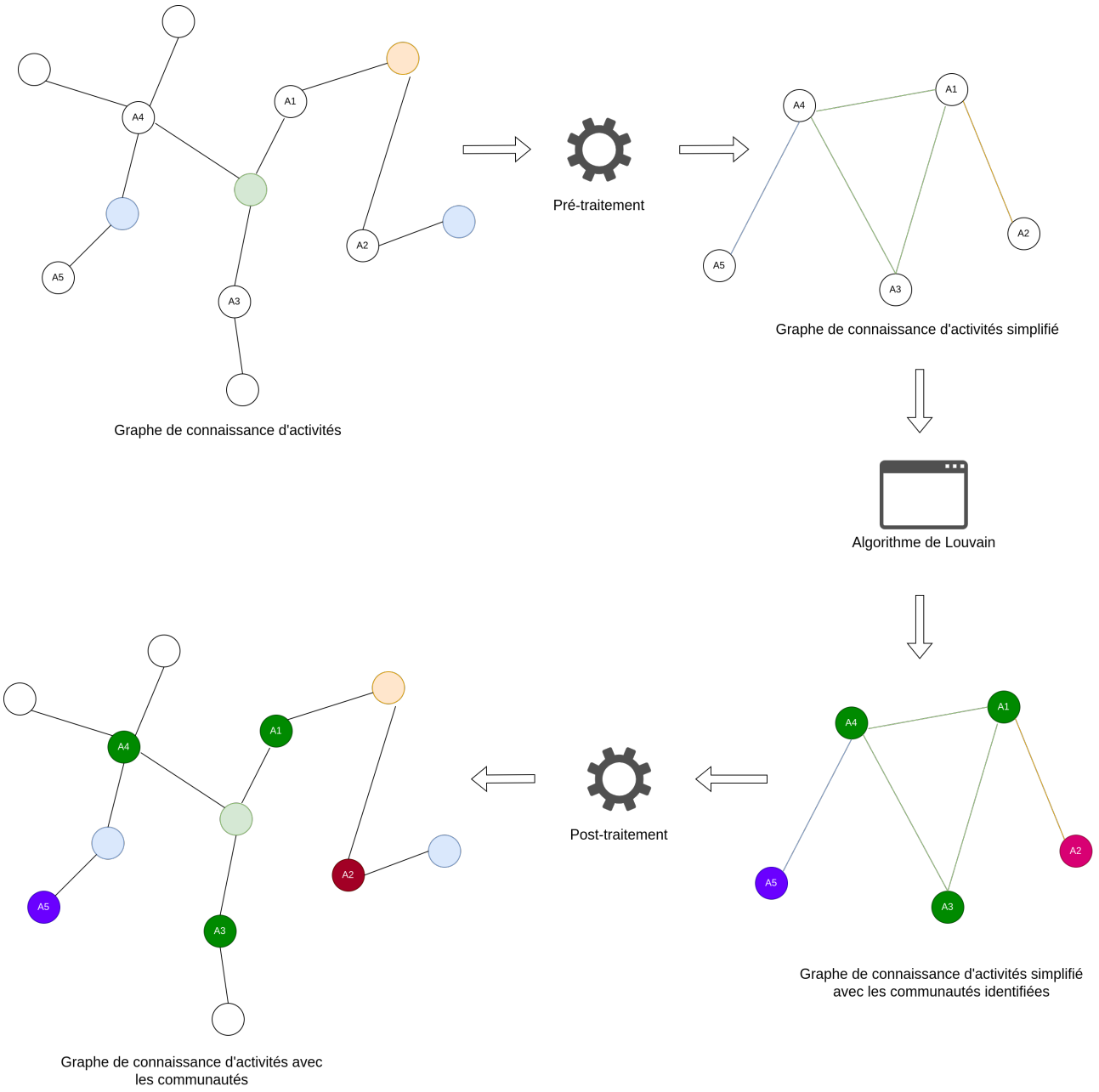


FIGURE 2 – Approche de détection de communauté pour les graphes de connaissance d'activités : 1) Pré-traitement

— $\delta(c_i, c_j)$ est la fonction delta de Kronecker, qui vaut 1 si i et j appartiennent à la même communauté, et 0 sinon.
— $\frac{k_i k_j}{2m}$ représente la probabilité attendue d'une arête entre les nœuds i et j dans un graphe aléatoire ayant les mêmes degrés.
— Initialisation : chaque nœud du graphe est initialement assigné à sa propre communauté.
— Fusion des communautés : l'algorithme parcourt chaque nœud et tente de l'assigner à la communauté de ses voisins, maximisant l'augmentation de la modularité.
— Itération : les communautés sont progressivement fusionnées jusqu'à atteindre une optimisation maximale de la modularité. Ce processus est répété de manière itérative pour créer une hiérarchie de communautés.

3.2.3 Post-traitement

Cette étape de l'approche proposée permet de transformer le graphe de connaissance d'activités simplifié avec les communautés détectées au graphe de connaissance d'activité initiale, mais en conservant les différentes communautés identifiées. Le post-traitement est aussi l'étape qui permettra de faire l'interprétation ou de comprendre le graphe de connaissance d'activités de départ

en fonction de la sémantique de l'étape de *pré-traitement*. L'algorithme 2 illustre cette étape.

Algorithm 2 Post-traitement

Require: Graphe de connaissance d'activités (AKG) simplifié avec les différentes communautés
Require: Liste des communautés
Ensure: AKG : G
Ensure: Communautés :C ▷ Liste des communautés obtenues de l'algorithme de Louvain

```
for Activité :A ∈ G do
    for c ∈ C do
        if A ∈ c then
            Mettre à A l'étiquette de sa communauté
        end if
    end for
end for
```

Résult : Graphe : G ▷ Graphe de connaissance d'activités avec les activités étiquetés en fonction de leur communauté

4 Cas d'application

4.1 Présentation de données

Les données utilisées dans l'application de l'approche proposée sont un graphe de connaissance d'activités d'enseignement construit d'une extension du core activity ontology (CAO) et contenant les activités comme les cours, les travaux dirigés, les projets et les travaux pratiques. Les instances de ces activités sont reliées aux instances faisant partie de leur contexte comme les ressources, les lieux, les motivations, les objectifs, les heures de début et de fin, les personnes, les activités.

4.2 Résultats

Le graphe de connaissance utilisé pour l'application de notre approche contient un total de 341 nœuds représentant des activités et des instances, ainsi que 505 arêtes représentant les relations entre ces entités. La figure 3 représente le graphe de connaissance d'activités de notre étude.

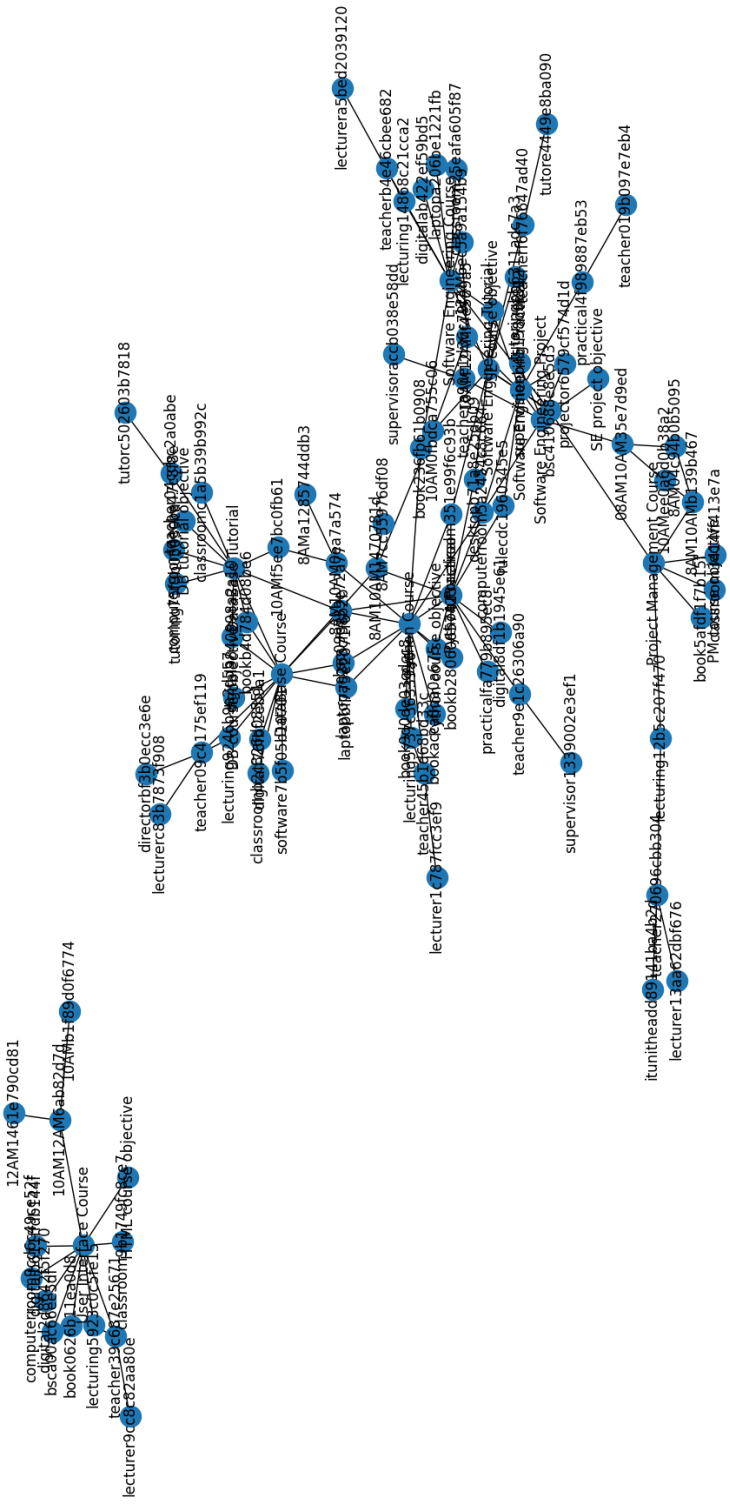


FIGURE 3 – Graphe de connaissance d'activités pivoté pour une meilleure lisibilité en colonne unique.

4.2.1 Pré-traitement

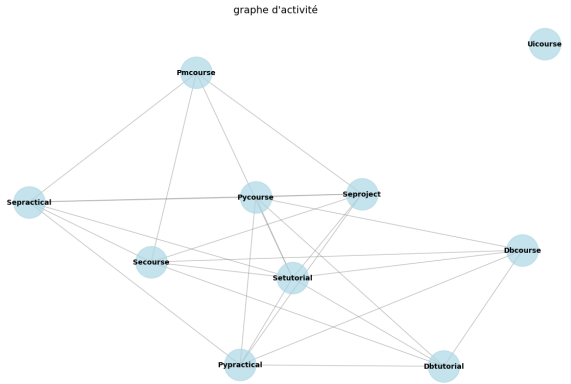


FIGURE 4 – Graphe de connaissance contenant uniquement les activités extraites du graphe de connaissances d'activités de la figure 3

Le graphe de connaissance simplifié d'activités de la figure 4 illustre les relations essentielles entre les différentes activités pédagogiques après la simplification du graphe de connaissances initial. Les dix nœuds représentent les activités principales ("Pmcourse", "Pycourse", "Setutorial", etc.), tandis que les arêtes indiquent les connexions significatives maintenues entre ces activités. La simplification conserve la structure fondamentale tout en réduisant la complexité, permettant d'identifier clairement les liens critiques entre les différentes composantes du système d'apprentissage. Cette représentation épurée facilite l'analyse des relations clé entre les activités pédagogiques et leur organisation en communautés distinctes.

4.2.2 Détection des communautés

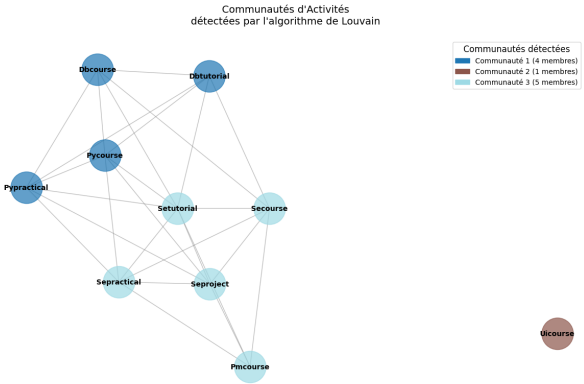


FIGURE 5 – Détection des communautés dans le graphe de connaissance d'activités simplifié de la figure 4

L'application de l'algorithme de Louvain sur le graphe de connaissance d'activités simplifié a permis d'identifier trois communautés distinctes, comme illustrées par la Figure 5. La première communauté comprend 4 membres (Dbcourse, Dbtutorial, Pypractical, Pycourse), représentant les activités liées aux bases de données et à la programmation Python. La deuxième communauté, composée d'un seul membre (Uicourse), suggère une activité indépendante d'interface utilisateur. La troisième communauté regroupe 5 membres (Pmcourse, Seproject, Sepractical, Secourse, Setutorial) centrés sur les activités de gestion de projet et de génie logiciel. Cette détection de communautés révèle une organisation naturelle des activités pédagogiques autour de domaines d'expertise spécifiques.

4.2.3 Post-traitement

La figure 6 correspond à la représentation du graphe de connaissance d'activité en tenant compte des communautés auxquelles appartient chaque activité. Elle permet aussi de voir les ressources partagées entre les activités de la même ou de communautés différentes.

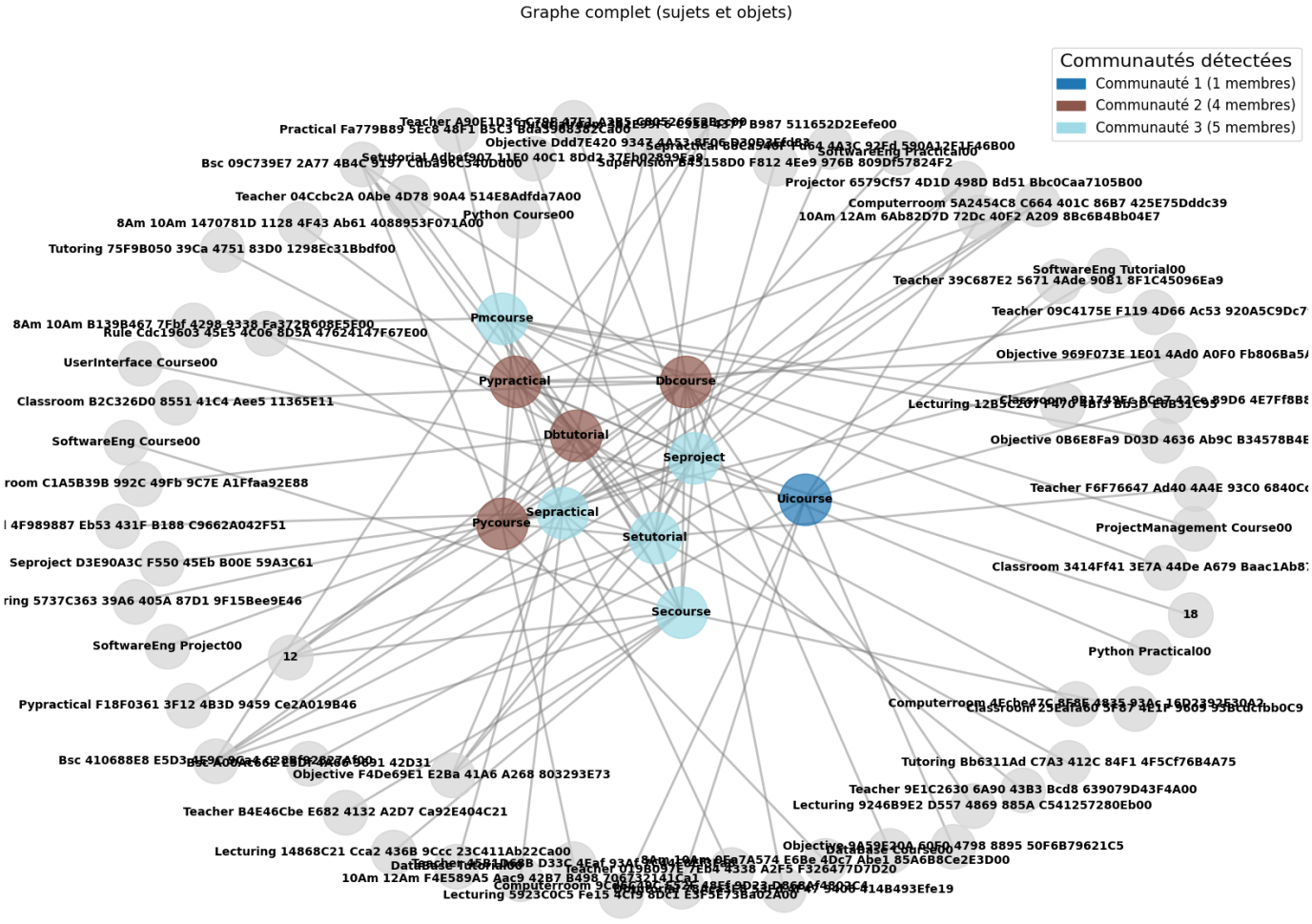


FIGURE 6 – Graphe de connaissance d'activités final obtenu en tenant compte du graphe d'activité simplifié indiquant les communautés de la Figure 5.

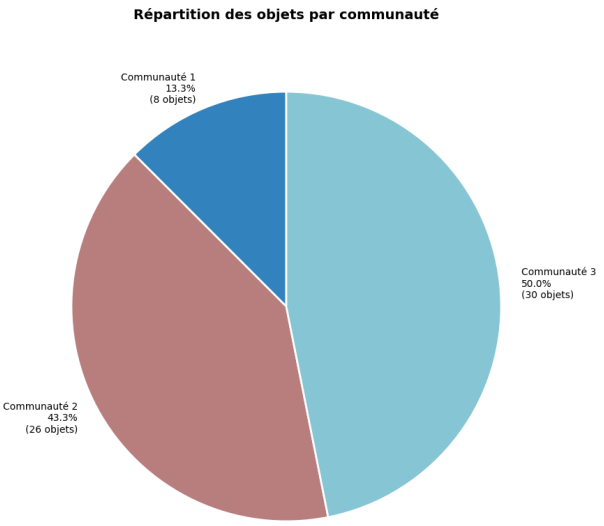


FIGURE 7 – Répartition des communautés détectées sous forme de dendrogramme dans un graphe éducatif.

Le diagramme circulaire de la Figure 7 présente la distribution des objets au sein de trois communautés distinctes. La communauté 3 domine avec 50% de la totalité des objets (30 objets), suivie de la communauté 2 représentant 43,3% (26 objets), tandis que la communauté 1 constitue la plus petite fraction avec 13,3% (8 objets). Cette répartition démontre une concentration significative des objets dans les communautés 2 et 3, qui regroupent ensemble plus de 90% des ressources, suggérant une distribution asymétrique mais potentiellement optimisée pour les besoins du système.

5 Conclusion

Dans cette étude, nous avons démontré l'importance de la détection de communautés dans les graphes de connaissances d'activités pour les organisations. Pour parvenir à la détection de communautés sur les graphes de connaissance d'activités, nous avons proposé une approche

constituée de trois (03) étapes : (1) un pré-traitement du graphe de connaissance d'activité qui permet de simplifier sa structure du graphe de connaissance d'activités initial tout en préservant une sémantique, (2) l'application de l'algorithme de Louvain pour détecter les communautés et (3) un post-traitement pour intégrer les communautés détectées dans le graphe de connaissance d'activités initial. Une application dans le domaine de l'éducation permet d'illustrer l'utilisation de notre approche et de montrer l'importance de la détection des communautés pour la compréhension des graphes de connaissance d'activités dans les organisations.

L'algorithme de Louvain s'est révélé efficace pour regrouper des activités similaires, offrant ainsi une visualisation claire des relations entre activités et facilitant l'analyse des informations pour des prises de décision dans une organisation. Dans notre cas d'étude dans le domaine de l'enseignement, les résultats obtenus soulignent la pertinence de cette méthode pour des applications concrètes.

Cette étude marque la première étape de l'abstraction et de la transformation d'un graphe de connaissance d'activités en une représentation plus digeste pour faciliter la prise de décision dans un domaine par les non-spécialistes. La suite de ce travail consiste d'une part à permettre la possibilité de spécifier la sémantique du type de communautés que l'on souhaite voir, de créer plusieurs niveaux de communautés, permettant de visualiser les communautés de communautés sur des graphes de connaissance d'activités plus complexes et d'autre part à permettre l'interrogation des connaissances inter et intra communautés, dans le but d'inférer des connaissances.

Références

[1] Majid Hameed Ahmed, Sabrina Tiun, Nazlia Omar, and Nor Samsiah Sani. Short text clustering algorithms, application and challenges : A survey. *Applied Sciences*, 13(1) :342, 2022.

[2] Vincent Blondel, Jean-Loup Guillaume, and Renaud Lambiotte. Fast unfolding of communities in large networks : 15 years later. *Journal of Statistical Mechanics : Theory and Experiment*, 2024(10) :10R001, 2024.

[3] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics : theory and experiment*, 2008(10) :P10008, 2008.

[4] Cécile Bothorel. *Analyse de Réseaux Sociaux : Détection et Qualification de communautés*. PhD thesis, Université de Bretagne Occidentale (UBO), 2023.

[5] CNAM. Cours sur les graphes et les communautés, 2024.

[6] Jean Delahousse. Graphe de connaissance, ontologie, et vocabulaires contrôlés, 2022. Consulté le 29 Août 2024.

[7] Hossein Hajibabaei, Vahid Seydi, and Abbas Koochari. A motif-based probabilistic approach for community detection in complex networks. *Journal of Intelligent Information Systems*, pages 1–19, 2024.

[8] Abdul Majeed and Ibtisam Rauf. Graph theory : A comprehensive survey about graph theory applications in computer science and social networks. *Inventions*, 5(1) :10, 2020.

[9] Cédric Pruski, Louis Deladiennée, Emmanuel Scolan, and Marcos Da Silveira. Une plateforme de management des connaissances pour le domaine des ressources spatiales. In *Extraction et Gestion des Connaissances : Actes de la conférence EGC'2023*, volume 39. BoD-Books on Demand, 2023.

[10] Naw Safrin Sattar and Shaikh Arifuzzaman. Scalable distributed louvain algorithm for community detection in large graphs. *The Journal of Supercomputing*, 78(7) :10275–10309, 2022.

[11] Serge Sonfack Sounchio, Thierry Coudert, Bernard Kamsu Foguem, Laurent Geneste, Cédrick Beler, and Sina Namakiaraghi. Organizations' interpersonal activity knowledge representation. In *HHAI 2023 : Augmenting Human Intellect*, pages 254–262. IOS Press, 2023.

[12] Serge Sonfack Sounchio, Laurent Geneste, Bernard Kamsu-Foguem, Cédrick Béler, Sina Namaki Araghi, and Muhammad Raza Naqvi. An enterprise architecture for interpersonal activity knowledge management. In *Iberoamerican Knowledge Graphs and Semantic Web Conference*, pages 66–81. Springer, 2023.

[13] Serge Sonfack Sounchio, Trawina Halguieta, Baudelaire Ismael Nguekeu Tankeu, Laurent Geneste, and Bernard Kamsu-Foguem. Interpersonal activity knowledge graph for organizations. *Article*, 2023.

[14] Serge Sonfack Sounchio, Halguieta Trawina, Baudelaire Ismael Tankeu Nguekeu, Laurent Geneste, and Bernard Kamsu-Foguem. Activity knowledge graph (akg).

[15] Jiajing Wu. Construct a knowledge graph for china coronavirus (covid-19) patient information tracking. *Risk Management and Healthcare Policy*, pages 4321–4337, 2021.

[16] Bibao Yao, Peijie Zhu, Zhuand Ma, Kun Gao, and Xuezao Ren. A constrained louvain algorithm with a novel modularity. *Applied Sciences*, 13(6) :4045, 2023.

[17] Silvia Zaoli, Giovanni Scaini, and Lorenzo Castelli. Community detection for air traffic networks and its application in strategic flight planning. *Sustainability*, 13(16) :8924, 2021.

Couplage d'Approches LLM et SLM pour le Déploiement de Solutions d'Extraction d'Entités Nommées

Samuel Kierszbaum[1], Nicolas Heulot[2]

[1]Airbus Protect
[2]IRT SystemX

samuel.kierszbaum@airbus.com

Résumé

La reconnaissance d'entités nommées (NER) requiert de gros volumes de données annotées qui sont complexes et coûteux à obtenir dans des domaines spécialisés. Cette étude compare deux approches adaptées à ce contexte : une méthode few-shot exploitant un grand modèle de langage (LLM) et une approche hybride combinant des annotations LLM avec un fine-tuning sur un petit modèle (SLM). Nos résultats confirment l'intérêt d'une approche hybride pour permettre de déployer à l'échelle des systèmes NER nécessitant très peu d'exemples annotés manuellement en entrée.

Mots-clés

TAL, NER, LLM.

Abstract

Named Entity Recognition (NER) effectiveness is often limited by the scarcity of annotated data, in particular in specialized domains. This study compares two NER approaches designed for low-resource settings : (1) a few-shot approach leveraging a large language model (LLM) for annotation and (2) a hybrid method that combines LLM-generated annotations with fine-tuning on a small language model (SLM). Our findings suggest that hybrid strategies can alleviate the challenges of manual annotation while maintaining high-quality entity recognition.

Keywords

NLP, NER, LLM.

1 Introduction

La reconnaissance des entités nommées (NER) est une tâche de traitement du langage naturel (TAL) qui classifie les entités nommées dans le texte, telles que des personnes, des organisations, des lieux, etc. Les travaux récents utilisent un modèle de langage tel que BERT (Bidirectional Encoder Representational Transformer) [7] affiné pour reconnaître des entités spécifiques à un domaine. Cependant, cet ajustement (*fine-tuning*) requiert d'avoir accès à un grand nombre d'exemples étiquetés pour donner de bons résultats. La collecte d'une grande quantité de données de haute qualité pour la tâche NER est très difficile et coûteuse, ce qui limite l'applicabilité de cette tâche.

Dans cet article, nous proposons une approche basée sur l'utilisation d'un LLM pour l'augmentation des données d'annotation, afin de réduire les coûts d'étiquetage manuel et de permettre l'adaptation rapide à des entités spécifiques d'un domaine. Les coûts d'utilisation d'un LLM à l'inférence étant très important, nous proposons de conserver une approche basée sur l'ajustement d'un modèle BERT, car cette dernière permet un passage à l'échelle en termes de nombres de documents traités. En utilisant le LLM comme pseudo-annotateur et l'ajustement d'un modèle BERT sur des annotations générées par le LLM, nous cherchons à évaluer si les performances peuvent égaler celle d'un apprentissage supervisé traditionnel à base de données annotées manuellement.

2 Travaux Antérieurs

La qualité de la reconnaissance d'entités nommées en utilisant de petits modèles de langage (SLMs) dépend fortement à la fois de la qualité des données annotées et de leur quantité. L'obtention de jeux de données avec une haute qualité des annotations demeure un défi majeur, en particulier dans les domaines spécialisés [20, 24].

Pour atténuer ces limitations, de nombreuses approches ont été explorées dans la littérature. L'une d'elles est l'amélioration itérative, qui vise à accroître l'efficacité de l'annotation en s'appuyant sur des méthodes avancées, des outils spécialisés et des techniques semi-automatisées [10, 12, 3]. Ces approches intègrent souvent des mécanismes d'interaction humaine, permettant d'affiner progressivement la qualité des données annotées. On notera certaines de ces approches sont rendues facilement accessibles et reutilisables via des frameworks comme FLAIR [2].

La supervision distante est une approche qui utilise des heuristiques basées sur des règles pour annoter automatiquement les données, réduisant ainsi le besoin d'annotation manuelle [15, 19]. Bien que cette méthode permette de générer efficacement des jeux de données annotés à grande échelle, son efficacité est souvent limitée par la précision et la capacité de généralisation des règles prédéfinies.

La reconnaissance d'entités nommées en apprentissage par faible nombre d'exemples (*few-shot* NER) est une alternative prometteuse car les grands modèles de langage ont démontré une robustesse face aux contextes à faibles

ressources avec des approches comme GliNER [25] par exemple, notamment en termes de disponibilité de données annotées, y compris dans les domaines spécialisés [14, 9].

Bien que l'apprentissage par faible nombre d'exemples (*few-shot learning*) ait déjà été appliqué directement aux tâches de reconnaissance d'entités nommées, son potentiel pour la génération de jeux de données annotés dans un cadre de supervision distante reste peu exploré. Des travaux récents, tels que [18, 21], exploitent les capacités de génération des grands modèles de langage afin de produire des jeux de données annotés à partir de zéro, en générant à la fois les phrases et leurs annotations correspondantes.

Dans notre étude, nous explorons une approche hybride en utilisant l'annotation basée sur les grands modèles de langage en apprentissage par faible nombre d'exemples (*few-shot*) comme étape intermédiaire avant le *fine-tuning*. Notre approche vise à générer des données annotées en exploitant des données existantes mais non annotées, ce qui la distingue des méthodes déjà explorées dans la littérature. Bien que ces dernières soient particulièrement prometteuses dans des contextes où les données disponibles sont limitées, notre méthode semble mieux adaptée aux situations où une quantité importante de données brutes est disponible, mais sans annotations.

3 Méthode

Afin d'évaluer l'impact de l'annotation automatique par les grands modèles de langage sur la reconnaissance d'entités nommées, nous comparons trois approches distinctes. Ces approches représentées dans la la figure 1 ont été choisies pour mesurer les bénéfices et les limites de l'utilisation des LLMs dans un contexte à faibles ressources.

Approche 1 : L'approche *few-shot* NER basée sur les LLM dans un contexte à faibles ressources, où le contexte à faibles ressources fait référence à la disponibilité limitée de données annotées manuellement (30 documents annotés).

Approche 2 : La seconde approche suit le paradigme classique du *fine-tuning* d'un petit modèle de langage sur des données annotées manuellement, ce qui constitue une référence pour comparer l'impact de l'annotation automatique utilisée pour les approches 1 et 3.

Approche 3 : L'approche hybride, également conçue pour un contexte à faibles ressources. Dans un premier temps, la méthode *few-shot* NER basée sur les LLM est utilisée pour générer des données annotées. Ces nouvelles données sont ensuite combinées avec les données annotées manuellement utilisées pour l'apprentissage *few-shot* afin de constituer un jeu de données, qui est ensuite utilisé pour le *fine-tuning* d'un SLM.

En comparant ces trois approches, nous cherchons à voir la viabilité de notre approche hybride pour la tâche de reconnaissance d'entités nommées dans des contextes spécialisés avec peu de données annotées manuellement.

Dans les sections suivantes, nous détaillons le protocole expérimental utilisé pour comparer ces trois approches. Nous commençons par une description du jeu de données, suivie d'un aperçu des procédures d'entraînement spécifiques à chaque approche, afin d'assurer une compréhension approfondie de leur mise en œuvre.

3.1 Jeu de Données

Nous considérons ici le jeu de données AeroBERT-NER [1][22]. Ce jeu de données est composé de 1 432 phrases en anglais issues du domaine de l'ingénierie des exigences en aérospatiale. Chaque phrase est annotée pour la reconnaissance d'entités nommées selon le schéma d'étiquetage BIO, avec des entités réparties en cinq catégories :

- SYS : systèmes et matériels
- VAL : valeurs numériques
- ORG : entreprises et organisations
- DATETIME : expressions de date et d'heure
- RES : ressources documentaires

Comme expliqué dans la section 3.2.2, nous nous concentrons exclusivement sur l'entité SYS pour diverses raisons. Le corpus comprend un nombre total de 1855 entités SYS. Parmi le corpus, 999 phrases contiennent au moins une mention de l'entité système. Les 433 phrases restantes sont conservées dans notre corpus afin de vérifier que nos approches ne créent pas de faux positifs.

Comme nous allons le voir dans les sections 3.3 et 3.4, nous procédons pour les approches utilisant un SLM à une classique validation croisée en 5 partitions. Les phrases sans mentions d'entités SYS sont réparties de manière homogènes parmi les 5 partitions.

Cette validation croisée permet une évaluation robuste, en agrégeant les performances sur les différentes partitions, donc sur l'ensemble du corpus à l'exception des 30 exemples annotés. Nous calculons la performance du modèle LLM de la même manière, sur l'ensemble du jeu de donnée à l'exception des 30 exemples. Les résultats et la méthode d'évaluation sont présentés avec plus de détails dans la section 3.5 le tableau 1.

3.2 Approche 1 - Few-Shot NER LLM

Dans cette section, nous décrivons le protocole expérimental de notre approche *few-shot* NER utilisant un grand modèle de langage dans un contexte à faibles ressources. La conception de cette approche repose sur plusieurs choix méthodologiques clés, chacun ayant des implications sur les performances du modèle, l'efficacité computationnelle et la facilité de mise en œuvre. Dans ce qui suit, nous présentons ces choix et en expliquons la justification.

3.2.1 Choix du Modèle

Pour notre étude, nous avons choisi GPT-4 [16], car il s'agit d'un des modèles les plus fréquemment étudiés dans la littérature et il démontre de manière constante des performances à l'état de l'art pour les tâches de reconnaissance d'entités nommées en apprentissage par faible nombre d'exemples.

D'autres modèles alternatifs existent et, bien que potentiellement moins performants, ils peuvent présenter d'autres avantages en termes de transparence, coût et flexibilité de

1. Il est disponible sur Hugging Face à l'adresse suivante : https ://huggingface.co/datasets/archanatikayatray/aeroBERT-NER.

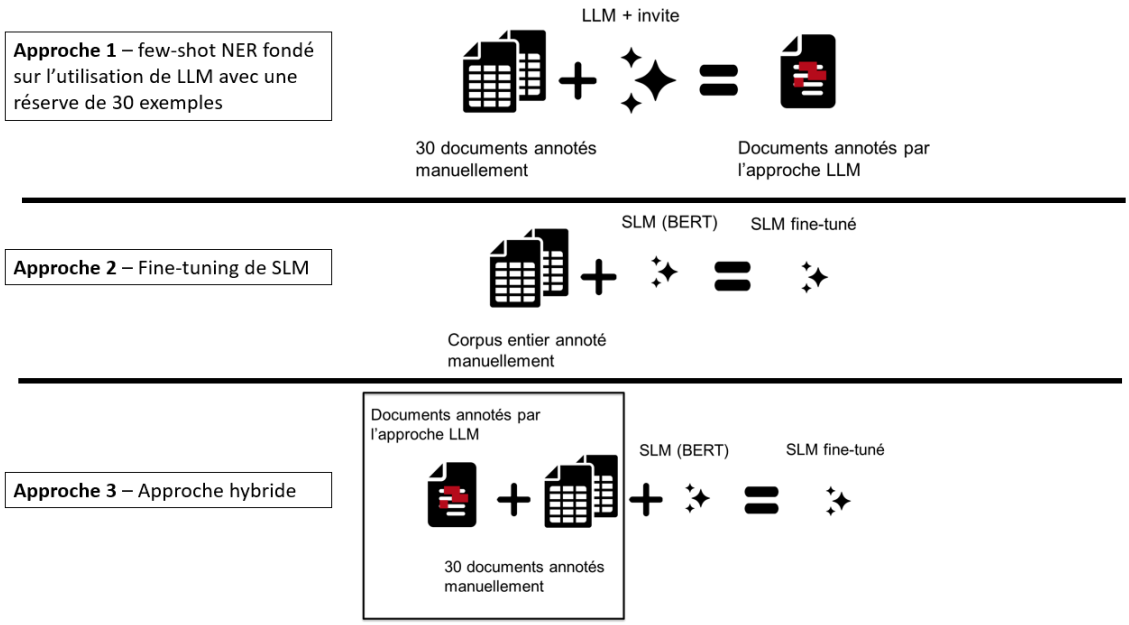


FIGURE 1 – Synthèse des trois approches comparées

déploiement. En particulier, contrairement à GPT-4, qui est un modèle propriétaire à boîte noire, des alternatives open source comme Mixtral [11] ou DeepSeek [6] offrent une meilleure transparence et peuvent être facilement affinés à des domaines spécifiques. De plus, ces modèles sont souvent moins coûteux, et certains peuvent être déployés localement, évitant ainsi les frais liés aux API et réduisant la dépendance aux fournisseurs externes.
Malgré ces considérations, notre choix de GPT-4 est motivé par ses performances supérieures. Étant donné le rôle critique des capacités du modèle dans l'apprentissage avec peu d'exemples, utiliser un modèle hautement performant permet d'établir une référence robuste pour la génération d'annotations afin d'évaluer à la fois notre approche de reconnaissance d'entités nommées en apprentissage par faible nombre d'exemples et l'approche hybride.

3.2.2 Extraction Multi-Entités vs. Mono-Entité

Une décision clé dans la conception d'un système de reconnaissance d'entités nommées basé sur les LLM concerne le choix entre extraire plusieurs entités simultanément ou une seule entité à la fois. Bien que la littérature ne fournisse pas de preuves solides en faveur d'une approche plutôt que l'autre, ces deux stratégies présentent des avantages et des compromis distincts.
L'extraction multi-entités permet d'effectuer une seule inférence par phrase, mais elle complexifie la sélection des exemples, car il est nécessaire d'assurer une représentation équilibrée de toutes les catégories d'entités dans le prompt. L'extraction mono-entité, en revanche, simplifie la structure du prompt et la sélection des exemples, mais elle nécessite plusieurs inférences par phrase (une par type d'entité) ainsi qu'un post-traitement pour consolider les résultats.
Nous avons opté pour l'approche mono-entité, qui présente plusieurs avantages dans notre contexte. Elle réduit la longueur de la section de définition des entités dans le prompt, ce qui diminue la consommation de tokens et les coûts induits. Elle simplifie la sélection des exemples, puisque chaque prompt ne traite qu'un seul type d'entité. Afin d'éviter le post-traitement nécessaire pour fusionner les résultats des différentes entités, nous avons restreint notre étude à une seule catégorie d'entités. Nous avons choisi de nous concentrer sur l'entité SYS, car elle est la plus spécifique au domaine. En mettant l'accent sur les entités SYS, nous nous assurons que nos approches sont évaluées sur l'aspect le plus complexe et spécialisé du jeu de données, là où les modèles généralistes sont les plus susceptibles d'échouer.

3.2.3 Structure du Prompt

Nous proposons d'utiliser un prompt suivant une structure couramment utilisée pour la reconnaissance d'entités nommées en *few-shot learning* avec les LLM, s'inspirant de travaux antérieurs tels que [23, 9].
Le prompt est composé des éléments suivants :
Rôle du système : Définit le modèle comme un assistant spécialisé en NER.
Lignes directrices : Spécifie les règles à suivre pour assurer une annotation cohérente des entités.
Exemples *few-shot* : Illustre la manière dont les entités doivent être annotées.
Phrase test : Contient la phrase à analyser pour l'inférence.

Le contenu des sections Rôle du système et Lignes directrices a été affiné par essais et erreurs à l'aide de LLM plus petits et moins coûteux, afin d'optimiser la clarté et la cohérence de la reconnaissance des entités.

Un prompt complet est présenté ci-dessous :

```
System: You are an AI-assistant tasked with
    identifying and annotating hardware terms
    related to spacecraft, satellites, and
    aeronautical systems in a given text. These
    terms should be enclosed within double "at"
    symbols (@@) at the beginning and double hash
     symbols (##) at the end.

**Guidelines**:
1. Identify Hardware Systems and Components:
Focus on highlighting terms related to physical
    spacecraft, satellites, aeronautical systems,
     or their components. This includes specific
    names (e.g., "CubeSats," "Deep Space Network
    ") and technical terms (e.g., "landing gear,"
     "fuel system," "airplane," "satellite").
    Only hardware systems and their components
    should be annotated.

2. Annotation Format:
Enclose each identified hardware system or
    component within @@ and ##. For example, the
    term "CubeSats" should be annotated as
    @@CubeSats##.

3. Consistency:
Ensure that all instances of similar terms are
    annotated consistently throughout the
    document. For example, if "fuel system" is
    annotated in one sentence, ensure that all
    instances of "fuel system" are similarly
    annotated.

4. Annotation of Compound Terms:
Annotate only the aerospace-related hardware
    systems or components within a compound term.
     Do not enclose the entire phrase unless it
    consists solely of technical elements.
For example, in the phrase "turbine engine
    powered airplane," only @@turbine engine##
    and @@airplane## should be annotated, as "
    powered" is not part of the system or
    component.

5. Contextual Understanding:
Annotate based on relevance to aerospace hardware
     technology. Terms not directly related to
    hardware aerospace systems, components, or
    technologies should not be annotated.

6. Avoid Over-annotation:
Do not annotate people, titles, regulations, or
    operational terms:
Do not annotate roles, personnel, or
    organizational titles, even if they are part
    of aerospace organizations (e.g., "
    administrator," "NASA Administrator," "NASA,"
     "NSF," "Planet Labs").
Avoid annotating regulations, legal references,
    or general phrases (e.g., "requirements of
    part 34," "Sections 25-1181").
Avoid annotating general operational terms like "
    flight," "landing," "takeoff," "stall," "
    surge," "flameout," and "navigation" unless
    they are directly part of a specific
    aerospace hardware technology.
Finally, avoid general scientific terms (e.g., "
    snow," "science," "jamming," "ice accretion")
    .
These are not related to aerospace hardware
    systems and should be left unannotated.

Human: Failure of structural elements of the drag
     limiting systems need not be considered if
    the probability of this kind of failure is
    extremely remote.
AI: Failure of structural elements of the @@drag
    limiting systems## need not be considered if
    the probability of this kind of failure is
    extremely remote.
Human: It must be shown by analysis or test, or
    both, that each operable reverser can be
    restored to the forward thrust position.
AI: It must be shown by analysis or test, or both
    , that each operable @@reverser## can be
    restored to the forward thrust position.
Human: This is an example of input sentence with
    2 example before.
```

3.2.4 Sélection du Réservoir d'Exemples Disponibles

Dans cette étude, nous imposons une contrainte sur le nombre d'exemples annotés pouvant être utilisés pour constituer le réservoir à partir duquel les exemples du prompt seront sélectionnés. Cette contrainte est motivée par l'hypothèse que nous opérons dans un contexte à faibles ressources, où l'accès aux données annotées spécifiques au domaine est limité. Par conséquent, nous limitons ce réservoir à 30 exemples annotés.

À notre connaissance, peu de travaux antérieurs se sont intéressés à la stratégie optimale de sélection des exemples pour constituer ce réservoir dans le cadre du *few-shot* NER basé sur les LLM. Bien que la recherche ait largement exploré la sélection des exemples à inclure dans le prompt, les critères déterminant quels exemples doivent être annotés et ajoutés au réservoir restent encore peu étudiés.

3.2.5 Mécanisme de Sélection des Exemples

Le mécanisme de sélection des exemples à inclure dans le prompt joue un rôle crucial dans les performances du modèle, en particulier dans notre contexte à faibles ressources. En effet, la littérature indique qu'une sélection d'exemples de haute qualité à partir d'un petit réservoir de 30 instances surpasse une approche de prélèvement aléatoire dans un ensemble de données bien plus vaste [1].

Ces résultats renforcent l'hypothèse selon laquelle, dans les contextes à faibles ressources, une curation minutieuse des exemples et une structuration rigoureuse du prompt peuvent compenser les limitations imposées par la taille réduite du jeu de données. En sélectionnant stratégiquement les exemples, il est ainsi possible d'optimiser les performances de l'apprentissage *few-shot*, même lorsque la disponibilité des données est fortement restreinte.

Il est également souligné dans [14] que l'efficacité des grands modèles de langage dépend fortement de la sélection rigoureuse des exemples utilisés pour l'inférence. De plus, la stratégie de sélection optimale est spécifique à chaque jeu de données : différentes approches donnent des résultats variables en fonction des caractéristiques du corpus.

Parmi les diverses stratégies de sélection d'exemples décrites dans la littérature, on distingue :

Le prompting statique : approche la plus simple, où un même ensemble d'exemples est utilisé systématiquement pour toutes les entrées.

La sélection aléatoire : qui introduit de la variabilité à chaque étape d'inférence.
Les méthodes basées sur la similitude : où les exemples sont sélectionnés en fonction de leur pertinence par rapport à la phrase d'entrée.

D'autres approches plus avancées ont également été proposées, telles que le guidage de la sélection des exemples avec un score de complexité [1], ou la sélection des exemples les plus pertinents grâce à des *embeddings* au niveau des entités [23]. Mais ces techniques requièrent un volume de données annotées important en contradiction avec notre contexte à faibles ressources.
Dans notre étude, nous avons choisi d'utiliser la mesure de similitude la plus courante dans la littérature : les *embeddings* de phrases avec calcul par cosinus de similitude. Plus précisément, nous utilisons le modèle all-MiniLM-L6-v2 [2] de sentence transformers [17], comme dans [1], pour le calcul du score de similitude entre phrases.
Bien que cette méthode offre un mécanisme de récupération simple et efficace sur le plan computationnel, cette approche est limitée par la similitude au niveau des phrases qui ne correspond pas toujours à la pertinence au niveau des entités nommées [23]. Par exemple, on peut avoir une similarité sémantique importante entre une phrase dont le sujet correspond à une entité nommée et la même phrase dont le sujet est un pronom faisant référence à cette entité. L'utilisation de cette deuxième phrase comme exemple est peu pertinente pour l'annotation des entités. Dans notre étude, cette limitation ne nous affecte pas car la réserve d'exemples disponibles ne contient que des phrases comportant au moins une occurrence de l'entité cible (SYS).

3.2.6 Format de Sortie

Le format de sortie en *in-context learning (ICL)* influence également le comportement du modèle. Différentes stratégies de formatage existent :
Le format utilisé dans [13] structure les sorties sous forme de dictionnaires (ex. : 'Chemical' : ['apomorphine'], 'Disease' : ['hypothermia']). Toutefois, cette approche ne fournit pas la position des entités dans la phrase, ce qui peut poser des problèmes d'extraction lorsque plusieurs mentions similaires apparaissent.
Une alternative est l'annotation en format BIO, utilisée dans [1], où chaque token est étiqueté comme Beginning (B), Inside (I) ou Outside (O) d'une entité.
D'autres formats, tels que le BMES tagging ou l'extraction basée sur la position des entités, ont été testés mais n'offrent pas de résultats supérieurs [23].
Alternativement, des symboles spéciaux peuvent être utilisés pour mettre en évidence les entités au sein du texte (par exemple, "Carcinoma @@ductal de mama derecha"), une méthode employée avec succès dans plusieurs travaux [14, 23, 9].
Le Chain-of-Thought (CoT) prompting, exploré dans [4], constitue une autre approche visant à améliorer l'interprétabilité. Cependant, l'intégration du raisonnement CoT introduit une charge supplémentaire pour les équipes d'annotation manuelle et complique la mise en œuvre.
Dans notre approche, nous avons privilégié la simplicité et l'efficacité en adoptant le format à base de symboles spéciaux [23, 9], pour plusieurs raisons.
Ce format permet de distinguer efficacement les différentes occurrences de l'entité cible sans accroître la complexité de l'annotation. Il offre l'avantage d'une conversion simple et réversible avec le format BIO, qui constitue à la fois le format original du corpus et celui requis pour le SLM dans l'approche hybride. Pour passer du format @@... au format BIO, il suffit de repérer les tokens encadrés par les balises spéciales : le premier token précédé de @@ reçoit l'étiquette B-SYS, les suivants jusqu'à sont étiquetés I-SYS, et tous les autres tokens sont marqués O. Inversement, pour convertir un corpus BIO en format à balises, on détecte les séquences d'étiquettes B-SYS/I-SYS consécutives et on encadre les tokens correspondants avec @@ au début et à la fin. Ce choix octroie une flexibilité appréciable dans les chaînes de traitement.

3.3 Approche 2 - Fine-tuning

L'approche traditionnelle de *fine-tuning* suit la méthodologie standard d'entraînement des modèles basés sur les transformeurs pour les tâches de reconnaissance d'entités nommées. Plus précisément, nous affinons un modèle BERT, à l'instar de l'étude [22], en utilisant des données annotées manuellement issues du domaine des exigences aérospatiales.
Afin d'assurer une évaluation robuste, nous adoptons une validation croisée en 5 partitions (5-fold cross-validation). À chaque itération, l'ensemble de données est divisé en cinq sous-ensembles, dont quatre sont utilisés pour l'entraînement et un pour la validation. Il est important de noter que les 30 instances sélectionnées pour l'approche NER en apprentissage par faible nombre d'exemples sont systématiquement incluses dans l'ensemble d'entraînement, garantissant ainsi leur présence dans les cinq partitions.
Nous explorons le même espace d'hyperparamètres que celui recommandé dans l'article original sur BERT pour le *fine-tuning* [8] (*Batch size* : 16, 32, *Learning rate* : 2e-5, 3e-5, 5e-5) Le modèle est entraîné sur 5 époques, avec des évaluations périodiques pour suivre ses performances. Plus précisément, la performance du modèle est évaluée à chaque tiers d'époque sur l'ensemble de validation. À l'issue de l'entraînement, nous conservons la meilleure performance obtenue au cours de ces évaluations. Dans la suite, les hyperparamètres ayant permis d'obtenir les meilleures performances l'affinage des différentes approches sont : *batch size* : 16 et *learning rate* : 5e-5.

3.4 Approche 3 - Hybride LLM SLM

L'approche hybride combine la méthode NER en apprentissage par faible nombre d'exemples basée sur un modèle de langage à grande échelle avec le paradigme de *fine-tuning*. Plutôt que de s'appuyer uniquement sur des données annotées manuellement pour le *fine-tuning*, nous utilisons la

2. le modèle est disponible ici : https ://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

TABLE 1 – Évaluation des résultats pour les différentes approches

Approche	F1 Score	Exact	Partial	Missing	Spurious
fine-tuning	**0.90**	**1563**	179	**113**	**238**
LLM	0.83	1362	244	249	242
Hybride	0.85	1408	**256**	191	332

méthode NER basée sur un LLM en *few-shot* pour générer des instances annotées, qui servent ensuite de données d'entraînement pour le processus d'affinage.

Plus précisément, nous reproduisons l'ensemble de données utilisé dans l'approche d'affinage avec une validation croisée en 5 partitions, mais avec une modification essentielle : à l'exception des 30 instances annotées manuellement, qui sont conservées dans les cinq jeux d'entraînement, chaque partition d'entraînement est constituée d'instances annotées par le LLM plutôt que par des annotateurs humains. Hormis cette substitution, la procédure de *fine-tuning* reste inchangée, en maintenant la même architecture de modèle et la même procédure d'exploration et de sélection d'hyperparamètres.

3.5 Évaluation

Pour évaluer la performance, nous utilisons la bibliothèque Python nervaluate [5], qui permet une analyse détaillée des performances en reconnaissance d'entités nommées. Plus précisément, nous rapportons les métriques suivantes :
Score F1 : Moyenne harmonique de la précision et du rappel, fournissant une mesure globale de la capacité du modèle à identifier correctement les entités tout en minimisant les faux positifs et les faux négatifs.
Nombre d'entités manquantes (Missing Count) : Nombre de mentions d'entités présentes dans la vérité terrain mais non prédites par le modèle, mettant en évidence les erreurs liées au rappel.
Nombre d'entités superflues (Spurious Count) : Nombre de prédictions d'entités ne correspondant à aucune entité dans la vérité terrain, reflétant les faux positifs.
Nombre de correspondances exactes (Exact Match Count) : Nombre d'entités prédites correspondant exactement aux annotations de la vérité terrain, à la fois en termes de frontières et d'étiquettes. Cette métrique stricte évalue la capacité du modèle à identifier précisément les entités sans aucune déviation.
Nombre de correspondances partielles (Partial Match Count) : Nombre de cas où une entité prédite chevauche partiellement une entité de la vérité terrain sans correspondance exacte. Cette métrique tient compte des prédictions approximativement correctes, offrant une évaluation plus nuancée que les seules correspondances exactes.

4 Résultats

Nous observons que l'approche traditionnelle de *fine-tuning* sur des données manuellement annotées surpasse les deux autres approches, tandis que l'approche hybride dépasse légèrement la méthode basée uniquement sur le LLM (voir Table 1).

Nos résultats indiquent que l'approche hybride se rapproche des performances de la méthode traditionnelle. Cela permet ainsi de réduire considérablement l'effort d'annotation qui est particulièrement précieuse dans des contextes à faibles ressources, où l'annotation manuelle est coûteuse et chronophage. Les performances des modèles BERT fournissent une indication indirecte de la qualité de l'annotation, les ensembles de données bien annotés contribuant généralement à de meilleures performances des modèles. Cela suggère que les LLMs peuvent être exploités pour accélérer le processus d'annotation. En automatisant ou en augmentant le processus d'annotation, les LLMs ont le potentiel de réduire la dépendance à un effort humain intensif tout en maintenant une annotation de haute qualité, accélérant ainsi le développement de jeux de données annotés dans des environnements à ressources limitées.

Bien que l'approche hybride surpasse la méthode basée sur un LLM dans notre cas, l'écart de performance reste relativement faible. Toutefois, le choix entre ces deux approches ne doit pas se limiter à des critères de performances. D'autres facteurs peuvent être pris en compte.

Premièrement, l'utilisation des LLMs introduit un risque d'hallucination, un problème absent dans la phase d'inférence fondé sur le modèle BERT de la méthode hybride. Un autre aspect critique est le stockage et le coût computationnel : alors que les méthodes d'annotation basées sur les LLMs nécessitent souvent des modèles de grande envergure, exigeant d'importantes ressources de stockage et de mémoire, l'approche hybride exploite les LLMs uniquement pour générer l'ensemble d'entraînement. Par la suite, seul BERT ou un autre modèle de langage de taille réduite est utilisé, diminuant ainsi les besoins computationnels et de stockage à long terme. Cette efficacité est particulièrement pertinente dans les environnements où les contraintes de calcul et de stockage peuvent affecter la faisabilité du déploiement d'une solution de reconnaissance d'entités nommées.

Une autre distinction clé réside dans le compromis entre précision et rappel, comme le reflètent les métriques de nombre d'entités manquantes (Missing Count) et d'entités superflues (Spurious Count). L'approche hybride identifie un plus grand nombre d'entités mais génère également un nombre plus élevé de prédictions superflues, ce qui suggère une inclination vers le rappel au détriment de la précision. À l'inverse, l'approche basée sur les LLMs se montre plus conservatrice.

Le choix entre ces méthodes doit donc être guidé par les exigences opérationnelles spécifiques, en fonction de la

priorité accordée soit à un rappel plus élevé, soit à la réduction des faux positifs. Par ailleurs, les coûts computationnels et les contraintes de stockage doivent également être pris en compte, car ces facteurs peuvent influencer de manière significative la faisabilité de chaque approche selon l'environnement de travail.

4.1 Gestion des Hallucinations

Les sorties générées par le LLM peuvent contenir des hallucinations [26] car le modèle génère des réponses qui ne correspondent pas à l'intention de l'utilisateur. Dans notre cas, les instances d'hallucination sont facilement repérables en comparant la sortie du LLM (un document annoté, auquel on soustrait les annotations pour notre analyse) à son entrée (le document à annoter). Ainsi, nous avons identifié quatre cas d'hallucinations, affectant soit la ponctuation, soit la casse des caractères. Nous les reproduisons ci-dessous :

- **Modification de la casse** :
  - **Entrée 1** : **cubesat** attitude determination techniques have significantly advanced in the past decade , with many of the techniques found on larger spacecraft now also available on CubeSats .
  - **Sortie 1** : **CubeSat** attitude determination techniques have significantly advanced in the past decade , with many of the techniques found on larger spacecraft now also available on CubeSats .
  - **Entrée 2** : **cubesat** instrument builders are also reimagining their instruments based on commercial off-the-shelf (cots) parts .
  - **Sortie 2** : **CubeSat** instrument builders are also reimagining their instruments based on commercial off-the-shelf (cots) parts .

Explication : Le LLM a modifié la casse de "cubesat" en "CubeSat". On constate que cette erreur peut s'expliquer par l'inconsistance au niveau de la capitalisation de ce terme dans le corpus, particulièrement visible dans l'Entrée 1, où l'on trouve les deux manières d'écrire le terme dans la même phrase.

- **Ponctuation** :
  - **Entrée 1** : each fuel storage system must be designed to prevent significant loss of stored fuel from any vent system due to fuel transfer between fuel storage or supply **systems , or** under likely operating conditions .
  - **Sortie 1** : each fuel storage system must be designed to prevent significant loss of stored fuel from any vent system due to fuel transfer between fuel storage or supply **systems, or** under likely operating conditions .
  - **Entrée 2** : the exhaust system , including exhaust heat exchangers for each powerplant or auxiliary power **unit , must** provide a means to safely discharge potential harmful material .
  - **Sortie 2** : the exhaust system, including exhaust heat exchangers for each powerplant or auxiliary power **unit, must** provide a means to safely discharge potential harmful material .

Explication : Suppression d'espace superflu après la virgule, probablement due à un ajustement automatique du modèle pour respecter la convention typographique standard assimilée lors de son apprentissage de la modélisation du langage.

Ces modifications sont mineures et n'affectent que 0,4% du jeu de données. Nous avons décidé de ne pas les prendre en compte. Ainsi, le *fine-tuning* de l'approche hybride a été réalisé sur l'ensemble des données annotées produites par le LLM, y compris celles affectées par des hallucinations.

4.2 Limitations

Pour des raisons de praticité, notre étude s'est concentrée exclusivement sur un seul type d'entité dans un seul ensemble de données. Étendre l'analyse à plusieurs types d'entités et à divers ensembles de données permettrait d'évaluer plus largement les approches proposées et leur capacité de généralisation. Également, il aurait été intéressant d'explorer comment le choix du LLM impacte la performance de notre approche hybride, par exemple en utilisant des modèles comme GPT-3.5 ou un modèle open-source type Mistral. De plus, notre implémentation de l'approche basée sur les LLMs n'a pas exploré de manière exhaustive toutes les techniques d'optimisation possibles, notamment l'auto-vérification (*self-verification*), comme démontré dans [23]. Cette technique introduit une étape d'inférence supplémentaire où le modèle réévalue et affine ses prédictions, renforçant ainsi sa robustesse et sa fiabilité. L'intégration de telles améliorations pourrait encore optimiser les performances de l'approche basée sur les LLMs et réduire les erreurs potentielles.

5 Conclusion

Dans cet article, nous avons comparé une approche classique à une nouvelle approche hybride utilisant un LLM pour une tâche de reconnaissance d'entités nommées. Notre approche hybride se base sur un LLM pour l'augmentation des données d'annotation en partant d'un faible nombre d'exemples, puis l'utilisation de ces données pour affiner un SLM dans le but d'éviter les coûts prohibitifs d'utilisation d'un LLM à l'inférence. Nous avons comparé ces approches sur un jeu de données d'exigences en aérospatiale. Nos résultats, même si ils restent préliminaires, indiquent que l'approche hybride se rapproche des performances d'une approche classique avec toutes des données annotées manuellement (F1 > 0.80). Ainsi l'intégration des LLMs dans le processus d'annotation peut aider à réduire les coûts de déploiement à l'échelle de solutions d'extraction d'entités nommées sur des domaines spécifiques.

5.1 Travaux futurs

Nos résultats suggèrent que l'intégration des LLMs dans le processus d'annotation offre des avantages significatifs. Étant donné la complexité et le coût associés à la construction de corpus annotés de haute qualité, nous estimons qu'une exploration approfondie de cette approche est justifiée. Plus précisément, l'exploitation des LLMs pourrait rationaliser le flux de travail d'annotation en facilitant l'affinement itératif des consignes d'annotation et en assurant une cohérence accrue entre les ensembles de données.
En outre, les LLMs pourraient être utilisés pour appliquer rétroactivement des consignes mises à jour à des corpus précédemment annotés, garantissant ainsi leur alignement avec l'évolution des standards d'annotation. Il serait intéressant d'approfondir ces perspectives afin d'optimiser et d'élargir le rôle des LLMs dans l'annotation des données.

6 Remerciements

Ce travail a obtenu le soutien du gouvernement français dans le cadre du programme "France 2030", au sein de l'Institut de Recherche Technologique SystemX.

Références

[1] Rishabh Adiga, Lakshminarayanan Subramanian, and Varun Chandrasekaran. Designing informative metrics for few-shot example selection, 2024.

[2] Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. FLAIR : An easy-to-use framework for state-of-the-art NLP. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, 2019.

[3] Sachi Angle, Pruthwik Mishra, and Dipti Mishra Sharma. Automated error correction and validation for POS tagging of Hindi. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, Hong Kong, 1–3 December 2018. Association for Computational Linguistics.

[4] Dhananjay Ashok and Zachary C. Lipton. Promptner : Prompting for named entity recognition, 2023.

[5] David Batista and Matthew Antony Upson. nervaluate, October 2020.

[6] DeepSeek-AI. Deepseek-v3 technical report, 2024.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics : human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert : Pre-training of deep bidirectional transformers for language understanding, 2019.

[9] Á. García-Barragán, A. González Calatayud, O. Solarte-Pabón, et al. Gpt for medical entity recognition in spanish. *Multimedia Tools and Applications*, 2024.

[10] Nancy Ide, Christian Chiarcos, Manfred Stede, and Steve Cassidy. *Designing Annotation Schemes : From Model to Representation*, pages 73–111. Springer Netherlands, Dordrecht, 2017.

[11] Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of experts, 2024.

[12] Jan-Christoph Klie, Bonnie Lynn Webber, and Iryna Gurevych. Annotation error detection : Analyzing the past and present for a more coherent future, 2022. arXiv preprint.

[13] Mingchen Li, Yang Ye, Jeremy Yeung, Huixue Zhou, Huaiyuan Chu, and Rui Zhang. W-procer : Weighted prototypical contrastive learning for medical few-shot named entity recognition, 2023.

[14] Mingchen Li and Rui Zhang. How far is language model from 100

[15] Shifeng Liu, Yifang Sun, Bing Li, Wei Wang, and Xiang Zhao. Hamner : Headword amplified multi-span distantly supervised method for domain specific named entity recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05) :8401–8408, Apr. 2020.

[16] OpenAI. Gpt-4 technical report, 2024.

[17] Nils Reimers and Iryna Gurevych. Sentence-bert : Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.

[18] Joan Santoso, Patrick Sutanto, Billy Cahyadi, and Esther Setiawan. Pushing the limits of low-resource NER using LLM artificial data generation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics : ACL 2024*, pages 9652–9667, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[19] Jingbo Shang, Liyuan Liu, Xiang Ren, Xiaotao Gu, Teng Ren, and Jiawei Han. Learning named entity tagger using domain-specific dictionary, 2018. arXiv preprint.

[20] Amber C Stubbs. *A methodology for using professional knowledge in corpus annotation*. Brandeis University, 2013.

[21] Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. Does synthetic data generation of llms help clinical text mining ?, 2023.

[22] Archana Tikayat Ray, Olivia J Pinon-Fischer, Dimitri N Mavris, Ryan T White, and Bjorn F Cole. aerobert-ner : Named-entity recognition for aerospace requirements engineering using bert. In *AIAA SCITECH 2023 Forum*, page 2583, 2023.

[23] Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. Gpt-ner : Named entity recognition via large language models, 2023.

[24] Qiang Wei, Amy Franklin, Trevor Cohen, and Hua Xu. Clinical text annotation–what factors are associated with the cost of time ? In *AMIA Annual Symposium Proceedings*, volume 2018, page 1552. American Medical Informatics Association, 2018.

[25] Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. GLiNER : Generalist model for named entity recognition using bidirectional transformer. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (Volume 1 : Long Papers)*, pages 5364–5376, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[26] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. Siren's song in the ai ocean : A survey on hallucination in large language models, 2023.

Technologies d'assistance pour les personnes malvoyantes basées sur la vision: avancées, limites et perspectives

Aela Le Sommer, Panagiotis Papadakis, Christophe Lohr

IMT Atlantique, Lab-STICC, UMR CNRS 6285, team RAMBO, F-29238 Brest, France

Résumé

Bien que les méthodes actuelles d'image captioning, de video summarization et de Visual Question-Answering soient des solutions prometteuses pour aider les personnes malvoyantes et aveugles dans leur vie quotidienne, elles se heurtent à de nombreux problèmes (biais, hallucinations, pertinence des descriptions et réponses générées, etc.). Cet article tente donc de dresser un aperçu de l'état de l'art des technologies basées sur la vision et le langage pour les personnes malvoyantes, en identifiant les principales problématiques et en proposant des pistes d'exploration.

Mots-clés

Malvoyants, Visual Question-Answering, Image Captioning, Technologies d'assistance, Vision par ordinateur

Abstract

Although current methods of image captioning, video summarization and Visual Question-Answering are promising solutions to help visually impaired and blind people in their daily lives, they face many problems (bias, hallucinations, relevance of the descriptions and answers generated, etc.). This article therefore attempts to provide an overview of the state-of-the-art of vision-language-based technologies for visually impaired people by identifying the main issues and proposing exploration avenues.

Keywords

Visually impaired, Visual Question-Answering, Image Captioning, Assistive technologies, Computer vision

1 Introduction

Le handicap peut prendre différentes formes. Il peut être moteur, sensoriel, cognitif ou bien psychique. Quelles que soient ses formes, il demande des aménagements pour garantir l'égalité des chances et l'inclusion de chacun. Dans leur vie quotidienne, les personnes aveugles et malvoyantes sont confrontées à de nombreux défis. En effet, elles rencontrent des difficultés considérables lorsqu'il s'agit de se déplacer dans des environnements intérieurs et extérieurs inconnus. Il leur est souvent difficile de comprendre l'environnement qui les entoure, surtout dans des environnements extérieurs qui sont en constante évolution. Même évoluant dans un environnement familier, ces personnes font face à des difficultés, notamment lorsqu'il s'agit de localiser un objet ou d'évaluer des risques et des dangers potentiels.

Les technologies d'aide au déplacement pour les personnes malvoyantes couvrent un large éventail de solutions, allant des cannes intelligentes aux dispositifs de guidage par GPS, en passant par les assistants vocaux et les capteurs ultrasons pour détecter les obstacles. L'article [26] dresse un large panorama de ces innovations. Les progrès récents en intelligence artificielle (IA) ont permis le développement de technologies basées sur la vision par ordinateur, capables d'analyser l'environnement et de fournir des retours en temps réel. Ces technologies offrent une perception plus riche et contextuelle de l'environnement, allant au-delà de la détection d'obstacles. Dans ce but, de nombreuses technologies basées sur l'image captioning (IC), la video summarization (VS) et le Visual-Question Answering (VQA) couplées avec des technologies de text-to-speech (TTS) ont vu le jour pour aider les malvoyants à se déplacer de manière plus confiante dans l'environnement qui les entoure (c.f. Sec. 2).

Les tâches d'IC, de VS et de VQA ont respectivement pour but de décrire le contenu d'une image avec du texte, de générer un court résumé du contenu d'une vidéo, et de répondre à des questions basées sur une image. La tâche de TTS, quant à elle, permet de convertir un texte en paroles, ce qui est particulièrement utile, par exemple, pour adapter les approches basées sur l'IC aux besoins des malvoyants.

Initialement, ces trois tâches étaient traitées avec des méthodes basées sur des règles et des approches statistiques. L'essor du deep learning a transformé ces approches en intégrant des architectures neuronales, notamment des réseaux de neurones convolutionels et récurrents. Avec l'arrivée des Transformers, de nouvelles approches ont vu le jour, permettant d'améliorer la compréhension sémantique, la précision et la richesse des descriptions en combinant efficacement vision et langage. Enfin, suite à l'avènement des grands modèles de langage (LLM), de nombreux modèles initialement orientés uniquement sur le texte sont progressivement devenus multimodaux en intégrant le traitement de nouvelles modalités comme les fichiers pdf, les images, les vidéos, etc. Des modèles de vision-language (VLM) sont alors apparus. Ces modèles composés généralement d'un encodeur de vision, d'un module de connexion de modalité et d'un LLM sont capables d'apprendre de manière simultanée à partir d'images et de textes.

L'évolution récente rapide des techniques de descriptions d'images ou de vidéos et des techniques qui permettent aux utilisateurs de poser des questions sur des images ou des vidéos offre des perspectives prometteuses pour aider les

App. Mobile	Reconnaissance d'objets	Lecture de textes	Description de scène	VQA	Reconnaissance faciale	Nb Téléchargements
Be My Eyes (Be My AI)			✓	✓		1M+*
Lookout	✓	✓	✓	✓		500k+*
TapTapSee	✓	✓				500k+*
Envision	✓	✓	✓	✓	✓	100k+*
Seeing AI		✓	✓	✓	✓	100k+*
Supersense	✓	✓				50k+*
Aipoly	✓	✓				-
Oorion	✓	✓				-
VizWiz				✓		-
VoiceVision			✓			-

* sur *Google Play* en février 2025

TABLE 1 – Exemples d'applications mobiles d'assistance basée sur la vision pour les personnes malvoyantes.

personnes malvoyantes à se déplacer. Dans ce contexte, ce travail tente d'offrir un aperçu des technologies basées sur la vision et le langage pour les personnes malvoyantes (Sec. 2), en identifiant les principales problématiques rencontrées dans la littérature (Sec. 3) et en proposant des pistes d'exploration pour y répondre (Sec. 4).

2 Technologies basées sur la vision et le langage pour les malvoyants

Les technologies basées sur la vision et le langage ont connu des avancées significatives ces dernières années leur permettant de renforcer l'autonomie des personnes malvoyantes. Ces solutions exploitent des techniques de vision par ordinateur et de traitement du langage naturel qui répondent à un ou plusieurs besoins : décrire l'environnement, répondre à des questions sur une image, identifier des objets ou encore lire du texte à voix haute.

Les applications mobiles jouent un rôle clé dans l'accessibilité à ces technologies grâce à leur faible coût et leur portabilité. Actuellement, sur le marché, plusieurs applications mobiles destinées aux personnes malvoyantes utilisent des modèles d'IA pour analyser l'environnement via l'appareil photo et fournir des descriptions audio. Le Tableau 1 donne un aperçu des fonctionnalités et de la popularité de ces applications. Be My Eyes est l'application la plus téléchargée. Elle se distingue des autres par sa fonctionnalité de mise en relation avec des volontaire humains, en plus de l'IA, offrant une assistance plus fiable et personnalisée.

Certains dispositifs physiques complètent ces applications. Parmi eux, *OrCam MyEye*, un dispositif sous forme d'une mini-caméra à fixer sur une monture de lunette, peut lire du texte, reconnaître des visages et des objets. D'autres solutions, comme les lunettes intelligentes dotées de caméras et d'IA telles qu'*Envision Glasses*, permettent de trouver des objets et des personnes, de décrire des scènes et permettent aux utilisateurs de poser des questions sur des images.

En recherche, avec l'explosion récente des performances des modèles d'IC, de VS et de VQA, de nombreux travaux basés sur ces modèles ont vu le jour pour aider les personnes malvoyantes dans leur vie quotidienne et notamment pour les aider à se déplacer de manière plus confiante dans des environnements inconnus. Les techniques d'IC et le VQA, associées à des modules de TTS, sont largement utilisées pour assister les personnes malvoyantes. Ces technologies permettent respectivement de générer des descriptions des images en temps réel et de répondre à des questions par rapport à une image, facilitant ainsi la compréhension de l'environnement. Dans le domaine de l'aide au déplacement, plusieurs approches ont été proposées, notamment des systèmes embarqués dans des lunettes intelligentes [24] ou des applications mobiles [6]. Ces solutions reposent sur des architectures neuronales avancées, comme les LSTM [9] et des Transformers [14]. Plus récemment, l'émergence des VLMs [12] et des modèles de langage multimodaux pré-entraînés [28] a permis d'accroître les performances des systèmes. Ces modèles permettent non seulement de générer des descriptions plus détaillées, mais aussi d'analyser des scènes plus complexes en tenant compte du contexte. Cependant, ces modèles étant sujet à des hallucinations (c.f. Sec. 3.2), certains auteurs intègrent des modules complémentaires (OCR, détection d'objets, correction d'angle,...) afin d'améliorer leur fiabilité pour qu'ils puissent être utilisés par des personnes malvoyantes [6, 12].

Comme mentionné précédemment, les systèmes actuels d'assistance pour les personnes malvoyantes reposent sur des modèles d'IC et de VQA, couplés à un module de TTS pour transformer les descriptions textuelles en audio. Bien que ces approches aient montré leur efficacité, ces approches restent limitées par leur architecture en plusieurs étapes. Cette séparation peut entraîner une perte d'information et un délai de traitement plus important. Pour pallier ces limites, une nouvelle tâche a émergé : l'image-to-speech end-to-end. Contrairement aux systèmes traditionnels, ces modèles convertissent directement l'image en parole [18, 7]. Cependant, bien qu'innovante, cette tâche présente plusieurs limites. Tout d'abord, l'absence d'une étape textuelle intermédiaire complique la vérification et l'évaluation des modèles. De plus, l'entraînement de ces modèles nécessite de grandes quantités de données annotées associant directement image et audio, ce qui peut être très difficile et couteux à obtenir.

Cependant, les technologies présentées précédemment restent toutefois limitées à l'analyse d'images statiques. L'exploitation de flux vidéo joue un rôle crucial pour aider les personnes malvoyantes à naviguer dans des environnements complexes. En effet, en prenant en compte la dimension temporelle, elle permet de capter les changements dans

un environnement dynamique. Cette capacité est particulièrement précieuse pour les malvoyants qui doivent continuellement s'adapter aux changements de leur environnement et anticiper d'éventuels dangers. Malheureusement, peu de travaux exploitent des flux vidéo pour les aider à se déplacer plus sereinement dans leur environnement. Certains chercheurs s'intéressent néanmoins à cette approche. Par exemple, afin de capturer plus largement l'environnement qui entoure les personnes malvoyantes, [25] introduit VIEW-QA, un nouvel ensemble de données de Video Question Answering capturé à l'aide d'une caméra portable égocentrique à 360 degrés. D'autres encore, comme [22], développent des systèmes intelligents portatifs basés sur le flux vidéo pour aider les personnes malvoyantes dans leur quotidien, leur permettant, entre autres, d'identifier les obstacles présents sur leur chemin.

3 Identification de problématiques majeures

Actuellement, les approches proposées pour aider les malvoyants basées sur l'IC et le VQA font face à de nombreuses problématiques qu'il semble nécessaire de surmonter, ou de moins d'atténuer pour améliorer leur fiabilité. Cette section aborde quatre problématiques majeures, classées par ordre de priorité : (i) les biais des modèles, (ii) les hallucinations, (iii) la qualité des photos prises par des personnes malvoyantes et (iv) la complexité d'évaluation des modèles d'IC et de VQA.

3.1 Biais

Malgré les avancées majeures dans les domaines de l'IC et du VQA, ces modèles restent très sensibles aux données sur lesquelles ils sont entraînés. Ils sont donc susceptibles d'hériter des biais provenant de ces ensembles de données. L'article [30] classe ces biais en trois catégories :

1. **Biais de labellisation** : Ils découlent des déséquilibres présents dans les annotations des jeux de données. Cette catégorie fait référence aux classes ou aux mots qui apparaissent plus fréquemment que d'autres ou aux mots qui apparaissent ensemble de manière récurrente dans le jeux de données. Par exemple, si les annotations mentionnent souvent "voiture" mais rarement "vélo", le modèle pourrait ne pas signaler un cycliste approchant d'un passage piéton. Les modèles VQA ont tendance à exploiter les régularités statistiques entre les occurrences de réponses et certains schémas dans la question. Bien qu'ils soient conçus pour fusionner les informations des modalités visuelles et textuelles, ils répondent souvent en n'utilisant que la question.
2. **Corrélations trompeuses faites par les modèles** : Certains mots et éléments visuels (objets, arrière-plan, etc.) qui apparaissent régulièrement ensemble peuvent orienter la réponse sans que le modèle ne regarde en détail la question et l'image [5]. Par exemple, si un jeu de données associe régulièrement les passages piétons aux feux tricolores, un modèle pourrait annoncer automatiquement la présence d'un feu dès qu'il détecte un passage piéton, ce qui pourrait entraîner des situations dangereuses pour les malvoyants.

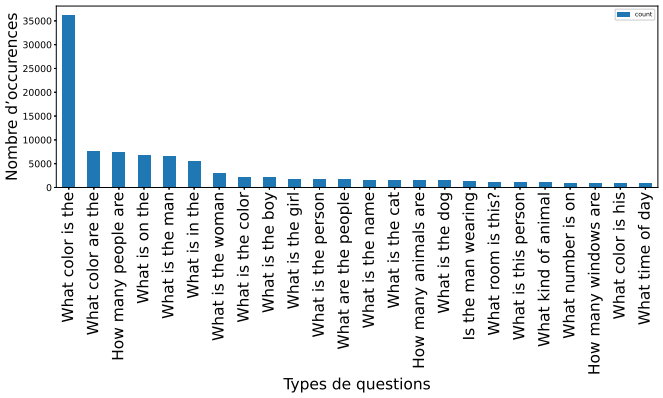


FIGURE 1 – Types de questions les plus posés dans VQAv2

3. **Biais sociaux** : Les paires image-texte peuvent contenir des préjugés humains qui entraînent des réponses discriminatoires ou stéréotypées.

D'autre part, à l'exception du jeu de données VizWiz [6], la majorité des autres jeux de données d'IC et de VQA n'ont pas été spécifiquement conçus pour répondre aux besoins des malvoyants. Par conséquent, ils contiennent des données qui ne sont pas toujours adaptées au développement de technologies destinées à cette population. En effet, les questions de nombreux jeux de données de VQA comme VQAv2 [1] impliquent une certaine connaissance préalable de la scène, ce qui, dans le cas d'applications pour les personnes malvoyantes, semble peu compatible. Par exemple, dans VQAv2, le type de question le plus fréquent est : "what color is the ..." (c.f. Figure 1), impliquant que l'utilisateur sache déjà qu'un certain objet est présent dans l'image.

3.2 Hallucinations

Les modèles d'IC et de VQA sont sujets à des hallucinations [19]. Dans ces domaines, les hallucinations correspondent à des cas où le contenu visuel d'une image et le texte généré pour la décrire ou répondre à une question sur cette image sont en inadéquation. De telles hallucinations affectent considérablement la fiabilité de ces modèles et, dans le cadre d'applications pour les personnes malvoyantes peuvent même s'avérer dangereuses.
Les hallucinations dans les modèles d'IC et de VQA peuvent provenir de plusieurs sources. Elles peuvent venir des biais dans les données d'entraînement, de l'incapacité des encodeurs de vision à ancrer avec précision les images, du désalignement entre les modalités visuelles et textuelles mais aussi pour les VLMs des LLMs, etc. Dans la littérature, plusieurs travaux ont tenté, en travaillant sur ces différentes sources d'hallucinations, de les atténuer. L'article [19] fait en particulier un état de l'art sur les hallucinations dans les VLMs dans lequel les auteurs présentent les différentes sources ainsi que diverses méthodes permettant d'atténuer ces hallucinations. Ces approches tournent principalement autour de l'optimisation des données d'apprentissage, du perfectionnement de divers modules et du post-traitement des sorties générées.
L'évaluation des hallucinations dans les modèles d'IC et

de VQA est une tâche complexe en raison de leur nature générative, qui peut entraîner l'ajout, la suppression ou la reformulation de détails par rapport à une sortie de référence. L'un des défis majeurs réside dans la difficulté d'établir, à l'aide des métriques existantes, une distinction claire entre les éléments fidèlement reformulés et ceux résultant d'hallucinations. En effet, ces modèles peuvent générer des descriptions plausibles mais incorrectes, rendant l'analyse encore plus complexe, notamment lorsque les différences sont subtiles ou contextuelles. C'est pourquoi, pour évaluer les hallucinations d'un modèle de VQA ou d'un VLM, certaines métriques comme CIEM [1], POPE[1] ou FGHE[1] formulent l'évaluation de l'hallucination comme une tâche de classification binaire qui invite les modèles à répondre par "Oui" ou par "Non". Cependant, les métriques comme CIEM, POPE, FGHE, CHAIR[1] ou encore NOPE[1] se focalisent sur la présence d'objets. À part certaines métriques comme FGHE et CIEM, elles ne prennent pas en compte les relations entre objets ni les attributs, ce qui peut conduire à des évaluations partielles. Les métriques reposant sur des modèles pré-entraînés, comme FAITHScore[1], utilisent des modèles de langage et de vision pour détecter les incohérences. Cependant, elles peuvent hériter des biais des modèles utilisés. Les métriques mentionnées ci-dessus comptent parmis les plus connues pour évaluer les hallucinations mais il en existent encore bien d'autres. Le dépôt GitHub[1] recense une grande partie d'entre-elles.

3.3 Qualité des images

Comme décrit dans la section précédente, de nombreuses technologies essaient de décrire et de répondre à des questions à propos d'une image prise directement par une personne malvoyante. Cependant, la plupart du temps, les images prises par ces personnes sont de mauvaise qualité et ne capturent souvent que partiellement la cible visée. En effet, de par leur déficience visuelle, il leur est très difficile, voire impossible, d'inspecter visuellement les images capturées et d'en garantir ainsi une certaine qualité et pertinence. Malheureusement, ces images de mauvaise qualité peuvent amener les modèles à générer des réponses peu fiables. La plupart des modèles actuels génèrent des réponses directes sans évaluer la suffisance des informations. En effet, peu de modèles d'IC et de VQA sont capables d'indiquer qu'ils ne peuvent pas répondre à la question en raison de la qualité de l'image ou de l'absence d'indices visuels nécessaires pour répondre à la question [28, 23].
Le tableau 2 met en évidence la diversité des jeux de données utilisés pour l'IC et le VQA, en termes de sources d'images et de types de scènes capturées. Il souligne notamment le caractère unique de VizWiz qui est l'un des seuls jeux de données contenant des images capturées directement par des personnes malvoyantes, contrastant avec les autres jeux de données, dont la plupart contiennent des images bien cadrées et de bonne qualité. Cependant, dans des applications réelles, l'hypothèse de bonne qualité des images ne tient plus. Un modèle entraîné uniquement sur des images de bonne qualité risque donc de mal généraliser sur ces cas réels et de fournir des réponses erronées. Paradoxalement, pour obtenir de bonnes performances en IC et en VQA, il est crucial de disposer de descriptions riches et précises, ce qui, dans le cas d'application pour les malvoyants, peut s'avérer compliqué. En effet, les images prises par ces dernières sont souvent de faible qualité, rendant l'annotation plus difficile et parfois moins détaillée.

1. https ://github.com/lhanchao777/LVLM-Hallucinations-Survey

3.4 Difficultés d'évaluation des modèles

L'évaluation des modèles d'IC et de VQA est particulièrement complexe en raison de la nature générative de ces modèles. En effet, une même image peut être décrite de multiples façons, et un modèle peut produire des réponses très différentes en fonction du contexte, de la formulation de la question ou même de biais présents dans les données d'entraînement. Les métriques classiques pour évaluer ces modèles comme ROUGE, BLEU, METEOR ou CIDEr [8] qui mesurent la correspondance entre la sortie du modèle et une ou plusieurs références humaines sans toujours capturer la pertinence sémantique ni l'utilité réelle d'une prédiction, semblent donc peu adaptées à cette nature générative. Pour pallier ces limites, d'autres métriques sont alors apparues. SPICE [8] par exemple évalue les descriptions en analysant leur contenu sémantique et leur relation avec la structure de l'image. BERTScore [8], basé sur des représentations de texte obtenues via des modèles de langage, permet de mieux mesurer la similarité sémantique entre une légende générée et une référence humaine. CLIPScore [8], quant à lui, exploite le modèle multimodal CLIP pour évaluer dans quelle mesure une légende est visuellement pertinente par rapport à l'image. Malgré ces avancées, ces métriques ne prennent pas en compte les biais du modèle, les hallucinations ou l'adéquation aux besoins des utilisateurs finaux. L'évaluation humaine reste donc indispensable. Une approche hybride combinant métriques traditionnelles et évaluation humaine permet d'obtenir un meilleur aperçu des performances réelles de ces modèles [13, 27].

4 Pistes d'exploration

Dans cette section, nous proposons quelques pistes d'exploration visant à atténuer les problématiques mentionnées dans la section précédente (Sec. 3). Compte tenu de la complexité de la tâche et des risques associés, nos futurs travaux viseront d'abord à comprendre la scène de manière la plus exhaustive possible et à informer simplement l'utilisateur, plutôt que de tenter de le guider directement.

4.1 Réduction de biais via des données synthétiques

Pour atténuer les biais évoqués dans la section 3.1, une approche consiste à générer des données synthétiques. En effet, générer de telles données permet de contourner les problèmes de confidentialité associés aux données du monde réel, d'équilibrer les jeux de données, de tester des scénarios rares et de réduire les coûts de collecte. Pour ce faire, on pourrait utiliser des plateformes de simulations 3D [4, 3] comme ThreeDWorld ou bien des modèles génératifs et de diffusions [21, 20]. Cependant, il semblerait que les images

Jeux de données	Tâches	Source des images	Nb images	Nb descript°/quest°	Apports potentiels pour des modèles destinés aux malvoyants
VizWiz [11]	IC, VQA	Photo prises par des malvoyants	+ 39k / +32k	+ 195k / +32k	Questions et images réelles issues d'utilisateurs malvoyants
MSCOCO [11]	IC	Flickr	+123k	+ 616k	Grandes diversités d'environnement (intérieur, extérieur, objets courant), Diversité des annotations
Flickr30k [11]	IC	Flickr	+ 31k	+ 158k	Images d'activités, d'événements et de scènes de la vie quotidienne
Conceptual Captions [11]	IC	Pages Web	+ 3.3M	+ 3.3M	Très grand jeux de données, Grande variété de types d'images (ex : images naturelles, de produits, dessins animés, dessins, etc.)
NoCaps [11]	IC	OpenImages (validation + test)	+ 15k	+ 166k	Conçu pour tester la généralisation sur de nouveaux objers, ce qui peut être utile pour décrire des scènes inconnues
Visual Genome [11, 1]	IC, VQA	YFCC100M ∩ MS-COCO	+ 108k	5.4M descriptions de régions / 1.7M	Très dense au niveau des annotations : descriptions de régions, objets, attributs, relations, graphes de régions, graphes de scènes et VQA
GQA [1]	VQA	Visual Genome	+ 113k	+ 22M	Exploite les représentations sémantiques des questions et les graphes de scènes pour répondre à la question, permettant un raisonnement structuré
VQAv2 [1]	VQA	COCO	+ 204k	+ 1.1M	Biais réduit : Chaque question est associée à une paire d'images similaires qui donnent lieu à deux réponses différente
OK-VQA [1]	VQA	COCO	+ 14k	+ 14k	Images du monde réel nécessitant des connaissances externes pour répondre aux questions
TDIUC [17]	VQA	MS-COCO + Visual Genome	+ 167k	+ 1.6M	Grande diversité de type de questions (12), comprenant des questions absurdes pour forcer le système à raisonner sur le contenu de l'image.

TABLE 2 – Principaux jeux de données d'IC et de VQA sur des scènes du monde réel

synthétiques générées par des modèles d'IA induisent des hallucinations en quantité plus élevée et avec une distribution plus uniforme que les images naturelles [10].

4.2 Distinction des niveaux de gravité des hallucinations

Dans les cas d'application pour les personnes malvoyantes, il serait intéressant de pouvoir distinguer différents niveaux de gravité des hallucinations générées. En effet, pour une personne malvoyante, que le modèle n'identifie pas correctement la couleur d'un lampadaire dans une description de scène a un impact limité ; par contre, que le modèle ne reconnaisse pas une voiture qui arrive lorsque la personne souhaite traverser la route est beaucoup plus grave.

4.3 Sélection d'images

Bien que certaines études [6, 29] aient montré un certain succès sur la tâche d'IC dans de mauvaises conditions, il semble que cette amélioration ait ses limites. En effet, il semble difficile pour un modèle de générer des descriptions ou des réponses utiles sur des images de mauvaises qualités. Dans les cas d'application pour les malvoyants il semblerait plus judicieux que le système notifie l'utilisateur de la mauvaise qualité de sa photo et l'aide à en prendre une nouvelle plutôt que de prendre le risque de générer une description inexacte qui peut amener l'utilisateur à prendre une mauvaise décision.

Pour pallier ce problème, certains articles évaluent la qualité de l'image avant de la décrire ou de répondre à une question [28, 23]. Si l'image est jugée de bonne qualité, le modèle génère une description ou répond à la question, sinon l'utilisateur est informé des défauts détectés et est invité à prendre une nouvelle photo. Cependant, un tel procédé dans la vie quotidienne peut s'avérer long et laborieux.

Une autre piste qui pourrait être explorée consiste à demander à l'utilisateur de capturer l'environnement qui l'entoure en prenant une série de photos, puis d'utiliser un modèle pour décrire la scène ou répondre à ses questions sur ces images. De cette façon, on pourrait limiter le nombre de photos de mauvaise qualité qui devront être reprises. En combinant les techniques de VQA appliquées à une série d'images d'une même scène [2] et de sélection de photos parmi une série d'images (techniques spécialisées dans l'évaluation de la qualité esthétique des images, qui recherchent la meilleure photo parmi une série de photos)[15, 16], on pourrait envisager un modèle en deux étapes. D'abord, il filtrerait les images en ne conservant que les images de bonne qualité puis il répondrait aux questions en se focalisant sur les images de la scène portant les informations nécessaires pour y répondre.

4.4 Prise en compte de l'utilité des réponses dans l'évaluation des modèles

Dans les cas d'application pour les malvoyants, l'évaluation des performances d'un modèle ne devrait pas se limiter à des scores automatiques, mais devrait également prendre en compte l'utilité des descriptions et des réponses générées [12]. En effet, une légende peut être objectivement correcte tout en étant peu utile ou peu informative pour un utilisateur malvoyant. Ici, l'objectif n'est pas seulement d'obtenir des sorties précises, mais également de s'assurer qu'elles apportent une réelle valeur ajoutée aux utilisateurs.

Références

[1] M. Agrawal, A. Jalal, and H. Sharma. A review on vqa : Methods, tools and datasets. In *International Conference on Computer Science and Emerging Technologies*, 2023.

[2] A. Bansal, Y. Zhang, and R. Chellappa. Visual Question Answering on Image Sets. In *Computer Vision – ECCV*, 2020.

[3] P. Cascante-Bonilla, K. Shehada, J. Smith, S. Doveh, D. Kim, R. Panda, G. Varol, A. Oliva, V. Ordonez,

R. Feris, and L. Karlinsky. Going Beyond Nouns With Vision & Language Models Using Synthetic Data. In *IEEE/CVF International Conference on Computer Vision*, 2023.

[4] P. Cascante-Bonilla, H. Wu, L. Wang, R. Feris, and V. Ordonez. Sim VQA : Exploring Simulated Environments for Visual Question Answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[5] C. Dancette. *Shortcut Learning in Visual Question Answering*. PhD thesis, Sorbonne Université, 2023.

[6] P. Dognin, I. Melnyk, Y. Mroueh, I. Padhi, M. Rigotti, J. Ross, Y. Schiff, R. Young, and B. Belgodere. Image Captioning as an Assistive Technology : Lessons Learned from VizWiz 2020 Challenge. *Journal of Artificial Intelligence Research* , 73, 2022.

[7] J. Effendi, S. Sakti, and S. Nakamura. End-to-End Image-to-Speech Generation for Untranscribed Unknown Languages. *IEEE Access*, 9, 2021.

[8] S. Elguendouze. *Explainable Artificial Intelligence approaches for Image Captioning*. PhD thesis, Université d'Orléans, 2024.

[9] R. Faurina, A. Jelita, A. Vatresia, and I. Agustian. Image captioning to aid blind and visually impaired outdoor navigation. *IAES International Journal of Artificial Intelligence*, 12(3), 2023.

[10] Y. Gao, J. Wang, Z. Lin, and J. Sang. AIGCs Confuse AI Too : Investigating and Explaining Synthetic Image-induced Hallucinations in Large Vision-Language Models. 2024.

[11] T. Ghandi, H. Pourreza, and H. Mahyar. Deep learning approaches on image captioning : A review. *ACM Comput. Surv.*, 56(3), 2023.

[12] Y. Hao, F. Yang, H. Huang, S. Yuan, S. Rangan, J. Rizzo, Y. Wang, and Y. Fang. A Multi-Modal Foundation Model to Assist People with Blindness and Low Vision in Environmental Interaction. *Journal of Imaging*, 10(5), 2024.

[13] Y. Hao, F. Yang, H. Huang, S. Yuan, S. Rangan, J. Rizzo, Y. Wang, and Y. Fang. A Multi-Modal Foundation Model to Assist People with Blindness and Low Vision in Environmental Interaction. *Journal of Imaging*, 10(5), 2024.

[14] R. Harshitha, B. Lakshmipriya, and V. Krishnamurthy. TransEffiVisNet – an image captioning architecture for auditory assistance for the visually impaired. *Multimedia Tools and Applications*, 2024.

[15] J. Huang, Y. Gong, Y. Shi, X. Zhang, J. Zhang, and Y. Yin. Focusing on Subtle Differences : A Feature Disentanglement Model for Series Photo Selection. *IEEE Transactions on Multimedia*, 26, 2024.

[16] J. Huang, Y. Gong, L. Zhang, J. Zhang, L. Nie, and Y. Yin. Modeling Multiple Aesthetic Views for Series Photo Selection. *IEEE Transactions on Multimedia*, 26, 2024.

[17] K. Kafle and C. Kanan. An analysis of visual question answering algorithms. In *IEEE International Conference on Computer Vision*, 2017.

[18] M. Kim, J. Choi, S. Maiti, J. Yeo, S. Watanabe, and Y. Ro. Towards Practical and Efficient Image-to-Speech Captioning with Vision-Language Pre-Training and Multi-Modal Tokens. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2024.

[19] H. Liu, W. Xue, Y. Chen, D. Chen, X. Zhao, K. Wang, L. Hou, R. Li, and W. Peng. A Survey on Hallucination in Large Vision-Language Models. 2024.

[20] Z. Liu, H. Liang, X. Huang, W. Xiong, Q. Yu, L. Sun, C. Chen, C. He, B. Cui, and W. Zhang. SynthVLM : High-Efficiency and High-Quality Synthetic Data for Vision Language Models. 2024.

[21] F. Ma, Y. Zhou, F. Rao, Y. Zhang, and X. Sun. Image Captioning with Multi-Context Synthetic Data. *AAAI Conference on Artificial Intelligence*, 38(5), 2024.

[22] R. Meghana and M. Kulkarni. Smart system with descriptive video service and spoken dialogue system for visually impaired individuals. In *International Symposium on VLSI Design and Test*, 2024.

[23] K. Ohata, S. Kitada, and H. Iyatomi. Feedback is Needed for Retakes : An Explainable Poor Image Notification Framework for the Visually Impaired. In *IEEE International Conference on Smart Communities : Improving Quality of Life Using ICT, IoT and AI*, 2022.

[24] C. Rane, A. Lashkare, A. Karande, and Y. Rao. Image Captioning based Smart Navigation System for Visually Impaired. In *International Conference on Communication information and Computing Technology*, 2021.

[25] I. Song, M. Joo, J. Kwon, and J. Lee. Video Question Answering for People with Visual Impairments Using an Egocentric 360-Degree Camera. 2024.

[26] Y. Thoo, M. Jeanneret Medina, J. Froehlich, N. Ruffieux, and D. Lalanne. A large-scale mixed-methods analysis of blind and low-vision research in acm and ieee. In *International ACM SIGACCESS Conference on Computers and Accessibility*, 2023.

[27] P. Xu, W. Shao, K. Zhang, P. Gao, S. Liu, M. Lei, F. Meng, S. Huang, Y. Qiao, and P. Luo. LVLM-eHub : A Comprehensive Evaluation Benchmark for Large Vision-Language Models. 2023.

[28] B. Yang, L. He, K. Liu, and Z. Yan. VIAssist : Adapting Multi-modal Large Language Models for Users with Visual Impairments. 2024.

[29] L. Yu, M. Nikandrou, J. Jin, and V. Rieser. Quality-agnostic image captioning to safely assist people with vision impairment. In *International Joint Conference on Artificial Intelligence*, 2023.

[30] B. Zhu and H. Zhang. Debiasing vision-language models for vision tasks : a survey. *Frontiers of Computer Science*, 19(1), 2025.

Détection Automatique des Traînées Astronomiques avec YOLO – Une Approche Exploratoire pour la Connaissance du Domaine Spatial

Loshan RASAN[1], Sonimith HANG[1], Xhesika LACI[1], Binbin XU[2]

[1]Department of Computer Science & Artificial Intelligence, IMT Mines Ales
[2]EuroMov Digital Health in Motion, Univ. Montpellier, IMT Mines Ales

loshan.rasan@mines-ales.org, sonimith.hang@mines-ales.org, xhesika.laci@mines-ales.org, binbin.xu@mines-ales.fr

Résumé

L'identification des traînées transitoires est cruciale pour la Connaissance du Domaine Spatial, mais les méthodes classiques sont coûteuses et inadaptées au temps réel. Nous explorons YOLOv8m pour cette tâche, en l'entraînant sur le StreaksYoloDataset (2 388 images annotées, télescope Stellina). L'approche prend en compte les variations de bruit, d'éclairement et de champ de vision. Les résultats montrent un bon équilibre entre rapidité et précision (mAP@50-95 : 0,90), bien que des améliorations soient nécessaires pour limiter les faux positifs et détecter les traînées faibles.

Mots-clés

détection de traînées astronomiques, YOLO, Connaissance du Domaine Spatial, Veille Spatiale, suivi des satellites, détection d'objets

1 Introduction

L'intensification des activités spatiales accentue le besoin de Connaissance du Domaine Spatial (Veille Spatiale), car les satellites, les débris spatiaux et les rayons cosmiques perturbent les observations astronomiques. Une détection efficace de ces traînées transitoires est donc essentielle pour la surveillance spatiale et l'astronomie.

Les approches classiques de détection des traînées astronomiques reposent principalement sur des méthodes telles que le seuillage [1], la transformée de Hough [2], la transformée de Radon [3], la détection de contours [4, 5], la soustraction d'images [6] etc. Bien que ces techniques soient efficaces dans des cas spécifiques, elles présentent plusieurs limitations. Le seuillage, par exemple, est rapide mais très sensible au bruit et aux variations d'intensité, rendant la détection imprécise. La transformée de Hough est performante pour les traînées linéaires mais souffre d'un coût computationnel élevé, limitant son application en temps réel. De même, les méthodes de détection de contours dépendent fortement des paramètres choisis et sont inefficaces pour les traînées diffuses ou à faible contraste. La soustraction d'images, bien qu'efficace pour détecter des objets transitoires, nécessite une image de référence propre et peut être perturbée par des variations instrumentales.

Toutefois, ces méthodes traditionnelles atteignent rapidement leurs limites face à la complexité croissante des observations astronomiques. Cette contrainte a conduit au développement de nouvelles approches basées sur l'apprentissage profond et automatique, notamment les réseaux de neurones, qui ont démontré une performance supérieure dans la détection automatique des traînées. Varela et al. (2019) ont montré la supériorité des réseaux de neurones convolutionnels (CNN) sur les méthodes classiques pour la détection de traînées [7]. Pöntinen et al. (2020) ont évalué StreakDet, démontrant son efficacité pour l'extraction automatique de traînées dans des images simulées d'Euclid [8].

Ces modèles restent néanmoins coûteux en calcul et peu adaptés aux contraintes du temps réel. Dans cette étude, nous proposons une approche basée sur YOLOv8m [9], un détecteur rapide et optimisé pour l'identification automatique des traînées astronomiques, offrant une solution performante pour l'astronomie et la surveillance spatiale.

L'objectif de cette étude est de développer un modèle automatisé de détection des traînées astronomiques avec une haute précision, tout en garantissant un compromis optimal entre rapidité et fiabilité. Pour cela, nous évaluons et validons notre approche sur le StreaksYoloDataset [10], un ensemble d'images annotées permettant de tester la robustesse du modèle. Nous analysons également ses limites, notamment en ce qui concerne les sous-détections et les faux positifs, afin d'identifier des axes d'amélioration. Enfin, cette étude vise à fournir un outil évolutif, capable de répondre aux besoins croissants de la communauté astronomique en matière de surveillance et d'analyse des traînées transitoires.

2 Méthodologie

2.1 Dataset

L'efficacité d'un modèle d'apprentissage profond dépend fortement de la qualité et de la diversité des données utilisées pour son entraînement. Dans cette étude, nous utilisons le *StreaksYoloDataset* [10], un ensemble d'images astrono-

miques brutes capturées entre mars 2022 et février 2023 avec des télescopes intelligents Stellina, dans la région du Luxembourg. Ce dataset contient des images avec une résolution de 640 x 640 pixels, prises à l'aide d'un télescope Stellina, qui utilise un doublet à faible dispersion (aperture de 80 mm, focale de 400 mm, ratio focal de f/5) et un capteur CMOS Sony IMX178 d'une résolution de 6,4 millions de pixels. Les images sont annotées avec la position des traînées, qui représentent des satellites, des débris spatiaux ou des rayons cosmiques.

Le dataset est structuré selon le format YOLO, avec des fichiers séparés pour les images en format JPEG (compression minimale) et des fichiers texte contenant les annotations des positions des traînées. Ces fichiers sont compatibles avec des outils d'entraînement de pointe et des logiciels graphiques comme MakeSense [11]. Le *StreaksYoloDataset* permet d'entraîner des modèles de détection pour la surveillance du domaine spatial (SDA), et offre ainsi une solution adaptée aux équipements astronomiques accessibles au public.

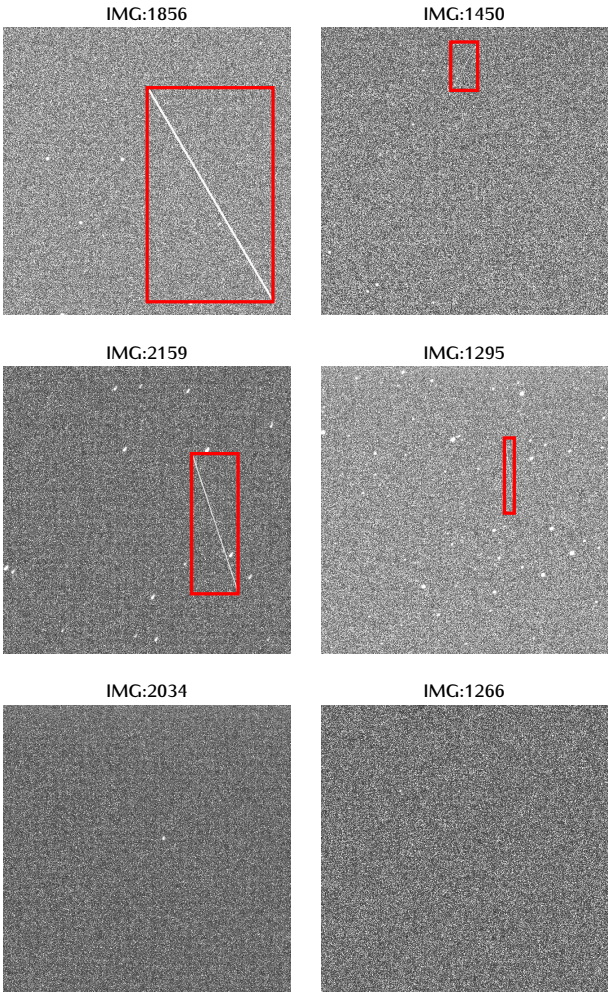


FIGURE 1 – Exemples d'images dans le StreaksYoloDataset. Les "streaks" sont annotés en rectangle rouge. Un "streak" pourrait être absent dans certaines images.

Pour améliorer la robustesse du modèle, ce dataset couvre une large variabilité des conditions d'observation : ajustement du temps d'exposition en function de la luminosité, diversité des niveaux de bruit selon les conditions atmosphériques, prise en compte de l'éclairement variable, et capture sur des champs de vision variés, allant des zones denses en étoiles aux régions plus vides.

2.2 Modèle – YOLOv8

Les traînes (streaks) observés dans les images astronomiques résultent de passage des objets dans l'espace / l'atmosphère très bruité. Cela rend leur identification particulièrement complexe. Ce problème s'inscrit dans le cadre plus large de la détection d'objets en apprentissage automatique avec les réseaux de neurones profonds (DNN). Parmi les architures courantes, on y retrouve notamment les R-CNN (Region-based Convolutional Neural Network) [12] et ses variants qui offrent une haute performance mais sont très coûteux en calcul ; ainsi que les SSD (Single Shot MultiBox Detector) [13] qui sont plus rapides mais souvent moins performants sur la détection des objets de petite taille.
Les images astronomiques sont particulièrement volumineuses : les images de très haute résolution sont très rarement compressées pour éviter toute perte d'informations. Dans cette étude, nous optons pour YOLO (You Only Look Once), une famille de détecteurs en état de l'art réputés pour leur compris entre la vitesse et la précision. YOLOv8m publié en janvier 2023 [9], en particulier, se distingue par son architecture optimisée et son efficacité. YOLOv8 surpasse d'autres modèles de la même famille (YOLOv5, YOLOv6, YOLOv7), en offrant une meilleure performance tout en réduisant la taille du modèle et la latence (Fig. 2).

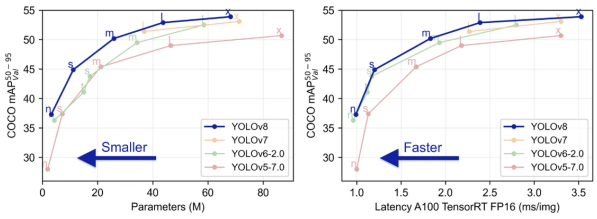


FIGURE 2 – Comparaison de YOLOv8 avec les versions précédentes de YOLO (YOLOv5, YOLOv6, YOLOv7) en termes de précision (mAP) par rapport à la taille du modèle (gauche) et de vitesse d'inférence (droite). Source [14]

Un avantage clé de YOLOv8m réside dans sa facilité de fine-tuning, grâce aux bibliothèques bien documentées d'Ultralytics, permettant une adaptation rapide à des ensembles de données spécifiques. Les principales caractéristiques de YOLOv8m, la variante utilisée dans cette étude, sont les suivantes :

— 25,9 millions de paramètres, répartis sur 295 couches.
— Coût computationnel : 79,1 GFLOPs.
— Optimisé pour l'inférence, offrant une précision élevée tout en restant léger.

2.3 Configuration de l'Entraînement

Bien que les images astronomiques diffèrent des images naturelles utilisées pour l'entraînement des modèles YOLO, la structure des traînées (streaks) présente des similarités avec des formes simples telles que des lignes ou des rectangles très édroites sur un fond complexe. Le fine-tuning d'un modèle pré-entraîné sur un large corpus d'images permet ainsi de tirer parti des représentations apprises tout en adaptant le modèle aux spécificités des traînées astronomiques. Toutefois, l'un des inconvénients de cette approche est que le modèle conserve une architecture relativement lourde. Une réduction du nombre de paramètres, en sélectionnant un modèle plus compact, pourrait améliorer son efficacité sur des dispositifs à ressources limitées.

Hyperparamètres et Protocole d'Entraînement

L'entraînement du modèle a été effectué sur 75 époques avec un batch-size de 32 et une résolution d'image de 640×640 pixels. L'optimisation repose sur SGD (Stochastic Gradient Descent) avec un taux d'apprentissage initial de 0.001, un momentum de 0.9 et une pénalisation L2 (weight decay) de 0.0005 pour éviter le sur-apprentissage.

Pour améliorer la généralisation, une technique de label smoothing (0.05) a été appliquée pour atténuer la surconfiance du modèle. De plus, l'entraînement a été accéléré grâce à l'Automatic Mixed Precision (AMP), permettrant une réduction significative de l'utilisation de mémoire. Les poids du modèle ont été sauvegardés dans un répertoire projeté pour assurer un suivi efficace des expériences.

Cette configuration vise à maximiser l'apprentissage du modèle tout en équilibrant précision et efficacité computationnelle.

Validation du Modèle et Seuil de Confiance

L'impact du seuil de confiance sur la précision, le rappel et le F1-score a été analysé sur l'ensemble de validation, avec un seuil IoU de 0.70 assurant une correspondance rigoureuse entre les prédictions et les annotations (Fig. 3). Un seuil de 0.4 a été retenu pour réduire les faux positifs tout en maintenant un bon rappel. Ce paramètre reste ajustable en fonction des exigences spécifiques de l'application.

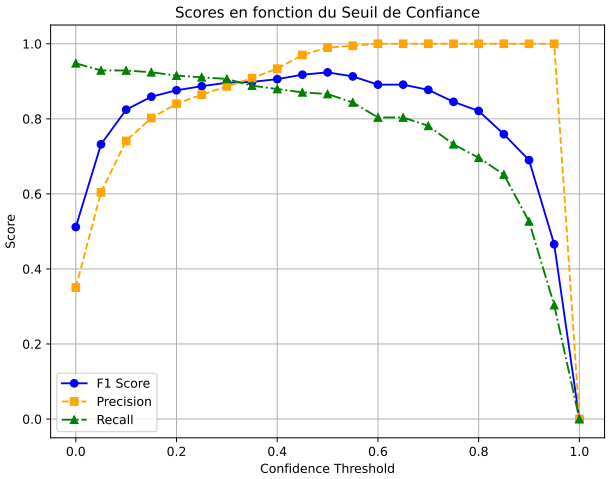


FIGURE 3 – Impact du seuil de confiance sur la précision, le rappel et le F1-score (validation, IoU = 0.70).

Le modèle a ensuite été évalué sur l'ensemble de test.

2.4 Évaluation des Performances

Nous évaluons le modèle à l'aide de plusieurs métriques, notamment la mAP@50-95, la précision, le rappel et le F1-score. La mAP@50-95 mesure la précision moyenne sur plusieurs seuils IoU (0.50 à 0.95), équilibrant précision et rappel pour une évaluation fiable [15, 16].
La précision (precision) représente la proportion de détections correctes parmi l'ensemble des prédictions, tandis que le rappel (*recall*) quantifie la capacité du modèle à détecter toutes les traînées existantes. Le F1-score est la moyenne harmonique de la précision et du rappel, offrant une mesure globale de la performance [16]. Enfin, la spécificité reflète la capacité du modèle à éviter les fausses détections.
Les métriques utilisées suivent les standards établis dans le domaine de la détection d'objets, notamment ceux définis dans les travaux du PASCAL VOC Challenge [15]. De plus, des analyses comparatives approfondies ont été menées sur ces métriques pour évaluer leur pertinence dans différents contextes de détection [16, 17].

2.4.1 Métrique et Seuils IoU

L'*Intersection over Union* (IoU) mesure le chevauchement entre la boîte de prédiction et la vérité terrain [18] :

$$\text{IoU} = \frac{|B_p \cap B_g|}{|B_p \cup B_g|} \tag{1}$$

où B_p et B_g sont respectivement la boîte de prédiction et la boîte de vérité terrain.
Un seuil IoU optimal de 0.7 a été retenu pour équilibrer précision et rappel.

Seuil IoU	Effet
0.50	Plus de détections, risque de faux positifs
0.75	Bon compromis précision-rappel
0.95	Haute précision, risque de faux négatifs

TABLE 1 – Impact des seuils IoU sur la détection.

3 Résultats et Discussion

3.1 Courbe de Perte

Pour évaluer l'évolution de l'apprentissage du modèle YOLOv8m, la Figure 4 présente la variation des pertes d'entraînement et de validation, ainsi que l'évolution du mAP50-95, au fil des époques.

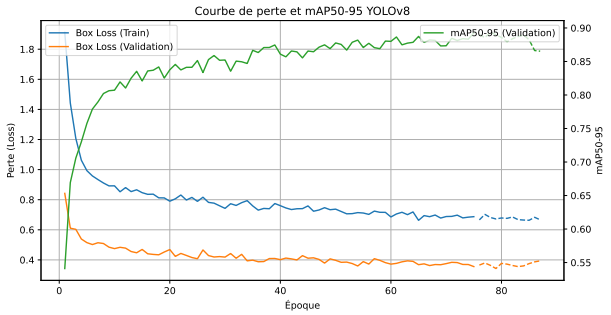


FIGURE 4 – Évolution de la perte (Loss) et du mAP50-95 pour l'entraînement et la validation de YOLOv8m.

La courbe de perte de validation reste inférieure à celle de l'entraînement et commence à converger de manière significative sur l'ensemble d'entraînement à partir d'environ 60 époques. La perte d'entraînement diminue rapidement avant de se stabiliser, tandis que la perte de validation suit une évolution similaire, témoignant d'un bon apprentissage du modèle. Parallèlement, le mAP50-95 progresse rapidement avant d'atteindre un plateau, indiquant une amélioration notable de la précision du modèle.
En vue des résultat, 75 époques est un bon compromis.

3.2 Performance du Modèle

Le modèle YOLOv8m a été évalué sur un ensemble de test de 333 images, atteignant des performances élevées tout en maintenant une faible complexité computationnelle. Les principales métriques sont résumées dans le Tableau 2.

Métrique	Score
Rappel	0.93
Précision	0.88
F1-Score	0.90
Spécificité	0.80
mAP@50-95	0.90

TABLE 2 – Performance du modèle YOLOv8m sur le StreaksYoloDataset.

Le modèle atteint une précision de $0,88$ et le rappel $0,93$ avec un seuil IoU fixé à $0,70$. Cette performance se traduit par un F1-score de $0,90$ qui indique un bon équilibre entre la capacité du modèle à détecter les traînées et à limiter les fausses détections. La spécificité de $0,80$ montre également que le modèle YOLOv8m parvient à éviter un excès de fausse-positive, un aspect essentiel pour des applications astronomiques où la fiabilité des détections est primordiale. La mAP@50-95 de $0,90$ confirme la stabilité du modèle sur une large plage de seuils IoU, ce qui démontre ainsi sa cohérence en termes de détection.

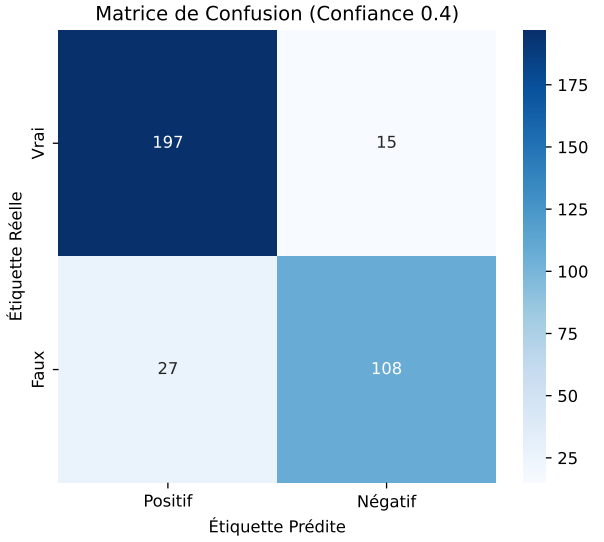


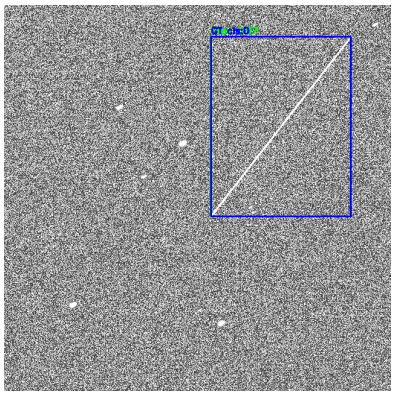
FIGURE 5 – Matrice de confusion du modèle avec un seuil de confiance de 0.4 sur le dataset de test

3.3 Analyse des Faux Positifs et Faux Négatifs

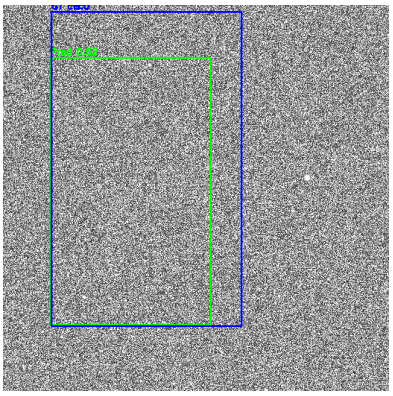
Une analyse détaillée des erreurs a été réalisée à l'aide d'une matrice de confusion (Figure 5) avec un seuil de confiance de 0.4 (Fig. 3), sélectionné à partir du dataset de validation pour optimiser la précision et le rappel.
Sur les données de test, 42 sur les 333 images ne sont pas classées correctement :

- Faux Positifs (FP) : 27 faux positifs ont été détectés, réduisant légèrement la précision. Ces erreurs sont principalement dues à :
  - La confusion avec des étoiles brillantes ou des rayons cosmiques.
  - Des artefacts optiques ou du bruit du capteur.
  - Une tendance du modèle à détecter des traînées dans des zones à fort contraste.
- Faux Négatifs (FN) : Le modèle a omis 15 traînées réelles, affectant légèrement le rappel. Ces omissions sont attribuées à :
  - Des traînées de faible intensité, difficilement discernables.
  - Des occultations partielles causées par d'autres objets célestes.
  - Une variabilité en épaisseur et en longueur des traînées rendant leur détection plus complexe

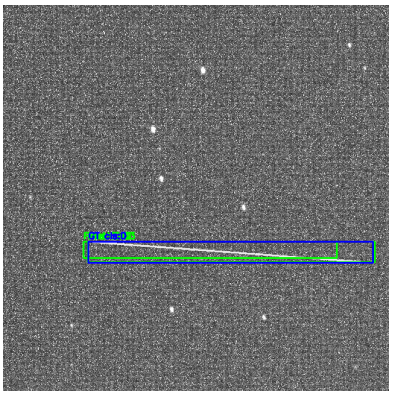
Pour mieux analyser les performances du modèle, nous présentons six cas illustrant différents scénarios de détection des traînées. Ces exemples permettent d'identifier les points forts du modèle ainsi que ses principales limitations :



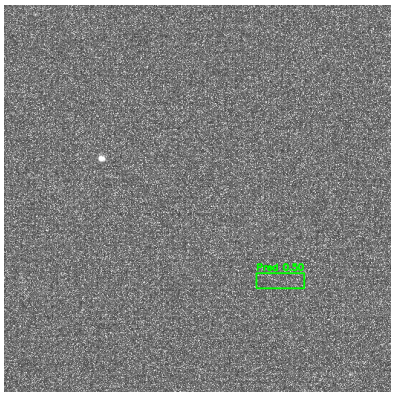
(a) Exemple de détection précise où la boîte prédite (vert) est alignée avec la vérité terrain (bleu).



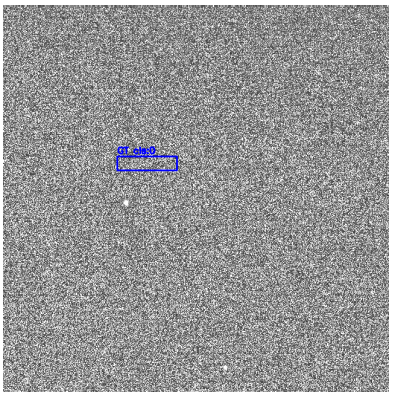
(b) Exemple de sous-détection où la boîte prédite couvre une zone inférieure à la traînée réelle.



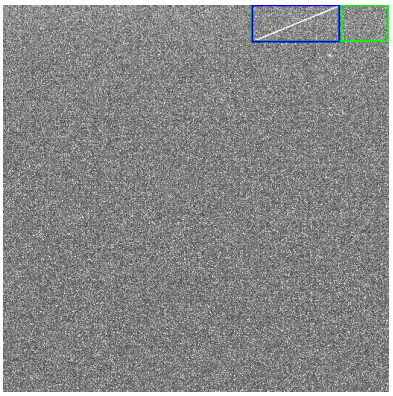
(c) Exemple de mauvais alignement entre la boîte prédite (vert) et la vérité terrain (bleu).



(d) Exemple de faux positif où le bruit de fond est mal classifié comme une traînée.



(e) Exemple de détection manquée où la traînée n'est pas reconnue par le modèle.



(f) Exemple de mauvaise classification où un objet céleste est détecté comme une traînée.

FIGURE 6 – Exemples de détection de traînées

- **Cas 1 : Détection Précise et Alignement Optimal** Dans certains cas, la boîte englobante prédite correspond étroitement à la vérité terrain, illustrant la capacité du modèle à localiser précisément les traînées. Un exemple est présenté dans la Figure 6a.

- **Cas 2 : Sous-Détection avec Boîte Réduite** Le modèle peut parfois sous-estimer la région de la traînée, générant une boîte plus petite que la vérité terrain (Figure 6b). Cela peut être due à des variations de luminosité ou de longueur des traînées, rendant la détection plus difficile.

- **Cas 3 : Mauvais Alignement de la Détection** Dans certains cas, la boîte englobante prédite ne recouvre pas entièrement la traînée réelle (Figure 6c), ce qui entraîne une détection partielle ou imprécise.

- **Cas 4 : Détection d'Artefacts de Bruit** Le modèle identifie parfois des artefacts de bruit de fond comme des traînées, générant des faux positifs (Figure 6d). Une meilleure prétraitement des images ou un ajustement du seuil de confiance pourrait limiter ce phénomène.

- **Cas 5 : Détection Manquée** Certaines traînées ne sont pas détectées, notamment lorsqu'elles présentent un faible contraste avec le fond (Figure 6e).

- **Cas 6 : Mauvaise Classification d'Objets** Dans de rares cas, le modèle confond un objet céleste brillant avec une traînée, classifiant incorrectement une structure lumineuse comme un objet linéaire (Figure 6f).

3.4 Évolution des métriques selon la taille du jeu d'entraînement

Afin d'évaluer la robustesse de notre méthode de fine-tuning ainsi que de dégager des recommandations pratiques quant à l'usage optimal du cadre proposé, nous avons étudié l'impact de la taille du jeu d'entraînement sur les performances du modèle. Pour cela, nous avons divisé l'ensemble des données d'entraînement en cinq sous-ensembles égaux à l'aide d'une validation croisée (K-Fold avec $K = 5$), en maintenant un jeu de validation constant.

Le modèle YOLOv8 a été entraîné cinq fois (avec 20 epochs seulement), avec une quantité croissante de données d'entraînement : $\frac{1}{5}$, $\frac{2}{5}$, ..., jusqu'à l'intégralité des données disponibles ($\frac{5}{5}$). À chaque étape, le modèle est réinitialisé afin de garantir l'indépendance des résultats. Les performances sont ensuite évaluées sur le même jeu de validation, selon plusieurs métriques standard : précision, rappel, F1-score, spécificité, et mAP@50 :95.
L'ensemble de notre dataset contient 1 722 images. Les résultats expérimentaux montrent une nette amélioration des performances lorsque la taille du jeu d'entraînement passe de 344 images (soit 20%) à 690 images (soit 40%). Au-delà de 690 images, les courbes de performances tendent à se stabiliser, suggérant que les bénéfices d'un apport supplémentaire de données deviennent marginaux. Cela indique qu'un compromis raisonnable peut être trouvé entre le coût d'annotation et les performances du modèle.
Il est à noter que les performances pourraient encore être améliorées en adaptant le nombre d'epochs pour chaque volume de données, afin d'optimiser pleinement l'entraînement. Toutefois, nous avons conservé un nombre fixe d'epochs (20) pour toutes les expériences, afin de maintenir un cadre de comparaison cohérent.
Cette analyse permet de mieux comprendre la sensibilité du modèle à la quantité de données disponibles, et constitue une base utile pour orienter les futures campagnes d'acquisition et d'annotation.

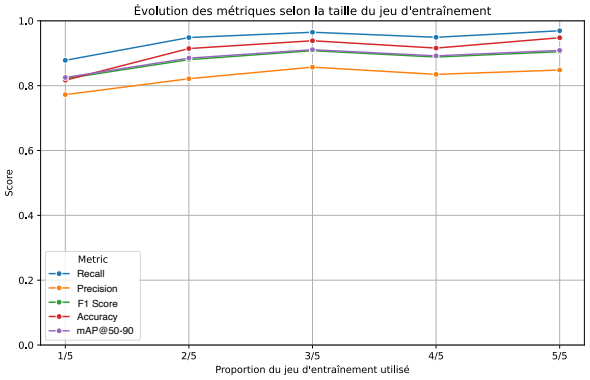


FIGURE 7 – Évolution des métriques globales (rappel, précision, F1-score, accuracy et mAP@50 :95) en fonction de la taille du jeu d'entraînement

3.5 Perspectives et Améliorations Futures

Bien que performant, le modèle peut être optimisé pour une meilleure adaptation aux contraintes astronomiques. Plusieurs axes d'amélioration sont envisageables :

— **Diversification des données** : Intégrer des traînées faibles, occultées et de morphologies variées pour améliorer la généralisation.

— **Ajustement dynamique du seuil de confiance** : Adapter dynamiquement le seuil pour réduire les faux positifs sans altérer le rappel.

— **Optimisation du post-traitement** : Améliorer la suppression non maximale (*NMS*) pour affiner les prédictions et limiter les détections redondantes.

— **Exploitation temporelle** : Intégrer une analyse multi-image pour renforcer la robustesse face aux artefacts transitoires.

— **Exploration et comparaison d'architectures de détection** : L'approche actuelle repose sur YOLOv8m en raison de son bon compromis entre rapidité et précision. Toutefois, d'autres architectures telles que *SSD (Single Shot MultiBox Detector)*, *RetinaNet* ou *DETR (Detection Transformer)* pourraient également être adaptées à la détection de traînées astronomiques. Une comparaison systématique avec ces modèles constitue une perspective pertinente pour affiner l'évaluation. Par ailleurs, l'étude d'architectures plus récentes, comme *YOLOv11*, intégrant des mécanismes d'attention ou une fusion multi-échelle des caractéristiques, pourrait contribuer à une meilleure détection des traînées subtiles ou complexes, tout en maintenant une efficacité computationnelle.

— **Collaboration interdisciplinaire** : Renforcer les partenariats avec des astronomes pour intégrer des connaissances métier spécifiques (par exemple, les contraintes physiques des traînées) dans la conception du modèle, afin d'aligner les prédictions avec les attentes du domaine.

— **Modèles plus légers** : Entraîner un modèle compact sans recourir aux poids pré-entraînés sur le jeu de données COCO (Common Objects in Context), afin d'éviter le transfert de caractéristiques inadaptées issues de domaines visuels génériques (tels que les objets du quotidien). Cette approche vise à réduire la complexité du modèle, à éliminer les biais hors domaine et à optimiser les performances en inférence, notamment dans des environnements contraints en ressources computationnelles.

Ces améliorations renforceront la précision et l'efficacité du modèle, tout en garantissant une adaptation efficace aux contraintes opérationnelles des missions astronomiques, qu'elles soient professionnelles ou amateurs.

Accessibilité et Impact sur la Recherche Astronomique

L'inférence rapide sur des GPU de moyenne gamme facilite l'adoption de cette approche, aussi bien par les observatoires professionnels que par les projets de science participative. Cette accessibilité favorise une collaboration internationale pour la surveillance du ciel et contribue à la préservation des données astrophysiques.
L'optimisation du modèle repose sur l'utilisation de GPU comme le Tesla T4 (15 Go de VRAM) et l'Automatic Mixed Precision (AMP), permettant un traitement rapide et efficace tout en réduisant la charge mémoire. Le modèle atteint un temps d'inférence de 11,6 ms par image, rendant l'approche adaptée aux applications en temps réel de détection des traînées. Ces optimisations la rendent viable sur

des plateformes accessibles, assurant une grande réactivité pour la surveillance spatiale.
Pour le finetuning, il est possible d'utiliser Google Colab, offrant un environnement cloud avec GPU pour les utilisateurs disposant de ressources limitées, et facilitant ainsi l'accès à la technologie de détection.
Cette approche améliore l'accessibilité à la surveillance spatiale et contribue à la préservation des données astrophysiques.
En automatisant la détection et la filtration des traînées, notre modèle réduit les interférences artificielles, permettant aux chercheurs d'optimiser leur temps d'analyse et d'améliorer la qualité des observations. Cette approche s'inscrit ainsi dans une démarche de veille spatiale, essentielle pour la protection des infrastructures spatiales et la recherche astronomique.

3.6 Perspectives et Défis Ouverts

Une détection multi-classes en temps réel permettra-t-elle une surveillance totalement autonome du ciel ? Comment la fusion de capteurs et le traitement avancé des formes pourraient-ils améliorer la détection des traînées faibles ou courbes ? Ces questions illustrent le potentiel d'innovation de cette approche, tant pour la protection des satellites que pour l'exploration du cosmos.
En combinant apprentissage profond et amélioration des méthodes d'observation, ce pipeline démontre que la détection des traînées peut être rapide, précise et accessible, ouvrant ainsi la voie à une surveillance plus agile et efficace du ciel nocturne.

4 Conclusion

Cette étude a exploré la détection automatique des traînées transitoires dans un ciel de plus en plus encombré. En combinant le StreaksYoloDataset et l'architecture YOLOv8m, nous avons développé une approche rapide et précise, adaptée aux contraintes du temps réel.
Nos résultats montrent que le modèle offre un bon équilibre entre précision et rappel, mais montrent également des marges d'amélioration, notamment dans la détection des traînées faibles et la réduction des faux positifs. Des améliorations, telles que l'optimisation du post-traitement, l'intégration d'informations temporelles et l'exploitation de nouvelles architectures pourraient renforcer la robustesse du modèle.
Ces travaux ouvrent ainsi la voie à des méthodes plus efficaces pour la surveillance du ciel, contribuant à la recherche astronomique et à la *Veille Spatiale*.

Références

[1] E. Bertin and S. Arnouts, "Sextractor : Software for source extraction," *Astronomy and Astrophysics Supplement Series*, vol. 117, p. 393–404, June 1996.

[2] P. Rautiainen and A. M. Mel'nik, "N-body simulations in reconstruction of the kinematics of young stars in the galaxy," *Astronomy and Astrophysics*, vol. 519, p. A70, Sept. 2010.

[3] G. Nir, B. Zackay, and E. O. Ofek, "Optimal and efficient streak detection in astronomical images," *arXiv preprint arXiv :1806.04204*, 2018.

[4] C. Hollitt and M. Johnston-Hollitt, "Feature detection in radio astronomy using the circle hough transform," *Publications of the Astronomical Society of Australia*, vol. 29, no. 3, p. 309–317, 2012.

[5] D. e. a. Kim, "Astride : Automated streak detection for astronomical images." Astrophysics Source Code Library, 2016.

[6] B. Zackay, E. O. Ofek, and A. Gal-Yam, "Proper image subtraction—optimal transient detection, photometry, and hypothesis testing," *The Astrophysical Journal*, vol. 830, p. 27, Oct. 2016.

[7] A. e. a. Varela, "Application of convolutional neural networks for streak detection in wide-field images," in *Proceedings of the Asteroid, Comets, and Meteors Conference*, p. E89, 2019.

[8] M. e. a. Pöntinen, "Detection of solar system objects using streakdet in euclid simulated images," *Astronomy & Astrophysics*, vol. XXX, p. XXX, 2020.

[9] G. Jocher, J. Qiu, and A. Chaurasia, "Ultralytics YOLO," Jan. 2023.

[10] O. Parisot, "Streaksyolodataset : Labeled raw astronomical images for streaks detection," *Zenodo*, 2023.

[11] P. Skalski, "make-sense." Online Documentation, 2019.

[12] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 1, pp. 142–158, 2016.

[13] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, *SSD : Single Shot MultiBox Detector*, p. 21–37. Springer International Publishing, 2016.

[14] Ultralytics, "Yolov8 : Real-time object detection." Online Documentation, 2023.

[15] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[16] R. Padilla, S. Netto, and E. Da Silva, "A comparative analysis of object detection metrics with a companion open-source toolkit," in *2020 International Conference on Pattern Recognition (ICPR)*, pp. 2911–2918, 2020.

[17] S. Kashyap, "Mean average precision (map) in object detection." Learn OpenCV Blog, 2018.

[18] B. J. Tatiana Merkulova, "Evaluation framework for image segmentation algorithms," April, 2025.

Architecture Multimodale Robuste : Vers un Système Exécutif Inspiré du Cerveau Humain

Sébastien Grand [1], Aurélie Montarnal [1], Guillaume Pouget[1], Charles Piffault[1], Bruno Mériaux[2], Frederick Benaben[1,3]

[1] Centre de Génie Industriel, IMT Mines Albi
[2] EPSI Radar
[3] ISyE, Georgia Institute of Technology

Résumé

Les modèles d'apprentissage unimodaux, bien que performants, manquent de polyvalence en environnements dynamiques. Cet article de positionnement examine les défis des architectures multimodales : apprentissage des représentations, fusion adaptative et résilience aux variations. Des stratégies de généralisation et d'adaptation de domaine sont explorées ainsi que des approches neuro-symboliques inspirées de Kahneman. Enfin, un système exécutif combinant inhibition sélective, attention modulée et manipulation conceptuelle est proposé, favorisant un raisonnement adaptatif et un résultat interprétable, adapté aux applications critiques en milieux complexes.

Mots-clés

Apprentissage multimodal, Neuro-symbolique, Environnement dynamique, Adaptation, Généralisation, Raisonnement

Abstract

Unimodal learning models, although powerful, lack versatility in dynamic environments. This position paper examines the challenges of multimodal architectures : learning representations, adaptive fusion, and resilience to variation. Strategies for generalisation and domain adaptation are explored, as well as neuro-symbolic approaches inspired by Kahneman. Finally, an executive system combining selective inhibition, modulated attention, and conceptual manipulation is proposed, favouring adaptive reasoning and interpretable results, adapted to critical applications in complex environments.

Keywords

Multimodal learning, Neuro-symbolic, Dynamic environment, Adaptation, Generalization, Reasoning

1 Introduction

Les modèles de vision unimodaux ont certes démontré des aptitudes impressionnantes dans diverses tâches comme la classification d'images [19], la détection d'objets [17] ou en description d'images [16], mais ils demeurent fondamentalement peu versatiles et très spécialisés. Leur capacité à maintenir leurs performances face à des contextes et environnements changeants n'est nullement garantie, créant ainsi une limitation significative pour leur déploiement dans des scénarios réels. C'est dans cette perspective que les modèles multimodaux émergent comme une approche prometteuse, exploitant plusieurs sources de données hétérogènes pour enrichir substantiellement la compréhension du monde. Si l'intégration de nouvelles modalités sensorielles (radar, lidar, données thermiques, etc.) permet d'améliorer la généralisation des modèles, leur déploiement en environnement ouvert soulève néanmoins de nombreux défis complexes. Ces défis sont particulièrement liés aux environnements dynamiques. Un environnement dynamique se caractérise par des distributions de données non stationnaires qui évoluent au fil du temps, en raison de variations météorologiques, d'illuminations, etc. Un environnement dynamique est également marqué par des changements de contexte opérationnel, ainsi que par l'émergence de nouvelles classes non vues lors de l'entraînement. À l'inverse, les cadres statiques supposent que les distributions d'entraînement et de test sont identiques et indépendamment distribuées (i.i.d). Dans un monde réel dynamique, les modèles doivent maintenir une robustesse en continu face aux changements. De plus, la frugalité des modèles devient une nécessité, les architectures actuelles étant très coûteuses en ressources (données, calcul). Ainsi, trois questions fondamentales émergent pour répondre à ces défis : (i) de quelle manière concevoir des espaces de représentation – qu'ils soient communs ou spécifiques – permettant d'optimiser la synergie entre les différentes sources de données ? Ceci implique un équilibre délicat entre redondance informationnelle et spécificité des modalités, tout en développant des mécanismes de fusion de données adaptatifs ; (ii) face à des variations de distribution inhérentes tant à l'hétérogénéité multimodale qu'aux contextes dynamiques, comment garantir la robustesse et le maintien des performances des modèles ? Cette problématique touche au cœur de la fiabilité des systèmes multimodaux en conditions réelles ; (iii) comment élaborer des architectures capables de produire un raisonnement à partir de données multimodales ? Les architectures neuro-symboliques constituent une approche prometteuse, elles visent à combiner un système rapide de

reconnaissance de motifs avec un système plus lent basé sur des règles, pour le raisonnement [22]. Ces approches sont directement inspirées des travaux de Kahneman [11] sur la cognition. De plus, l'introduction d'un système exécutif inspiré des mécanismes de contrôle cognitif du cerveau humain [15], permettrait un raisonnement frugal et intrinsèquement explicable, tout en gérant dynamiquement les problématiques précédentes en fonction du contexte d'opération. Cette exploration des modèles multimodaux en environnement dynamique ouvre ainsi des perspectives prometteuses pour le développement de systèmes d'intelligence artificielle plus robustes, adaptatifs et naturellement alignés avec la complexité du monde réel. Ainsi, les trois premières sections s'intéressent à l'état de l'art (i) de l'apprentissage de représentation et de la fusion en condition de multimodalité et d'environnement dynamique ; (ii) des différentes stratégies pour faire face à la variabilité des données en environnement dynamique ; (iii) des approches neuro-symboliques et leurs avantages sont présentés. Dans cette perspective, la dernière section propose une réflexion sur le développement d'un système exécutif pour les systèmes multimodaux intelligents. Cette section explore les caractéristiques nécessaires pour un tel système, en mettant l'accent sur la frugalité et l'explicabilité des processus cognitifs.

2 Multimodalité

Les modèles multimodaux, bien que visant à combler les lacunes des modèles unimodaux, présentent néanmoins leurs propres défis. L'apprentissage des représentations se complexifie car chaque modalité possède sa propre structure, dimensionnalité et sémantique. Il faut créer un espace représentationnel captant simultanément les spécificités de chaque modalité et leurs corrélations, tout en préservant l'information pertinente et en filtrant le bruit. La fusion de données multimodales soulève également des questions sur le moment et la méthode optimale d'intégration. En environnement dynamique particulièrement, cette fusion doit être adaptative, permettant une sélection optimale des caractéristiques selon le contexte.

2.1 Apprentissage de Représentation

L'apprentissage des représentations peut se décliner en trois axes principaux (Figure 1) : les représentations jointes, coordonnées et hybrides. Cette dernière approche vise à pallier les limitations des deux premières en combinant leurs forces respectives.

L'apprentissage de représentations jointes consiste à fusionner les différentes modalités dès l'encodage afin d'apprendre une représentation commune dans un espace latent partagé [14]. Cette approche permet de capturer les interactions complexes entre les modalités, ce qui est particulièrement bénéfique lorsque les signaux sont fortement corrélés ou complémentaires. Elle est souvent mise en œuvre à l'aide de réseaux multi-branches ou d'opérations de fusion comme la concaténation, l'attention croisée ou la multiplication bilinéaire. Toutefois, cette stratégie peut souffrir d'une perte d'information propre à chaque modalité, notamment lorsque les modalités sont très différentes en nature ou en granularité. Elle est également plus sensible au bruit ou aux données manquantes dans l'une des modalités, ce qui peut dégrader la qualité de la représentation commune. Enfin, il se peut également qu'une modalité prenne le dessus sur les autres durant l'apprentissage, dégradant les capacités de généralisation du modèle [10].

L'apprentissage de représentations coordonnées , contrairement aux représentations jointes, encode chaque modalité à l'aide d'un encodeur spécifique, puis les représentations respectives de chaque modalité sont alignées [16]. L'alignement peut être réalisé à l'aide de pertes contrastives, de correspondances explicites ou d'autres contraintes d'alignement sémantique, favorisant ainsi l'émergence d'une sémantique partagée entre les modalités sans les fusionner directement. Cependant, cette approche peut échouer à aligner efficacement des modalités hétérogènes ou déséquilibrées, surtout sans supervision explicite. Elle néglige aussi les interactions croisées riches, mieux exploitées par des méthodes de fusion plus intégrées. Enfin, les informations spécifiques à chacune des modalités peuvent être effacées par le processus d'alignement.

L'apprentissage de représentations hybrides pallie les limitations des deux approches précédentes. [5] sépare explicitement les caractéristiques partagées et spécifiques à chaque modalité dans deux espaces distincts, via des tâches contrastives, de régularisation ou des mécanismes de traduction. Ainsi, les composantes communes sont projetées dans un espace partagé, tandis que les caractéristiques spécifiques demeurent dans leur espace latent propre. De fait, cette approche permet aussi bien de capturer les synergies inter-modales que de préserver les informations intra-modales.

2.2 La fusion adaptative en environnement dynamique

La variabilité des environnements et la fiabilité des capteurs peuvent affecter la qualité des données, qui peut varier selon les scénarios. Ce constat motive un nouveau paradigme d'apprentissage multimodal : la fusion adaptative. Cette approche vise à s'adapter aux variations de qualité des données et à sélectionner les informations pertinentes pour la tâche. Trois axes de fusion émergent alors : la fusion basée sur des règles, la fusion basée sur la quantification des incertitudes et la fusion basée sur l'attention.

La fusion basée sur des règles est une approche dite *top-down*, qui s'appuie sur des connaissances utilisateur et la tâche à exécuter. Par exemple, le contexte opérationnel (conditions météorologiques, illumination, etc.) peut être exploité afin de guider la fusion en attribuant plus ou moins de poids à une modalité selon le contexte [3]. En environnement changeant, il peut être particulièrement difficile de définir avec exhaustivité les règles de prise de décision.

La fusion basée sur l'incertitude peut être employé dès lors que des règles ou heuristiques ne sont pas simples à expliciter. Ainsi, plusieurs méthodes existent, dont l'une

d'elles propose de se baser sur l'évaluation du niveau d'entropie. Une entropie élevée implique une incertitude élevée et inversement [26]. Cependant, si cette approche est frugale, elle s'avère peu flexible et peut être difficile à mettre en place face à des contextes changeants.

La fusion basée sur l'attention [21] représente une approche prometteuse pour la fusion adaptative. Deux stratégies principales émergent : (i) fusionner les modalités dès l'entrée du modèle en identifiant dynamiquement l'importance de chaque source de données, permettant ainsi de filtrer et d'intégrer les informations les plus pertinentes [23] ; (ii) traiter chaque modalité via des modèles spécifiques (unimodaux) avant de les fusionner après traitement, au moyen d'un mécanisme d'attention [24]. Cette dernière approche permet d'accorder plus d'importance aux caractéristiques fiables qu'aux données bruitées. Cependant, la fonction d'attention telle que présentée dans [21] est très coûteuse en calcul.

La multimodalité répond aux limites des modèles unimodaux mais présente des défis en environnement dynamique. Les approches de représentation (jointes, coordonnées, hybrides) intègrent l'hétérogénéité des données, tandis que les mécanismes de fusion adaptative (règles, incertitudes, attention) optimisent l'utilisation des modalités selon le contexte. Ces avancées favorisent des systèmes plus robustes face aux fluctuations de qualité et aux changements contextuels. La section suivante explore des stratégies de généralisation et d'adaptation essentielles pour maintenir les performances face aux distributions changeantes et aux contextes imprévus.

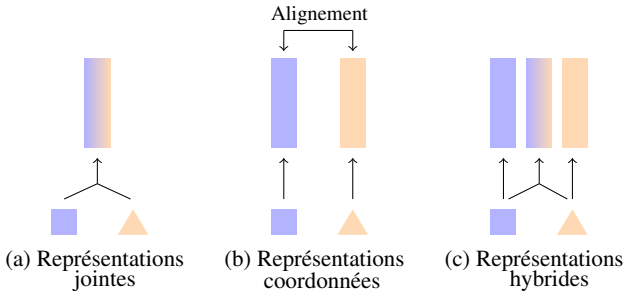


FIGURE 1 – Illustration des differents types de représentations - ■, ▲ : sont deux différentes modalités

3 Variations de distributions : Détecter, Généraliser et s'Adapter

Une hypothèse fondamentale mais en pratique fausse (en partie) en apprentissage machine concerne la distribution des données d'entraînement et de test. En effet, les données d'entraînement et de test sont supposées indépendantes et ayant la même distribution (i.i.d). Ceci est en pratique faux, compte tenu de l'évolution des environnements dynamiques (conditions météorologiques, changements d'illumination, changements d'environnements, etc.). DeLiVER [25] et MUSES [2] sont deux jeux de données présentants

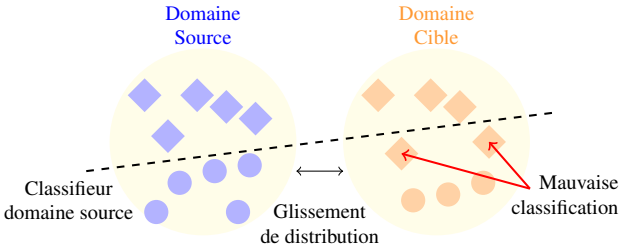


FIGURE 2 – Illustration du problème de glissement de domaine.

de tels challenges de glissement de domaine. La Figure 2 illustre ce phénomène de glissement. Cette section explore différentes stratégies permettant d'être résilient face aux variations de distributions. Elle couvre la détection des données hors distribution, la généralisation de domaine, puis l'adaptation de domaine et enfin l'adaptation à l'inférence. Ces méthodes sont encore peu étudiées dans un cadre multimodal, et sont cruciales pour les systèmes critiques intégrant diverses sources de données.

3.1 Détecter les données hors-distribution

Étant donné la nature changeante des environnements dynamiques, l'apparition de données hors distribution est inévitable. Ces données peuvent résulter d'un glissement de distribution, par exemple, lorsqu'un modèle est entraîné sur des images prises par temps ensoleillé et qu'il est ensuite testé sur des images prises sous la pluie. Elles peuvent également provenir de nouvelles classes jamais vues auparavant, comme un système de surveillance vidéo entraîné à reconnaître des voitures et des camions, mais qui rencontre des bus lors de son déploiement. Il est crucial d'identifier ces données pour éviter de compromettre les décisions du modèle. La plupart des méthodes de détection des données hors distribution ont été développées pour des cas unimodaux. Cependant, [6] soutient que la multimodalité améliore cette détection. En effet, la divergence des prédictions entre modalités est plus marquée pour les données hors distribution que pour celles appartenant à la distribution d'entraînement. L'idée est que, pour les données dans la distribution, les différentes modalités devraient s'accorder sur leur prédiction, tandis que pour les données hors distribution, elles devraient diverger. Pour faire face aux changements de distribution, la généralisation de domaine est une méthode qui permet un apprentissage robuste, renforçant le modèle face à ces variations.

3.2 Généralisation de domaine pour une robustesse accrue

La généralisation de domaine vise à apprendre des représentations invariantes à partir de données provenant de multiples domaines sources. Cela permet de garantir une performance stable face aux variations de distribution, sans nécessiter d'informations spécifiques sur un domaine cible. [5] propose d'apprendre à reconnaître sept activités différentes en utilisant des données RVB et de flux optique (*op-*

tical flow). Ces activités sont réalisées par des humains, des animaux et des personnages de dessins animés, représentant ainsi trois domaines distincts. Grâce à cette diversité, les représentations apprises deviennent plus riches et robustes, améliorant ainsi la capacité de généralisation du modèle.

3.3 S'Adapter à la nouveauté et aux changements

L'adaptation de domaine ajuste un modèle pré-entraîné sur un domaine source à un domaine cible spécifique. Contrairement à la généralisation de domaine, l'adaptation de domaine nécessite des données du domaine cible pour réduire l'écart de distribution. L'adaptation de domaine non supervisée et l'adaptation sans accès aux données sources sont deux approches réalistes dans des contextes réels.

L'adaptation de domaine non supervisée (UDA) est une approche pragmatique où les données du domaine cible sont disponibles mais non étiquetées. Dans les environnements réels, les conditions peuvent évoluer rapidement, rendant les données étiquetées obsolètes. L'UDA permet au modèle de s'adapter à ces changements sans nécessiter de nouvelles données étiquetées. Dans le contexte de la reconnaissance d'émotion, [27] démontre que lors de l'adaptation de domaine, les différences de distribution entre les domaines source et cible rendent difficile l'équilibrage de l'alignement des domaines entre les différentes modalités. La méthode proposée permet d'apprendre des représentations indépendantes optimales pour chaque modalité. Ensuite, elle ajuste dynamiquement les poids des différentes modalités pendant l'entraînement, garantissant ainsi un équilibre de chaque modalité dans les représentations.

L'Adaptation de domaine non-supervisée sans accès au domaine source (SF-UDA) constitue une approche plus exigeante où l'adaptation utilise uniquement un modèle pré-entraîné et des données cibles non annotées, particulièrement utile pour les données sensibles ou soumises à des restrictions d'accès. [12] exploite un classifieur préalablement entraîné sur le domaine source pour générer des *"impressions de domaine"*, c'est-à-dire des échantillons synthétiques représentant la distribution du domaine source. Ces échantillons synthétiques sont ensuite utilisés pour adapter le modèle au domaine cible sans nécessiter l'accès aux données sources réelles.

3.3.1 Adaptation à l'inférence

Les approches d'adaptation de domaine nécessitent souvent un entraînement préalable, ce qui limite leur flexibilité face aux changements rapides de distribution. L'adaptation au temps de test ou à l'inférence permet d'ajuster un modèle durant l'inférence, sans réentraînement complet ni accès préalable aux données cibles. En multimodalité, [24] module l'attention entre les modalités de façon auto-adaptative pour permettre une fusion fiable et adopte une fonction objectif évaluant l'entropie, pour une adaptation multimodale robuste. Le problème de biais de fiabilité dans les scénarios multimodaux consiste à gérer les divergences d'informations entre différentes modalités résultant des variations de distribution intra-modales.

En conclusion, gérer les variations de distribution est crucial pour des modèles robustes évoluant en environnements dynamiques, que ce soit par le biais de la généralisation ou de l'adaptation.

4 Raisonnement : Au delà de la reconnaissance de motifs

Les modèles d'intelligence artificielle actuels excellent dans la reconnaissance de motifs complexes à travers différentes modalités, mais cette capacité révèle aussi leurs limites fondamentales. Ces modèles maîtrisent l'apprentissage de corrélations statistiques entre données d'entrée et de sortie, mais peinent à établir des relations causales. À l'opposé, le raisonnement humain implique souvent la décomposition de problèmes complexes en sous-problèmes plus simples, comme lors de la planification. Les récents développements des grands modèles de langage (LLM) ont ravivé le débat sur leurs capacités de raisonnement, sans qu'un consensus n'émerge. En s'appuyant sur la théorie du double processus de Kahneman [11], ces systèmes semblent principalement ancrés dans un mode de pensée rapide et intuitif (Système 1), mais largement dépourvus des capacités de réflexion délibérée et analytique (Système 2) caractéristiques du raisonnement humain avancé. Plus récemment, [8] décrit un *Système 3* associant les deux systèmes de Kahneman à un système métacognitif. [15] propose une approche neuro-symbolique bio-inspirée couplant les 3 systèmes précédemment définis afin de créer un modèle capable d'adaptation et efficace face à des situations nouvelles.

4.1 Du traitement de l'information au raisonnement structuré

4.1.1 Désenchevêtrement, compositionalité et causalité

Pour être déployés efficacement en environnement réel, les modèles multimodaux doivent dépasser la simple reconnaissance de motifs et accéder à une compréhension structurelle des données. Trois propriétés complémentaires constituent le socle d'un raisonnement avancé, [22, 15] énoncent et discutent de celles-ci, au travers de *l'apprentissage de représentations désenchevêtrée*s, de la *compositionnalité* et de la *causalité*.

Le désenchevêtrement des représentations [7] permet d'identifier et d'isoler des facteurs indépendants dans les données, facilitant l'interprétabilité et la robustesse face aux variations non pertinentes.

La compositionalité offre la capacité de manipuler des concepts [20] et de les recombiner de manière flexible, améliorant la généralisation et les performances du modèle.

Les relations causales [18] dépassent les simples corrélations statistiques pour comprendre les liens de cause à effet entre variables, renforçant la robustesse face aux biais et la capacité de généralisation.
Ces propriétés complémentaires constituent le socle d'un raisonnement avancé, capable d'opérer efficacement dans des environnements complexes et dynamiques. La section

suivante discute de la théorie de Kahneman et de sa possible application en apprentissage automatique [1].

4.1.2 Architecture cognitive : Systèmes 1, 2 et 3

La théorie cognitive de Kahneman, articulée autour de deux systèmes complémentaires, offre un cadre conceptuel pertinent pour l'apprentissage automatique. Le Système 1, rapide et automatique, s'apparente aux modèles de reconnaissance de motifs comme les réseaux convolutifs. Le Système 2, plus lent et délibéré, se rapproche du raisonnement symbolique. Ces systèmes fonctionnent en synergie : lorsque le *système 1* rencontre des difficultés, le *système 2* prend le relais. Les découvertes du *système 2* peuvent ensuite être intégrées au *système 1* par la pratique répétée, transformant progressivement une tâche complexe en processus automatique. Plus récemment, [9] met en lumière un troisième système de contrôle cognitif permettant, par un mécanisme d'inhibition et d'activation, le raisonnement déductif face à des tâches nouvelles et complexes. La capacité à générer des idées nouvelles exige d'abord d'inhiber les solutions spontanées issues du *système 1*. Par exemple, [4] montre par l'analyse de l'activité cérébrale que les adolescents et adultes emploient le contrôle inhibitif pour comparer deux fractions. [15] pose que le développement d'un tel contrôle cognitif couplé à des approches neuronales et symboliques, constitue une voie bio-inspirée prometteuse pour des machines capables de raisonner face à des situations nouvelles et complexes. Les discussions dans les sections suivantes s'appuient sur ces travaux et propositions.

4.2 Approches Neuro-Symboliques : vers une intégration fonctionnelle

Les approches neuro-symboliques offrent un cadre prometteur pour implémenter ces capacités de raisonnement avancées. Elles combinent la puissance d'apprentissage statistique des réseaux neuronaux avec la précision et l'interprétabilité des systèmes symboliques. Concrètement, cela se traduit par des architectures hybrides [13] où les réseaux neuronaux extraient des représentations pertinentes des différentes modalités, tandis que les modules symboliques (graphes de connaissances, logique propositionnelle, logique du premier ordre, etc.) manipulent ces représentations selon des règles formelles pour produire des inférences explicables et des résultats interprétables. Cette hybridation facilite l'incorporation de connaissances préalables et de contraintes logiques dans le processus d'apprentissage, améliorant la généralisation et la robustesse face aux changements de distribution.

5 Vers un système exécutif pour les systèmes multimodaux intelligents

Le cerveau humain structure sa perception de manière abstraite et raisonne en manipulant des concepts sémantiques de haut niveau. Lorsqu'il perçoit une scène, il peut sélectionner de façon adaptative les sources de données importantes à traiter et composer son raisonnement en fonction de la tâche à effectuer.

Actuellement, un tel système exécutif supervisant des sous-modèles de façon adaptative, en sélectionnant les sources de données pertinentes pour produire une représentation robuste face aux changements, reste peu exploré.
Ainsi, le développement de modèles plus intelligents, à l'image du cerveau humain, passe par la mise en place d'un système méta-cognitif de haut niveau.
Ce système exécutif :

1. Permet de contrôler le raisonnement et de sélectionner de manière optimale les sources de données en fonction de la tâche à effectuer, car toutes les sources de données ne sont pas pertinentes dans certaines situations.
2. Est versatile, capable de s'adapter dans des situations changeantes tout en conservant ses performances.
3. Doit pouvoir produire un résultat basé sur des concepts sémantiques de haut niveau, de sorte que ce même résultat soit interprétable.

Une approche de traitement hiérarchique semble appropriée pour un modèle multimodal plus intelligent.

Des modèles spécifiques légers permettent un traitement préliminaire appliqué à toutes les sources de données, permettant une évaluation rapide de leur pertinence potentielle.

Un mécanisme de contrôle inhibitoire bloque complètement le traitement approfondi des sources jugées non pertinentes, permettant d'économiser considérablement les ressources computationnelles. Si cela n'est pas possible, une pondération respective des modalités pourra être envisagée de manière à accorder plus d'importance aux modalités les plus robustes en fonction du contexte.

Un mécanisme d'attention sélective filtre, parmi les sources pertinentes, les caractéristiques spécifiques au sein de chaque modalité, conservant uniquement les attributs informatifs qui sont projetés dans un espace de représentation unifié.

Extraction de concepts sémantiques À partir de la représentation unifiée, une extraction de concepts sémantiques est effectuée. Ces concepts servent à établir des liens entre les objets, facilitant ainsi la création de structures compositionnelles.
Cette démarche méthodique permet d'éviter une exploration exhaustive de l'ensemble des sources, réduisant ainsi de manière significative le coût computationnel. Le principe de sélection d'une ou plusieurs modalités, sans recourir à l'ensemble complet, confère à l'approche une robustesse accrue face aux modalités manquantes ou corrompues. Enfin, l'analyse conceptuelle des résultats offre à l'utilisateur un moyen d'en évaluer la fiabilité de manière interprétable.

6 Conclusion

Cet article se positionne à l'intersection des enjeux liés à la multimodalité et à l'évolution des modèles en environnement dynamique. Il défend une approche bio-inspirée, fondée sur des mécanismes cognitifs du

cerveau, comme solution prometteuse pour intégrer efficacement des sources de données hétérogènes. Les travaux futurs porteront sur l'implémentation des différentes composantes énoncées en section 5 ainsi que sur leur évaluation expérimentale. En conjuguant apprentissage automatique et inspiration cognitive, cette démarche ouvre des perspectives prometteuses pour la conception de systèmes intelligents capables de s'adapter aux exigences des environnements en constante évolution.

Note : Certaines parties ont été réécrites à l'aide de modèles de langages.

Références

[1] Grady Booch and al. Thinking Fast and Slow in AI. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[2] Tim Brödermann and al. Muses : The multi-sensor semantic perception dataset for driving under uncertainty. In *European Conference on Computer Vision (ECCV)*, 2024.

[3] Tim Broedermann and al. CAFuser : Condition-aware multimodal fusion for robust semantic perception of driving scenes, 2025.

[4] Mathieu Cassotti and al. Inhibitory control as a core process of creative problem solving and idea generation from childhood to adulthood. *New directions for child and adolescent development*, 2016.

[5] Hao Dong and al. SimMMDG : A simple and effective framework for multi-modal domain generalization. In *Advances in Neural Information Processing Systems*, 2023.

[6] Hao Dong and al. MultiOOD : Scaling out-of-distribution detection for multiple modalities. In *Advances in Neural Information Processing Systems*, 2024.

[7] Irina Higgins and al. Towards a definition of disentangled representations, 2018.

[8] Olivier Houdé and al. Measuring inhibitory control in children and adults : brain imaging and mental chronometry. *Frontiers in Psychology*, 2014.

[9] Olivier Houdé and al. Evidence for an inhibitory-control theory of the reasoning brain. *Frontiers in Human Neuroscience*, 2015.

[10] Yu Huang and al. Modality competition : What makes joint training of multi-modal network fail in deep learning ? (Provably). In *Proceedings of the 39th International Conference on Machine Learning*. PMLR, 2022.

[11] Daniel Kahneman. *Thinking, fast and slow*. 2011.

[12] V. Kurmi and al. Domain impression : A source data free domain adaptation method. *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021.

[13] Liulei Li and al. Logicseg : Parsing visual semantics with neural logic learning and reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[14] Jiasen Lu and al. Vilbert : Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, 2019.

[15] Aurelie Montarnal and al. Future AI in crisis management : Proposing a bio-inspired, neuro-symbolic architecture. In *Proceedings of the 21st ISCRAM Conference*, 2024.

[16] Alec Radford and al. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*. PMLR, 2021.

[17] Joseph Redmon and al. You only look once : Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[18] Bernhard Schölkopf and al. Toward Causal Representation Learning. *Proceedings of the IEEE*, 2021.

[19] Karen Simonyan and al. Very deep convolutional networks for large-scale image recognition. In *The Forth International Conference on Learning Representations*, 2015.

[20] Adam Stein and al. Towards compositionality in concept learning. In *Proceedings of the 41st International Conference on Machine Learning*. PMLR, 2024.

[21] Ashish Vaswani and al. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

[22] Wenguan Wang and al. Towards data-and knowledge-driven ai : A survey on neuro-symbolic computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[23] Yikai Wang and al. Multimodal Token Fusion for Vision Transformers . In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[24] Mouxing Yang and al. Test-time adaption against multi-modal reliability bias. In *The Twelfth International Conference on Learning Representations*, 2024.

[25] Jiaming Zhang and al. Delivering arbitrary-modal semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[26] Xiaohui Zhang and al. Multimodal Representation Learning by Alternating Unimodal Adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[27] Xinxin Zhang and al. Amanda : Adaptively Modality-Balanced Domain Adaptation for Multimodal Emotion Recognition. In *Findings of the Association for Computational Linguistics ACL 2024*.

Session HNIA-APIA : Réflexions et perspectives sur les enjeux de traçabilité des données pour les Humanités numériques

Repenser les collections patrimoniales par le prisme de l'IA 2025

Emmanuelle Bermès[1], Marion Charpier[1]

[1] École nationale des chartes - PSL, Centre Jean-Mabillon

6 février 2025

Résumé

Le projet TORNE-H explore l'intégration de l'IA dans le traitement documentaire et scientifique des collections muséales. Centré sur le fonds Henrot du musée des Arts décoratifs (MAD), il vise à optimiser l'inventaire, l'indexation et la valorisation des œuvres grâce à des techniques avancées d'analyse d'images. Porté par MAD, l'École nationale des chartes et la BnF, il cherche à repenser la coopération humain-machine et à développer des processus reproductibles pour d'autres institutions patrimoniales.

Mots-clés

Vision par ordinateur, Collections patrimoniales, Musée, Coopération humain-machine

Abstract

The TORNE-H project explores the integration of AI in the documentary and scientific processing of museum collections. Focused on the Henrot collection at the musée des Arts décoratifs (MAD), it aims to optimize the inventory, indexing, and promotion of artworks through advanced image analysis techniques. Led by MAD, the École nationale des chartes, and the BnF, the project seeks to rethink human-machine collaboration and develop reproducible processes for other heritage institutions.

Keywords

Computer vision, Heritage collections, Museum, Human-machine cooperation

1 Introduction

Six cent quarante-trois années : c'est le temps estimé nécessaire pour réaliser l'inventaire et la description des près de 700 000 œuvres non inventoriées du seul département des dessins, papiers peints et photographies du Musée des Arts Décoratifs de Paris (MAD). Face à ce constat, le musée a pris contact en 2023 avec l'équipe Humanités numériques de l'École nationale des chartes (ENC) afin d'examiner si l'intelligence artificielle (IA) et la vision par ordinateur pouvaient contribuer à améliorer la découvrabilité des collections et à répondre à l'obligation de l'inventaire réglementaire.

De cette rencontre est né un projet commun, réunissant l'expertise du MAD sur les collections, l'ENC comme force de travail alliant connaissances en histoire de l'art et compétences en vision par ordinateur, ainsi que la Bibliothèque nationale de France (BnF) pour son savoir-faire en matière de structuration et de traitement de données patrimoniales[2].

Cette première collaboration fructueuse a incité les différentes institutions à poursuivre leurs travaux dans un cadre élargi, donnant naissance en 2024 au projet TORNE-H, financé pour un an par le Ministère de la Culture. En s'appuyant sur la maturité actuelle des modèles de vision par ordinateur, le projet vise à étudier les cas d'usage spécifiques du musée et à analyser comment l'automatisation de la description d'images peut influencer l'organisation des compétences et des tâches au sein de l'institution.

Dans cet article, nous présentons les premiers résultats obtenus avec la collection Jean Royère, les développements prévus avec la collection Henrot, ainsi que les perspectives pour l'opérationnalisation du processus à travers différents cas d'usage identifiés.

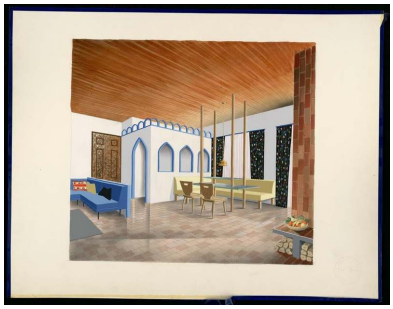
2 La collection Royère : l'art du *fine-tuning*

L'origine du projet d'IA au sein du MAD repose sur la collection spécifique du designer Jean Royère, dont le fonds d'archives, conservé au sein du département des dessins, papiers peints et photographies, a été donné par l'artiste au musée en 1980.

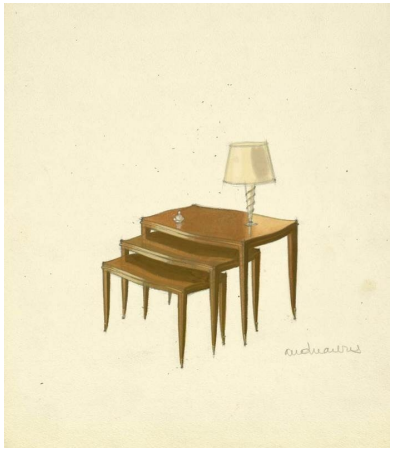
2.1 A propos de la collection

Jean Royère (1902-1981) est l'une des figures de proue de l'avant-garde française des années 1950. Artiste autodidacte, il quitte un emploi stable à l'âge de 29 ans pour devenir architecte d'intérieur, avec sa propre agence à partir de 1942[5, 4]. Anticipant l'essor du design biomorphique, des séries comme les chaises et fauteuils Œuf ou le canapé Ours-Polaire se vendent aujourd'hui plusieurs centaines de milliers d'euros aux enchères, atteignant parfois le million.

Le fonds de dessins est constitué d'environ 18 000 objets répartis en quatre grandes catégories : les grandes gouaches (environ 400), les petites gouaches (1400), les calques de vue d'ensemble, classés par commanditaire (5500) et les calques d'exécution (plus de 10 500) (Fig. 1). Les grandes gouaches représentent le plus souvent des projets complets pour une pièce spécifique, permettant de rendre compte de la manière dont Jean Royère concevait l'articulation entre les différents objets. Les petites gouaches se présentent majoritairement sous la forme d'un objet unique réalisé au



(a) Exemple de grande gouache



(b) Exemple de petite gouache

(c) Exemple de calque d'exécution

FIGURE 1 – Exemples issus du fonds de dessins Royère

crayon de graphite et à la gouache. Parfois les variations d'un même objet sont représentées sur une seule gouache. C'est en particulier le cas des lampes dont le nombre d'abat-jours peut varier. Les petites gouaches peuvent également représenter une mise en contexte de l'objet principal, un bureau avec la chaise qui l'accompagne et un luminaire, par exemple. Les calques d'exécution sont constitués de deux grandes catégories. La première représente les différents modèles (fauteuils, canapés, luminaires, etc.) avec l'ensemble des dimensions et des vues variées : de face de profil et de ¾. La seconde catégorie est constituée des dessins de projet en cours de conception. Ces descriptions techniques des objets sont exécutées au crayon de graphite.

À l'heure où nous écrivons ces lignes, aucune de ces 18 000 pièces n'est décrite dans la base de données des collections du musée, rendant la recherche d'un design spécifique particulièrement laborieuse. La complexité des motifs, l'absence de descriptions détaillées et le volume massif d'éléments rendent l'identification manuelle longue et fastidieuse. Ainsi, la collection Royère constitue un cas d'utilisation très spécifique, centré sur les besoins des conservateurs du musée, et un corpus à la fois suffisamment homogène et particulièrement exigeant pour les modèles de vision par ordinateur les plus récents. Alors que ces modèles excellent dans la description globale de dessins en couleur, tels que les grandes gouaches, identifier un meuble spécifique au sein de cet ensemble demeure un véritable défi. Les calques d'exécution sont encore plus rares pour les modèles d'IA, qui ont été formés sur des bases de données iconographiques standard, issues d'Internet ou de corpus tels que ImageNet. Par conséquent, pour générer des descriptions réellement utiles dans le cadre du cas d'usage Royère, un réglage fin est nécessaire.

2.2 Un processus itératif

La nature du fonds de dessins Royère et les contraintes spécifiques dont il fait l'objet nous ont amenées à préférer l'utilisation de modèle supervisé. Afin de créer des modèles les plus robustes qui soient, nous avons choisis d'utiliser l'algorithme YOLO (You Only Look Once) créé par Ultralytics, dans sa version 9. Ce modèle a été conçu pour effectuer des tâches de détection d'objets : identifier et localiser des objets spécifiques dans une image. Le modèle génère des boîtes de délimitation autour des objets détectés, ce qui permet une classification rapide et efficace de chaque élément visible. Grâce à ce modèle, nous sommes en mesure d'identifier avec précision les meubles figurant dans chacun des dessins de Royère, de créer des boîtes englobantes pour isoler chaque objet dans l'image et de générer des métadonnées décrivant les objets, ou annotations, sur la base d'une ontologie simple de 23 concepts, qui répond aux contraintes techniques de la vision par ordinateur, tout en s'appuyant sur la classification créée par l'artiste lui-même.

Le traitement du fonds Royère, depuis la constitution des *datasets* successifs, suit un *workflow* conçu pour simplifier le traitement et permettre l'entraînement de modèles robustes de manière itérative. La première phase d'annotation du corpus a permis la création d'une vérité terrain riche de 159 images. Afin d'optimiser cette tâche, nous avons exploré différentes stratégies, notamment la déformation d'images, dans le but d'augmenter la taille des ensembles de données de vérité terrain grâce à des méthodes automatisées. La transformation de perspective applique un algorithme qui simule le type de déformation induit par la perspective, un aspect particulièrement pertinent pour notre corpus, notamment lorsqu'il s'agit d'identifier un meuble dans une vue de pièce. Lors de l'entraînement, nous avons choisi de ne pas utiliser un jeu de données de test, préférant déployer le modèle sur un nouveau jeu de données d'une centaine d'images. Les corrections apportées sont récupérées afin de générer les métriques d'évaluation du modèle et d'établir une nouvelle vérité terrain : les données corrigées servant ainsi à créer un nouveau *dataset*. Ce *workflow* permet de réduire le temps de traitement et de relancer de manière itérative de nouvelles sessions d'entraînement jus-

Classes	Nb TP	Nb FP	Nb FN	Precision	Rappel	Score F1	Support
Armchair	60	18	36	0.769	0.625	0.690	96
Bar	3	4	2	0.43	0.60	0.50	5
Bed	24	10	10	0.706	0.706	0.706	34
Bookcase	36	9	15	0.800	0.706	0.750	51
Buffet	21	11	13	0.656	0.618	0.636	34
Chair	88	14	42	0.863	0.677	0.759	130
Fireplace	13	3	5	0.812	0.722	0.765	18
Lamp	96	6	42	0.941	0.696	0.800	138
Mirror	2	1	5	0.67	0.29	0.40	7
N°_Model	58	4	6	0.935	0.906	0.921	64
Rectangle_stamp	28	1	0	0.966	1.000	0.982	28
Round_stamp	62	0	1	1.000	0.984	0.992	63
Sofa	26	7	24	0.788	0.520	0.627	50
Sponsor	75	4	5	0.949	0.938	0.943	80
Stool	9	4	3	0.69	0.75	0.72	12
Table	96	14	46	0.873	0.676	0.762	142
Wardrobe	10	0	7	1.000	0.588	0.741	17
curtain	30	1	11	0.968	0.732	0.833	41
plant	83	3	14	0.965	0.856	0.907	97
rug	50	5	7	0.909	0.877	0.893	57
Overall	870	119	294	0.8797	0.7474	0.8082	1164

FIGURE 2 – Model inference results `Royere_202231030_x_i640_e100_b8_w24`

qu'à l'obtention d'un modèle suffisamment robuste.

2.3 Premiers résultats et perspectives

Le traitement des images à l'aide de YOLOv9 a permis un enrichissement automatisé des métadonnées, tout en réduisant significativement le temps d'indexation. Nous présentons ici les métriques issues du projet : True Positive / False Positive / False Negative / Score 1 et support (ensemble des éléments réellement présents - TP + FN). Les mesures obtenues (rappel, précision et score F1) lors des corrections nous ont permis de suivre l'évolution du modèle, tant pour chaque classe déterminée que pour l'ensemble des éléments. Comme le montre le tableau d'évaluation ci-dessus (*Fig. 2*), même après plusieurs itérations d'ajustement, certains objets restent difficiles à identifier. Cela concerne en particulier les objets les moins représentés, mais aussi ceux dont la forme est similaire, comme le met en évidence la matrice de confusion (*Fig. 3*) : le modèle peine à distinguer une chaise d'un fauteuil ou à détecter la présence d'une lampe. En revanche, certains éléments sont très bien prédits, notamment la présence de texte dans l'image.

Le degré de précision que nous recherchons pour répondre aux attentes du personnel du musée n'est pas encore atteint. Certaines classes, en particulier celles essentielles à la compréhension de la collection, échappent encore à une détection précise.

3 Pistes alternatives : une exploration des modèles de vision par ordinateur les plus récents

Alors que les modèles de vision par ordinateur deviennent de plus en plus précis et que de nouveaux modèles multimodaux voient le jour, la nécessité d'un tel processus de mise au point manuel, long et fastidieux, peut être remise en question. Pourrions-nous nous appuyer sur des modèles plus puissants et plus récents pour éviter cette étape ? L'exploration et l'évaluation de différents modèles est l'une des tâches de notre projet. Afin d'appliquer le prototype de

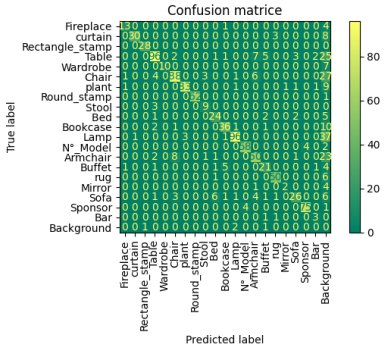


FIGURE 3 – Confusion matrix.

workflow de Royère à d'autres collections, nous voulons adapter l'approche itérative pour tenir compte de l'évolution des modèles d'IA.

3.1 CLIP : une approche zéro-shot

L'approche que nous avons développée avec l'utilisation de YOLO permet de générer des mots-clés utiles pour l'indexation ou le signalement des collections, mais ne permet pas, dans notre implémentation actuelle, une exploration visuelle ou une interrogation en langage naturel de la collection. Or, le personnel scientifique du musée a exprimé le besoin de parcourir les résultats de manière autonome, en visualisant les objets détectés et en formulant des requêtes simples.

Pour répondre à ces attentes, nous avons exploré une approche complémentaire avec CLIP (Contrastive Language-Image Pre-training), un modèle développé par OpenAI qui associe images et descriptions textuelles en langage naturel[8]. CLIP permet d'interroger une collection par des requêtes simples du type « chaise rouge » ou « applique murale », sans nécessiter d'annotations manuelles ni d'entraînement supervisé. Ce modèle, basé sur des paires d'images et de légendes, est dit « zéro-shot », ce qui signifie qu'il peut effectuer une tâche sans formation spécifique, grâce à l'utilisation d'*embeddings*. En représentant les vecteurs associés à la fois aux images et aux légendes dans un espace multidimensionnel, CLIP devient capable d'associer correctement des images et des descriptions textuelles qu'il n'a jamais vues lors de son entraînement.

Afin de tester cette approche dans le cadre du projet Royère, nous avons pu réutiliser un prototype d'implémentation de CLIP, doté d'une interface web, créé à la BnF. Cette approche, bien que prometteuse, présente des limites liées encore une fois à la complexité de la collection Jean Royère et aux besoins d'interrogation très spécifiques du personnel du musée : certains objets échappent encore à une détection précise et la recherche des styles spécifiques que Jean Royère a mis en œuvre dans ses dessins sous forme de motifs, tels que « fauteuil œuf » ou « style croisillon », ne donne pas les résultats escomptés. Cela témoigne de la nécessité d'affiner les modèles, aussi puissants soient-il, dès lors qu'il s'agit de données patrimoniales spécifiques.

3.2 Autres perspectives

Au-delà des déformations en perspective, d'autres méthodes de transformation d'image, comme la transformation affine, peuvent être testées afin d'améliorer la robustesse des modèles. Nous explorons également l'intégration de déformations dynamiques au sein de notre workflow, dans le but d'augmenter davantage le volume de données d'apprentissage et d'optimiser l'efficacité de l'entraînement.

D'autres modèles comme Florence2 de Microsoft[11], SAM2[9] et DinoV2[6] de Meta, ainsi que des combinaisons de différents modèles, sont des pistes potentielles pour améliorer encore le processus. Dans le cadre de TORNE-H, nous souhaitons à la fois évaluer les résultats de ces modèles sans ajustement spécifique (fine-tuning) et itérer avec notre processus de fine-tuning. Ainsi, nous disposerons d'un panel de modèles à comparer et pourrons compléter notre étude par une dernière étape : confronter les résultats de l'annotation automatisée à une évaluation qualitative réalisée par le personnel du musée, une approche similaire à celle entreprise par la Bibliothèque Nationale d'Australie[3].

4 TORNE-H : inverser les processus

TORNE-H signifie « Traitement d'Objets par Reconnaissance Numérique en Environnement Humain » et constitue également l'anagramme inversée du nom « Henrot ». Ce projet vise à poursuivre le travail amorcé avec la collection Royère en l'étendant à un autre fonds conservé par le département des Dessins, Papiers peints et Photographies, en mettant un accent particulier sur l'opérationnalisation de l'IA dans le musée et son impact sur les tâches manuelles effectuées par le personnel. Notre hypothèse est que l'intégration de l'IA dans le musée permettrait d'inverser le processus de numérisation et de description des collections.

4.1 Le fonds Henrot, éléments pour une généralisation

Paul Henrot (1908-1986) était un photographe d'architecture moderne et d'architecture d'intérieur, ayant développé un style unique. Artiste oublié, il est pourtant central dans le domaine de l'architecture moderne et du design d'intérieur, comme en témoignent les noms de ses commanditaires, au premier rang desquels Le Corbusier, mais aussi Jean Prouvé, Michel Roux-Spitz, ainsi que des entreprises telles que Saint-Gobain, Citroën, EDF, Kodak, etc. Il a réalisé de nombreux reportages lors d'événements comme la Foire de Paris, le Salon des arts ménagers, le Salon des artistes décorateurs, l'Exposition internationale de 1937 à Paris, l'exposition « Surréalisme » rue de la Boétie en 1938, l'Exposition internationale de New York en 1939, ainsi que la Libération de Paris en 1944. Henrot a également collaboré avec des magazines tels que *Architecture d'aujourd'hui, Décor d'aujourd'hui, Plaisir de France, Urbanisme*. Il a travaillé avec des créateurs comme Simone Prouvé, Louis Sognot, André Arbus, César et Raphaël. Sa formation en architecture a affiné son regard. Ses photographies

FIGURE 4 – Paul Henrot, *Escalier de l'IRSID (Institut de recherche de la sidérurgie) à Saint-Germain-en-Laye*, 1953, négatif souple. Don Marcelle Henrot, 1987 © DR © Photo : MAD Paris / Christophe Dellière

se distinguent par une expression particulière reposant sur des angles et des points de vue originaux (contre-plongées, compositions diagonales, goût du détail parfois proche de l'abstraction)[1].

En 1987, un fonds de 430 000 photographies de Paul Henrot a été donné au MAD par sa veuve, avec les droits d'exploitation. Ce fonds, exceptionnel par son volume, sa complétude et les sujets photographiés, comprend plus de 12 500 rouleaux de négatifs majoritairement en noir et blanc et quelques-uns en couleur. Il couvre cinquante ans d'histoire des arts et de la vie culturelle, de 1933 à 1982. Ce fonds est unique : seuls quelques tirages et négatifs sont conservés dans d'autres institutions.

Travailler avec l'IA et la vision par ordinateur sur cette collection photographique soulève des défis spécifiques : la détection de motifs et de formes récurrents, mettant en lumière la naturalisation de l'architecture ; l'exploration de l'esthétique organique dans les photographies de Paul Henrot ; l'identification et la géolocalisation de monuments historiques.

4.2 Opérationnaliser l'IA dans le musée

Le projet TORNE-H explore également une nouvelle manière d'intégrer l'IA dans les musées en examinant les transformations introduites par ces technologies et en développant des flux de traitement automatisés adaptés au contexte patrimonial. En analysant un large corpus de photographies, nous souhaitons démontrer la valeur ajoutée de l'IA et explorer un nouveau paradigme, qui remet en question le processus traditionnel de conservation en organisant le traitement computationnel des collections avant le travail scientifique traditionnel, et en fournissant une description automatisée initiale avant l'analyse humaine.

4.3 Collecte des cas d'usage

L'une des premières tâches du projet TORNE-H a été de collecter des cas d'usage formels auprès du personnel du musée. Nous avons utilisé une adaptation du *Library of Congress AI Planning Framework*[7] pour mener des entretiens avec six acteurs clés du musée, dont les rôles étaient

les suivants :

— Responsable d'activité ;
— Chargé(e)s de conservation ;
— Gestion des acquisitions, inventaire, dépôts et coordination de l'inventaire ;
— Direction générale du MAD ;
— Administration des bases de données.

Ces entretiens ont permis de mettre en avant les limitations des pratiques de travail actuelles. En particulier, ils ont révélé l'existence d'un passif important d'œuvres encore non inventoriées ou non numérisées, ce qui limite leur connaissance, freine leur consultation et entrave leur diffusion sur le portail des collections. Cette situation a fait émerger, de manière unanime parmi les personnes sondées, une volonté d'utiliser l'IA pour faciliter l'inventaire dans la base de données du musée, considérée comme une avancée bénéfique.

5 Applications potentielles de l'IA

Les entretiens menés ont révélé de nombreux cas d'usage pour l'intégration de l'IA au sein du musée. Afin de mieux structurer ces besoins et de comprendre les attentes des professionnels, nous les avons regroupés en deux grandes catégories.

5.1 Inventaire

L'inventaire est l'un des domaines où l'IA pourrait avoir un impact significatif en facilitant l'indexation des œuvres et la gestion des archives historiques. Son utilisation permettrait d'optimiser le travail des équipes tout en améliorant l'accessibilité des collections.

Aide à l'indexation L'IA pourrait automatiser certaines tâches fastidieuses en proposant une pré-indexation des œuvres. Elle pourrait générer des descriptions sommaires et attribuer des mots-clés pertinents afin de faciliter leur identification et leur recherche dans la base de données. Une autre possibilité réside dans la création automatique de fiches d'inventaire, permettant d'attribuer systématiquement un numéro à chaque œuvre et de réduire le temps consacré aux tâches administratives. Enfin, l'IA pourrait enrichir ces fiches en intégrant des informations issues de bases de données existantes ou de documents historiques.

Compléter l'historique des œuvres L'IA pourrait également contribuer à l'amélioration des connaissances sur les œuvres en facilitant la numérisation et l'exploitation des archives historiques. Ce processus inclurait la numérisation des registres conservés et leur intégration dans la base de données afin de centraliser les informations aujourd'hui dispersées sous format papier. L'utilisation de modèles de reconnaissance de texte manuscrit (*Handwritten Text Recognition - HTR*) permettrait d'extraire automatiquement des informations clés, telles que les noms, dates ou références, pour retrouver les donateurs et compléter l'historique des collections[10]. En croisant ces données avec d'autres sources, l'IA pourrait ainsi améliorer la fiabilité des informations disponibles.

5.2 Exploration visuelle des collections

L'IA pourrait considérablement enrichir l'exploration des collections en facilitant l'identification d'œuvres similaires, en optimisant la gestion des médias et en assistant les équipes dans leurs recherches de provenance.

Rattacher les métadonnées L'exploitation de bases de données vectorielles permettrait d'améliorer la catégorisation des œuvres et l'enrichissement des métadonnées. La reconnaissance visuelle offrirait la possibilité d'identifier des œuvres similaires à partir d'images, tandis que l'IA pourrait associer automatiquement des informations telles que la date ou la période de création, l'artiste ou le lieu de production en comparant ces images avec des bases de données existantes.

Rattacher les médias L'IA pourrait rationaliser la gestion des médias liés aux collections en évitant la duplication d'informations inutiles. L'identification et la suppression des doublons d'images présentes dans la base de données et sur les serveurs internes, grâce à des algorithmes dédiés, optimiseraient l'organisation des fichiers multimédias. Une meilleure gestion des ressources assurerait également une interconnexion fluide entre la base de données du musée et le portail de diffusion des collections.

Aide à la recherche de provenance L'IA pourrait renforcer la fiabilité des recherches de provenance, un enjeu central pour la gestion des acquisitions et la lutte contre le trafic d'œuvres d'art. Lors des campagnes d'acquisition, elle faciliterait la vérification des œuvres en s'appuyant sur des bases de données internationales recensant les objets volés ou spoliés. En automatisant la comparaison avec ces bases externes, elle permettrait d'accélérer les vérifications et d'améliorer la robustesse des conclusions des experts.

5.3 Une technologie bien accueillie

L'intégration de l'IA au sein du musée des Arts décoratifs de Paris apparaît comme une opportunité intéressante pour optimiser la gestion des collections, enrichir la documentation, et améliorer l'accessibilité des œuvres pour les chercheurs et le grand public. A ce titre, aucune des personnes interrogées n'a formulé de crainte quant à l'utilisation de l'IA. Au contraire, les différents cas d'usage identifiés montrent un intérêt particulier pour l'automatisation de tâches répétitives liées à l'inventaire, l'indexation des œuvres et la recherche de provenance. L'IA pourrait également jouer un rôle clé dans l'exploration visuelle des collections et la modernisation de l'expérience des visiteurs grâce à des interfaces plus intuitives et interactives.

Les professionnels du musée reconnaissent les bénéfices que pourrait apporter une telle transformation, notamment en termes de gain de temps et d'amélioration de la qualité des données. Toutefois, des précautions sont nécessaires pour garantir la fiabilité des résultats, éviter les biais dans l'indexation, protéger la confidentialité des informations et s'assurer que les infrastructures techniques puissent supporter l'augmentation du volume de données traitées. L'IA est ainsi perçue comme un outil complémentaire qui, bien encadré, permettrait d'assister les équipes sans remettre en

cause l'expertise humaine essentielle à la validation des informations.

6 Conclusion

Pour conclure, les projets Royère et TORNE-H, encore en cours, visent à démontrer la capacité de la vision par ordinateur à effectuer des tâches de haute précision sur des collections spécifiques. En affinant des modèles de pointe et en concevant des flux de travail adaptés, nous avons pu mettre en évidence la valeur ajoutée de cette approche sur mesure pour répondre aux besoins du personnel du musée dans des cas d'usage très spécifiques.

Notre projet cherche à alléger la charge de l'annotation manuelle en combinant l'IA avec d'autres approches d'automatisation, afin de construire un flux de travail éthique qui, à court et moyen terme, permettra la création de milliers de références dans la base de données du musée pour des œuvres encore non cataloguées.

Il convient toutefois de nuancer l'intérêt et la volonté d'adopter l'IA parmi les personnels du MAD. Si les entretiens menés dans le cadre du projet TORNE-H révèlent une disposition favorable à l'utilisation de ces technologies, celle-ci repose avant tout sur la recherche d'une solution viable aux défis liés au traitement des collections. Plus que l'IA en tant que telle, c'est avant tout une réponse efficace au passif à traiter qui suscite l'intérêt des acteurs du musée.

Les prochaines étapes du projet doivent encore explorer l'industrialisation de ce processus pour le MAD, en examinant à la fois l'évolution du système d'information du musée et l'organisation du travail humain en coopération avec le flux de traitement automatisé. La mise en place d'une interface optimisée pour l'exploitation des modèles d'IA et la formation des personnels constitueront des enjeux cruciaux pour assurer une intégration pérenne de ces nouveaux outils. La conception d'un système de gestion des ressources numériques (DAM) pour stocker les images et les annotations sera également un élément clé dans cette perspective.

Avec notre partenaire, la BnF, et potentiellement d'autres musées parisiens intéressés, nous envisageons également de tester la réutilisabilité de nos flux de travail et modèles sur d'autres collections et institutions. Cela implique de valider les scénarios d'usage dans des contextes variés, de déployer le processus itératif d'annotation et de fine-tuning sur de nouvelles collections, et de construire un cadre d'évaluation robuste qui consolidera les métriques de pertinence grâce à une validation scientifique des annotations par le personnel du musée.

Remerciements

Le projet Royère (2023-24) a été financé par un mécénat de la Fondation Jean Royère.

Le projet TORNE-H (2024-25) est financé par le programme FTNC du ministère français de la Culture. Nous remercions Bénédicte Gady, promotrice du projet au sein des Arts Décoratifs ; Natacha Grim, stagiaire sur le projet en 2024 ; Jean-Philippe Moreux, Chef de mission Intelligence artificielle à la BnF et expert sur le projet.

7 Bibliographie

Références

[1] Philippine Bergère. *Paul Henrot (1908-1986). Un photographe oublié au service du Mouvement moderne Le fonds Paul Henrot du musée des Arts décoratifs*. Mémoire de master, Faculté des Lettres de Sorbonne Université, Paris, 2022.

[2] Emmanuelle Bermès, Céline Leclaire, and Jean-Philippe Moreux. L'image comme particule élémentaire, ou les prémisses d'un changement d'échelle à la BnF. In *The Measurement of Images. Computational Approaches in the History and Theory of the Arts, sous la direction de Clarisse Bardiot et Emmanuel Château-Dutier*, Humanités numériques et science ouverte. Presses universitaires du Septentrion, 2023.

[3] Francis Crimmins. Evaluation of techniques that improve findability of historic images in a large and diverse corpus using AI vision models and embeddings, 2024.

[4] Jacques Lacoste. *Jean Royère*. Galerie Jacques Lacoste Galerie Patrick Seguin, Paris, 2012.

[5] Pierre-Emmanuel Martin-Vivier. *Jean Royère*. Norma, Paris, France, 2002.

[6] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2 : Learning robust visual features without supervision, 2024.

[7] Abigail Potter. Introducing the LC Labs Artificial Intelligence Planning Framework, 2023.

[8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision, February 2021.

[9] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2 : Segment Anything in Images and Videos, 2024. Version Number : 2.

[10] Dominique Stutzmann. Automatische Texterkennung (ATR) in der Mediävistik : Werkstattbericht zu den Projekten Himanis und Home und neue Perspektiven für Historiker :innen. In *Forschungskolloquium*

- Fachbereich Geschichts- und Kulturwissenschaften - Freie Universität Berlin, Berlin, Germany, December 2023.

[11] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2 : Advancing a unified representation for a variety of vision tasks. *arXiv preprint arXiv :2311.06242*, 2023.