Biais de confirmation, de cadrage et de position dans les LLMs *

Liana Ermakova[1] Anton Firsov[2] Jaap Kamps[3]
(1) HCTI, Université de Bretagne Occidentale, 29200 Brest, France
(2) Université de Perm, 614000, Perm, Russie
(3) University of Amsterdam, 1012 WP Amsterdam, Pays Bas
liana.ermakova@univ-brest.fr, antfirsov@gmail.com, kamps@uva.nl

RÉSUMÉ

Les LLMs présentent des capacités remarquables de génération et de raisonnement, mais leurs productions reflètent souvent des biais cognitifs systématiques analogues à ceux observés dans le jugement humain. Cet article examine trois formes de biais interdépendantes : le biais de confirmation, le biais de position et le biais de cadrage. À travers une série d'expériences de prompting contrôlées des LLMs ouverts (Qwen, Mistral, Gemma, Olmo et LLaMA), nous montrons que les LLMs ont tendance à renforcer les prémisses intégrées dans les requêtes des utilisateurs (biais de confirmation), à favoriser les éléments initiaux ou saillants d'un prompt (biais de position), et à modifier leurs conclusions selon que l'entrée est formulée de manière positive ou négative (biais de cadrage). Nos résultats peuvent contribuer à améliorer les pratiques d'ingénierie de prompt, à renforcer les protocoles d'évaluation et à soutenir un usage responsable des LLM dans l'enseignement, la recherche et la prise de décision.

ABSTRACT

Confirmation, Framing, and Position Biases in LLM Responses

Large Language Models (LLMs) exhibit remarkable generative and reasoning capabilities, yet their outputs often reflect systematic cognitive biases analogous to those observed in human judgment. This paper investigates three interrelated forms of bias : confirmation bias, position bias, and framing bias. Through a series of controlled prompting experiments, we demonstrate that LLMs tend to reinforce the premises embedded in user queries (confirmation bias), favor initial or prominent elements within a prompt (position bias), and vary their conclusions depending on the positive or negative framing of the input (framing bias). We analyze these effects across different open LLMs : Qwen, Mistral, Gemma, Olmo, and LLama. These insights can inform better prompt engineering practices, strengthen evaluation benchmarks, and support the responsible use of LLMs in education, research, and decision-making.

MOTS-CLÉS : LLM, biais de confirmation, biais de position, biais de cadrage.

KEYWORDS: LLM, confirmation bias, position bias, framing bias.

ARTICLE ACCEPTÉ À : CHIIR '26 : 2026 ACM SIGIR Conference on Human Information Interaction and Retrieval .
URL : https://dl.acm.org/doi/10.1145/3786304.3787879

*. L. Ermakova a bénéficié d'un financement de l'ANR dans le cadre du projet ANR-22-CE23-0019-01 et du programme « France 2030 » (réf. ANR-19-GURE-0001). J. Kamps est soutenu par l'Organisation néerlandaise pour la recherche scientifique (NWO NWA #1518.22.105), l'Université d'Amsterdam (programme AI4FinTech) et ICAI (AI for Open Government Lab).