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

Étude des stéréotypes de genre dans les LLM à l'aide des Déterminants Sociaux de la Santé

Trung Hieu Ngo[1] Adrien Bazoge[2]
Solen Quiniou[1] Pierre-Antoine Gourraud[2] Emmanuel Morin[1]
(1) Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France
(2) Nantes Université, CHU Nantes, Clinique des données, INSERM, CIC 1413, F-44000 Nantes, France
{prénom.nom}@ls2n.fr, {prénom.nom}@chu-nantes.fr

RÉSUMÉ

Les grands modèles de langage (LLM) affichent d'excellentes performances dans les tâches de traitement automatique des langues (TAL). Cependant, ils propagent souvent les biais inhérents à leurs données d'entraînement, ce qui peut avoir des conséquences importantes dans des domaines sensibles comme la santé. Si les méthodes d'évaluation actuelles mesurent bien les biais liés à des déterminants sociaux de la santé (DSS) pris isolément, comme le genre ou l'origine ethnique, elles négligent souvent les interactions entre ces facteurs et manquent de contextualisation. La présente étude s'intéresse aux biais des LLM en analysant les relations entre le genre et d'autres DSS au sein de dossiers de patients français. À travers une série d'expériences, nous montrons que les stéréotypes intégrés dans les modèles peuvent être mis en évidence à partir des DSS fournis en entrée, et que les LLM s'appuient sur ces préjugés pour prendre des décisions genrées. Ces résultats suggèrent qu'évaluer les interactions entre les différents DSS apporterait un éclairage complémentaire aux approches existantes pour mesurer les performances et les biais des LLM.

ABSTRACT

Investigating Gender Stereotypes in Large Language Models via Social Determinants of Health

Large Language Models (LLM) excel in Natural Language Processing (NLP) tasks, but they often propagate biases embedded in their training data, which is potentially impactful in sensitive domains like healthcare. While existing benchmarks evaluate biases related to individual social determinants of health (SDoH) such as gender or ethnicity, they often overlook interactions between these factors and lack context-specific assessments. This study investigates bias in LLMs by probing the relationships between gender and other SDoH in French patient records. Through a series of experiments, we found that embedded stereotypes can be probed using SDoH input and that LLMs rely on embedded stereotypes to make gendered decisions, suggesting that evaluating interactions among SDoH factors could usefully complement existing approaches to assessing LLM performance and bias.

MOTS-CLÉS : Évaluation des biais, applications dans le domaine de la santé, TAL biomédicale.

KEYWORDS: Model bias/fairness evaluation, healthcare applications, clinical NLP.

ARTICLE ACCEPTÉ À : The 19th Conference of the European Chapter of the Association for Computational Linguistics.
URL : https://arxiv.org/abs/2603.09416/

Actes des 33ème Conférence sur le Traitement Automatique des Langues Naturelles.
Volume 2 : articles déjà publiés, pages 17–17.
Cette œuvre est mise à disposition sous licence Attribution 4.0 International.