EvalLLM 2025 : présentation de la deuxième édition de l'atelier Évaluation des LLM et des challenges

Nihel Kooli[1] Vincent Claveau[1] Julianne Flament[1] Lorenzo Gerardi[1]
Maxime Poulain[1] François Delon [2] Hugo Moquet [2]
(1) AMIAD, Ministère des Armées
(2) SSA, Ministère des Armées
prenom.nom@def.gouv.fr

RÉSUMÉ

La deuxième édition de l'atelier EvalLLM, organisée par l'AMIAD, poursuit l'objectif d'explorer les méthodes d'évaluation des grands modèles de langue (LLM) génératifs, en particulier pour le français et les domaines spécialisés. L'atelier a réuni chercheurs et industriels autour de communications scientifiques et de deux challenges : un défi d'extraction d'information few-shot dans le domaine de la santé et un défi de fine-tuning de LLM sur des textes du domaine de la défense. Les travaux présentés abordent l'évaluation des LLM sous des angles variés : création de benchmarks, nouvelles métriques, biais culturels, robustesse, et utilisation des LLM comme juges. Les résultats des challenges soulignent à la fois les avancées méthodologiques et les difficultés persistantes dans l'adaptation et l'évaluation de ces modèles en français et en contexte métier.

ABSTRACT

EvalLLM 2025 : presentation of the second edition of the workshop on LLM evaluation and its challenges

The second edition of the EvalLLM workshop, organized by AMIAD, aims to explore methods for evaluating large generative language models (LLMs), particularly for French and specialized domains. The workshop brought together researchers and industry professionals around scientific presentations and two challenges : a few-shot information extraction challenge in the field of healthcare and an LLM fine-tuning challenge on texts from the field of defense. The work presented addresses the evaluation of LLMs from various angles : benchmark creation, new metrics, cultural biases, robustness, and the use of LLMs as judges. The results of the challenges highlight both methodological advances and persistent difficulties in adapting and evaluating these models in French and in a domain-specific context.

MOTS-CLÉS : Grands modèles de langue, évaluation, benchmark, extraction d'information, fine-tuning, français, défense, domaine de la santé.

KEYWORDS: Large Language Models, evaluation, benchmark, information extraction, fine-tuning, French, defence, health domain.

ARTICLE : **Accepté à** Atelier EvalLLM, conjoint aux conférences CORIA-TALN 2025.

1 Introduction

Les grands modèles de langue (LLM) génératifs se démocratisent et s'intègrent dans des chaînes de traitements de plus en plus complexes, offrant une grande variété de cas d'usage. L'évaluation de ces objets protéiformes pose cependant des problèmes méthod : les benchmarks existants sont largement anglo-centrés (aussi bien en matière de langue que de culture), parfois eux-mêmes issus de LLM anglo-centrés (benchmarks synthétiques), et ne couvrent pas l'ensemble des langues, des domaines de spécialité et des usages de ces LLM. Les lien entre évaluation et LLM sont aussi à examiner dans l'autre sens, l'usage de grands modèles de langue comme système d'évaluation (eg. *LLM as a judge*) étant de plus en plus répandu.

Partant de ce constat, et suite au succès de la première édition en 2024 (voir (Kooli *et al.*, 2024)), l'AMIAD (Agence Ministérielle pour l'IA de Défense, Ministère des Armées) a de nouveau organisé l'atelier EvalLLM. Comme la précédente édition, il a eu vocation à rassembler les chercheurs du domaine autour de présentations de travaux sélectionnés sur appel à soumission (section 2) et de deux challenges mettant en pratique l'usage et l'évaluation de LLM sur des données en français (sections 3 et 4).

2 Atelier

2.1 Thématiques

L'atelier a vocation à réunir les chercheuses et chercheurs, industriels et académiques, s'intéressant aux multiples facettes de l'évaluation des LLM génératifs sur des domaines de spécialité ou sur des langues autres que l'anglais. L'appel à communication sollicitait des propositions d'articles sur tous les travaux relevant de ce périmètre.

Cela incluait notamment les recherches concernant :

- l'évaluation de modèles de fondation, fine-tunés ou de systèmes complets (RAG par exemple)
- la création ou adaptation de benchmarks, pour du français ou autres langues d'intérêt, qu'elles soient bien ou peu dotées, en domaine général ou spécialisé, ou pour des langues bruitées ou non standard (eg. réseaux sociaux, commandes vocales…)
- l'évaluation sur des tâches de TAL (traduction, résumé, extraction d'information…)
- l'adaptation des méthodologies d'évaluation existantes aux systèmes génératifs
- les dimensions éthiques, biais, *privacy*, alignement culturel ou législatif
- les dimensions de performances en temps de calcul, mémoire, frugalité énergétique
- l'évaluation avec des utilisateurs, ergonomie, aspects cognitifs
- l'évaluation de modèles multimodaux (eg. texte-image, texte-parole…)

2.2 Soumissions

Les articles reçus en réponse à l'appel ont été soumis au comité de programme de l'atelier dont la composition est détaillée en annexe A et que nous remercions pour le travail de qualité effectué dans un temps très contraint. Suite aux retours du comité, ce sont 22 articles qui ont été retenus, dont 8 ont

été présentés à l'oral et 14 en posters [1].

Ces articles, issus de travaux académiques ou industriels, abordent le problème du couplage entre LLM et évaluation sous une grande variété d'angles. Tout d'abord, les tâches considérées couvrent une grande partie des tâches classiques du TAL : résumé automatique, reconnaissance d'entité, systèmes de question-réponse, transcription de la parole, analyse de reviews, classification de texte, génération de texte. Cela illustre l'utilisation "couteau-suisse" qui est maintenant faite des LLM. D'autre part, les travaux décrits dans les articles portent sur des domaines de spécialité ou des cadres d'application également variés : domaine de l'hydraulique (Vartampetian *et al.*, 2025), du renseignement (Aubertin *et al.*, 2025), de la médecine (Servan *et al.*, 2025), des ressources humaines (Rozera *et al.*, 2025), de la géographie (Decoupes & Guille, 2025), de la génération de code (Moughit & Hafidi, 2025; Perez *et al.*, 2025a), de la désinformation (Séjourné *et al.*, 2025), de l'analyse de discours politiques (Perez *et al.*, 2025b) et des sciences humaines et sociales (Vallet & Suignard, 2025).

Certains articles ont présenté des travaux sur les méthodologies d'évaluation, en se focalisant sur des tâches usuelles des LLM ou sur les métriques. Ainsi, de nouvelles approches d'évaluation pour le résumé ont été proposées (Herserant & Guigue, 2025). Similairement, (El Yagouby *et al.*, 2025) ont introduit une nouvelle métrique basée sur de multiples passes, utile pour l'évaluation de la génération de code par exemple. (Gatti~Pinheiro *et al.*, 2025) propose un cadre de méta-évaluation des métriques pour en mesurer la fiabilité. L'évaluation de systèmes de RAG est abordée par (Martinon *et al.*, 2025), qui proposent un protocole et un jeu d'évaluation combinant annotations humaines et LLM-Juge.

D'autres aspects sur les capacités des LLM ont également été étudiés au travers de plusieurs soumissions. (Decoupes & Guille, 2025) se sont ainsi intéressés aux connaissances intégrées aux LLM. La question des biais culturels des modèles et de leur mesure est développée par (Valette, 2025). Les capacités des petits modèles (SML) sur une tâche de classification de textes sont examinées par Vallet & Suignard (2025). La modalité audio des modèles est étudiée (Gibier *et al.*, 2025) sur de l'*Audio Question Answering* pour évaluer des tâches de description de scènes audio. Les techniques de génération de sorties structurées sont comparées par (Séjourné *et al.*, 2025).

Sur la question des protocoles et métriques d'évaluation, il est intéressant de remarquer qu'une bonne part des articles étudient ou utilisent le concept de LLM-as-a-judge, c'est-à-dire l'emploi d'un LLM pour évaluer une sortie de LLM (Barkar *et al.*, 2025; Grina & Kalashnikova, 2025; Martinon *et al.*, 2025).

Le développement de ressources d'évaluation reste un enjeu fort du domaine et un intérêt central de l'atelier. Plusieurs travaux présentent de nouveaux jeux de données pour l'évaluation de différentes tâches sur des domaines variés : des articles annotés avec des entités d'intérêt pour le renseignement (Aubertin *et al.*, 2025), des appels d'offre annotés en entités (Grina & Kalashnikova, 2025), des questions-réponses sur la due-diligence (en droit des sociétés, l'analyse préalable à une décision d'investissement, fusion ou acquisition) (Martinon *et al.*, 2025), et un jeu d'évaluation pour le domaine de l'hydraulique (Vartampetian *et al.*, 2025). Des questions connexes sur les données sont également étudiées, telles que l'emploi de datasets synthétiques (Jourdain & Hellal, 2025) ou la pertinence de jeux de données existants (Gibier *et al.*, 2025).

1. Il n'est pas fait de distinction de qualité entre les modes de présentation

	Apprentissage	Test
nb de documents	40 documents + guide d'annotation	200 documents
nb de documents avec un évènement	35 documents	103 documents
nb de phrases	811 phrases	2931 sentences
nb de tokens	18959 tokens	67446 tokens
nb d'évènements	61 évènements	215 évènements
max du nb d'évènements par document	4 évènements	10 évènements
nb de documents avec >1 évènement	17 docs	53 documents

TABLE 1 – Statistiques sur les données pour le challenge en extraction d'information

3 Challenge en extraction d'information few-shot

Un premier challenge d'évaluation de LLM par la tâche a été proposé dans le cadre de cet atelier. Il s'agit d'extraction d'information en français dans le domaine de la santé. Ce challenge s'inscrit dans un contexte few-shot où seuls un guide d'annotation et un petit ensemble de documents annotés sont fournis. Les résultats obtenus par les participants — que ce soit à l'aide de grands modèles de langage (LLM), d'approches plus traditionnelles ou de méthodes hybrides combinées à différents traitements — ont permis de mieux situer les apports des systèmes fondés sur les LLM.

3.1 Tâche, données et métriques

Les données utilisées pour ce challenge sont représentatives de celles analysées par le Centre d'épidémiologie et de santé publique des armées (CESPA ; organisme rattaché au Service de santé des armées) dans ses missions de veille sanitaire internationale de défense. Le guide d'annotation transmis aux participants a été conçu par des praticiens de santé pour répondre aux besoins spécifiques de cette activité. Il définit les principes généraux, décrit les différents types d'entités à annoter (21 types d'entités) et détaille les attributs associés aux événements sanitaires. Les participants ont disposé de 40 documents annotés, accompagné du guide d'annotation, pour la phase d'apprentissage. L'évaluation a été menée sur un corpus de 200 documents supplémentaires. Les corpus documentaires utilisés sont exclusivement en langue française et contiennent des documents issus de sources journalistiques et, dans une moindre mesure, de sources institutionnelles (OMS, Santé Publique France...). Ces documents intègrent différentes entités d'intérêt métier (maladies infectieuses, agents pathogènes et agents du spectre nucléaire-radiologique, chimique et explosif) pouvant être discontinues, ainsi que des évènements sanitaires représentés par des tuples d'entités, où chaque entité correspond à un attribut de l'évènement (voir 1). Les noms des attributs sont indiqués par des relations sémantiques entre un évènement et les entités qui le composent.

La Table 1 présente différentes caractéristiques des corpus d'apprentissage et de test. La distribution des entités par type et des évènements par type d'entité centrale pour ces mêmes corpus est illustrée en figure 2. Cette dernière montre un déséquilibre entre les données d'apprentissage et celles d'évaluation.

Les métriques d'évaluation employées pour comparer les différentes approches sont listées ci-dessous.

- ScoreEntity : macro-F1 mesure des entités
- ScoreEvent : moyenne sur tous les évènements des macro-F1 des attributs, en considérant que c'est correct pour les attributs si la liste des occurrences trouvées est une sous-partie de la liste

FIGURE 1 – Exemple d'un document annoté en entités et en évènements dans le domaine de la veille sanitaire

des occurrences dans la vérité terrain

— ScoreDoc : moyenne sur tous les documents des macro-F1 des évènements, en considérant qu'un évènement est correct si son élément central est trouvé

— ScoreAvg : définit par l'équation (1)

$$ScoreAvg = moyenne(scoreEntity, moyenne(scoreEvent, scoreDoc)) \quad (1)$$

— ScoreFinal : définit par l'équation (2)

$$ScoreFinal = ScoreAvg + 0.1 * ScoreMoyen * \begin{cases} 1 & \text{si publication des codes-sources} \\ 0 & \text{sinon} \end{cases} \quad (2)$$

3.2 Participation

Sept équipes ont participé au challenge avec 19 runs soumis au total. Ces runs explorent diverses approches :

— fine-tuning de modèles pré-entraînés (RoBERTa, GliNER…),
— stratégies de prompting de plusieurs LLM (GPT4, LLaMA, Gemma…),
— augmentation de données d'apprentissage,
— post-traitement à base de règles,
— utilisation de bases de connaissances.

Parmi les sept équipes participantes, cinq ont publié leur code-source. La liste des équipes, les détails des runs pour chacun des participants et les liens vers les codes-sources sont détaillés ci-dessous :

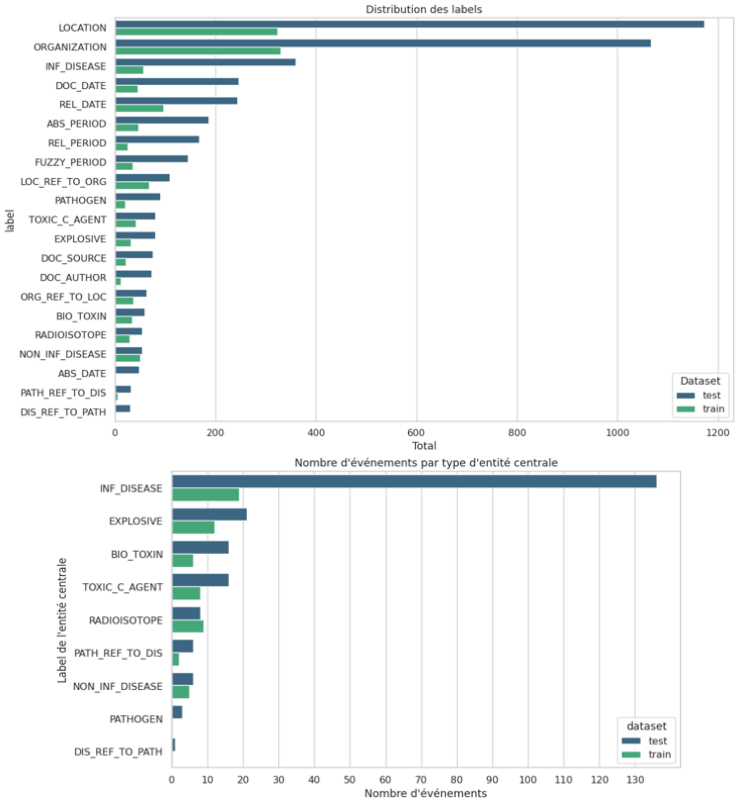

FIGURE 2 – Distribution des entités par type de label - distribution des évènements par type d'entité centrale

— EMVISTA
  — Run 1 : Un modèle NER composé d'un CamemBERT fine tuné NER avec une couche Biaffine. La prédiction des événements repose entièrement sur la détection des entités et se fait à base de règles.
  — Run 2 : Même(s) modèle(s) que pour la run 1 avec injection de données générées (avec GPT-4o) en plus des données de train du challenge.
  — Run 3 : Même modèle de NER, même données que pour la run 2 mais le système de détection d'événements a été "amélioré" à l'aide de OpenAI o4-mini/ GPT-4o.
— Magadir : Univ. de Marseille + LS2N (Univ de Nantes) (Boudraa & Belfathi, 2025)
  — Run 1 : Stratégie de sélection basée sur la densité des entités (density-based selection, k = 8), avec chaîne de raisonnement pour la détection d'événements.
  — Run 2 : Stratégie de sélection basée sur la diversité des entités (diversity-based selection, k = 4), avec chaîne de raisonnement pour la détection d'événements.
  — Run 3 : Stratégie de sélection basée sur la diversité des entités (diversity-based selection, k = 4), sans chaîne de raisonnement pour la détection d'événements.

  Les codes-sources sont disponibles sur https://github.com/hossambd/magadir_evalLLM2025
— Kairntech (Deturck, 2025)
  — Run 1 : Détection d'entité avec le LLM "GPT-4.5" et un prompt reprenant les instructions du guide d'annotation plus les annotations "train", puis détection d'événement avec le LLM "Llama-4-Maverick" et un prompt conçu comme précédemment.
  — Run 2 : Détection d'entité avec le LLM "Llama-4-Maverick" et un prompt reprenant les instructions du guide d'annotation plus les annotations "train", puis détection d'événement avec le LLM "Llama-4-Maverick" et un prompt conçu comme précédemment.
  — Run 3 : Détection d'entité avec "GPT-4.5" et un prompt reprenant les instructions du guide d'annotation plus les annotations "train", puis détection d'événements avec une intersection des sorties des LLM « Llama-4-Maverick » et « GPT-4.1 » et un prompt conçu comme précédemment.
— TALN : Univ. de Marseille + LS2N + Inetum (Belmadani *et al.*, 2025)
  — Run 1 : Emploi de GPT-4.1 en in-context learning (ICL) en fournissant 10 exemples les plus similaires (via similarité cosinus) et un résumé du guide d'annotation dans le prompt système. Pour la détection d'évènements, même approche que pour le NER : GPT-4.1 avec ICL, 10 exemples similaires, et un résumé du guide en prompt système.
  — Run 2 : un modèle gliner-biomed-large-v1.0 affiné sur des données synthétiques générées. Une fois les prédictions effectuées, celles-ci sont vérifiées et corrigées si besoin par un LLM ici GPT-4.1. Extraction d'évènement identique au Run 1.
  — Run 3 : un modèle NER affiné à partir de LLaMA-3.1-8B-Instruct sur les données générées. Le modèle utilise également les 10 exemples les plus similaires en ICL. Extraction d'évènement identique au Run 1.

  Les codes-sources sont disponibles sur https://github.com/ikram28/EvalLLM2025.git
— Listellm25 : CEA List (Armingaud *et al.*, 2025)
  — Run 1 : le modèle GLiNER adapté avec les données d'entraînement et une stratégie d'augmentation des données axée sur les entités et sur non leur contexte. Pour l'extraction des événements, un modèle encodeur-décodeur de classification de relations candidates entre événement central et événement associé + heuristiques de fusion.
  — Run 2 : pour le NER, ensemble avec vote de différents modèles GLiNER (avec et sans

différentes méthodes d'augmentation de données) + modèles de type encodeur XLM-RoBERTa + utilisation de ressources dictionnairiques. Pour l'extraction des événements, un modèle encodeur-décodeur de classification de relations candidates entre événement central et événement associé + heuristiques de fusion.
  - Run 3 : pour le NER, ensemble avec vote de différents modèles GLiNER (avec sans différentes méthodes d'augmentation de données) + modèles de type encodeur XLM-RoBERTa. Pour l'extraction des événements, un modèle encodeur-décodeur de classification de relations candidates entre événement central et événement associé sans intégrer le type des entités + heuristiques de fusion.

  Les codes-sources sont disponibles sur https://github.com/cea-list-lasti/evalllm2025-extractioninformation.
- INRIA Mission Défense et Sécurité (Soutrenon *et al.*, 2025)
  - Run 1 : prompting du modèle Llama 3.3 (70 B).
  - Run 2 : prompting du modèle Gemma 3 (27 B).
  - Run 3 : modèle CamemBERT bio GLiNER pour une première phase de NER avec une liste simplifiée de labels. Ces labels ont ensuite été affinés à l'aide du modèle Mistral 7B, suivi d'un post-traitement automatique des sorties NER. Llama 3.2 1B utilisé pour l'extraction des événements avec également une étape de post-traitement automatisé.

  Les codes-sources sont disponibles sur https://github.com/LucieBader/Challenge_ et https://github.com/pauline-soutrenon/challenge_evalllm2025_combined_approach
- TIBS : LITIS (Univ. de Rouen) (Haddag *et al.*, 2025)
  - Run 1 : emploi de l'API OpenRouter avec des prompts précis pour extraire des entités nommées qui sont ensuite parsées en dictionnaire, regroupées par synonymie via une seconde requête LLM, puis utilisées pour annoter des événements selon un schéma JSON structuré. Le système impose des contraintes strictes. Les codes-sources sont disponibles sur https://github.com/titusse3/EvalLLM2025.

3.3 Baselines

Pour définir un point de comparaison, une baseline a été proposée par l'AMIAD, entraînée et évaluée dans les conditions du challenge. Les jeux de données utilisés sont les mêmes que ceux fournis aux participants.

La baseline pour l'extraction d'entités nommées se fonde sur une approche standard de classification de tokens avec étiquetage IOB. Un modèle XLM-RoBERTa (Conneau *et al.*, 2019) est affiné et une couche de *Conditional Random Field* (CRF) est ajoutée pour assurer la cohérence de la séquence d'étiquettes.

L'extraction des événements utilise une méthode proposée par Diniz *et al.* (2025) : il s'agit d'aborder la tâche comme une tâche de classification de textes, dans lesquels on identifie une paire d'entités en ajoutant des balises. Cette classification est réalisée de manière binaire, selon l'existence ou non d'une relation entre ces deux entités. Le modèle affiné est le même XLM-RoBERTa que pour l'extraction d'entités.

Pour l'inférence, une étape d'identification des synonymes est proposée par calcul de similarité sur les représentations hors contexte des entités obtenues avec un modèle BERT plus léger (Devlin *et al.*, 2018). Les événements sont ainsi extraits en regroupant toutes les entités qui ont une relation avec

	ScoreEntity	ScoreEvent	ScoreDoc	ScoreAvg
Baseline	44,95	5,32	23,58	29,70
Baseline_NER_golden	N/A	15,79	37,22	63,25

TABLE 2 – Résultats des baselines en reconnaissance d'entités nommées et détection d'évènements de veille sanitaire

Équipe	ScoreEntity	ScoreEvent	ScoreDoc	Moy.	Score final final	Empreinte carbone
Listellm25	**66,75**	5,37	43,78	**45,66**	**50,23**	124
TALN	61,53	15,02	43,74	45,46	50,01	>>9 (inférence)
Magadir	47	**28,17**	48,97	42,78	47,06	4 700
Kairntech	50	26,08	**54,51**	45,15	45,15	103 000
TIBS	31,12	0,23	13,45	18,98	20,88	386,86
INRIA	13,02	4	17,42	11,86	13,05	217
EMVISTA	9,87	0,12	5,57	6,36	6,36	10

TABLE 3 – Résultats du meilleur run pour chaque participant en reconnaissance d'entités nommées et détection d'évènements sanitaires et mesure de l'empreinte carbone en gCO2e

une même entité centrale.

Les performances de la baseline sont présentées dans la Table 2. De plus, une évaluation de l'extraction d'événements a été réalisée à partir des entités de la vérité du jeu de test, afin de pouvoir mesurer la propagation des erreurs entre les deux étapes d'extraction. On constate que le score pour les événements augmente de 5,32 à 15,79. Néanmoins, ce score reste faible et témoigne de la difficulté de la tâche d'extraction d'événements dans notre challenge. L'empreinte carbone est de 37,60 gCO2e en apprentissage et 1,79 gCO2e en inférence

3.4 Résultats et analyse métier

Les résultats du meilleur run pour chaque participant sont détaillés dans la Table 3.

L'analyse des résultats fournis par les participants illustre la complexité de la tâche, en particulier lorsqu'il s'agit d'extraire des entités avec des définitions métier spécifiques ou d'identifier les évènements sanitaires associés.

3.4.1 Complexité liée à la détection des entités d'intérêt métier

Un des défis techniques majeurs réside dans la complexité du vocabulaire médical : les textes peuvent contenir des abréviations, des acronymes, une terminologie spécialisée, et des termes dont le sens varie selon le contexte. La présence d'entités discontinues contribue également à la difficulté de la tâche, en particulier dans un contexte few-shot où les exemples annotés sont rares. Un autre facteur de complexité se situe dans l'usage parfois imprécis ou erroné de cette terminologie scientifique dans les sources journalistiques, qui sont majoritaires dans ce challenge. Ces documents peuvent contenir des termes mal employés, des confusions entre concepts proches, ou des simplifications excessives,

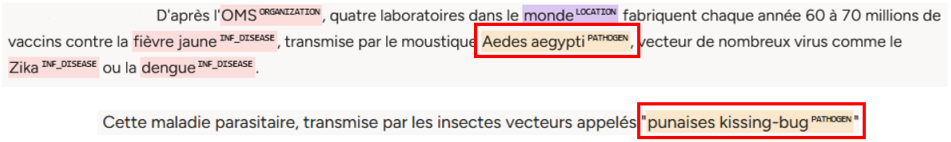

FIGURE 3 – Segments de texte (en rouge) incorrectement identifiés comme entités de type «PATHOGEN»

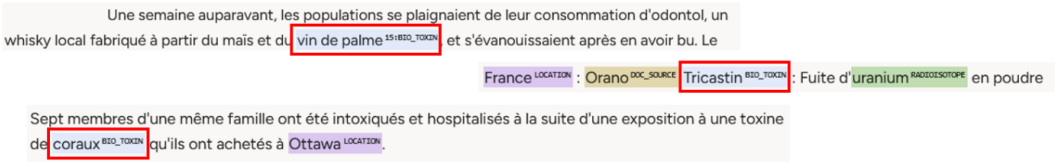

FIGURE 4 – Détections erronées d'entités de type « BIO-TOXIN »

ce qui complique l'identification correcte des entités. Dans les runs des participants, les vecteurs (moustiques, punaises, tiques) sont parfois considérés à tort comme des agents pathogènes. Cette confusion pourrait s'expliquer par un contexte d'emploi proche dans le langage courant, et donc dans les corpus d'apprentissage des modèles de fondation, ce qui rend la désambiguïsation difficile, comme en témoigne la Figure 3. De nombreux faux positifs sont également associés aux toxines biologiques, une classe sous-représentée dans le corpus d'apprentissage du challenge (figure 4).

Une partie des erreurs constatées provient directement de la rédaction des documents, en particulier de l'usage de raccourcis, de métaphores ou d'abus de langage. Dans la figure ci-dessous (figure 5), des noms de maladies sont improprement utilisés pour désigner les virus qui en sont à l'origine. La phrase fait mention d'analyses en laboratoire destinées à détecter la présence de l'agent pathogène, et non de la maladie. Ce type d'ambiguïté souligne que la prise en compte du contexte sémantique nécessite une expertise fine, tant sur le plan biomédical que linguistique, afin d'éviter les erreurs d'interprétation.

Certaines règles établies dans le guide d'annotation ne sont pas systématiquement assimilées par les modèles utilisés dans le cadre de ce challenge. Certains termes génériques tels que « virus », « bactérie » ou « parasite » sont détectés à tort et introduisent un bruit informationnel (Figure 6).

Ainsi, dans ce challenge, l'identification correcte des entités d'intérêt pourrait relever à la fois d'une approche ontologique, avec des classes de concepts, d'une approche sémantique permettant de prendre en compte la polysémie et la synonymie, fréquentes dans le corpus, et d'une approche syntaxique

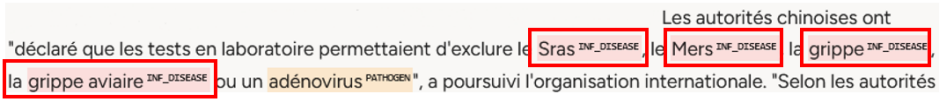

FIGURE 5 – Segments de texte (en rouge) correctement détectés mais associés à la mauvaise catégorie d'entité (« INF-DISEASE » au lieu de « PATHOGEN »)

FIGURE 6 – Entités correctement détectées mais contrevenant aux règles établies dans le guide d'annotation

et linguistique, permettant d'identifier des abus ou des erreurs de langage, en associant par exemple des types d'entités à des patterns spécifiques. La capacité des modèle de langue à prendre en compte cette approche hybride, permettant de saisir les nuances du texte et de réinterpréter le vocabulaire en fonction du contexte et du locuteur est encore lacunaire, même pour les LLM (Suravee *et al.*, 2025).

3.4.2 Complexité liée à la détection des évènements sanitaires

Un événement est un objet d'information complexe caractérisé par de nombreuses propriétés. Dans le cadre de ce challenge, un évènement peut être considéré comme un objet multi-tokens discontinu ou comme un graphe d'entités. En traitement automatique du langage, et plus spécifiquement dans le domaine de l'extraction d'information, il existe de très nombreuses définitions de la notion d'évènement, sans qu'aucune ne fasse consensus (Sprugnoli & Tonelli, 2017). Le domaine spécifique de la veille sanitaire internationale ne fait pas exception (Delon *et al.*, 2024). Cette multiplicité de définitions complexifie la tâche d'extraction : il faut repenser la méthode et les modèles pour chaque domaine métier, mais également pour chaque modification structurelle du concept d'évènement. Dans ce challenge, un événement sanitaire, pour être pleinement caractérisé, peut intégrer des dimensions spatio-temporelles multiples (ex : propagation transfrontalière d'un agent pathogène, mobilité des populations affectées, etc.). De même, plusieurs évènements distincts peuvent avoir un recouvrement spatio-temporel, ce qui rend plus difficile encore l'identification unitaire de chacun d'eux. Pour un même temps et un même lieu, la dichotomisation entre deux évènements nécessite une compréhension fine de la situation décrite dans le texte, requérant parfois des connaissances médicales, et une bonne compréhension des règles du guide d'annotation. Le même problème se retrouve également en cas de recouvrement du sujet de l'évènement. Dans l'exemple ci-dessous (Figure 7), si l'élément central est identique pour les deux évènements sanitaires, chacun renvoie à une épidémie distincte, associée à des contextes géographiques et temporels différents. Le premier fait référence à un événement récent, tandis que le second évoque une épidémie survenue dans le passé. L'analyse des runs, qui relève une dégradation des performances de détection avec le nombre d'évènements à détecter, souligne également cette difficulté de séparation des évènements.

3.4.3 Bénéfices attendus

Sur le plan opérationnel, l'extraction d'informations à partir de sources ouvertes (presse généraliste, publications institutionnelles, blogs spécialisés, réseaux sociaux, etc.) constitue une étape essentielle de la veille sanitaire de défense [2]. Les résultats limités des différents runs dans la tâche d'extraction d'évènement doivent être mis en perspective avec les excellents résultats dans la classification binaire

2. Early detection, assessment and response to acute public health events: implementation of early warning and response with a focus on event-based surveillance: interim version

Évènement n°1 : « Entité d'intérêt (maladie : **Ebola**) – Élément associé à l'entité d'intérêt (agent pathogène : **virus ebola**) – Lieu (**République Démocratique du Congo**) – Date relative (vendredi) »

Évènement n°2 : « Entité d'intérêt (maladie : **Ebola**) – Élément associé à l'entité d'intérêt (agent pathogène : **virus ebola**) – Lieu (**Afrique de l'Ouest**) – Lieu (**Nigeria**) – Période absolue (**2014**) »

FIGURE 7 – Exemple de document annoté comprenant deux évènements sanitaires

des documents, c'est-à-dire la capacité à déterminer si un document contient ou non au moins un évènement sanitaire. La mise en œuvre de cette capacité à filtrer les documents d'intérêt au sein d'un flux de documents dans un système d'information de veille permettrait non seulement de réduire la charge cognitive des analystes humains, mais aussi de garantir une couverture plus large et continue du spectre informationnel.

4 Challenge de fine tuning de LLM

4.1 Organisation

Le second challenge proposé dans le cadre de l'atelier portait sur le fine-tuning de LLM pour produire un modèle adapté à un domaine particulier. En l'occurrence, il s'agit du domaine de la défense, riche en vocabulaire, en sigles et en connaissances métier. L'objectif était de faire émerger les meilleures pratiques et techniques pour l'adaptation de modèle, les hyper-paramètres essentiels, et de mesurer l'impact des données et les coûts associés.

Des textes du domaine ont été fournis aux participants ; ils sont issus de sources publiques ou internes non sensibles. Ils sont de thématiques variées autour du domaine de la défense : ils décrivent l'organisation du ministère des Armées, les procédures au sein de différents services, les grands concepts et les doctrines des armées, l'état de la recherche en géopolitique, etc. L'ensemble totalisait environ 100 millions de mots, et les participants pouvaient, s'ils le souhaitaient, collecter d'autres textes.

Deux modèles de base étaient proposés pour le fine-tuning, soit le Mistral-7B v0.3 [3], soit le Mistral-Small-24B (3.0) [4]. Pour permettre de comparer au mieux, nous distinguons les runs utilisant uniquement les données fournies (données fermées) de ceux utilisant des données supplémentaires (données ouvertes). Trois *runs* pour chaque catégorie et chaque modèle de base pouvaient être soumis.

4.2 Evaluation

Les performances des modèles produits sont mesurées selon plusieurs axes : non-régression sur le domaine général, performances sur le domaine défense, tendance à l'hallucination et coût en équivalent carbone. À l'exception du coût carbone, fourni par les participants, ces évaluations de performances ont été effectuées par l'AMIAD sur un ensemble de jeux d'évaluation développés au sein du Ministère des Armées.

La non-régression est évaluée avec MMLU (Hendrycks *et al.*, 2021) (mesurée par *accuracy*) et FrenchBench (Faysse *et al.*, 2024) (par *accuracy* ou ROUGE selon les tâches), couvrant ainsi des connaissances générales, en anglais et en français. L'adaptation au domaine est mesurée au travers de plusieurs tâches :

— résumé de documents, sur des documents longs et courts, mesuré par BERTscore ;
— QCM : plus de 500 questions sur différents domaines de la défense, mesurées par le taux de bonnes réponses (*accuracy*) ;
— titrage d'articles d'actualité du domaine, mesuré par BERTscore ;

3. https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3
4. https://huggingface.co/mistralai/Mistral-Small-24B-Instruct-2501

— Questions-réponses sur un petit ensemble de questions, jugées comme des connaissances absolument essentielles du domaine de la défense et du Ministère des Armées, mesurées par BERTscore et par des expressions régulières.

Enfin, la tendance à l'hallucination est mesurée par différents biais : d'une part, un jeu de questions (questions sur des points qui ne peuvent pas être connus du modèle, sur des services ou des personnes du ministère des Armées inventés, etc.) auxquelles il est attendu que le LLM réponde qu'il ne sait pas. Nous mesurons cela à l'aide d'expressions régulières et de détection d'entités. Nous mesurons aussi la perplexité (sur les mots) des modèles sur des textes du domaine ; il est attendu qu'elle reste proche ou légèrement inférieure à celle du modèle de base.

Quelques exemples de ces jeux d'évaluation ont été donnés aux participants pour information. Le jeu d'évaluation complet a vocation à rester interne au Ministère des Armées de par la présence de données sensibles et pour éviter des phénomènes de contamination qui biaiseraient de futures évaluations.

Enfin, le classement final pour désigner les vainqueurs du challenge est déterminé par la moyenne pondérée des ratios d'écart par rapport aux performances du modèle initial. Les pondérations permettent de donner plus d'importance aux tâches mesurant la non-régression, puis aux tâches d'adaptation et enfin une importance moindre aux tâches de mesure de l'hallucination.

4.3 Participation

Sept équipes ont participé au challenge d'adaptation de LLM, avec des équipes industrielles (Airbus, Orange, Ouest-France), académiques (CEA, Inria, CNRS), ou d'étudiants (Rousseau *et al.*, 2025; Boulanger *et al.*, 2025; Innocenzi *et al.*, 2025; Kouhoue, 2025; Rabuel & Duval, 2025). Ce sont ainsi 34 runs qui ont été soumis.

Pour ces *runs*, les participants ont exploré beaucoup de voies pour adapter les modèles. Parmi les éléments principaux distinguant les approches, on peut citer :

— l'emploi de données supplémentaires à celles fournies (Innocenzi *et al.*, 2025; Rousseau *et al.*, 2025) ;
— la continuation du préentraînement (CPT pour *Continuous Pre-Training*) vs. l'entraînement par instructions ;
— plusieurs participants ont testé l'entraînement par LoRA (ou QLoRA), avec différents rangs (Boulanger *et al.*, 2025; Innocenzi *et al.*, 2025; Kouhoue, 2025; Rabuel & Duval, 2025) ; Innocenzi *et al.* (2025); Rousseau *et al.* (2025) a également testé l'adaptation complète des poids du modèle (*full finetuning*).
— (Boulanger *et al.*, 2025) a utilisé la fusion de modèles, que ce soit entre différentes versions de modèles entraînés ou avec le modèle initial ;

L'ensemble des participants a proposé un ou plusieurs *runs* pour le *fine-tuning* du modèle 7B et seule une équipe a proposé un *fine-tuning* du modèle 24B (Kouhoue, 2025).

4.4 Résultats

La figure 8 présente les résultats du meilleur run de chaque équipe participante (en ratio par rapport au modèle de base). Le classement, selon la mesure globale présentée précédemment, donne le podium de runs suivant :

1. Orange – Ouest-France (données ouvertes) : 0.1220
2. Orange – Ouest-France (données ouvertes) : 0.1189
3. Orange – Ouest-France (données fermées) : 0.1177
4. CEA (données fermées) : 0.0912
5. CEA (données fermées) : 0.0900
6. Airbus (données ouvertes) : 0.0818

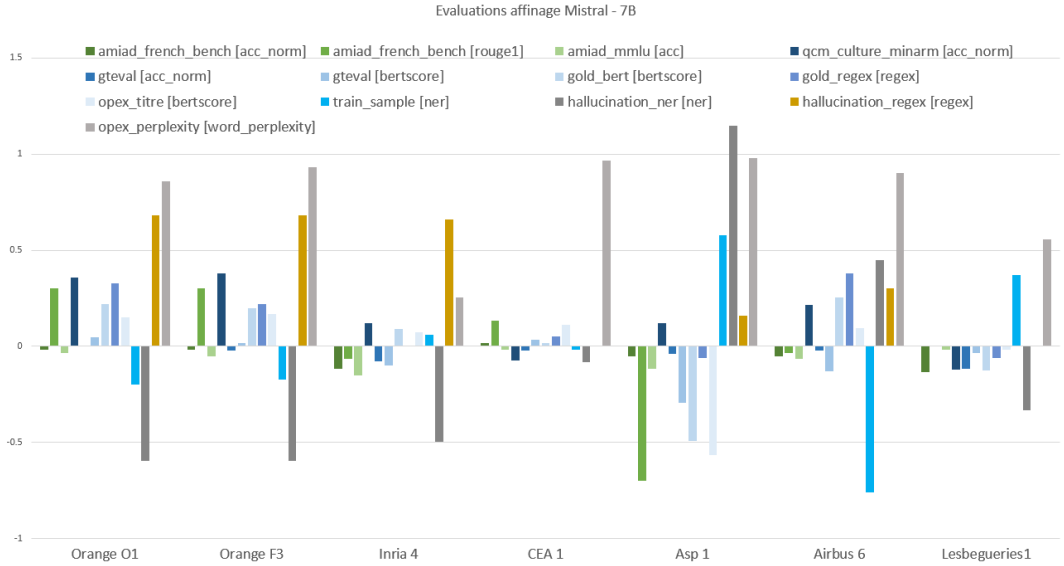

FIGURE 8 – Résultats des meilleurs *runs* de chaque participant pour le fine-tuning du modèle 7B, exprimés en écart au modèle initial.

Plusieurs points méritent d'être relevés. Tout d'abord, l'adaptation au domaine de la défense a tendance à dégrader les modèles sur le domaine général, à l'exception très notable des runs de Rousseau *et al.* (2025) qui montrent même une amélioration sur le français (FrenchBench). Ensuite, l'adaptation au domaine est réussie de manière très variable : allant de gains légers à des performances en baisse par rapport au modèle d'origine. À ce titre, la stratégie de fusion du CEA rend les modèles évalués très similaires, en terme de performances, au modèle initial, montrant peu de régression, peu d'hallucinations supplémentaires, mais aussi peu d'adaptation. Enfin, sur l'ensemble des runs, on observe que l'hallucination est corrélée à la force de l'adaptation ; certaines équipes ont cependant réussi à la contenir (Innocenzi *et al.*, 2025). On remarque au final une grande place à l'amélioration : la tâche d'adaptation de modèle, où les attendus ne sont pas seulement de reprendre un style langagier mais d'incorporer des connaissances, reste difficile.

Les participants avaient pour instruction de tracer le coût carbone de leurs expérimentations, sur toute leur chaîne, du pré-traitement des données jusqu'à l'apprentissage des modèles. Comme on peut le voir en figure 9, les résultats de coût carbone rapportés par les participants donnent des indications convergentes sur le coût énergétique de ce type de tâche. L'entraînement par (Q)LoRA d'un petit modèle tel que le Mistral 7B consomme de 280 g à 1,7 kg équivalent carbone, et jusqu'à 3,2 kg (environ 15 km en voiture thermique) pour les modèles ayant nécessité plusieurs entraînements et/ou

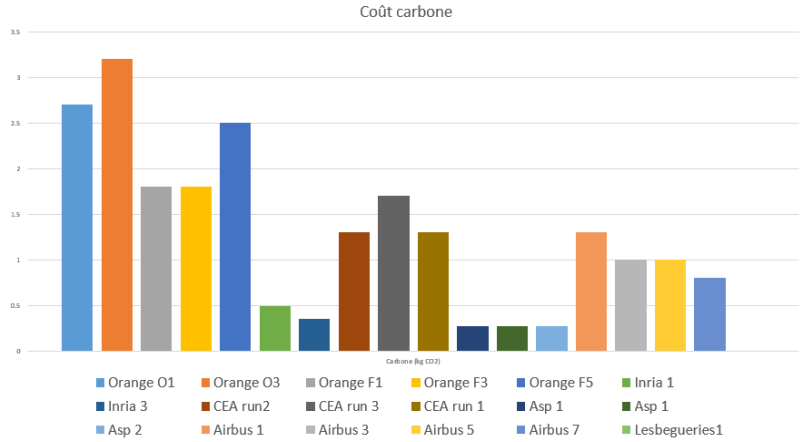

FIGURE 9 – Coûts carbones des principaux runs, en kg eq. CO2. Attention, dû à des disparités d'outils de mesure, les résultats ne sont pas parfaitement comparables.

du *full-finetuning*. Pour aller plus en détail, Boulanger *et al.* (2025) et (Innocenzi *et al.*, 2025) détaillent le coût de chaque étape de leurs entraînements, montrant notamment que le coût de préparation des données (génération d'instructions) pèse très lourd dans le coût total.

5 Conclusion

La seconde édition de l'atelier EvalLLM confirme la dynamique croissante de la communauté francophone autour des problématiques d'évaluation des grands modèles de langue, à la fois en nombre de soumissions et en termes d'audience durant l'atelier. Les échanges scientifiques et les deux challenges organisés ont permis d'illustrer la richesse des approches explorées, mais aussi la complexité persistante de cette évaluation, pourtant essentielle notamment pour les domaines spécialisés et les langues autres que l'anglais.

Les travaux présentés lors de l'atelier soulignent l'intérêt de concevoir des protocoles d'évaluation adaptés, fondés sur des données représentatives, bien annotées et culturellement situées. L'usage des LLM comme juges ouvre des perspectives prometteuses, tout en posant de nouvelles questions de fiabilité, de biais et de transparence. Les défis techniques identifiés — extraction d'évènements complexes, adaptation fine des modèles à des contextes métiers, mesure des hallucinations ou des performances multimodales — montrent que l'évaluation est un vaste champ de recherche.

Les travaux du challenge d'extraction d'information ont mis en évidence la difficulté d'exploiter pleinement les capacités des LLM dans des contextes à faibles ressources, tandis que le challenge de fine-tuning a montré que l'adaptation au domaine entraîne souvent un compromis entre spécialisation, stabilité et coût énergétique. Les résultats sur ces deux challenges de l'AMIAD et du SSA sont encourageants mais montrent une large marge d'amélioration, dès lors que l'on traite de domaines spécialisés.

Enfin, EvalLLM a de nouveau été un lieu d'échanges pour la communauté académique et industrielle,

favorisant la création et le partage de benchmarks ouverts, de bonnes pratiques d'évaluation et le développement de ressources en langue française ou dans d'autres langues moins dotées. L'évaluation des LLM demeure un chantier essentiel pour garantir leur fiabilité, leur équité et leur utilité dans les applications réelles.

Remerciements

L'organisation de cet atelier a été rendue possible grâce au soutien du Ministère des Armées. Nous tenons également à remercier les membres du comité scientifique de l'atelier pour leur travail de relecture très complet, malgré le court délai accordé. Nos remerciements vont ensuite au comité d'organisation de TALN pour leur confiance et leur aide pour la mise en place de cet atelier et la gestion des actes. Enfin, nous remercions les auteurs des soumissions, les participants des deux challenges et notre orateur invité Louis Martin (Mistral AI) pour leurs contributions scientifiques.

Références

ARMINGAUD R., PEUVOT A., BESANÇON R., MAURER C., SEMMAR N., FERRET O. & SOUIHI S. (2025). Cea-list@evalllm2025 extraction d'information : des llm mais sans modèle décodeur. In *Actes du challenge EvalLLM 2025 à la conférence CORIA-TALN 2025*, Marseille, France : AMIAD.

AUBERTIN L., GADEK G., SÉRASSET G., PRIEUR M., VUTH N., GRILHERES B., SCHWAB D. & LOPEZ C. (2025). POPCORN-RENS : un nouveau jeu de données en français annoté en entités d'intérêts sur une thématique sécurité et défense. In *Actes de CORIA-TALN-RJCRI-RECITAL 2025. Actes de l'atelier Évaluation des modèles génératifs (LLM) et challenge 2025 (EvalLLM)*, p. 1–10, Marseille, France : Association pour le Traitement Automatique des Langues.

BARKAR A., CHOLLET M., LABEAU M., BIANCARDI B. & CLAVEL C. (2025). Décoder le pouvoir de persuasion dans les concours d'éloquence : une étude sur la capacité des modèles de langues à évaluer la prise de parole en public. In *Actes de CORIA-TALN-RJCRI-RECITAL 2025. Actes de l'atelier Évaluation des modèles génératifs (LLM) et challenge 2025 (EvalLLM)*, p. 77–90, Marseille, France : Association pour le Traitement Automatique des Langues.

BELMADANI I., HASHEMI P. N., SEBBAG T., FORTIER G., QUINIOU S., MORIN E., DUFOUR R. & FAVRE B. (2025). Llm, au rapport ! extraction d'informations médicales entre prompting, fine-tuning et post-correction. In *Actes du challenge EvalLLM 2025 à la conférence CORIA-TALN 2025*, Marseille, France : AMIAD.

BOUDRAA H. & BELFATHI A. (2025). Exploiter le prompting pour l'extraction d'information à partir de textes médicaux français avec peu de ressources. In *Actes du challenge EvalLLM 2025 à la conférence CORIA-TALN 2025*, Marseille, France : AMIAD.

BOULANGER H., EVAN DUFRAISSE O. F. & SOUIHI S. (2025). CEA-List@EvalLLM2025 finetuning : adaptation par apprentissage continu, instruction et fusion de modèles. In *Actes du challenge EvalLLM 2025 à la conférence CORIA-TALN 2025*, Marseille, France : AMIAD.

CONNEAU A., KHANDELWAL K., GOYAL N., CHAUDHARY V., WENZEK G., GUZMÁN F., GRAVE E., OTT M., ZETTLEMOYER L. & STOYANOV V. (2019). Unsupervised cross-lingual representation learning at scale. *CoRR*, **abs/1911.02116**.

DECOUPES R. & GUILLE A. (2025). étude des déterminants impactant la qualité de l'information géographique chez les llms : famille, taille, langue, quantization et fine-tuning. In *Actes de*

CORIA-TALN-RJCRI-RECITAL 2025. Actes de l'atelier Évaluation des modèles génératifs (LLM) et challenge 2025 (EvalLLM), p. 108–119, Marseille, France : Association pour le Traitement Automatique des Langues.

DELON F., BEDUBOURG G., BOUSCARRAT L., MEYNARD J.-B., VALOIS A., QUEYRIAUX B., RAMISCH C. & TANTI M. (2024). Infectious risk events and their novelty in event-based surveillance : new definitions and annotated corpus. *Lang. Resour. Evaluation*, **59**, 277–295.

DETURCK K. (2025). Kairntech à evalllm 2025. In *Actes du challenge EvalLLM 2025 à la conférence CORIA-TALN 2025*, Marseille, France : AMIAD.

DEVLIN J., CHANG M., LEE K. & TOUTANOVA K. (2018). BERT : pre-training of deep bidirectional transformers for language understanding. *CoRR*, **abs/1810.04805**.

DINIZ N., KOOLI N., CHASSEUR L. & SOUTRENON P. (2025). Participation de l'équipe Défense au défi TextMine'25 en extraction de relations dans des bulletins de renseignement. In *TextMine 2025 - Atelier du Groupe de travail sur la fouille de texte*, Strasbourg, France. HAL : hal-04974411.

EL YAGOUBY M. A., ZEKROUM M., LAHMADI A., GHOGHO M. & FESTOR O. (2025). Evaluating llms efficiency using successive attempts on binary-outcome tasks. In *Actes de CORIA-TALN-RJCRI-RECITAL 2025. Actes de l'atelier Évaluation des modèles génératifs (LLM) et challenge 2025 (EvalLLM)*, p. 120–126, Marseille, France : Association pour le Traitement Automatique des Langues. Évaluation de l'efficacité des LLMs à l'aide de tentatives successives sur des tâches à résultat binaire.

FAYSSE M., FERNANDES P., GUERREIRO N. M., LOISON A., ALVES D. M., CORRO C., BOIZARD N., ALVES J., REI R., MARTINS P. H., CASADEMUNT A. B., YVON F., MARTINS A. F. T., VIAUD G., HUDELOT C. & COLOMBO P. (2024). Croissantllm : A truly bilingual french-english language model.

GATTI~PINHEIRO G., GHARSALLAH S., ROBALDO A., TOKAREVA M., GUENDOUZ I., TRONCY R., PAPOTTI P. & MICHIARDI P. (2025). Peut-on faire confiance aux juges ? validation de méthodes d'évaluation de la factualité par perturbation des réponses. In *Actes de CORIA-TALN-RJCRI-RECITAL 2025. Actes de l'atelier Évaluation des modèles génératifs (LLM) et challenge 2025 (EvalLLM)*, p. 228–252, Marseille, France : Association pour le Traitement Automatique des Langues.

GIBIER M., DUROSELLE R., SERRANO P., BOËFFARD O. & BONASTRE J.-F. (2025). évaluation de la description automatique de scènes audio par la tâche d'audio question answering. In *Actes de CORIA-TALN-RJCRI-RECITAL 2025. Actes de l'atelier Évaluation des modèles génératifs (LLM) et challenge 2025 (EvalLLM)*, p. 164–177, Marseille, France : Association pour le Traitement Automatique des Langues.

GRINA F. & KALASHNIKOVA N. (2025). évaluation de la robustesse des llm : Proposition d'un cadre méthodologique et développement d'un benchmark. In *Actes de CORIA-TALN-RJCRI-RECITAL 2025. Actes de l'atelier Évaluation des modèles génératifs (LLM) et challenge 2025 (EvalLLM)*, p. 151–163, Marseille, France : Association pour le Traitement Automatique des Langues.

HADDAG E., VERDIERE T. R., MENAD S., MEDEIROS G. H. A. & SOUALMIA L. F. (2025). Tibs@evalllm : extraction d'entités et d'événements dans des documents francophones par prompt engineering. In *Actes du challenge EvalLLM 2025 à la conférence CORIA-TALN 2025*, Marseille, France : AMIAD.

HENDRYCKS D., BURNS C., BASART S., ZOU A., MAZEIKA M., SONG D. & STEINHARDT J. (2021). Measuring massive multitask language understanding. In *Proceedings of ICLR* : OpenReview.net.

HERSERANT T. & GUIGUE V. (2025). Allsummedup : un framework open-source pour comparer les métriques d'évaluation de résumé. In *Actes de CORIA-TALN-RJCRI-RECITAL 2025. Actes de l'atelier Évaluation des modèles génératifs (LLM) et challenge 2025 (EvalLLM)*, p. 11–21, Marseille, France : Association pour le Traitement Automatique des Langues.

INNOCENZI L., OLIVERI U., ANTHOINE L., GADEK G. & CASSART C. (2025). Challenge finetuning evalllm : Spécialisation de grands modèles de langue sur le domaine de la défense. In *Actes du challenge EvalLLM 2025 à la conférence CORIA-TALN 2025*, Marseille, France : AMIAD.

JOURDAIN L. & HELLAL S. (2025). Générer pour mieux tester : vers des datasets diversifiés pour une évaluation fiable des systèmes de question answering. In *Actes de CORIA-TALN-RJCRI-RECITAL 2025. Actes de l'atelier Évaluation des modèles génératifs (LLM) et challenge 2025 (EvalLLM)*, p. 204–227, Marseille, France : Association pour le Traitement Automatique des Langues.

KOOLI N., FLAMENT J., DUTREY C., DINIZ N. & CLAVEAU V. (2024). EvalLLM 2024 : présentation de l'atelier Evaluation des LLM et du Challenge en extraction d'information few-shot. In *EvalLLM2024 - Atelier sur l'évaluation des modèles génératifs (LLM) et challenge d'extraction d'information few-shot*, Toulouse, France : AMIAD, Ministère des Armées. HAL : hal-04926863.

KOUHOUE J. M. (2025). Fine-tuning des modèles mistral 7b et 24b pour le domaine de la défense à l'aide d'un adaptateur qlora. In *Actes du challenge EvalLLM 2025 à la conférence CORIA-TALN 2025*, Marseille, France : AMIAD.

MARTINON G., LORENZO~DE~BRIONNE A., BOHARD J., LOJOU A., HERVAULT D. & BRUNEL N. (2025). Vers une évaluation rigoureuse des systèmes rag : le défi de la due diligence. In *Actes de CORIA-TALN-RJCRI-RECITAL 2025. Actes de l'atelier Évaluation des modèles génératifs (LLM) et challenge 2025 (EvalLLM)*, p. 291–308, Marseille, France : Association pour le Traitement Automatique des Langues.

MOUGHIT I. & HAFIDI I. (2025). Amélioration et automatisation de la génération des cas de tests logiciels à l'aide du modèle llama. In *Actes de CORIA-TALN-RJCRI-RECITAL 2025. Actes de l'atelier Évaluation des modèles génératifs (LLM) et challenge 2025 (EvalLLM)*, p. 22–35, Marseille, France : Association pour le Traitement Automatique des Langues.

PEREZ J., CONRAD A. & ELKOUSSY L. (2025a). évaluation pédagogique du code à l'aide de grands modèles de langage. une étude comparative à grande échelle contre les tests unitaires. In *Actes de CORIA-TALN-RJCRI-RECITAL 2025. Actes de l'atelier Évaluation des modèles génératifs (LLM) et challenge 2025 (EvalLLM)*, p. 188–201, Marseille, France : Association pour le Traitement Automatique des Langues.

PEREZ J., PELLET A. & PUREN M. (2025b). évaluation automatique du retour à la source dans un contexte historique long et bruité. application aux débats parlementaires de la troisième république française. In *Actes de CORIA-TALN-RJCRI-RECITAL 2025. Actes de l'atelier Évaluation des modèles génératifs (LLM) et challenge 2025 (EvalLLM)*, p. 138–150, Marseille, France : Association pour le Traitement Automatique des Langues.

RABUEL W. & DUVAL C. (2025). [aspirants data scientists] : Rapport fine-tuning. In *Actes du challenge EvalLLM 2025 à la conférence CORIA-TALN 2025*, Marseille, France : AMIAD.

ROUSSEAU I., PERROUX C., ADAM P., GIRAULT T., DELPHIN-POULAT L., VEYRET M., LECORVÉ G. & DAMNATI G. (2025). O_FT@EvalLLM2025 : étude comparative de choix de données et de stratégies d'apprentissage pour l'adaptation de modèles de langue à un domaine. In *Actes du challenge EvalLLM 2025 à la conférence CORIA-TALN 2025*, Marseille, France : AMIAD. HAL : hal-05140334.

ROZERA E., MELLOULI-NAUWYNCK N., LEGUIDE P. & MORCOMBE W. (2025). Des prompts aux profils : Evaluation de la qualité des données générées par llm pour la classification des soft skills. In *Actes de CORIA-TALN-RJCRI-RECITAL 2025. Actes de l'atelier Évaluation des modèles génératifs (LLM) et challenge 2025 (EvalLLM)*, p. 91–107, Marseille, France : Association pour le Traitement Automatique des Langues.

SÉJOURNÉ K., FOUCHER M., LATA A. & LEBRATY J.-F. (2025). évaluation comparative de la génération contrainte vs. du post-parsing pour l'analyse de contenu par llms : étude sur le corpus euvsdisinfo. In *Actes de CORIA-TALN-RJCRI-RECITAL 2025. Actes de l'atelier Évaluation des modèles génératifs (LLM) et challenge 2025 (EvalLLM)*, p. 127–137, Marseille, France : Association pour le Traitement Automatique des Langues.

SERVAN C., GROUIN C., NÉVÉOL A. & ZWEIGENBAUM P. (2025). Comment évaluer un grand modèle de langue dans le domaine médical en français ? In *Actes de CORIA-TALN-RJCRI-RECITAL 2025. Actes de l'atelier Évaluation des modèles génératifs (LLM) et challenge 2025 (EvalLLM)*, p. 51–67, Marseille, France : Association pour le Traitement Automatique des Langues.

SOUTRENON P., BADER L. & CHASSEUR L. (2025). Participation de l'équipe inria défense et sécurité au défi evalllm 2025 en reconnaissance d'entités nommées et extraction de relations dans le domaine biomédical. In *Actes du challenge EvalLLM 2025 à la conférence CORIA-TALN 2025*, Marseille, France : AMIAD.

SPRUGNOLI R. & TONELLI S. (2017). One, no one and one hundred thousand events : Defining and processing events in an inter-disciplinary perspective. *Natural Language Engineering*, **23**(4), 485–506. DOI : 10.1017/S1351324916000292.

SURAVEE S., SCHMIDT C. O. & YORDANOVA K. (2025). The challenge of performing ontology-driven entity extraction in real-world unstructured textual data from the domain of dementia. In *Proceedings of Recent Advances in Natural Language Processing*, p. 1205–1214. DOI : 10.26615/978-954-452-098-4-139.

VALETTE M. (2025). Culture et acculturation des grands modèles de langue. In *Actes de CORIA-TALN-RJCRI-RECITAL 2025. Actes de l'atelier Évaluation des modèles génératifs (LLM) et challenge 2025 (EvalLLM)*, p. 68–76, Marseille, France : Association pour le Traitement Automatique des Langues.

VALLET S. & SUIGNARD P. (2025). Evaluation de petits modèles de langues (slm) sur un corpus de sciences humaines et sociales (shs) en français. In *Actes de CORIA-TALN-RJCRI-RECITAL 2025. Actes de l'atelier Évaluation des modèles génératifs (LLM) et challenge 2025 (EvalLLM)*, p. 178–187, Marseille, France : Association pour le Traitement Automatique des Langues.

VARTAMPETIAN M., FABRE D., MULHEM P., JOUBERT S. & SCHWAB D. (2025). Supergpqa-hce-fr : un corpus spécialisé en français pour le domaine hydraulique et le génie civil. In *Actes de CORIA-TALN-RJCRI-RECITAL 2025. Actes de l'atelier Évaluation des modèles génératifs (LLM) et challenge 2025 (EvalLLM)*, p. 253–276, Marseille, France : Association pour le Traitement Automatique des Langues.

A Annexe 1 : Comités

Comité d'organisation :

- Vincent Claveau, AMIAD, Rennes, vincent.claveau@def.gouv.fr
- Julianne Flament, AMIAD, Rennes
- Lorenzo Gerardi, AMIAD, Rennes
- Nihel Kooli, AMIAD, Rennes, nihel.kooli@def.gouv.fr
- Maxime Poulain, AMIAD, Rennes

Comité scientifique :

- Rachel Bawden, Inria
- Lucie Chasseur, Inria mission Défense et Sécurité
- Olivier Ferret, CEA-List
- Vincent Guigue, AgroParisTech, UMR MIA-Paris-Saclay
- Damien Nouvel, INALCO
- Didier Schwab, LIG
- Gilles Sérasset, LIG
- Aurélie Névéol, LISN - CNRS
- Fabian Suchanek, Télécom Paris, Institut polytechnique de Paris
- François Yvon, ISIR - CNRS