

« Les femmes ne font pas de crise cardiaque ! »

Étude des biais de genre dans les cas cliniques synthétiques en français

Fanny Ducel¹ Nicolas Hiebel¹ Olivier Ferret² Karën Fort³ Aurélie Névéol¹

(1) Université Paris-Saclay, CNRS, LISN, 91400, Orsay, France

(2) Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France

(3) Université de Lorraine, CNRS, LORIA, F-54000 Nancy, France

¹prenom.nom@lisn.fr, ²olivier.ferret@cea.fr, ³karen.fort@loria.fr

RÉSUMÉ

De plus en plus de professionnels de santé utilisent des modèles de langue. Cependant, ces modèles présentent et amplifient des biais stéréotypés qui peuvent mettre en danger des vies. Cette étude vise à évaluer les biais de genre dans des cas cliniques générés automatiquement en français concernant dix pathologies. Nous utilisons sept modèles de langue affinés et un outil de détection automatique du genre pour mesurer les associations entre pathologie et genre. Nous montrons que les modèles sur-gènèrent des cas décrivant des patients masculins, allant à l'encontre des prévalences réelles. Par exemple, lorsque les invites ne spécifient pas de genre, les modèles génèrent huit fois plus de cas cliniques décrivant des patients (plutôt que des patientes) pour les crises cardiaques. Nous discutons des possibles dommages induits par les modèles de langue, en particulier pour les femmes et les personnes transgenres, de la définition d'un modèle de langue « idéal » et des moyens d'y parvenir.

ABSTRACT

"Women do not have heart attacks!" Gender Biases in Automatically Generated Clinical Cases in French

Healthcare professionals are increasingly including Language Models (LMs) in clinical practice. However, LMs have been shown to exhibit and amplify stereotypical biases that can cause life-threatening harm in a medical context. This study aims to evaluate gender biases in automatically generated clinical cases in French, on ten disorders. Using seven fine-tuned LMs and an automatic gender detection tool, we measure the associations between disorders and gender. We unveil that LMs over-generate cases describing male patients, creating synthetic corpora that are not consistent with documented prevalence for these disorders. For instance, when prompts do not specify a gender, LMs generate eight times more clinical cases describing male (vs. female patients) for heart attack. We discuss possible concerns about LM-induced harm, especially for women and transgender people, the definition of an "ideal" LM and the means to reach this goal.

MOTS-CLÉS : biais, stéréotype, modèle de langue, genre, cas cliniques, français.

KEYWORDS: bias, stereotype, language model, gender, clinical case, French.

ARTICLE : **Publié à NAACL 2025 Findings** (<https://aclanthology.org/2025.findings-naacl.398>)

Code : <https://github.com/FannyDucel/ClinicalCaseBias>.