CORIA-TALN 2025

20e Conférence en Recherche d'Information et Applications (CORIA)
32ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN)
27ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL)
Les 18e Rencontres Jeunes Chercheurs en RI (RJCRI)
(CORIA-TALN) [1]

Actes de CORIA-TALN-RJCRI-RECITAL 2025.

Actes de l'atelier Évaluation des modèles génératifs (LLM) et challenge 2025 (EvalLLM)

Frédéric BECHET, Adrian-Gabriel CHIFU, Karen PINEL-SAUVAGNAT, Benoit FAVRE, Eliot MAES, Diana NURBAKOVA (Éds.)

Marseille, France, 30 juin au 4 juillet 2025

1. https://coria-taln-2025.lis-lab.fr

Avec le soutien de

Préface

Atelier EvalLLM 2025

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 sérieux : 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 usages. La question de leur évaluation se pose en particulier pour le français et plus généralement pour des langues autres que l'anglais.

C'est dans ce contexte que l'AMIAD (Agence Ministérielle pour l'IA de Défense) a organisé cette seconde édition de l'atelier EvalLLM dans laquelle nous réunissons les chercheuses et chercheurs, industriels et académiques, du domaine.

Articles présentés

Une vingtaine de soumissions ont été retenues par le comité de programme, présentées sous forme de posters ou d'exposés oraux, sans distinction de qualité. Ces travaux couvrent les multiples facettes de l'évaluation des LLM mises en avant dans l'appel :

- 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...) ;
- ...

Challenges

Deux challenges ont également été proposés par l'AMIAD. Le premier visait à une évaluation des LLM par la tâche. Il s'agit d'extraction d'information (reconnaissance d'entités et identification d'évènements) dans le domaine de la santé, en français, dans un contexte few-shot où seuls étaient donnés le guide d'annotation et quelques documents annotés. Les participants étaient invités à évaluer les mises-en-oeuvre par des LLM ou par d'autres approches pour permettre de mettre en perspective les résultats des systèmes fondés sur les LLM.
Le deuxième challenge visait à évaluer l'intérêt du fine-tuning sur un domaine de compétences particulier. En l'occurrence, il s'agissait 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.

Invité

Enfin, nous avons la chance d'avoir Louis Martin en exposé invité. Chercheur en charge du post-training des modèles à Mistral.AI et précédemment à Meta, sa présentation se concentre sur la mise au point

de grands modèles de langage, en détaillant spécifiquement les processus et les avancées de Llama à Mistral. Elle couvre des aspects essentiels tels que la collecte de données, les pipelines d'alignement et l'utilisation de techniques telles que l'échantillonnage par rejet et l'optimisation directe des préférences (DPO) pour améliorer les performances du modèle et l'alignement sur les préférences humaines.

Vincent Claveau, Julianne Flament, Lorenzo Gerardi, Nihel Kooli, Maxime Poulain

Comités

Comité d'organisation

— Vincent Claveau, AMIAD, Ministère des armées
— Julianne Flament, AMIAD, Ministère des armées
— Lorenzo Gerardi, AMIAD, Ministère des armées
— Nihel Kooli, AMIAD, Ministère des armées
— Maxime Poulain, AMIAD, Ministère des armées

Comité de programme

— 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, Univ. Grenoble Alpes, LIG
— Gilles Sérasset, Univ. Grenoble Alpes, LIG
— Aurélie Névéol, LISN - CNRS
— Fabian Suchanek, Télécom Paris, Institut polytechnique de Paris
— François Yvon, ISIR - CNRS

Table des matières