LLM Génératif Zero/Few Shots ou Annotation Manuelle ? Retours d'Expériences du défi EvalLLM 2024

Maxime Prieur, Sylvain Verdy, Vuth Nakanyseth, Gilles Sérasset, Didier Schwab, Cédric Lopez

▶ **To cite this version:**

Maxime Prieur, Sylvain Verdy, Vuth Nakanyseth, Gilles Sérasset, Didier Schwab, et al.. LLM Génératif Zero/Few Shots ou Annotation Manuelle ? Retours d'Expériences du défi EvalLLM 2024. Atelier sur l'évaluation des modèles génératifs (LLM) et challence d'extraction d'information few-shot, Institut des sciences informatiques et de leurs interactions - CNRS Sciences informatiques [INS2I-CNRS], Jul 2024, Toulouse, France. hal-04678041

HAL Id: hal-04678041
https://hal.science/hal-04678041v1

Submitted on 26 Aug 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

LLM Génératif Zero/Few Shots ou Annotation Manuelle ? Retours d'Expériences du défi EvalLLM 2024

Maxime Prieur[1] Sylvain Verdy[2] Nakanyseth Vuth[3] Gilles Sérasset[3] Didier Schwab[3] Cédric Lopez [2]

(1) Airbus Defence and Space, 1 bd. Jean Moulin, 78990 Élancourt, France
(2) Emvista, 10 rue Louis Breguet, 34830 Jacou, France
(3) Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, Grenoble 38000, France
maxime.prieur@airbus.com, cedric.lopez@emvista.com,
nakanyseth.vuth@univ-grenoble-alpes.fr

RÉSUMÉ

Ce rapport détaille les approches testées dans le cadre de l'atelier EvalLLM 2024. Notre participation à l'atelier a eu pour intérêt d'apporter des éléments de réponses aux questions suivantes : dans quelle mesure les données annotées et développées dans le cadre du projet POPCORN peuvent-elles s'adapter à une nouvelle typologie de classes et à un nouveau guide d'annotation ? Quelles performances obtiennent des modèles supervisés entraînés sur des données simplement alignées avec cette nouvelle typologie ? Les contraintes de temps de développement (nous nous sommes limités à une participation de deux jours), environnementales et les performances obtenues par l'ensemble des systèmes participants au défi encouragent-elles l'utilisation des grands modèles de langage génératifs plutôt que des encodeurs entraînés pour la tâche souhaitée ?

ABSTRACT

Zero/Few Shots Generative LLM or Manual Annotation ? Feedback from the EvalLLM 2024 Challenge

This report details the approaches tested as part of the EvalLLM 2024 challenge. Our participation in the challenge aimed to provide answers to the following questions : to what extent can the annotated and developed data from the POPCORN project adapt to a new class typology and a new annotation guide ? What performance do supervised models trained on data simply aligned with this new typology ? Do the development time constraints (we limited ourselves to two days of participation), environmental factors, and the performances obtained by all the systems participating in the challenge encourage the use of large generative language models rather than encoders trained for the desired task ?

MOTS-CLÉS : Reconnaissance d'Entités Nommées, Apprentissage Supervisé - Alignement d'Ontologie.

KEYWORDS: Named Entity Recognition, Supervised Learning, Ontology Alignment.

1 Introduction

POPCORN (Peuplement OPérationnel de COnnaissances et Réseaux Neuronaux) est un projet de recherche collaboratif subventionné par l'Agence Innovation Défense qui vise à améliorer les

technologies d'Extraction d'Informations (EI) pour les services de renseignement. Ces services reçoivent de nombreux rapports décrivant des activités criminelles dans le monde entier. Analyser ces informations dans leur intégralité est impossible pour les agents opérationnels, et les rapports en langage naturel ne peuvent pas être directement assimilés par les systèmes d'information du renseignement. En raison des questions de sécurité, ces rapports ne peuvent pas être partagés avec le consortium de POPCORN, ce qui oblige le consortium à développer des modèles sans données terrain réelles.

Depuis janvier 2022, notre projet s'attaque à cette problématique en produisant des ensembles de données similaires à de véritables rapports, afin d'entraîner et d'évaluer des modèles d'EI. Nous nous concentrons principalement sur la Reconnaissance d'Entités (RE) et l'Extraction de Relations (ER).

Pour le défi actuel, les données fournies incluent des bulletins d'information et des blogs en français, annotés avec des entités pertinentes pour le renseignement (noms, lieux, organisations, mais aussi fonctions, équipements, etc.) ainsi que des déclencheurs d'événements. Étant donné la proximité de cette tâche avec les activités de recherche du projet POPCORN, nous avons décidé de participer à ce défi.

Malgré l'accent mis sur l'extraction d'information dans un contexte « few-shot », ce défi reste très ouvert. Nous avons donc choisi de participer pour répondre aux questions suivantes :

— Question 1 : Dans quelle mesure les données annotées et développées dans le cadre du projet POPCORN peuvent-elles s'adapter à une nouvelle typologie de classes et à un nouveau guide d'annotation ?
— Question 2 : Quelles performances obtiennent des modèles supervisés entraînés sur des données simplement alignées avec cette nouvelle typologie ? Comment se positionnent-ils par rapport à des LLM génératifs ?
— Question 3 : Les contraintes de temps de développement (nous nous sommes limités à une participation de deux jours), environnementales et les performances obtenues par l'ensemble des systèmes participants au défi encouragent-elles l'utilisation des grands modèles de langage (LLM) ou plutôt celle de modèles supervisés ?

La section 2 décrit succinctement le développement des données annotées dans le cadre du projet POPCORN ainsi que leur adaptation au défi par alignement des étiquettes « POPCORN »avec celles d'« EvalLLM ». Nous présenterons ensuite les expériences réalisées (section 3), dont les trois solutions soumises pour ce défi, et discuterons de leurs résultats à la lumière de nos questions de recherche (section 4).

2 Données

Les données utilisées dans le cadre de ce défi sont toutes issues du projet POPCORN. Elles sont réparties en deux jeux de données, l'un factice et l'autre réel, décrits ci-après.

2.1 Jeu de données factice

Le projet POPCORN a permis de concevoir et développer un jeu de données comprenant 2 000 rapports factices en français se rapprochant des rapports réels de renseignement. Ces rapports sont annotés en entités d'intérêts (nommées et non nommées) ainsi que leurs attributs et en relations. Ils

ont été rédigés et annotés par un prestataire selon des contraintes imposées par le consortium de POPCORN. L'approche adoptée est décrite en détail par (Giordano *et al.*, 2024). Le jeu de données est annoté selon 43 types d'entités hiérarchisés sur quatre niveaux, 18 types d'attributs hiérarchisés sur deux niveaux, 39 types de relations hiérarchisés sur un seul niveau. Dans le cadre de ce défi, nous avons utilisé les 2000 rapports en les ré-étiquetant automatiquement selon la typologie fournie.

Une partie du jeu de données POPCORN est hébergée et librement accessible sur un répertoire github [1].

2.2 Jeu de données réel

Suite à la création du premier dataset constitué de rapports factices, nous avons découvert le corpus Renseignor [2]. Ce corpus est un ensemble de transcriptions de bulletins d'actualité d'origine internationale. Plus riche en informations et surtout mentionnant des éléments ancrés dans le monde réel, nous avons entrepris l'annotation de ce corpus pour notre cas d'usage. Le contenu, légèrement différent des rapports POPCORN, a nécessité une adaptation de l'ontologie utilisée précédemment. Pour ce faire, nous avons étudié les transcriptions afin d'identifier les types d'événements récurrents. En appliquant la même méthodologie que précédemment, nous avons pu entamer l'annotation des événements sur un ensemble de 2000 rapports Renseignor.

Le jeu de données est à ce stade annoté uniquement en évènements selon une typologie constituée de 46 classes. Dans le cadre de ce défi, nous avons utilisé les 2000 rapports en les ré-étiquetant automatiquement selon la typologie fournie.

2.3 Alignement des ontologies

Dans le but de répondre à notre première question « *Dans quelle mesure les données annotées et développées dans le cadre du projet POPCORN peuvent-elles s'adapter à une nouvelle typologie de classes et à un nouveau guide d'annotation ?* »(section 1), nous avons opté pour une approche d'alignement des ontologies POPCORN et EvalLLM en trois étapes.

Alignement des classes Pour cela nous avons commencé par définir manuellement une surjection des classes d'entités et d'attribut POPCORN vers les classes EvalLLM, telle que définie dans la table 1.

Adaptation des frontières d'entités Les guides d'annotation de POPCORN et d'EvalLLM ont défini différemment les frontières des entités. Ainsi, le guide d'EvalLLM inclut les déterminants, quantités et autres termes (« depuis », « pendant », etc.) dans l'annotation des entités, alors que le guide de POPCORN les exclut. Aussi, l'adaptation de nos annotation passe par une modification des frontières des entités de POPCORN. Pour cela nous avons adopté une heuristique consistant à étendre les annotations existantes en identifiant les mots-clés à englober provenant d'une liste préétablie.

Séparation de classes Enfin, la dernière différence importante abordée provient de la différence de granularité des classes Civilian/Military/... de POPCORN et des classes Person/Function d'EvalLLM. Ainsi, lorsqu'une entité désigne une personne (au sens des classes de POPCORN),

1. https://github.com/Emvista/popcorn-dataset
2. https://cf2r.org/publications/lettre/renseignor/

Classes de POPCORN	Classes de EvalLLM
Materiel	Equipment
Event	Event
Category	Function
Group Of Individuals	Group
Material Reference	ID
Latitude	Location
Longitude	
Place	
Military Organization	Military Unit
Intergovernmental Organization	Organization
Nationality	
Non Governmental Organization	
Non Military Governmental Organization	
Civilian	Person*
Military	
Terrorist or Criminal	
N/A	Resource
N/A	Site
Time Exact	Time
Time Fuzzy	
Time Max	
Time Min	

TABLE 1 – Alignement des types d'entité et d'attribut de POPCORN avec ceux de EvalLLM 2024

par une référence à un rôle, EvalLLM choisi d'annoter l'entité par la classe Function si ce rôle est une fonction et Unknown sinon. Nous réglons ce problème par une liste fermée de fonctions connues.

L'ensemble du processus est illustré sur deux exemples dans la table 2.

Etape	**Exemple 1**	**Exemple 2**
Annotation POPCORN	Le [général]$_{Military}$	Le [témoin]$_{Civilian}$
Surjection	Le [général]$_{Person}$	Le [témoin]$_{Person}$
Adaptation des frontières	[Le général]$_{Person}$	[Le témoin]$_{Person}$
Séparation des classes	[Le général]$_{Function}$	[Le témoin]$_{Unknown}$

TABLE 2 – Deux exemples du processus d'adaptation des annotations POPCORN au défi EvalLLM sur deux cas complexes.

Enfin, deux différences dans la modélisation de la tâche de Reconnaissance d'Entités Nommées par POPCORN et EvalLLM n'ont pas pu été réconciliées : l'extraction des entités discontinues et l'extraction des entités imbriquées.

3 Approches explorées

Cette section décrit les méthodes employées dans le cadre de l'atelier EvalLLM 2024. Les solutions testées et soumises à la fin de la période d'évaluation relèvent principalement des techniques d'apprentissage supervisé. Une expérience a par ailleurs été menée en utilisant un LLM à titre de comparaison. L'entraînement supervisé a été rendu possible grâce à l'utilisation des jeux de données décrits précédemment, adaptés spécifiquement à la tâche. De manière similaire, il aurait également été possible d'utiliser le jeu de données DWIE-FR (Verdy *et al.*, 2023), qui comprend une ontologie de plus de 150 types d'entités.

3.1 Expérience 1 : adaptation de Biaffine avec lissage de frontières

La première approche proposée spécialise le modèle Biaffine avec lissage de frontières (*Boundary Smoothing*) décrit dans l'article (Zhu & Li, 2022). Le modèle Biaffine est un modèle qui altère la probabilité de considérer les segments *spans* voisins d'une entité détéctée. Cette méthode intervient lors du calcul de la fonction de coût en tant que technique de régularisation, ceci afin de réduire le problème de frontières mal annotées et d'essayer de lever l'ambiguïté dans le traitement des frontières.

Pour le modèle de langue, socle autour duquel s'articule l'approche, nous avons choisi d'utiliser la version de base du modèle CamemBERT (Martin *et al.*, 2020). Ce dernier étant pré-entraîné sur le jeu de données WikiNer-FR (Nothman *et al.*, 2013).

Le jeu de données utilisé pour cette expérience est le jeu de données factice.

La figure 1 montre les entités extraites par le modèle Biaffine spécialisé sur le jeu de données POPCORN aligné. Selon, le calculateur Green Algorithm [3], la spécialisation du modèle a engendré une empreinte carbone de 20.68 g de $CO2e$ tandis que l'empreinte de l'inférence ne s'élève qu'à 1.9×10^{-2} g de $CO2e$.

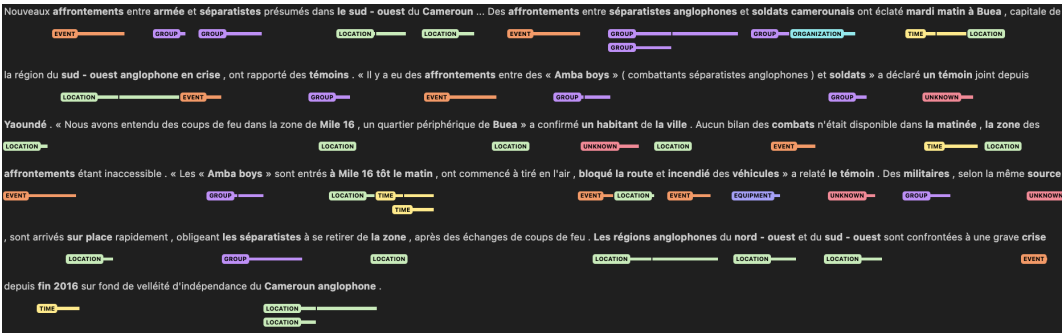

FIGURE 1 – Visualisation avec Spacy de la sortie de biaffine avec *boundary smoothing*

3. https://calculator.green-algorithms.org/

3.2 Expérience 2 : modèle unifié

Le deuxième fine-tuning a été appliqué sur le Modèle Unifié (MU) (Prieur *et al.*, 2024). Le Modèle Unifié est une approche conçue pour extraire depuis un texte les mentions, leurs types, les coréférences entre les mentions et les relations entre les entités afin de produire directement le graphe de connaissances du document tout en favorisant l'interaction entre les modalités d'intérêts.

Le modèle unifié comprend deux modules de classification qui utilisent en entrée la représentation vectorielle des tokens d'un texte, générée par un modèle de langue (CamemBERT-base dans notre cas). Le premier module a pour rôle d'identifier les segments représentant des entités dans le texte. Le second module prédit les interactions existantes entre toutes les paires de segments identifiées. Ces interactions concernent le type de l'entité, si la paire de span étudiée fait intervenir deux fois la même span, et dans le cas contraire, les relations ainsi que la co-référence entre ces segments.

Bien que le modèle actuel n'ait pas été entraîné sur les classes « Site » et « Resource », la taille du vecteur de prédiction inclut ces dernières, permettant ainsi leur future prédiction après une spécialisation sur des textes annotés, sans nécessiter le remplacement de la dernière couche linéaire.

Le jeu de données utilisé pour cette expérience est le jeu de données factice.

Le modèle a été implémenté sous pytorch et spécialisé selon un learning-rate de 1.10^{-4} avec un GPU de 46 Go pendant 6 minutes. Cette spécialisation a ainsi généré une empreinte carbone de 9.79g de $CO2e$. En inférence le modèle unifié traite un texte avec une vitesse moyenne de 0.22 secondes et génère 1.10^{-3} g de CO2e pour cette opération.

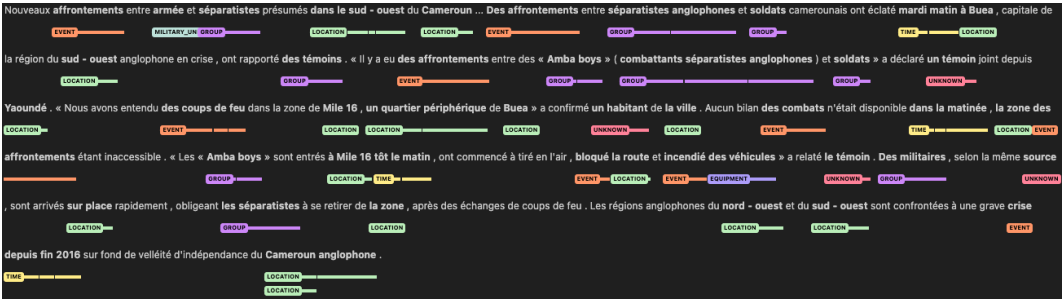

FIGURE 2 – Visualisation avec Spacy de la sortie du Modèle Unifié

3.3 Expériences 3 et 4 : fusion de exp. 1 *vs.* exp. 2 et d'un modèle d'extraction d'événements

Nous avons constaté que les modèles issus des deux premières expériences montraient des lacunes sur l'extraction d'événements, quand bien même ces derniers sont exprimés de façon explicite dans les textes. Nous avons entraîné un nouveau modèle en suivant le même processus que la première expérience mais cette fois sur les données du jeu réel et uniquement en se focalisant sur les événements. Ce modèle, noté M_{events} dans la suite, est utilisé dans le cadre de deux expériences :

— Expérience 3 : cette expérience consiste à utiliser les prédictions de M_{events} pour compléter les prédictions faites par le modèle issu de l'expérience 1. Les classes d'événement inférées

par M_{events} ont permis d'annoter les tokens qui n'avaient pas été annotés par le précédent modèle. Les tokens déjà annotés n'ont pas été modifiés, l'objectif étant d'améliorer le rappel.

— Expérience 4 : cette solution complète les prédictions du modèle de l'expérience 2 avec celles de M_{events}. Pour fusionner les prédictions des deux modèles en cas de conflit de label, ce sont les prédictions du modèle évènement qui sont gardées. Dans le cas d'un conflit avec un label « Group », on vérifie la présence d'un chiffre (« un », ..., « neuf ») précédent la prédiction et dans ce cas on procède étend l'annotation de le l'entité avec le label « Group »tout en gardant l'entité prédite comme « Event ».

M_{events} utilise l'architecture du modèle unifié et est entraîné sur 900 bulletins Renseignor avec un taux d'apprentissage de 1×10^{-4}. L'entraînement de M_{events} a duré 15 époques sur un modèle équivalent à celui de la deuxième expérience.

Comme on peut l'observer en comparant les figures 3.2 et 3.3, l'ajout des prédictions du modèle entraîné spécifiquement pour l'extraction des événements permet d'obtenir des résultats plus couvrant ce qui amène à penser que le rappel devrait être meilleur que pour la première approche.

FIGURE 3 – Visualisation avec Spacy de la fusion de Biaffine et du modèle d'Evénement

3.4 Expérience 5 : LLM Génératif avec Self-Consistency

En complément des modèles spécialisés, nous avons expérimenté une solution impliquant un LLM Génératif (la version spécialisée sur la langue française de Mistral-7B (Jiang *et al.*, 2023), Vigostral 7B [4]). L'extraction des entités avec Vigostral utilise la méthodologie dites (*K-Shot*), qui permet de fournir quelques exemples dans le prompt afin d'orienter le LLM sur la sortie attendue. Ces exemples sont tirés du jeu de données POPCORN. La sélection des exemples se fait en comparant la distance cosinus entre les représentations vectorielles du texte à traiter et celles des textes de POPCORN.

Étant donné que les LLM Génératifs produisent souvent des résultats variés et génèrent par essence des hallucinations, nous avons incorporé la méthode d'autoconsistence (*Self-Consistency*) (Wang *et al.*, 2022). Cela consiste à exécuter le prompt plusieurs fois et à fusionner ensuite les résultats les plus cohérents. Par exemple, lorsqu'une entité spécifique apparaît un certain nombre de fois, il est possible de la considérer comme correctement prédite. Dans nos travaux le seuil choisi a été de 40%.

Pour cette approche testée sur 2 GPU Quadro P6000, on mesure un temps d'inférence moyen par texte de 1 m 50 s et une empreinte carbone de 1.1 g de CO2.

4. https://huggingface.co/bofenghuang/vigostral-7b-chat

Les résultats obtenus en appliquant la méthode d'autoconsistance (cf. Fig. 4) semblent prometteurs. Toutefois, faute de temps, cette approche n'a pas pu être ni explorée plus en détail ni proposée à l'issue du défi. Nous prévoyons de le faire dans le futur.

FIGURE 4 – Visualisation avec Spacy de l'approche par LLM et Self-Consistency.

4 Résultats

Le défi EvalLLM permettait de soumettre trois résultats maximum. Nous avons décidé de soumettre les résultats des expériences 1, 2 et 4. Néanmoins, les organisateurs ont accepté de nous fournir les résultats des expériences 3 et 5 qui ne sont pas prises en compte dans le défi. Les résultats fournis par les organisateurs indiquent les macro précision, rappel et F1, ainsi que les micro précision, rappel et F1. Les métriques « macro »sont calculés en faisant la moyenne de la métrique obtenue pour chaque classe d'éléments tandis que les métriques « micro »sont des mesures obtenus en rassemblant l'ensemble des classes.

Ces métriques sont appliquées selon trois hypothèses notées :

- *STRICT* (cf. Table 3), c'est-à-dire que l'évaluation considère un alignement parfait entre les entités annotées automatiquement et manuellement.
- *No_DISC* (cf. Table 4), on ne considère pas les entités discontinues mais elle considère les frontières englobantes de leurs spans ;
- *Wo_Event* (cf. Table 5), les entités de type EVENT sont exclues de l'évaluation.

Les évaluations "micro" n'ont pas été fournies pour les deux hypothèses *No_DISC* et *Wo_Event*.

De façon générale, les meilleurs résultats ont été obtenus par l'expérience 4 de façon significative par rapport aux autres expériences. Cela semble indiquer que l'idée de considérer distinctement un modèle de reconnaissance d'entités (au sens classique du terme) et un modèle de reconnaissance d'événements semble pertinente. La méthode de fusion des résultats des deux modèles (une complétion naïve) mériterait d'être approfondie. Rappelons également que deux types d'entités (SITE et RESOURCE) n'ont pas été traités par nos modèles ce qui implique un nivellement des résultats par le bas.

Les organisateurs ont également fournis les F1 scores par type d'entité en mode STRICT, cf. Table 6. Quelle que soit l'expérience, l'étiquette PERSON obtient un F1 score autour de 90%. Tous les autres types ont un score au-dessous de 58 %. Ces résultats faibles s'expliquerait principalement par :

- Les guides d'annotations de POPCORN et d'EvalLLM qui diffèrent grandement sur l'identification des frontières des annotations qui sont plus larges dans le cadre de EvalLLM (notamment avec l'inclusion des articles des entités) ce qui a un impact important sur les

résultats.

— Un alignement des étiquettes POPCORN avec EvalLLM dont les définitions ne partageraient pas exactement la même sémantique.

Les organisateurs ont, au moment de l'écriture de ce rapport, indiqués que « le *Macro-F1 score* est de 37.14% par rapport à une médiane de scores de tous les participants d'environ 32 % »ce qui place le système de l'expérience 4 parmi les plus performants. À noter qu'il s'agit d'un système dont l'empreinte carbone de l'inférence est très faible par rapport à celle de l'expérience 5 qui implique un LLM.

TABLE 3 – Resultats en mode *STRICT*

	Macro précision	**Macro Rappel**	**Macro F1**	**Micro précision**	**Micro Rappel**	**Micro F1**
Expérience 1 (run 1)	33.76	30.46	29.52	44.28	25.96	32.73
Expérience 2 (run 2)	37.07	33.82	31.16	45.58	24.38	31.76
Expérience 3 (hors concours)	35.38	32.04	31.62	48.87	32.49	39.03
Expérience 4 (run 3)	**41.08**	**39.17**	**37.14**	**51.05**	**33.57**	**40.51**
Expérience 5 (hors concours)	22.93	13.96	16.76	18.90	5.58	8.62

TABLE 4 – Resultats en mode *No_DISC*

	Macro précision	**Macro Rappel**	**Macro F1**
Expérience 1 (run 1)	33.92	30.48	29.74
Expérience 2 (run 2)	37.20	33.84	31.32
Expérience 3 (hors concours)	35.56	32.06	31.83
Expérience 4 (run 3)	**41.25**	**39.17**	**37.34**
Expérience 5 (hors concours)	22.96	13.96	16.77

TABLE 5 – Resultats en mode *Wo_event*

	Macro précision	**Macro Rappel**	**Macro F1**
Expérience 1 (run 1)	32.31	32.42	30.93
Expérience 2 (run 2)	35.57	36.06	32.66
Expérience 3 (hors concours)	32.31	32.42	30.93
Expérience 4 (run 3)	**38.35**	**39.89**	**36.62**
Expérience 5 (hors concours)	23.59	14.90	17.71

5 Conclusion

Notre participation au défi EvalLLM a permis d'expérimenter une adaptation rapide (en deux jours) de nos approches développées dans le cadre du projet POPCORN. Les questions définies en introduction ne peuvent trouver de réponses complètes sans les résultats détaillés officiels de ce défi, ceux-ci étant fournis le jour de l'atelier. Cependant, nous pouvons avancer ce qui suit :

— question 1 : l'adaptation des données a été facilitée par le grand nombre de classes d'entités et d'attributs de l'ontologie POPCORN. Les classes de EvalLLM étant plutôt génériques, l'alignement des classes spécifiques de POPCORN vers les classes généralement plus génériques de EvalLLM n'a pas posé de problème. Un point négatif a été l'impossibilité de prendre en compte deux des classes de EvalLLM car elles ne correspondaient à aucune classe mère de l'ontologie POPCORN. D'après nous, c'est précisément là où un LLM aurait pu intervenir. En mettant de côté la manière d'annoter (par exemple avec ou sans déterminant) nous considérons que l'adaptation a été un succès.

TABLE 6 – F1 scores par type d'entité ; mode *STRICT*

	PERSON	FUNCTION	GROUP	ORGANIZATION	MILITARY_UNIT	LOCATION	SITE	EQUIPMENT	RESOURCE	TIME	EVENT	ID	UNKNOWN
Expérience 1	90.11	18.42	14.86	52.39	21.62	54.74	0	31.76	0	41.95	12.14	35.29	10.53
Expérience 2	90.32	16.55	24.32	41.72	16.07	**57.66**	0	36.01	0	49.6	11.68	28.57	32.56
Expérience 3	90.11	18.42	14.67	**52.39**	21.62	54.74	0	31.67	0	41.64	40.0	35.29	10.53
Expérience 4	**91.3**	**19.35**	**36.36**	40.28	**29.31**	57.64	0	**38.37**	0	**54.61**	**41.30**	35.29	**39.02**
Expérience 5	88.1	0	11.86	5.23	2.5	14.53	0	8.89	0	2.27	4.49	**80.0**	0

— question 2 : dans le cadre de nos expériences, les modèles supervisés obtiennent des résultats nettement meilleurs que les LLM génératifs, mais cela est une conclusion temporaire qui pourrait être revue lors de l'annonce des résultats des autres participants. Le système de l'expérience 4 obtient les meilleurs résultats ce qui encourage la distinction des tâches de reconnaissance d'entités et d'événements.

— question 3 : nos meilleurs modèles retournent une réponse en 20 ms, ce qui est beaucoup plus rapide qu'avec le LLM (1 min 50 s dans le cadre de l'expérience 5). L'entraînement de nos modèles engendrent une empreinte carbone de l'ordre de 9 et 20 g et 1.103 g de CO_2e en inférence, ce qui est faible par rapport aux LLM génératifs. De même l'impact carbone de nos meilleurs modèles est très faible en comparaison avec celui du LLM génératif utilisé dans l'expérience 5.

Suite à ce défi, il serait intéressant de procéder à une campagne similaire à EvalLLM mais avec une ontologie représentant davantage de concepts avec plusieurs nivaux de spécialisations, car cela permettrait de mieux comprendre l'impact du niveau de spécialisation sur les méthodes à base de LLM (sur l'étiquette PERSON par exemple, l'expérience 5 à base de LLM a obtenu un très bon score). En guise de perspectives, POPCORN s'intéresse à la capacité des modèles à extraire les relations entre les entités d'intérêts. Nous avons en ce sens lancé le défi TextMine'25 [5].

Références

GIORDANO B., PRIEUR M., VUTH N., VERDY S., COUSOT K., SÉRASSET G., GADEK G., SCHWAB D. & LOPEZ C. (2024). Popcorn : Fictional and synthetic intelligence reports for named entity recognition and relation extraction tasks. In *Proceedings of the 28th International Conference on Knowledge-Based and Intelligent Information Engineering Systems, KES 2024, Sevilla, Spain, 2024*.

JIANG A. Q., SABLAYROLLES A., MENSCH A., BAMFORD C., CHAPLOT D. S., DE LAS CASAS D., BRESSAND F., LENGYEL G., LAMPLE G., SAULNIER L., LAVAUD L. R., LACHAUX M.-A., STOCK P., SCAO T. L., LAVRIL T., WANG T., LACROIX T. & SAYED W. E. (2023). Mistral 7b.

MARTIN L., MULLER B., SUÁREZ P. J. O., DUPONT Y., ROMARY L., DE LA CLERGERIE É., SEDDAH D. & SAGOT B. (2020). Camembert : a tasty french language model. In D. JURAFSKY, J. CHAI, N. SCHLUTER & J. R. TETREAULT, Éds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, p. 7203–7219 : Association for Computational Linguistics. DOI : 10.18653/V1/2020.ACL-MAIN.645.

NOTHMAN J., RINGLAND N., RADFORD W., MURPHY T. & CURRAN J. R. (2013). Learning multilingual named entity recognition from wikipedia. *Artif. Intell.*, **194**, 151–175. DOI : 10.1016/J.ARTINT.2012.03.006.

5. https://textmine.sciencesconf.org/resource/page/id/10

PRIEUR M., DU MOUZA C., GADEK G. & GRILHERES B. (2024). Shadowfax : Harnessing textual knowledge base population. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Wshington, USA, July 23-27, 2023*, Washington, USA : ACM.

VERDY S., PRIEUR M., GADEK G. & LOPEZ C. (2023). DWIE-FR : Un nouveau jeu de données en français annoté en entités nommées. In C. SERVAN & A. VILNAT, Éds., *Actes de CORIA-TALN 2023. Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles, TALN 2023 - Volume 2 : travaux de recherche originaux - articles courts, Paris, France, June 5-9, 2023*, p. 63–72 : ATALA.

WANG X., WEI J., SCHUURMANS D., LE Q., HSIN CHI E. H. & ZHOU D. (2022). Self-consistency improves chain of thought reasoning in language models. *ArXiv*, **abs/2203.11171**.

ZHU E. & LI J. (2022). Boundary smoothing for named entity recognition. In S. MURESAN, P. NAKOV & A. VILLAVICENCIO, Éds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 7096–7108, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-long.490.