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

Modulation de la Copie du Contexte : le Rôle des Neurones d'Entropie dans la Gestion des Conflits entre Connaissances Paramétriques et Contextuelles

Zineddine Tighidet[1, 2] Andrea Mogini[1] Hedi Ben-younes[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

Face à des informations contextuelles contredisant leurs connaissances paramétriques, le comportement des grands modèles de langage (LLM) est inconstant, sans explication claire de la distribution des résultats. Des travaux récents ont identifié dans les Transformers autorégressifs des neurones d'entropie. Ces derniers affectent significativement l'entropie de sortie du modèle, avec un impact modéré sur le classement des tokens prédits. Cet article étudie l'hypothèse selon laquelle ces neurones inhibent le comportement de copie du contexte en examinant leur rôle dans la résolution des conflits entre informations contextuelles et paramétriques. Nous démontrons que les neurones d'entropie suppriment la copie du contexte dans divers LLM et que leur ablation altère considérablement le processus de génération. Ces résultats éclairent la dynamique interne des LLM face aux informations conflictuelles.

ABSTRACT

Context Copying Modulation : The Role of Entropy Neurons in Managing Parametric and Contextual Knowledge Conflicts

The behavior of Large Language Models (LLMs) when facing contextual information that conflicts with internal parametric knowledge is inconsistent, with no generally accepted explanation for the expected outcome distribution. Recent work identified a class of neurons in autoregressive transformers—entropy neurons—that significantly affect output entropy while having a moderate impact on predicted token ranking. We investigate the hypothesis that these neurons inhibit context copying by examining their role in resolving contextual and parametric conflicts. We demonstrate that entropy neurons suppress context copying across various LLMs, and their ablation significantly alters generation. These results enhance our understanding of LLM internal dynamics during conflicting information processing.

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

KEYWORDS: Interpretability, Transformers, Language Models Knowledge.

ARTICLE ACCEPTÉ À : In Findings of the Association for Computational Linguistics : EMNLP 2025, pages 20469–20481, Suzhou, China. Association for Computational Linguistics. (Tighidet *et al.*, 2025).
URL : https://aclanthology.org/2025.findings-emnlp.1116/

Actes des 33ème Conférence sur le Traitement Automatique des Langues Naturelles.
Volume 2 : articles déjà publiés, pages 35–36.
Cette œuvre est mise à disposition sous licence Attribution 4.0 International.

Références

TIGHIDET Z., MOGINI A., BEN YOUNES H., MEI J., GALLINARI P. & PIWOWARSKI B. (2025). Context copying modulation : The role of entropy neurons in managing parametric and contextual knowledge conflicts. In C. CHRISTODOULOPOULOS, T. CHAKRABORTY, C. ROSE & V. PENG, Éds., *Findings of the Association for Computational Linguistics : EMNLP 2025*, p. 20469–20481, Suzhou, China : Association for Computational Linguistics. DOI : 10.18653/v1/2025.findings-emnlp.1116.