



# 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)<sup>1</sup>*

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

Organisateurs



Soutiens académiques



Sponsors privés



# 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

<b>”POPCORN-RENS : un nouveau jeu de données en français annoté en entités d’intérêts sur une thématique ””sécurité et défense”””</b>	<b>1</b>
<i>Lucas Aubertin, Guillaume Gadek, Gilles Sérasset, Maxime Prieur, Nakanyseth Vuth, Bruno Grilheres, Didier Schwab, Cédric Lopez</i>	
<b>AllSummedUp : un framework open-source pour comparer les métriques d’évaluation de résumé</b>	<b>11</b>
<i>Tanguy Herserant , Vincent Guigue</i>	
<b>Amélioration et Automatisation de la Génération des Cas de Tests Logiciels à l’Aide du Modèle Llama</b>	<b>22</b>
<i>Imane Moughit , Imad Hafidi</i>	
<b>Approche générative de la conformation pragmatique : une étude de cas de l’analyse d’une conférence</b>	<b>36</b>
<i>Julien Perez, Idir Benouaret</i>	
<b>Comment évaluer un grand modèle de langue dans le domaine médical en français ?</b>	<b>51</b>
<i>Christophe Servan , Cyril Grouin, Aurélie Névéol, Pierre Zweigenbaum</i>	
<b>Culture et acculturation des grands modèles de langue</b>	<b>68</b>
<i>Mathieu Valette</i>	
<b>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</b>	<b>77</b>
<i>Alisa Barkar, Mathieu Chollet, Matthieu Labeau, Beatrice Biancardi, Chloé Clavel</i>	
<b>Des Prompts aux Profils : Evaluation de la qualité des données générées par LLM pour la classification des soft skills</b>	<b>91</b>
<i>Elena Rozera, Nédra Mellouli-Nauwynck, Patrick Leguide, William Morcombe</i>	
<b>Étude des déterminants impactant la qualité de l’information géographique chez les LLMs : famille, taille, langue, quantization et fine-tuning</b>	<b>108</b>
<i>Rémy Decoupes, Adrien Guille</i>	
<b>Evaluating LLMs Efficiency Using Successive Attempts on Binary-Outcome Tasks</b>	<b>120</b>
<i>Mohamed Amine El Yagouby, Mehdi Zekroum, Abdelkader Lahmadi, Mounir Ghogho, Olivier Festor</i>	
<b>Évaluation Comparative de la Génération Contrainte vs. du Post-Parsing pour l’Analyse de Contenu par LLMs : Étude sur le Corpus EUvsDisinfo</b>	<b>127</b>
<i>Kévin Séjourné, Marine Foucher, Alexandru Lata, Jean-Fabrice Lebraty</i>	
<b>Évaluation automatique du retour à la source dans un contexte historique long et bruité. Application aux débats parlementaires de la Troisième République française</b>	<b>138</b>
<i>Julien Perez, Aurélien Pellet, Marie Puren</i>	
<b>Évaluation de la Robustesse des LLM : Proposition d’un Cadre Méthodologique et Développement d’un Benchmark</b>	<b>151</b>
<i>Fares Grina, Natalia Kalashnikova</i>	

<b>Évaluation de la description automatique de scènes audio par la tâche d’Audio Question Answering</b>	<b>164</b>
<i>Marcel Gibier, Raphaël Duroselle, Pierre Serrano, Olivier Boëffard, Jean-François Bonastre</i>	
<b>Evaluation de petits modèles de langues (SLM) sur un corpus de Sciences Humaines et Sociales (SHS) en français</b>	<b>178</b>
<i>Sam Vallet, Philippe Suignard</i>	
<b>Évaluation pédagogique du code à l’aide de grands modèles de langage. Une étude comparative à grande échelle contre les tests unitaires</b>	<b>188</b>
<i>Julien Perez, Anton Conrad, Elkoussy Laïla</i>	
<b>Exploration de stratégies de prédiction de la complexité lexicale en contexte multilingue à l’aide de modèles de langage génératifs et d’approches supervisées.</b>	<b>202</b>
<i>Abdelhak Kelious</i>	
<b>Générer pour mieux tester : vers des datasets diversifiés pour une évaluation fiable des systèmes de Question Answering</b>	<b>204</b>
<i>Louis Jourdain, Skander Hellal</i>	
<b>Peut-on faire confiance aux juges ? Validation de méthodes d’évaluation de la factualité par perturbation des réponses</b>	<b>228</b>
<i>Giovanni Gatti Pinheiro, Sarra Gharsallah, Adèle Robaldo, Mariia Tokareva, Ilyana Guendouz, Raphaël Troncy, Paolo Papotti, Pietro Michiardi</i>	
<b>SuperGPQA-HCE-FR : un corpus spécialisé en français pour le domaine hydraulique et le génie civil</b>	<b>253</b>
<i>Markarit Vartampetian, Diandra Fabre, Philippe Mulhem, Sylvain Joubert, Didier Schwab</i>	
<b>Une Approche Linguistique pour l’Évaluation des Caractéristiques du Langage Parlé dans les Modèles Conversationnels</b>	<b>277</b>
<i>Oussama Silem, Maiwenn Fleig, Philippe Blache, Houda Oufaida, Leonor Becerra-Bonache</i>	
<b>Vers une évaluation rigoureuse des systèmes RAG : le défi de la due diligence</b>	<b>291</b>
<i>Grégoire Martinon, Alexandra De Brionne Lorenzo, Jérôme Bohard, Antoine Lojou, Damien Hervault, Nicolas Brunel</i>	