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

Le *code-switching* comme indicateur de biais dans les LLM : *"The consequences are not the same para nosotros"*

Fanny Ducel[1] Aurélie Névéol[1] Vidit Khazanchi[2] Loïc Leclere[2,3]
Arthur Pedrini[2,3] Léa Bouchet[3] Benjamin Caissial[3] Karën Fort[2,3]
(1) Université Paris Saclay, CNRS, LISN, Orsay, France
(2) CNRS, LORIA, F-54000 Nancy, France
(3) Université de Lorraine, F-54000 Nancy, France
fanny.ducel@universite-paris-saclay.fr

RÉSUMÉ

Le code-switching est une pratique répandue chez les locuteur·ices bilingues. Bien que son impact sur les performances des LLM ait fait l'objet d'études récentes, les biais des modèles sur le code-switching demeurent inexplorés. Nous étudions donc son potentiel comme indicateur implicite d'ethnicité permettant de mesurer les biais racistes ou xénophobes des modèles. Des textes générés avec des prompts avec/sans code-switching sont comparés, en Hinglish et en Spanglish, deux formes de code-switching omniprésentes dans les communautés indienne et hispanique. Les différences sémantiques sont capturées avec un arbre de décision utilisant des ressources linguistiques, des listes de stéréotypes, de l'étiquetage morpho-syntaxique et des classifieurs de sentiments. Plus de 84 000 paires de textes sont générées avec trois LLM. Environ 50 % des paires ne sont pas sémantiquement équivalentes, et 25 % pourraient porter préjudice aux communautés indienne ou hispanique. Cette étude montre que les LLM sont des vecteurs de biais, impactant négativement les communautés discriminées.

ABSTRACT

Code-switching as a Bias Indicator in LLM : "The consequences are not the same para nosotros"

Code-switching is a widespread linguistic practice among bilingual speakers. Recent studies have addressed its impact on downstream performance, but the potential biases that LLM may cause when prompted with code-switching remain uninvestigated. We study whether code-switching constitutes an implicit indicator of ethnicity that can be leveraged to unveil covert racist or xenophobic bias in LLM. Our methodology compares generated texts prompted with code-switching vs. with monolingual inputs. It is applied on Hinglish and Spanglish, two prevalent forms of code-switching in Indian and Hispanic communities. Building a decision tree, we tackle semantic differences with semantic resources, stereotypes lists, POS-tagging and sentiment classifiers. Over 84k text pairs are generated with 3 popular LLM. Around 50% of generated pairs are not semantically equivalent, and 25% exhibit potential for harm against the Indian or Hispanic community. Relying on sociological studies, we argue that bias and harms against socially discriminated communities have greater consequences.

MOTS-CLÉS : biais, stéréotype, code-switching, LLM, anglais, hindi, espagnol.

KEYWORDS: bias, stereotype, code-switching, LLM, English, Hindi, Spanish.

ARTICLE ACCEPTÉ À : The 15th biennial Language Resources and Evaluation Conference (LREC).
URL : https://inria.hal.science/hal-05529786

Actes des 33ème Conférence sur le Traitement Automatique des Langues Naturelles.
Volume 2 : articles déjà publiés, pages 1–1.
Cette œuvre est mise à disposition sous licence Attribution 4.0 International.