A Cross-Lingual Analysis of Bias in Large Language Models Using Romanian History

Matei-Iulian Cocu[1], Răzvan Cosmin Cristia[2], AND Adrian Marius Dumitran[3]

[1]*University of Bucharest*
cocu.matei24@yahoo.com
[2]*University of Bucharest*
cristiarazvan@gmail.com
[3]*University of Bucharest, Softbinator*
marius.dumitran@unibuc.ro

Abstract

Large Language Models have become a common tool used by many on a daily basis, but their apparent objectivity can mask significant underlying biases. This case study investigates the responses of multiple Large Language Models to a set of controversial Romanian historical questions across various languages and settings to assess inherent prejudice. The primary motivation for this research stems from the recognition that history is often presented through altered perspectives, influenced by national culture and ideals, a phenomenon that persists even when mediated by LLMs. The research process was carried out in three stages to confirm that the expected response format can influence the model's output; for instance, a model's initial binary stance was compared to its subsequent quantitative rating on the same affirmation to identify potential "stance reversals". The results reveal a fundamental lack of stability; while binary Yes/No consistency is relatively high (means 0.75–0.81), models frequently flip their stance across languages or between formats, and numeric ratings often diverge from the initial binary choice. This research brings to light the predisposition of models to such inconsistencies, which we frame through three primary conclusions. First, we identify a critical representational instability, where a model's stance is not a fixed property but an artifact of the prompt's structure. Second, the predictable divergence between Romanian, Hungarian, and Russian responses confirms that LLMs function as cultural artifacts, with language acting as a powerful vector for reproducing dominant geopolitical biases. Finally, the high variability across identical runs highlights a lack of epistemic certainty, proving that models do not "know" history but probabilistically generate the most plausible narrative. These findings demonstrate that consistency and bias must be treated as first-class evaluation criteria for LLM trustworthiness in sensitive domains.

Keywords: Romanian History, LLM Linguistic Bias, LLM Training and Assessment, Natural Language Processing, Digital Humanities