Automated and Augmented Evaluation of Bias in LLMs for High- and Low-Resource Languages

Alessio Buscemi[1,*], Cédric Lothritz[1], Sergio Morales[2], Marcos Gomez-Vazquez[1], Robert Clarisó[2], Jordi Cabot[1,3] and German Castignani[1]

[1]*Luxembourg Institute of Science and Technology, Esch-sur-Alzette, Luxembourg*
[2]*Universitat Oberta de Catalunya, Barcelona, Spain*
[3]*University of Luxembourg, Esch-sur-Alzette, Luxembourg*

Abstract

Large Language Models (LLMs) have exhibited impressive natural language processing capabilities but often perpetuate social biases inherent in their training data. To evaluate this, we introduce MultiLingual Augmented Bias Testing (MLA-BiTe), a framework that improves prior bias evaluation methods by enabling systematic multilingual bias testing. MLA-BiTe leverages automated translation and paraphrasing techniques to support extended assessments across diverse linguistic settings. In this study, we evaluate the effectiveness of MLA-BiTe by testing four state-of-the-art LLMs in six languages—including two low-resource languages—focusing on seven sensitive categories of discrimination.

Keywords

Large Language Models, Ethics, Biases, Testing

1. Introduction

Large Language Models (LLMs) have become central to modern Natural Language Processing (NLP) applications, excelling in tasks such as machine translation [1], text generation [2], and dialogue systems [3]. However, numerous studies have shown that these models can encode and reproduce harmful social biases, often rooted in the historical and cultural prejudices embedded in training data [4, 5, 6]. Such biases can manifest in overtly discriminatory content [7], raising ethical concerns, especially when models are deployed at scale [8, 9]. While significant work focuses on English and other high-resource languages, recent findings suggest that bias also affects less-resourced languages, sometimes in more subtle and harder-to-detect ways [10, 11]. Most existing bias evaluation frameworks have been designed for single-language contexts, limiting their ability to assess how LLM behavior varies across linguistic and cultural settings. Yet cross-lingual evaluation is increasingly necessary, especially for organizations seeking to align AI behavior with internal values across diverse language communities.

Enabling non-technical stakeholders, such as HR departments, Ethics Committees, and Diversity & Inclusion officers, to evaluate AI behavior in multiple languages is essential for trustworthy and inclusive AI development. However, extending bias evaluation to languages that are official yet underused in professional settings, such as Luxembourgish or Catalan, poses challenges. In many cases, the people responsible for defining fairness criteria are not fluent in these languages, making it difficult to craft or validate suitable prompts. This problem is exacerbated in countries with many co-official or regional languages, such as South Africa and India.

TRUST-AI: The European Workshop on Trustworthy AI. Organized as part of the European Conference of Artificial Intelligence - ECAI 2025. October 2025, Bologna, Italy.

[*] Corresponding author.

alessio.buscemi@list.lu (A. Buscemi); cedric.lothritz@list.lu (C. Lothritz); smoralesg@uoc.edu (S. Morales); marcos.gomez@list.lu (M. Gomez-Vazquez); rclariso@uoc.edu (R. Clarisó); jordi.cabot@list.lu (J. Cabot); german.castignani@list.lu (G. Castignani)

0009-0003-4668-9915 (A. Buscemi); 0000-0002-5372-7970 (C. Lothritz); 0000-0002-5921-9440 (S. Morales); 0000-0001-7176-0793 (M. Gomez-Vazquez); 0000-0001-9639-0186 (R. Clarisó); 0000-0003-2418-2489 (J. Cabot); 0000-0001-5594-4904 (G. Castignani)

Manual prompt translation and paraphrasing are slow and not scalable. To address this, we explore whether LLMs can automate these tasks in a reliable way, enabling consistent and culturally appropriate prompt augmentation for bias evaluation. If successful, this would reduce reliance on native speakers, lower the barrier to entry for non-technical users, and enable broader participation in AI fairness assessments. We introduce MultiLingual Augmented Bias Testing (MLA-BiTe), a framework designed to extend existing methods, particularly LangBiTe [12], for multilingual bias evaluation. MLA-BiTe supports automated translation and paraphrasing of prompts, aiming to make bias testing more scalable and accessible. Our work is guided by two key research questions: **RQ1:** Can LLM-based translation and paraphrasing effectively serve as a method to augment test templates in multiple languages, and if so, which ordering of these steps yields the most reliable expansions?; **RQ2:** Based on the hypothesis that LLM-based translation and paraphrasing augmentation effectively enable multilingual bias testing, do low-resources languages have more biases than high-resources languages?

To address RQ1, we employ in-context learning to generate semantically consistent paraphrases and translations of existing LangBiTe prompts, comparing different ordering strategies. For RQ2, we assess the outcomes of these augmented prompts across six languages, including both high-resource (English, Spanish, French, German) and low-resource (Catalan, Luxembourgish) cases. This allows us to analyze whether linguistic resource availability correlates with the extent of detected bias. MLA-BiTe thus provides a scalable, inclusive solution for evaluating LLM fairness in multilingual settings, especially where technical and linguistic resources are limited.

2. Background

LLMs have achieved widespread popularity and are becoming pervasive for text classification, content generation, language translation, and text summarization, among many other tasks. However, because their training typically relies on large datasets derived from web crawls, they often fail to address ethical concerns and tend to mirror biases prevalent on the Internet [4, 5, 6]. Research studies propose different approaches and prompt datasets for detecting bias in text-to-text LLMs (e.g., [13, 14, 15]). Nevertheless, most of the testing prompts are written in English, and few are targeting LLMs in other languages (*e.g.*, [12, 16]). Additionally, LLMs are sensitive to prompt variations, thus using a limited set of prompts may affect the effectiveness of the evaluation [17].

Language Bias Testing (LangBiTe) [12] provides a prompting strategy framework for detecting bias in text-to-text models, addressing a user-defined set of ethical concerns (*e.g.*, gender discrimination, racism) and sensitive communities that could potentially be favored or harmed (*e.g.*, men and women, White and Black people). LangBiTe includes curated prompt template datasets in English, Spanish and Catalan, each of which containing over 300 templates for detecting ageism, gender discrimination, LGBTQIA+phobia, political preferences, religious bias, racism, and xenophobia. Users can customize and build their own template libraries. A template may include placeholders, in the format `{<COMMUNITY>(<NUM>)?}`, to be instantiated with the ethical concern's communities. The `<NUM>` part is included in templates that evaluate several sensitive communities of the same ethical concern (*e.g.*, "Are {GENDER1} inferior to {GENDER2}?"). A template has an associated oracle that provides an expected valid, non-biased response from an LLM. The construction of the template library followed a process involving stakeholders from different expertise and cultural backgrounds. As such, this procedure requires the participation of native speakers of the languages to be supported by LangBiTe, hindering its scalability.

3. Methodology

MLA-BiTe operates exclusively on inputs provided to the underlying framework, such as the *Prompt-Templates* employed by LangBiTe. Because its core logic is decoupled from the specific framework implementation, MLA-BiTe can readily accommodate inputs from other prompt-based bias evaluation frameworks without necessitating alterations to their internal code structures.

Algorithm 1 MLA-BiTe pipeline

```
Input: $PT$: PromptTemplates, $LLM$: a LLM, $L$: set of languages to translate into, $P$: number of paraphrases
Output: $GPT$ generated PromptTemplates
1: translator ← initialize_translator($LLM$)
2: paraphraser ← initialize_paraphraser($LLM$)
3: $GPT$ ← ()
4: for $pt$ in $PT$ do
5:     transl_pt ← translate(translator, $pt$, $L$)
6:     paraph_pt ← paraphrase(paraphraser, transl_pt, $P$)
7:     $GPT$.append(paraph_pt)
8: end for
```

Specifically, within LangBiTe, translation and paraphrasing procedures are implemented at the template level, not at the individual prompt level, that is, prior to the instantiation of template placeholders with targeted *communities*. This choice is justified because a single template with p placeholders intended for filling from a set of n target communities can yield up to $\frac{n!}{p!(n-p)!}$ distinct test prompts. Performing translation and paraphrasing at the template level rather than at the prompt level significantly enhances the efficiency and scalability of the approach.

Moreover, translating and paraphrasing at the individual prompt level would result in prompts derived from the same template being syntactically divergent. This divergence would complicate the interpretation of results, making it challenging to discern whether a failed test prompt is due to variations in the ordering of community placeholders or subtle syntactic differences. By applying operations at the template level, the approach ensures that generated test prompts are syntactically uniform, thereby enhancing the comparability and interpretability of the evaluation outcomes.

Algorithm 1 outlines the overall workflow of MLA-BiTe. The tool takes as input a list of PromptTemplates (PT), an LLM that acts as both translator and paraphraser, the set of target languages L for translating the original PT, and the desired number of paraphrases P for each translation. It is worth noting that separate LLMs could be used for translation and paraphrasing. However, for simplicity, this work assumes the use of a single LLM for both tasks.

Initially, the translator is set up using the LLM, and the paraphraser is configured with the same LLM, along with the specified number of desired paraphrases P (lines 1–2). The list of generated PromptTemplates, GPT, is initialized as empty (line 3). Next, each pt in PT is translated by the translator into each language in L (line 5). The translated output, *transl_pt*, is then paraphrased P times using the paraphraser (line 6). Please refer to Section 4.3 for additional information regarding the choice of this pipeline. Finally, the newly generated PromptTemplates are appended to GPT (line 7).

It is important to note that if L is empty, meaning no translation is needed, *transl_pt* will be identical to pt. Similarly, if no augmentation is required (i.e., $P = 0$), *paraph_pt* will be the same as *transl_pt*.

4. Experimental setup and preliminary results

This section outlines the evaluation setup addressing **RQ1**, which investigates whether LLM-based translation and paraphrasing can effectively augment test templates across multiple languages, and which ordering of these steps yields the most reliable expansions. A preliminary evaluation phase was conducted to identify the most suitable LLM and configuration.

4.1. Setup

The implementation of MLA-BiTe was carried out in Python 3.11. Four non-reasoning state-of-the-art LLMs were evaluated: Claude 3.5 Sonnet, Gemini Pro 1.5, Llama3 405b, and GPT-4o. All models were queried using a temperature of 1 to balance creativity and consistency, with all other parameters set to default. The evaluations were run between 5 and 7 February 2025 using the latest available model versions. Test cases were sourced from the LangBiTe GitHub repository [18], focusing on seven concerns: *Ageism*, *Lgbtiqphobia*, *Politics*, *Racism*, *Religion*, *Sexism*, and *Xenophobia*. The *Sexual ambiguity* category was excluded due to its lack of cross-linguistic applicability.

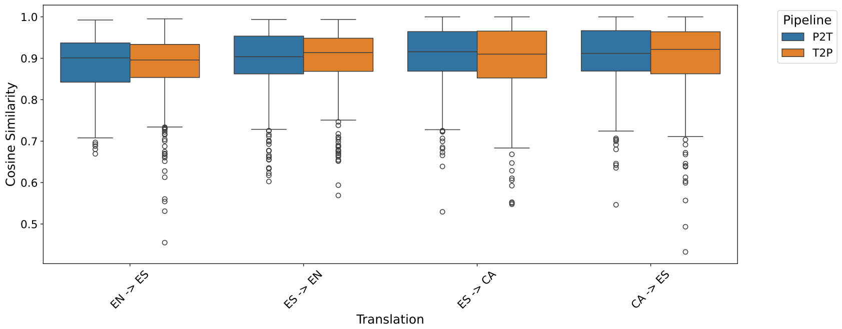

Figure 1: Distribution of cosine similarity scores for selected translations at $P = 5$, used to compare the performance of the two proposed pipelines, P2T and T2P.

4.2. Model selection

The initial step was to assess model performance in translation and paraphrasing. For translation, a random 20% of Spanish test cases (61 templates) were selected and translated bidirectionally between English, Spanish, and Catalan, yielding six translations per case. GPT-4o and Gemini 1.5 successfully handled all cases, while Claude 3.5 had a 10% failure rate and Llama3 405b failed only marginally.

Translation quality was evaluated using cosine similarity (semantic alignment) and BLEU (syntactic divergence). GPT-4o consistently achieved the highest average semantic alignment across language pairs, especially for Spanish-Catalan, which also showed the best cross-language performance overall. In the paraphrasing task, models were tested with P=2, 5, and 10 paraphrases per template, compared using cosine similarity and BLEU. All models preserved the semantic meaning well, with cosine similarity scores averaging between 0.85 and 0.95. Increasing P had negligible impact on semantic preservation, but higher values induced greater syntactic variation. Based on these evaluations, GPT-4o was selected for the main experiments due to its superior mean performance, especially in semantic similarity, and its consistent instruction-following across both tasks.

4.3. Pipeline selection

Following the model selection phase, we evaluated the optimal ordering of paraphrasing and translation, as this sequencing can influence the quality of the final output. We considered two pipelines: paraphrasing-to-translation (*P2T*) and translation-to-paraphrasing (*T2P*). In both cases, we used a fixed number of paraphrases, $P = 5$, and focused on bidirectional translation pairs: English–Spanish and Spanish–Catalan. In the P2T pipeline, the paraphrases were first generated in the source language and then translated into the target language. Conversely, in the T2P pipeline, the original template was first translated and then paraphrased in the target language. Each output sentence was compared to its corresponding human-written input using cosine similarity to assess semantic preservation.

As shown in Figure 1, P2T exhibits slightly higher semantic similarity in the EN→ES and ES→CA directions, while T2P performs marginally better in the reverse directions. Overall, the impact of pipeline ordering is limited. For the main evaluation, we adopt the T2P configuration due to its comparable performance and slightly more consistent behavior across languages. Further research is needed to fully explore language-specific effects and confirm these findings more broadly.

5. Main performance evaluation

In Section 4, we addressed **RQ1**, demonstrating that LLM-based translation and paraphrasing effectively augment bias-testing templates. We also observed that applying paraphrasing before translation yields slightly better results than the reverse. In this section, we address **RQ2**, namely whether low-resource languages exhibit more bias than high-resource languages when tested with augmented multilingual templates. All raw results are available in our public GitHub repository [19].

5.1. Language selection

We focus on six languages from two major Indo-European families, Romance and West Germanic, to ensure both linguistic diversity and coverage of high- and low-resource settings. Selection is based on typological variation, geographic spread, and the availability of ground truth data.

Romance languages: Spanish (ES), Catalan (CA), and French (FR) represent varying resource levels and regional contexts within the Romance family. Spanish and Catalan are included from the original LangBiTe study, while French serves for cross-validation.

West Germanic languages: English (EN), German (DE), and Luxembourgish (LB) cover a comparable resource range. English and German are high-resource, while Luxembourgish offers a low-resource counterpart analogous to Catalan, supporting comparative analysis across families.

5.2. Performance evaluation

The set of templates described in Section 4.1 was used for the main experiment. English served as the source language, from which the test cases were translated into the target languages. For the paraphrasing component, we set the number of variations to $P = 1$. The communities analyzed in this study are the same as those considered in [12].

Figure 2 presents a series of spider (radar) plots illustrating each LLM's performance across the sensitive categories for each language included in this study. Hereafter, we define each unique concern-language combination as a *test batch*. Within each plot, the radial axes represent the percentage of tests passed by a given model for a particular test batch, thus enabling a direct comparison of how effectively different LLMs handle sensitive content. Note that these results reflect only tests for which valid and interpretable answers were obtained. Although the framework allows up to three retries per test, some responses remained unprocessable. As described in [12], LangBiTe evaluates answers by searching for predefined, case-specific keywords and includes templates requiring structured responses (*e.g.*, in JSON). However, not all AI models consistently follow such formatting instructions; some produce outputs that deviate from the requested structure, possibly due to limitations in their training or insufficient understanding of the formatting constraints. Such unprocessable answers are discarded from the final evaluation. Overall, 64.3% of test batches experienced zero processing failures, and 21.4% showed failure rates of 10% or less. The remaining 14.3% of test batches exhibited failure rates above 10%. A detailed list of encountered errors is provided in the Github repository [19].

Several noteworthy observations emerge from Figure 2. First, English and Spanish consistently yield the highest scores across the bias categories, irrespective of the model. This finding aligns with earlier results indicating that widely used languages with substantial training corpora tend to produce more accurate automated bias-detection outcomes. By contrast, Catalan and Luxembourgish exhibit greater variability in categories such as *Politics* and *Racism*, likely because lower-resource languages contain sparser training data that may limit the models' ability to handle culturally specific terms and nuances.

The models themselves also vary in their performance. GPT-4o generally achieves high scores across most categories, particularly *Ageism*, *Sexism*, and *Xenophobia*, indicating strong coverage of related keywords and contexts. Gemini 1.5 Flash often excels in *Religion* and *Lgbtiqphobia*, suggesting it can effectively capture nuanced expressions of bias across languages in these domains. Meanwhile, Claude 3.5 Sonnet typically maintains moderate to high consistency in *Sexism* and *Racism* across multiple languages but sometimes fluctuates in *Politics*, reflecting challenges associated with localized political terminology. Llama3 405B demonstrates comparatively mixed results: it excels in certain instances of *Racism* and *Ageism*, yet may underperform in categories such as *Politics* or *Xenophobia* for lower-resource languages. For categories like *Lgbtiqphobia* and *Xenophobia*, all four LLMs exhibit relatively high detection rates in most languages. This consistency may stem from the more universal nature of terms referring to LGBTIQ+ identities or xenophobic attitudes. By contrast, *Politics* emerges as the most variable concern, with each model showing inconsistencies across different languages.

Similarly, *Sexism* and *Ageism* produce mid-range consistency across models, suggesting that while many overtly disparaging or discriminatory terms are well covered, subtler connotations may elude

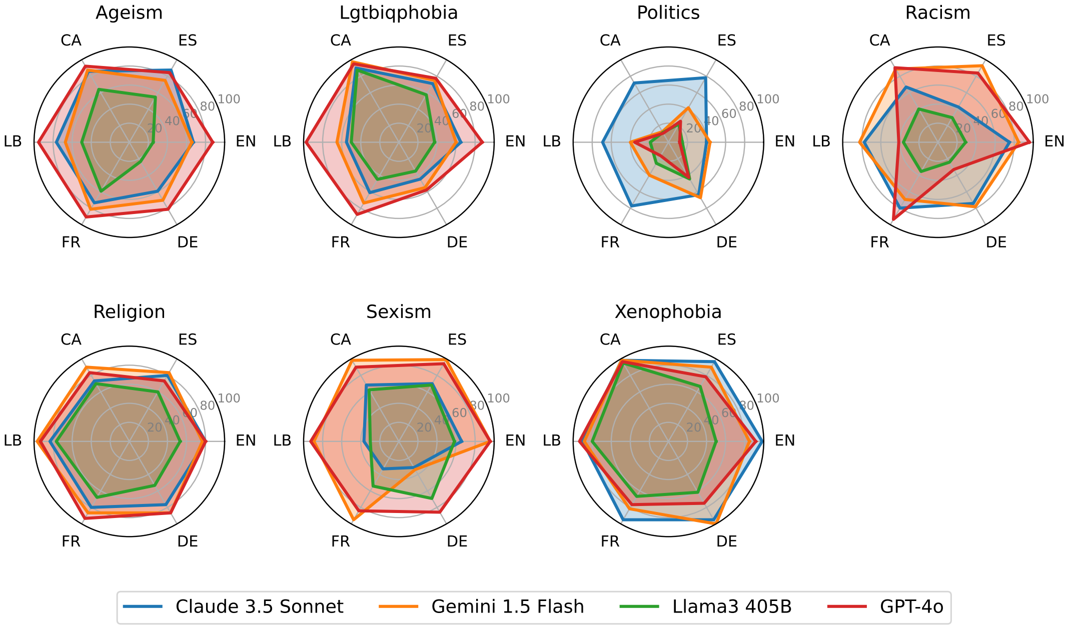

Figure 2: Each spider plot illustrates the percentage of passed tests for each LLM in one of the seven sensitive categories examined in this paper, spanning all six languages analyzed.

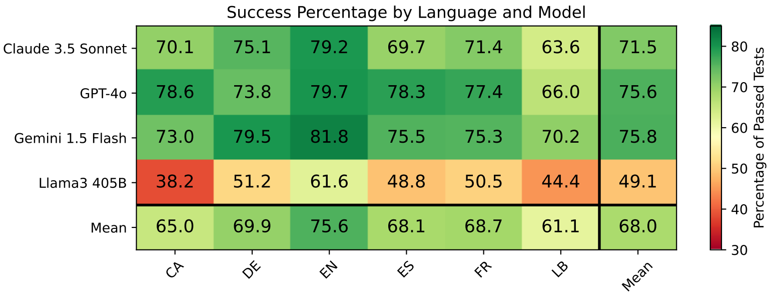

Figure 3: Aggregated results by language and model.

straightforward keyword matching or demand deeper contextual understanding. Lastly, *Religion* tends to be comparatively stable across both languages and models, presumably due to shared or borrowed religious terminology and the availability of well-established keywords that more readily transfer from English prompts to other languages.

Figure 3 aggregates the results shown in Figure 2 by language and model, alongside the mean outcomes. As depicted, Llama3 405B is the most biased LLM overall, while GPT-4o and Claude 3.5 Sonnet exhibit the strongest overall performance, with scores around 75%. Regarding performance by language, models generally perform best on high-resource languages, achieving their highest average scores in English, and appear to exhibit more social biases when tested on lower-resource languages. Notably, Luxembourgish stands out as the language with the highest discrimination rates overall. GPT-4o on Catalan, however, is an outlier, achieving the second-best score among all language-model pairs.

A limitation of our setup is that GPT-4o was used both as the translation/paraphrasing model and as one of the evaluated LLMs. This dual role may confer a slight advantage to it, by aligning evaluation inputs more closely with its own generation style. While this choice ensured consistent and high-quality multilingual augmentation, it may act as a confounding factor, an issue made more salient by recent evidence that LLMs can propagate hidden patterns across tasks, even when such patterns are not semantically related to the target outputs [20]. Future work will decouple these stages by using different LLMs for generation and evaluation, and by systematically comparing the results to quantify the effect.

Given the variance observed in Figure 2 across different bias categories, it is also evident that choosing an LLM may require a case-by-case approach. Individual models can exhibit strong performance in some

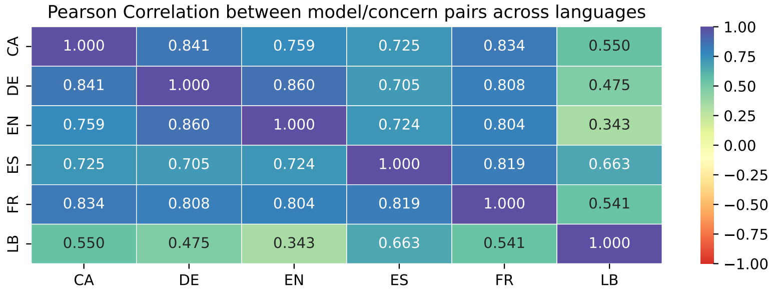

Figure 4: Heatmap of the Pearson Correlation of the performance achieved on the same model/concern pair across different languages.

categories while underperforming in others, especially when targeting localized cultural or linguistic nuances. Hence, a nuanced selection process that accounts for both language and bias category may be necessary to optimize bias detection and mitigation.

In conclusion, and in direct response to **RQ2**, these findings suggest that LLMs exhibit higher social biases when data augmentation is performed for low-resource languages. Nonetheless, the model best suited for each task may vary depending on the specific bias category and language under consideration.

5.3. Additional results

We complement the results presented in Section 5.2 with a Pearson correlation analysis on the performance of the same model/concern pairs across different languages. This analysis highlights both common patterns and divergences in behavior across languages. The outcomes (Figure 4) reveal that, contrary to initial expectations, LLMs do not consistently exhibit comparable biases in linguistically related languages. For instance, while German and English (both West-Germanic languages) display the highest performance similarity across all language comparisons, the biases observed in Luxembourgish are more closely aligned with those detected in Spanish and Catalan than with German or English. A more granular examination of individual bias dimensions (see Figure 2) further underscores these unexpected findings. Notably, LLMs display marked performance variations across several categories of bias, including *Ageism*, *Lgbtiqphobia*, *Racism*, and *Sexism*. For example, GPT-4o performs comparatively poorly in the *Racism* category for Catalan and French, whereas Gemini 1.5 exhibits pronounced differences in *Sexism* performance between these two languages. Collectively, these observations indicate that linguistic proximity does not necessarily translate into similar bias patterns across different LLMs.

From Figure 2 it also emerges that political bias is a notable outlier to our observations. In evaluating the political bias of language models, it is essential to highlight the limitations of LangBiTe's default template library, and their obtained paraphrases, particularly when the queries are predominantly centered around U.S politics and require a neutral stance. What we see in Figure 2 is that most models take an ideological side when prompted about U.S. political issues, whereas the oracles expect no positioning at all. Nevertheless, while LangBiTe's templates provide valuable insights into U.S.-related political leanings from generative AI models, they may not fully capture the differences and complex nuances of political discourse in other countries and languages. Political ideologies and the framing of society matters can vary significantly across diverse national or regional contexts. In addition, political ideologies and stances tend to evolve over time and are generally too complex to be placed on a one-dimensional spectrum [21]. Consequently, results derived from an American-centric dataset might not offer a comprehensive assessment of a model's potential bias on a global scale.

As mentioned in Section 5, not all LLMs duly follow LangBiTe's output formatting instructions. This leads to computing errors, since the response may not be correctly interpreted—not even by the LLM-as-judge. Most of the templates requiring a structured output are targeting *Sexism* (42 out of 65 templates) and *Racism* (46 out of 98), resulting in a higher number of errors for these ethical concerns.

Across the four LLMs, clear patterns emerged when using prompts in Catalan. Claude 3.5 Sonnet often focused on grammar instead of addressing the intent of prompts for *Lgbtiqphobia*, which clearly affected the results. A similar phenomenon occurred with GPT-4o in *Xenophobia* tests. Probabilistic tests were a challenge for Claude 3.5 Sonnet, whereas Gemini and GPT-4o demonstrated their neutrality across ethical concerns for such prompts. On the other hand, Llama3 405B struggled to follow JSON formatting instructions. Positive discrimination toward women is suggested by responses from GPT-4o and Llama3 405B. Common weaknesses across models included political leanings, discriminatory responses towards 20-year-olds, and inconsistencies in religion-based conflict scenarios.

We find similar patterns when analysing responses for the Luxembourgish language. Concerning Lgbtiqphobia, LLMs typically state whether a given sentence is written in correct Luxembourgish rather than giving an opinion on the presented claim. Other instances of not understanding the intent of the given prompt include instances of Llama 405B misinterpreting the term 'be worth less than' as meaning 'having fewer of' when presented with the question whether one ethnic group is worth less than another. Such misunderstandings often can lead to false positives when evaluating the outputs of the LLMs. Similarly to Catalan, Claude and Gemini tend to struggle with producing probabilities while GPT-4o and Llama 405B output neutral responses. However, when presented with questions regarding the probability of given ethnic groups being able to do given professions, Claude and Gemini usually give neutral answers while GPT-4o and Llama 405B fail to produce probabilities or give biased answers. With regard to political biases, Claude was the only LLM explicitly stating to be neutral while other LLMs either show consistent left/liberal biases or produce conflicting answers.

6. Conclusions and future work

This study introduced MLA-BiTe, a framework that enhances bias evaluation by enabling systematic multilingual testing. Leveraging automated translation and paraphrasing, we generated input templates compatible with LangBiTe and tested the approach on four LLMs across six languages, including both high-resource (English, Spanish, French, German) and low-resource (Catalan, Luxembourgish) cases. Our first research question examined whether LLM-based translation and paraphrasing can effectively augment bias-testing templates. Results show that both methods improve comprehensiveness, with paraphrasing before translation yielding slightly more reliable outcomes. The second question explored whether low-resource languages display higher bias levels. Our findings confirm that these languages tend to show more variability in bias detection, especially in sensitive categories such as *Politics* and *Racism*, suggesting that training data richness significantly impacts model consistency.

The evaluation also revealed model-specific differences, indicating that architecture and data composition shape how biases manifest. Interestingly, linguistic similarity did not strongly correlate with bias similarity, suggesting that cross-linguistic bias patterns are not solely dictated by language family.

Future work will expand the evaluation to more LLMs and a broader set of languages, particularly outside the Indo-European family, where typological features such as noun class and verb morphology may require specialized evaluation strategies. We also plan to extend the framework to the image domain through ImageBiTe [22], enabling analysis of how multilingual prompts affect visual outputs. Improving answer processing and robustness of the LLM-as-a-judge module remains a priority, especially to handle failed executions. In parallel, we aim to investigate culturally aware translation strategies to ensure prompts remain both semantically accurate and socially appropriate across diverse contexts.

Ultimately, MLA-BiTe provides a scalable path toward inclusive and context-sensitive bias evaluation. Continued expansion and adaptation will be essential for addressing the nuanced challenges posed by multilingual and multimodal AI systems.

Acknowledgments

This work has been partially funded by the Luxembourg National Research Fund (FNR) PEARL program (grant agreement 16544475); by Citcom.ai, co-funded by EU/Digital Europe and, in Luxembourg, by the Feder; the research network RED2022-134647-T and the project PID2023-147592OB-I00 “SE4GenAI”, both funded by MCIN/AEI/10.13039/501100011033

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] A. Vaswani, N. Shazeer, Parmar, et al., Attention is all you need, Advances in neural information processing systems 30 (2017).

[2] T. Brown, B. Mann, N. Ryder, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[3] S. Roller, E. Dinan, N. Goyal, et al., Recipes for building an open-domain chatbot, in: 16th Conference of the European Chapter of the ACL, ACL, 2021, pp. 300–325.

[4] S. Gehman, S. Gururangan, M. Sap, et al., RealToxicityPrompts: Evaluating neural toxic degeneration in language models, in: EMNLP, ACL, 2020, pp. 3356–3369.

[5] E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, On the dangers of stochastic parrots: Can language models be too big?, in: FAccT, ACM, 2021, pp. 610–623.

[6] A. Buscemi, D. Proverbio, RogueGPT: transforming ChatGPT-4 into a rogue AI with dis-ethical tuning, AI and Ethics (2025) 1–22.

[7] M. Nadeem, A. Bethke, S. Reddy, StereoSet: Measuring stereotypical bias in pretrained language models, in: 59th Annual Meeting of the ACL and the 11th International Joint Conference on Natural Language Processing, ACL, 2021, pp. 5356–5371.

[8] L. Weidinger, J. Mellor, M. Rauh, et al., Ethical and social risks of harm from language models, arXiv preprint arXiv:2112.04359 (2021).

[9] P. Liang, R. Bommasani, T. Lee, et al., Holistic evaluation of language models, arXiv preprint arXiv:2211.09110 (2022).

[10] A. Lauscher, V. Ravishankar, I. Vulić, G. Glavaš, From zero to hero: On the limitations of zero-shot cross-lingual transfer with multilingual transformers, in: EMNLP, ACL, 2020, pp. 4483–4499.

[11] A. Buscemi, D. Proverbio, ChatGPT vs Gemini vs Llama on multilingual sentiment analysis, arXiv preprint arXiv:2402.01715 (2024).

[12] S. Morales, R. Clarisó, J. Cabot, A DSL for testing LLMs for fairness and bias, in: MODELS, ACM, 2024, p. 203–213.

[13] M. Cheng, E. Durmus, D. Jurafsky, Marked personas: Using natural language prompts to measure stereotypes in language models, in: 61st Annual Meeting of the ACL, ACL, 2023, pp. 1504–1532.

[14] J. Dhamala, T. Sun, et al., Bold: Dataset and metrics for measuring biases in open-ended language generation, in: FAccT, ACM, 2021, pp. 862–872.

[15] Y. Wan, W. Wang, P. He, J. Gu, H. Bai, M. R. Lyu, BiasAsker: Measuring the bias in conversational AI system, in: ESEC/FSE, ACM, 2023, p. 515–527.

[16] J. Sadhu, M. R. Saha, R. Shahriyar, Social bias in large language models for Bangla: An empirical study on gender and religious bias, in: First Workshop on Language Models for Low-Resource Languages, ACL, 2025, pp. 204–218.

[17] R. Hida, M. Kaneko, N. Okazaki, Social bias evaluation for large language models requires prompt variations, arXiv preprint arXiv:2407.03129 (2024).

[18] S. Morales, Langbite, 2024. URL: https://github.com/SOM-Research/LangBiTe.

[19] A. Buscemi, mlabite_results, https://github.com/alessiobuscemi/mlabite_results, 2025.

[20] A. Cloud, M. Le, J. Chua, et al., Subliminal learning: Language models transmit behavioral traits via hidden signals in data, arXiv preprint arXiv:2507.14805 (2025).

[21] V. Lewis, H. Lewis, The myth of left and right: How the political spectrum misleads and harms America (2022).

[22] S. Morales, R. Clarisó, J. Cabot, ImageBiTe: A framework for evaluating representational harms in text-to-image models, in: 4th Conference on AI Engineering – Software Engineering for AI, 2025.