Actes de CORIA-TALN 2026
Nantes, France, 29 juin au 3 juillet 2026

Évaluation de l'adaptabilité des grands modèles de langage aux genres linguistiques attestés (AGLAGLA)

Ziyan Xu[1,2] Marina Seghier[1] Alice Millour[1]
Carlos-Emiliano González-Gallardo[2] Jean-Yves Antoine[2]
(1) LIASD, Université Paris 8, France
(2) LIFAT, Université de Tours, France
xzy1874@gmail.com, ms@up8.edu, am@up8.edu,
gonzalezgallardo@univ-tours.fr, jean-yves.antoine@univ-tours.fr

RÉSUMÉ

La question de l'adaptation des modèles au contexte linguistique, en particulier au genre textuel ou au domaine, a longtemps été centrale dans les techniques d'apprentissage automatique en TAL. Il est souvent considéré que les grands modèles de langue (GML) ont des capacités de généralisation telles qu'ils ne sont plus concernés par cette problématique. Leur capacité à s'adapter aux variations textuelles reste toutefois encore peu explorée, ce qui empêche de confirmer pleinement cette hypothèse. Notre étude aborde cette question à travers la tâche de reconnaissance d'entités nommées (REN) en français, menée sur NEM.fr, un corpus multigenre de référence que nous avons développé spécifiquement pour évaluer la robustesse des systèmes de REN dans des contextes linguistiques variés. Le corpus NEM.fr couvre onze types de textes, allant de la prose juridique et encyclopédique à la poésie, en passant par le discours politique, la parole spontanée et les échanges en ligne. Nous évaluons ici le modèle orienté raisonnement DeepSeek R1 selon six configurations de consignes (*zero-shot*, *one-shot* et *few-shot*, avec et sans raisonnement de type *chain-of-thought*), tout en maintenant constants le schéma d'annotation d'entités nommées, le format des *prompts* et la chaîne d'évaluation afin d'isoler le rôle de la variation du genre. Les performances sont mesurées à l'aide de métriques basées sur la F-mesure, en versions stricte et assouplie, en fonction des frontières des entités détectées. Les résultats montrent que les choix de *prompting* ont peu d'effet une fois que le modèle a appris le format de la tâche, mais que les différences de type de texte influent fortement sur les résultats : les scores de F1 assouplis vont d'environ 0,85 dans les genres formels à moins de 0,20 dans les genres informels. Même dans des conditions strictement contrôlées, le comportement des GMLs se révèle très sensible à la régularité textuelle et à la variation stylistique, ce qui met en évidence le genre textuel comme facteur clé de l'évaluation de la robustesse des modèles et suggère que les GMLs, eux aussi, demeurent concernés par cette question.

ABSTRACT

The issue of adapting models to the linguistic context, particularly to textual genre or domain, has long been central in machine learning techniques in NLP. It is often considered that large language models (LLM) have such generalization capabilities that they are no longer concerned with this problem. However, their ability to adapt to textual variation remains underexplored, preventing full confirmation of this hypothesis. Our study addresses this question through the task of named entity recognition (NER) in French, conducted on NEM.fr, a multi-genre reference corpus that we specifically developed to evaluate the robustness of NER systems across diverse linguistic contexts. The NEM.fr corpus

Actes des 33ème Conférence sur le Traitement Automatique des Langues Naturelles.
Volume 2 : articles déjà publiés, pages 18–19.

covers 11 text types, ranging from legal and encyclopedic prose to poetry, including political speech, spontaneous speech, and online exchanges. We evaluate the DeepSeek R1 reasoning-oriented model across six prompt configurations (zero-shot, one-shot, and few-shot, with and without chain-of-thought reasoning), while keeping the named entity annotation scheme, prompt format, and evaluation pipeline constant to isolate the role of genre variation. Performance is measured using the F1 measure, in both strict and fuzzy versions, depending on the boundaries of detected entities. The results show that prompt choices have little effect once the model has learned the task format, but that differences in text type strongly influence the results : fuzzy F1 scores range from about 0.85 in formal genres to less than 0.20 in informal genres. Even under strictly controlled conditions, LLM behavior proves highly sensitive to textual regularity and stylistic variation, highlighting textual genre as a key factor in evaluating model robustness and suggesting that LLMs, too, remain concerned with this issue.

MOTS-CLÉS : Reconnaissance d'entités nommées, Grands modèles de langue, Extraction d'informations.

KEYWORDS: Named Entity Recognition, Large Language Models, Information Extraction.

ARTICLE ACCEPTÉ À : LREC 2026 : The Fifteenth biennial Language Resources and Evaluation Conference, Palma, Mallorca, Spain, May 11-16, 2026..
URL : https://lrec.elra.info/lrec2026-main-183