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

De l'importance des formats : une évaluation critique des formats de sorties dans les amorces des Grands Modèles de Langues pour la compréhension de la parole et la REN

Pierre Lepagnol[1,2] Sahar Ghannay[1] Thomas Gerald[1] Christophe Servan[1,3]
Sophie Rosset[1]
(1) Université Paris-Saclay, CNRS, LISN, 91405, Orsay, France
(2) SCIAM, 75008, Paris, France
(3) AMIAD, Pôle Recherche, 91120, Palaiseau, France
firstname.lastname@lisn.fr, firstname.lastname@polytechnique.edu

RÉSUMÉ

Le format de sortie est un facteur souvent oublié lors de l'évaluation des grands modèles de langue (LLM) pour des tâches de remplissage de formulaire (slot-filling) ou de reconnaissance d'entités nommées (REN). Ce travail propose d'explorer l'impact des formats des structures des sorties générées par les LLM. Nous montrons que les résultats obtenus dépendent du format demandé (JSON, XML ou clé-valeur). Une étude est menée sur quatre tâches de compréhension de la parole et trois tâches de REN, avec treize LLM instruits à poids ouverts utilisant des amorces (prompts) et des analyseurs en sources ouvertes. Cette évaluation centrée sur les formats révèle des écarts significatifs de 2 à 46 points de F_1, selon les modèles et les corpus. Enfin, nous proposons une méthode élégante et peu impactante de sélection de la meilleure paire modèle-corpus en utilisant qu'une sous-partie du corpus de validation, ce qui permet de limiter le nombre d'essais.

ABSTRACT

Format Matters : A Critical Evaluation of Output Formats for Prompting LLMs in SLU and NER

Output format is often an unreported factor in LLM evaluations for structured NLP tasks such as Slot Filling or Named Entity Recognition. This work proposes to explore the impact of the output structured format generated by LLMs. We show that measured performance and reliability depend on the requested format (JSON, XML or inline Key-Values). A study is performed across four SLU and three NER benchmarks, considering 13 instruction-tuned open-weight LLMs, using standardized and open-source prompts and parsers. This format-specific evaluation reveals statistically significant swings of 2–46 F_1points, depending on model and dataset. Additionally, we propose a lightweight selection procedure to determine the best format per model-dataset combination using only a small development slice ; thus reducing trial-and-error in practice.

MOTS-CLÉS : évaluation, amorce de LLM, formats de sortie, compréhension de la parole, REN, grands modèles de langues.

KEYWORDS: evaluation, prompting, output formats, SLU, NER, large language models.

ARTICLE ACCEPTÉ À : The Fifteenth biennial Language Resources and Evaluation Conference (LREC 2026) .
URL : https://hal.science/hal-05546569

Actes des 33ème Conférence sur le Traitement Automatique des Langues Naturelles.
Volume 2 : articles déjà publiés, pages 11–11.