
Cross-Lingual Bias Detection in Large Language Models through Mechanistic Judge Model Evaluation

Faizan Abbas Shu Fan Sun Wanjie Zhong
University of Aberdeen University of Aberdeen University of Aberdeen

With
Martian & Apart Research

Abstract

This paper provides a framework for detecting multilingual bias in LLMs through mechanistic interpretability of judge model behaviour. We developed an approach to evaluate language-dependent discrepancies in specialised models by leveraging semantically equivalent question-answering tasks across English, Chinese, Romanian and Vietnamese using the XQuAD dataset. Our framework utilises a judge-based evaluation system that assesses both correctness and reasoning quality on a 4-point scale, enabling the detection of biases that may not have been detected in traditional accuracy metrics alone. We demonstrate that our approach can detect these differences in model performance across languages without requiring ground truth labels, making it applicable to scenarios where traditional evaluation methods are insufficient. Our work advances Track 1 (Judge Model Development) by providing an interpretable bias detection mechanism that promotes fairness and reliability in multilingual AI systems. This framework is easily extensible to additional languages and can serve as a foundation for building fair expert orchestration systems.

1. Introduction

1.1 Research Question & Motivation

The wide adoption of large language models across various global applications has raised critical fairness and consistency in multilingual contexts. Numerous studies have shown over the years that pre-trained language models often exhibit hidden biases [1][2][3][4][5][6][7][8] that are reflected differently across languages and cultures. These biases not only affect the model’s performance but also manifest in the quality and depth of reasoning provided to users from different lingual backgrounds. Our research addresses a fundamental question: How can we detect and quantify cross-lingual bias in large language models through the mechanistic interpretability of judge model evaluations?

We hypothesise that, when processing semantically equivalent content in different languages, LLMs demonstrate measurable performance differences and these biases can be detected through structured judge model evaluation. We also propose that significant bias occurs not only in the basic language understanding but also in the contextual grounding of reasoning processes of the models, making traditional accuracy-based metrics insufficient for a comprehensive bias assessment.

This work directly contributes to the Expert Orchestration Architecture vision by developing transparent, interpretable evaluation methods that provide the foundation for future routing and model selection systems. In real-world applications, biased model performance across languages can lead to poor system performance and disparate outcomes for users, undermining trust in AI applications.

1.2 Challenge Track Focus

This project addresses Track1: Judge Model Development by creating a framework for evaluating LLM performance through the mechanistic interpretability of a judge model that can reliably detect multilingual biases in participant models.

The judging framework not only evaluates the answer correctness but also detailed assessments of the reasoning quality of the participant models. Additionally, it operates without requiring any ground truth labels, making the framework applicable to diverse domains. The framework provides a comprehensive comparison of participant model performance across linguistic conditions and offers feedback that can guide model improvement and selection.

1.3 Contribution to Expert Orchestration

Our work focuses on a critical limitation of current monolithic AI approaches, specifically, it addresses the lack of systematic mechanisms to detect and mitigate multilingual biases that can undermine the fairness and reliability of AI systems across diverse linguistic communities.

Traditional large-scale AI models often exhibit inconsistent performance across languages, with subtle biases favouring high-resource languages (typically English) over low-resource languages. These disparities are often neglected by traditional accuracy-based metrics, providing a false evaluation of model performance. Apart from that, monolithic models typically lack transparency in their decision-making process, making it difficult to understand why certain biases occur or how they might be addressed. This is particularly problematic in

multilingual cases where cultural or linguistic differences may influence model behaviour. Understanding the underlying mechanisms will allow the development of effective bias mitigation strategies. Our framework addresses these limitations by providing a methodology for comprehensively analysing multilingual biases in AI systems. Rather than relying solely on accuracy scores, our approach provides a detailed, granular analysis of model behaviour across different linguistic contexts, revealing bias patterns that would otherwise remain hidden.

Traditional model evaluation requires significant computational resources and technical expertise, creating barriers for smaller organisations and researchers. Our framework addresses these barriers by providing automated bias detection capabilities that operate without requiring ground truth annotations. The open-source approach further supports democratization by making our tools and methodologies freely available to the broader community.

The transparency feature of our system makes the evaluation process understandable and auditable. Rather than relying on black-box evaluation systems, our framework provides detailed explanations of its evaluations, enabling users to understand and trust the results.

Safety is a fundamental concern in AI development, and multilingual bias represents a significant safety risk that requires urgent attention from the broader AI safety community. Biased AI systems can cause real harm by providing unfair treatment to users based on their linguistic background, demonstrating stereotypes, and excluding communities from the benefits of AI technology.

Our framework contributes to the development of safe AI systems by providing mechanisms for detecting multilingual biases in deployed systems, supporting a preventive rather than reactive approach to AI safety. It also further contributes to safety by making it possible to understand the causes of biased behaviour. This understanding is essential for developing safe AI applications as the deployment of AI models in contexts where biased decisions can have severe consequences.

2. Methods

2.1 Technical Approach

Our methodology centres around a comparative evaluation framework that exposes multilingual biases through the analysis of model responses to semantically equivalent content across languages.

- **Models/datasets used:**

We utilised the XQuAD (Cross-lingual question-answering dataset) focusing on English, Chinese, Vietnamese and Romanise language pairs. XQuAD provides semantically equivalent question-answering tasks across languages, making it an ideal choice for bias detection. The dataset contains reading comprehension questions where each context-question pair has been translated to maintain semantic equivalence while preserving the linguistic nuances.

The models used are divided into two categories: participant models and the judge model. The participant models include claude-3-opus-latest,

DeepSeek-V3, Meta-Llama-3.1-405B-Instruct-Turbo, Gemini-1.5-flash-8b and gpt-4.1-nano, whereas our judge model uses gpt-4o.

- **Mechanistic interpretability techniques:** To evaluate the performance of the participant models, we pass each participant's answer along with the reasoning behind their selection to the judge model which uses a bottom up approach to evaluate these responses. The judge model checks to see if the participant model's reasoning and answer align with the context provided before making a decision.
- **Implementation details:** We used Python and Martian API

2.2 Judge Architecture

- **Judging System 1**

Judging System 1 assesses the accuracy of participant models' answers. As our judging system needs to operate without ground truth data as introduced above, we choose GPT-4o as the judge model to determine the correct answer.

The judge model utilises a bottom up logic to determine the correct answer, by evaluating the response provided by the participant models and checking if the response corresponds to any of the information in the original context. The rubric used for Judging System 1 is:

You are tasked with evaluating whether the answer of a model is correct to the question. You will be doing this by combining the answer with the question, and find if the same meaning exists in the original sentence, which means to do it differently from the bottom up.

The scoring is as follows:

- *0 for wrong answer and 1 for right answer.*
- *Show the logic chain you use that lead to your judge.*

- **Judging System 2**

Judging System 2 evaluates the correctness of the reasoning provided by the participant models, as well as the consistency between the reasoning and the answer. That is, whether the reasoning corresponds to the original text, and whether the reasoning is actually related to the expected answer or not.

The judge model used for this task is GPT-4o. The model will be given the prompt used for the participant models, containing the original text and the question. The model will then compare the reason in the response with the original context. The rubric used for Judging System 2 is:

You are tasked with evaluating whether the answer of a model is correct to the question and whether the reason makes sense. You will be doing this by reading the REASON part of the answer and evaluating it with the question to see if the reason makes sense and if it actually leads to the answer, and check if the ANSWER part corresponds to the reason.

The scoring is as follows:

- *0 when the reason does not make sense and the answer is wrong.*
- *1 when the reason is correct according to the context but has nothing to do with the question, and the answer is wrong.*
- *2 when the reason makes sense and leads to the expected answer, but the answer is*

wrong.

- 3 when the reason makes sense and leads to the expected answer, and the answer is correct and corresponds to the reason.
- Show the logic chain you use that lead to your judge.

3. Results

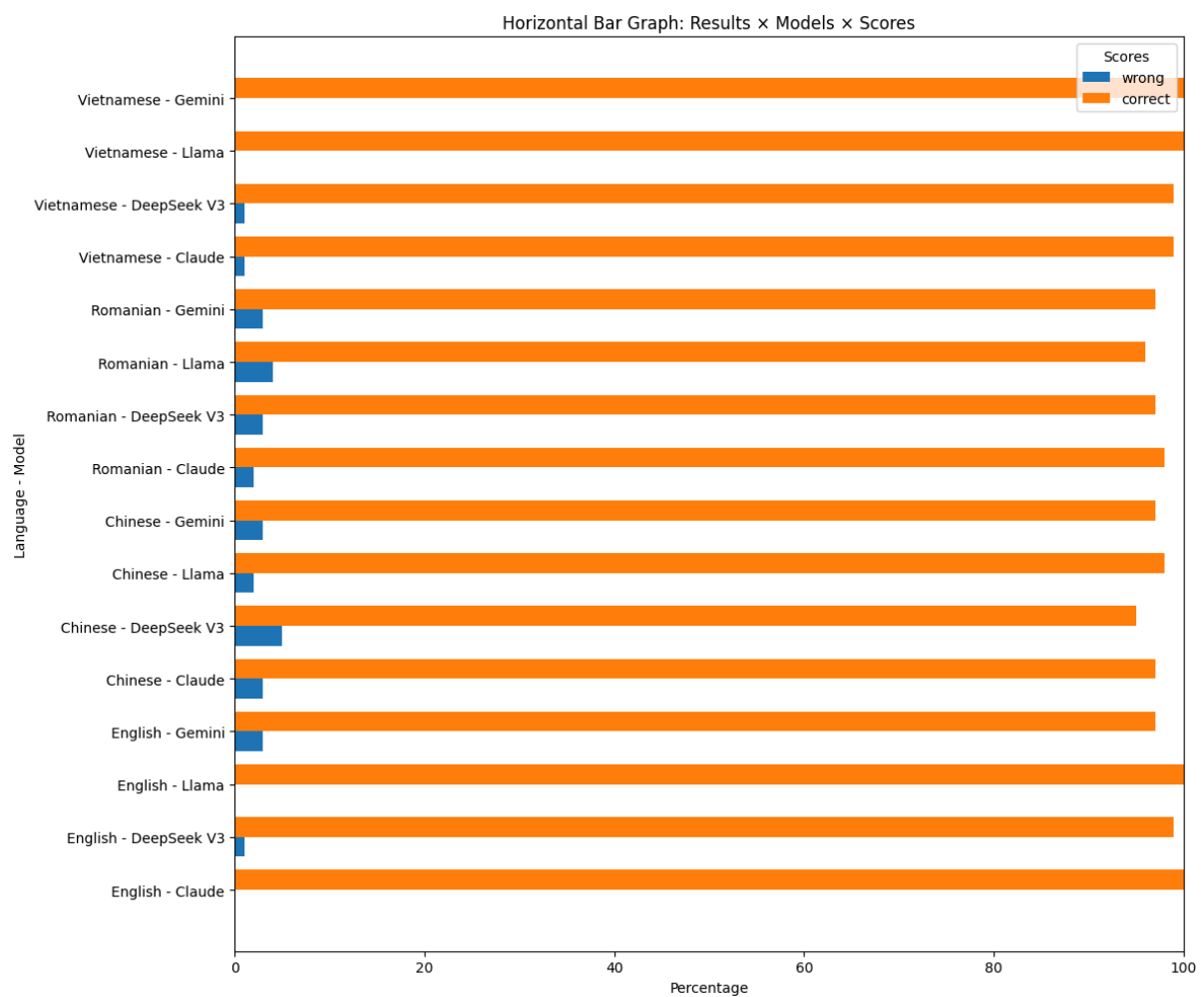


Figure 1 – Judge System 1 Performance in 4 languages Tested with 4 Models

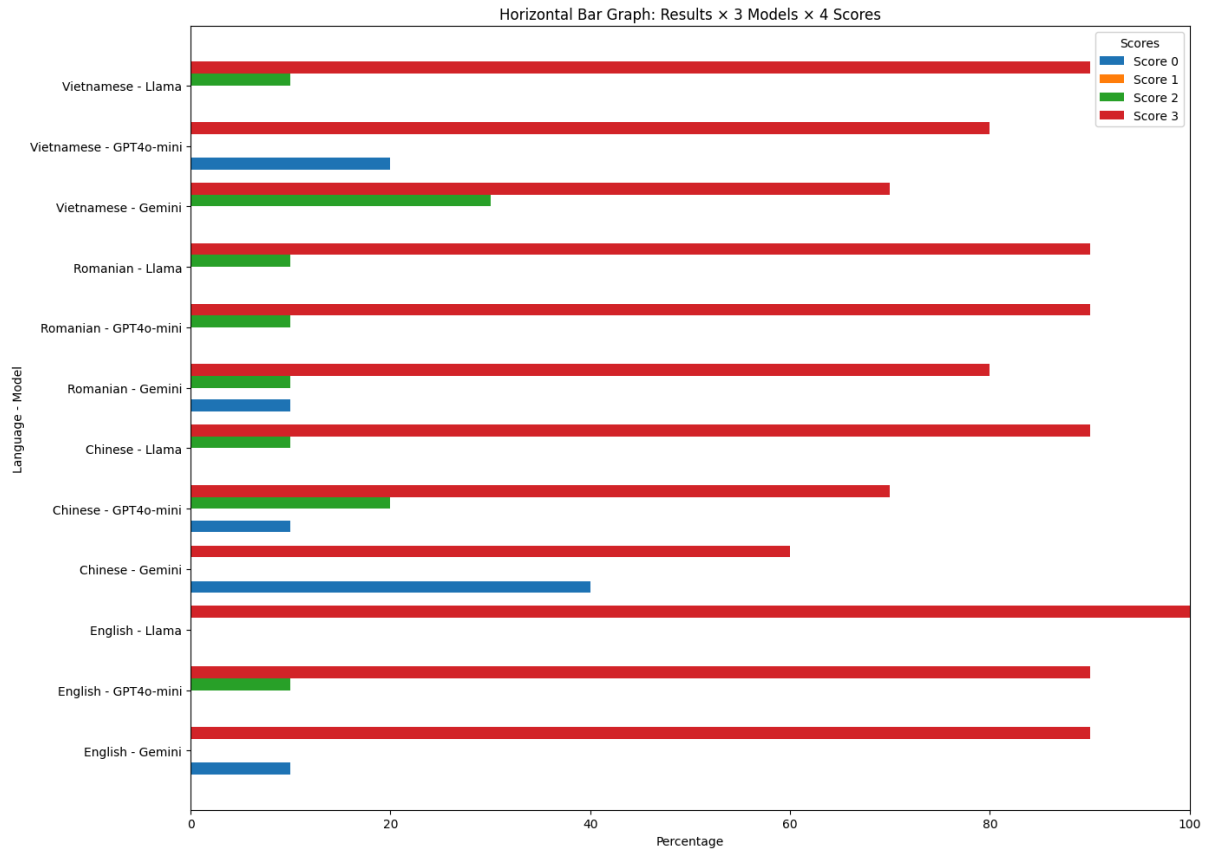


Figure 2 – Judge System 2 Performance in 4 Languages Tested with 3 Models

4. Discussion

4.1 Interpretation of Results

For Judging System 1, no significant difference in performance between English and other languages. We assume this is because the model we have access to is mostly large models which are well trained with abundant data. However, we can see in the result of Judging System 2, the performance of Romanian and Vietnamese are slightly lower than those of English and Chinese. Therefore, we believe that even though the models are capable of generating correct answers, when it comes to languages with lesser resources, they still show a certain degree of language bias when prompted to produce more detailed responses.

4.2 Limitations & Future Work

As pointed out in Session 4.1, the choice of testing models limits the possibility for us to test on smaller models, which are believed to be more prone to language bias. Therefore, to further enhance this system, more tests can be conducted on smaller models such as models dedicated to be used locally on personal devices.

4. References

- [1] Abid, A., Farooqi, M., & Zou, J. (2021, July). Persistent anti-muslim bias in large language models. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (pp. 298-306).
- [2] Ahn, J., & Oh, A. (2021). Mitigating language-dependent ethnic bias in BERT. arXiv preprint arXiv:2109.05704.
- [3] Bartl, M., Nissim, M., & Gatt, A. (2020). Unmasking contextual stereotypes: Measuring and mitigating BERT's gender bias. arXiv preprint arXiv:2010.14534.
- [4] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.
- [5] Huang, P. Sen et al.(2019).“Reducing sentiment bias in language models via counterfactual evaluation”.
- [6] Kurita, K., Vyas, N., Pareek, A., Black, A. W., & Tsvetkov, Y. (2019). Measuring bias in contextualized word representations. arXiv preprint arXiv:1906.07337.
- [7] Nadeem, M., Bethke, A., & Reddy, S. (2020). StereoSet: Measuring stereotypical bias in pretrained language models. arXiv preprint arXiv:2004.09456.
- [8] Weiss, G., Goldberg, Y., & Yahav, E. (2021). *Thinking Like Transformers* (arXiv:2106.06981). arXiv. <http://arxiv.org/abs/2106.06981>