Bias Detection in Multilingual LLMs

Abstract

This paper investigates Bias Detection in Multilingual LLMs. Based on 5 studies, we synthesize key findings and analyze methodologies. 5 experiments illustrate trends, challenges, and results. We highlight gaps in current approaches and propose potential directions for future research.

Introduction

The study of Bias Detection in Multilingual LLMs has gained significant attention in recent years. Previous research includes Ready to Translate, Not to Represent? Bias and Performance Gaps in Multilingual LLMs Across Language Families and Domains by Md. Faiyaz Abdullah Sayeedi, Md. Mahbub Alam, Subhey Sadi Rahman, Md. Adnanul Islam, Jannatul Ferdous Deepti, Tasnim Mohiuddin, Md Mofijul Islam, Swakkhar Shatabda; Guardians of Discourse: Evaluating LLMs on Multilingual Offensive Language Detection by Jianfei He, Lilin Wang, Jiaying Wang, Zhenyu Liu, Hongbin Na, Zimu Wang, Wei Wang, Qi Chen; Investigating Language and Retrieval Bias in Multilingual Previously Fact-Checked Claim Detection by Ivan Vykopal, Antonia Karamolegkou, Jaroslav Kopcan, Qiwei Peng, Tomas Javurek, Michal Gregor, Marián Simko. Despite these efforts, challenges remain in addressing bias, performance disparities, and limitations in experimental methodologies. This paper synthesizes findings from multiple studies, provides a comprehensive overview of experimental results, and identifies directions for improving model robustness and fairness.

Literature Review

Multilingual Large Language Models (LLMs) have been examined in 5 studies focusing on performance, bias, and fairness [1-5].

Performance disparities: Models often perform differently across high- and low-resource languages, revealing disparities.

Bias in data: LLMs inherit social and cultural biases from training datasets, influencing predictions for underrepresented groups.

Annotation bias: Annotation processes vary based on annotators' background, introducing cultural biases into datasets.

Offensive language detection: Detecting offensive language is challenging across languages due to cultural and linguistic variations.

Overall, these studies highlight limitations and provide potential directions for future multilingual LLM research.

Methodology and Experiments

To evaluate Bias Detection in Multilingual LLMs, 5 experiments were conducted using multiple LLMs, datasets, and languages.

Experiment 1: Ready to Translate, Not to Represent? Bias and Performance Gaps in Multilingual LLMs Across Language Families and Domains

- Dataset(s): Translation Tangles (unified framework and dataset), high-quality, bias-annotated dataset (1,439 translation-reference pairs based on human evaluations)
- Models: open-source LLMs
- Training Setup: Evaluation across 24 bidirectional language pairs and multiple domains.
- Metrics: translation quality metrics, fairness metrics, hybrid bias detection pipeline (rule-based heuristics, semantic similarity filtering, LLM-based validation)
- Key Results: Introduced Translation Tangles framework and dataset. Proposed a hybrid bias detection pipeline. Created a high-quality, bias-annotated dataset.

Experiment 2: Guardians of Discourse: Evaluating LLMs on Multilingual Offensive Language Detection

- Dataset(s): Augmented translation data
- Models: GPT-3.5, Flan-T5, Mistral
- Training Setup: Evaluation in both monolingual and multilingual settings. Examined impact of different prompt languages and augmented translation data.
- Metrics: multilingual offensive language detection performance
- Key Results: Evaluated multilingual offensive language detection capabilities of GPT-3.5, Flan-T5, and Mistral in English, Spanish, and German. Discussed the impact of inherent bias in LLMs and datasets on mispredictions related to sensitive topics.

Experiment 3: Investigating Language and Retrieval Bias in Multilingual Previously Fact-Checked Claim Detection

- Dataset(s): AMC-16K dataset
- Models: six open-source multilingual LLMs, multilingual embedding models
- Training Setup: Fully multilingual prompting strategy, translating task prompts into each language. Evaluation across 20 languages.
- Metrics: monolingual and cross-lingual performance, frequency of retrieved claims, retrieval performance
- Key Results: Uncovered disparities in monolingual and cross-lingual performance. Identified key trends based on model family, size, and prompting strategy. Highlighted persistent bias in LLM behavior. Revealed disproportionate retrieval of certain claims, leading to inflated performance for popular claims.

Experiment 4: RuBia: A Russian Language Bias Detection Dataset

- Dataset(s): RuBia (Russian Language Bias Detection Dataset) consisting of nearly 2,000 unique sentence pairs across 19 subdomains (gender, nationality, socio-economic status, diverse), created by volunteers and

validated by native-speaking crowdsourcing workers.
• Models: state-of-the-art or near-state-of-the-art LLMs
• Training Setup: Diagnostic evaluation.
• Metrics: bias detection
• Key Results: Introduced the RuBia dataset for Russian language bias detection. Conducted a diagnostic evaluation of LLMs to illustrate their predisposition to social biases using the new dataset.

Experiment 5: Bias in, Bias out: Annotation Bias in Multilingual Large Language Models

• Models: multilingual Large Language Models (LLMs)
• Training Setup: Proposed a comprehensive framework for understanding annotation bias, distinguishing instruction bias, annotator bias, and contextual/cultural bias. Outlined proactive and reactive mitigation strategies including diverse annotator recruitment, iterative guideline refinement, and post-hoc model adjustments.
• Metrics: inter-annotator agreement, model disagreement, metadata analysis, multilingual model divergence, cultural inference
• Key Results: Proposed a typology of annotation bias, a synthesis of detection metrics, and an ensemble-based bias mitigation approach adapted for multilingual settings. Provided an ethical analysis of annotation processes.

Results and Discussion

Performance varies significantly across tasks, datasets, and languages. High-resource languages consistently achieve better outcomes than low-resource languages. Bias patterns are observed based on dataset and model choices, emphasizing the need for mitigation strategies. Cross-lingual performance often reveals hidden disparities that monolingual evaluations might miss.

Conclusion

In conclusion, this paper presents a detailed analysis of Bias Detection in Multilingual LLMs. Based on 5 studies, trends, challenges, and methodologies were identified. 5 experiments revealed performance variations and bias patterns. These findings highlight the need for improved datasets, dynamic evaluation methods, and effective bias mitigation strategies, laying a foundation for future research in multilingual LLMs.

References

[1] Md. Faiyaz Abdullah Sayeedi, Md. Mahbub Alam, Subhey Sadi Rahman, Md. Adnanul Islam, Jannatul Ferdous Deepti, Tasnim Mohiuddin, Md Mofijul Islam, Swakkhar Shatabda. (2025). Ready to Translate, Not to Represent? Bias and Performance Gaps in Multilingual LLMs Across Language Families and Domains.
[2] Jianfei He, Lilin Wang, Jiaying Wang, Zhenyu Liu, Hongbin Na, Zimu Wang, Wei Wang, Qi Chen. (2024). Guardians of Discourse: Evaluating LLMs on Multilingual Offensive Language Detection.
[3] Ivan Vykopal, Antonia Karamolegkou, Jaroslav Kopcan, Qiwei Peng, Tomas Javurek, Michal Gregor, Marián Simko. (2025). Investigating Language and Retrieval Bias in Multilingual Previously Fact-Checked

Claim Detection.
[4] Veronika Grigoreva, Anastasiia Ivanova, I. Alimova, Ekaterina Artemova. (2024). RuBia: A Russian Language Bias Detection Dataset.
[5] Xia Cui, Ziyi Huang, Naeemeh Adel. (2025). Bias in, Bias out: Annotation Bias in Multilingual Large Language Models.