Actes de CORIA-TALN-RJCRI-RECITAL 2025
Marseille, France, 30 juin au 4 juillet 2025

Sondage des Modèles de Langue sur leur Source de Connaissance

Zineddine Tighidet[1, 2] Andrea Mogini[1]
Jiali Mei[1] Patrick Gallinari[2] Benjamin Piwowarski[2]
(1) BNP Paribas, Paris, France
(2) Sorbonne Université, CNRS, ISIR, F-75005 Paris, France
prénom.nom@{isir.upmc.fr, bnpparibas.com}

RÉSUMÉ

Les grands modèles de langue (GML) sont souvent confrontés à des conflits entre leurs connaissance interne (connaissance paramétrique, CP) et la connaissance externe fournie pendant l'inférence (connaissance contextuelle, CC). Comprendre comment les GML priorisent une source de connaissance par rapport à l'autre reste un défi. Dans cet article, nous proposons un nouveau cadre de sondage pour explorer les mécanismes régissant la sélection entre CP et CC dans les GML. En utilisant des prompts contrôlées conçues pour contredire la CP du modèle, nous démontrons que des activations spécifiques du modèle sont indicatives de la source de connaissance employée. Nous évaluons ce cadre sur divers GML de différentes tailles et démontrons que les activations des couches intermédiaires, en particulier celles liées aux relations dans l'entrée, sont cruciales pour prédire la sélection de la source de connaissances, ouvrant la voie à des modèles plus fiables capables de gérer efficacement les conflits de connaissances.

ABSTRACT

Probing Language Models on Their Knowledge Source

Large Language Models (LLMs) often encounter conflicts between their learned, internal (parametric knowledge, PK) and external knowledge provided during inference (contextual knowledge, CK). Understanding how LLMs models prioritize one knowledge source over the other remains a challenge. In this paper, we propose a novel probing framework to explore the mechanisms governing the selection between PK and CK in LLMs. Using controlled prompts designed to contradict the model's PK, we demonstrate that specific model activations are indicative of the knowledge source employed. We evaluate this framework on various LLMs of different sizes and demonstrate that mid-layer activations, particularly those related to relations in the input, are crucial in predicting knowledge source selection, paving the way for more reliable models capable of handling knowledge conflicts effectively.

MOTS-CLÉS : Interprétabilité, Transformers, Connaissance des modèles de langue .

KEYWORDS: Interpretability, Transformers, Language Models Knowledge.

ARTICLE : **Accepté à** EMNLP 2024 Workshop BlackboxNLP :
Lien ACL Anthology (Tighidet *et al.*, 2024) : https://aclanthology.org/2024.blackboxnlp-1.35/.

Actes des 32ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 2 : traductions d'articles publiés, pages 59–60.

Références

TIGHIDET Z., MEI J., PIWOWARSKI B. & GALLINARI P. (2024). Probing language models on their knowledge source. In Y. BELINKOV, N. KIM, J. JUMELET, H. MOHEBBI, A. MUELLER & H. CHEN, Éds., *Proceedings of the 7th BlackboxNLP Workshop : Analyzing and Interpreting Neural Networks for NLP*, p. 604–614, Miami, Florida, US : Association for Computational Linguistics. DOI : 10.18653/v1/2024.blackboxnlp-1.35.