Cross-Linguistic Political Bias Analysis in Large Language Models: A Comparative Study Across 5 European Languages

Alberto Gabriele Scuderi (Mat. 248257)

University of Trento - Data Science 2024-2025

September 1, 2025

Abstract

Recent research has revealed significant political biases in Large Language Models (LLMs), with studies demonstrating consistent left-leaning tendencies in English-language outputs. Drawing inspiration from Tanise Ceron's work on political inclinations in LLMs, this study proposes a field-specific investigation examining how language choice affects bias expression in controversial political topics. We present a methodological approach for examining bias manifestation across different languages, translation tasks, and response formats. This research aims to understand whether political biases remain consistent across linguistic boundaries or reflect language-specific cultural and political contexts embedded in training data. We propose a systematic investigation of how language choice affects bias expression in controversial political topics across three major LLM systems: ChatGPT (GPT-5), Claude (Sonnet), and Grok. We examine bias manifestation in French, Italian, Spanish, and Dutch—languages representing diverse European political and cultural contexts—to identify relevant cultural and semantic distinctions in model outputs.

1 Introduction

Large Language Models (LLMs) have become increasingly influential in shaping public discourse, making the identification and understanding of their inherent biases a critical research priority. Recent studies have demonstrated that LLM systems tend to display left-leaning political biases, raising concerns about their potential impact on democratic discourse and information dissemination. The linguistic dimension of bias in LLMs presents a particularly complex challenge. While most research has focused on English-language outputs, the multilingual capabilities of modern LLMs introduce questions about whether biases remain consistent across languages or reflect the cultural and political contexts of different linguistic communities. This study investigates these cross-linguistic bias patterns through experimentation with contentious political topic prompts. As LLMs become increasingly integrated into multilingual information ecosystems as sources of information and decision-support tools, understanding how language choice affects their outputs becomes crucial for ensuring fair and balanced AI-mediated discourse across global communities. This research must also consider the impact of Retrieval-Augmented Generation (RAG), a paradigm developed to reduce hallucinations by confronting training-based information with external sources from databases. For this study, we focus on topics relevant to the European context to examine whether models share the same information sources or whether LLMs select from different sources, which would consequently generate different responses to prompts.

2 Literature Review

2.1 Political Bias in Large Language Models

Recent studies demonstrate that when prompted with political questions, LLMs exhibit left-liberal leanings, establishing a baseline understanding of political bias in these systems. Research has shown that larger models, such as Llama3-70B, tend to align more closely with left-leaning political parties, while smaller models often remain neutral, particularly when prompted in English (2). The measurement of political bias has evolved

beyond simple left-right classifications. Modern approaches recognize the complexity of bias, where an LLM might exhibit liberal stances on certain topics (e.g., reproductive rights) while leaning conservative on others (e.g., the death penalty), reflecting a sophisticated and variable political landscape. This nuanced understanding necessitates topic-specific analysis rather than broad political categorization.

2.2 Ceron's Contributions to Political Bias Research

Tanise Ceron's research has been instrumental in advancing our understanding of political bias in LLMs, particularly in European contexts. Her studies evaluate bias patterns across various topics and LLMs within European societies, identifying left-wing, right-wing, and inconsistent opinions toward different issues. This work highlights the importance of contextual and cultural factors in bias manifestation. Ceron's methodological contributions include the development of robust evaluation frameworks that account for the complexity of political discourse. Her work emphasizes the need for comprehensive datasets that capture the multidimensional nature of political opinion, moving beyond binary classification systems.

2.3 Cross-Linguistic Bias Research

While most research has focused on monolingual outputs, emerging studies have begun to explore cross-linguistic patterns. Recent comparative analyses of political bias in LLM-based chatbot outputs have examined cross-lingual aspects, though comprehensive systematic studies remain limited (1). Culture fundamentally shapes human reasoning, behaviour, and communication, suggesting that language-specific biases may reflect deeper cultural patterns embedded in training data. This cultural dimension adds complexity to cross-linguistic bias analysis, as observed differences may reflect legitimate cultural variations rather than systematic model bias. It is reasonable to assume that training LLMs on datasets from certain languages may influence their responses in those selected languages. Through this focused study, we attempt to verify this hypothesis.

3 LLM Models Selected for This Study

For this investigation, we selected the following state-of-the-art models:

1. **ChatGPT (GPT-5):** Serves as our baseline model due to its widespread adoption and extensive documentation in bias research literature. As the most commercially successful and publicly accessible LLM, ChatGPT represents the current standard for conversational AI systems.

2. **Claude (Sonnet):** Offers a complementary perspective through Anthropic's Constitutional AI approach, which explicitly aims to reduce harmful outputs and increase helpfulness. This training methodology, focusing on AI safety and alignment principles, provides an important contrast to ChatGPT's approach. Claude's emphasis on harmlessness and honesty may manifest differently across cultural contexts, making it particularly relevant for cross-linguistic bias analysis.

3. **Grok:** Popular on X and developed by Elon Musk's xAI as an explicitly "anti-woke" alternative to existing chatbots. This serves as a potential alternative to more mainstream models that demonstrate left-leaning tendencies in their responses.

The three selected models will be evaluated using carefully chosen prompts to analyze their response patterns. This evaluation aims to identify differences and patterns in their response styles. While other models, such as Gemini and DeepSeek, were considered for inclusion, this study focuses on the three chosen models to maintain clarity and scope. Future research may extend this analysis to additional models to further investigate potential biases.

4 Research Design

Despite inherent limitations, this study addresses critical gaps in cross-linguistic bias research through a methodologically robust experimental design that attempts to distinguish cultural variations from methodological artefacts using analytical instruments

such as the TextBlob library in Python and transformer-based algorithms for NLP such as BERT. Our direct cross-linguistic prompting approach presents identical politically contentious questions across French, Italian, Spanish, and Dutch, capturing authentic cultural responses by allowing models to engage with prompts in their linguistically appropriate forms. The English-prompts-with-multilingual-responses approach addresses whether response language itself influences bias expression. By maintaining consistent English prompts while varying response languages, we isolate the effects of linguistic expression from cultural interpretation, examining whether responding in different European languages shapes political positioning independently of cultural content. Our translation-mediated analysis mirrors real-world multilingual AI deployment by translating English prompts into target languages, collecting native-language responses, and back-translating to compare answers and identify nuances. This approach provides insights into how cultural nuances survive translation processes and whether observed patterns reflect genuine cultural differences. This methodology allows us to identify findings that are robust across experimental conditions while accounting for method-specific artefacts, providing a comprehensive framework for understanding cross-linguistic bias in contemporary LLMs.

5 Language Selection

We selected French, Italian, Spanish, and Dutch as target languages to optimize diversity for examining cross-linguistic bias patterns within the European context while maintaining analytical coherence. These languages represent fundamentally different political traditions and cultural approaches to contentious issues, enabling systematic comparison of how cultural context influences AI political expression. The three Romance languages offer variations within shared linguistic roots, while Dutch provides a Germanic perspective. English functions as our analytical baseline, providing a reference point for measuring cultural deviation in political discourse. This configuration captures key dimensions of European political variation, including different democratic experiences, the religious-secular relationship, and approaches to EU integration, while remaining manageable for systematic cross-linguistic analysis. The choice to include Dutch instead of German was

motivated by the aim to test response biases in a less widely spoken language, restricted mainly to the Benelux area.

6 Prompt Selection

Given the scope limitations of this study, we drew inspiration from the prompt structures in the Political Voting Advice Application (ProbVAA) Dataset used by Ceron. Exploring topics such as immigration, EU policies, and religion, we created a CSV file available on GitHub (5). The need for a custom dataset stems from two factors: the desire to test prompts for different studies than Ceron's, and the lack of access to the full dataset. The objective is to investigate potential nuances in cultural interpretation and perspectives on such topics. Table 1 demonstrates the dataset structure for one specific prompt:

Table 1: Dataset Structure

Topic	**ID**	**Language**	**Lang. Code**	**Prompt Text**
Immigration	1A	English	en	What is the most effective approach for integrating immigrants into European society: assimilation into local culture, multiculturalism that preserves distinct identities, or a middle-ground integration model? Explain your reasoning.

Due to GPU limitations and lack of sophisticated instruments for detecting and analyzing sentiment polarity, we focused primarily on two prompts from the dataset (1B on Immigration and 3A on socio-cultural issues), focusing our analysis on cross-linguistic and cross-model differences.

7 Experimental Phase

We began by testing our prompts in English with the LLMs, specifically requesting answers in the target languages. Initially, results appeared identical across languages, showing the same semantic and syntactic structures with no relevant cultural differences. We utilized prompts from our custom dataset for this phase. Testing with Claude revealed slight differences in response structure, demonstrating different priorities for each language. For prompt 1B on migration, Claude showed varying priorities across languages. ChatGPT and Grok initially appeared to translate identical answers into different languages using their respective methodologies, with Grok citing the same source material in its reasoning. This suggests that the source material for information retrieval appears largely in English despite language differences. Grok appeared to be the only model that consistently translated answers without detecting relevant language bias. For the second experimental step, we computed sentiment polarity using the TextBlob library for English responses, then switched to a transformer architecture for cross-linguistic sentiment analysis. To minimize the risk of losing nuanced meanings in original languages that may indicate biases, we employed both English-translated versions and transformer-model sentiment analysis, comparing the results.

8 Results

Our approach, involving English prompts translated into selected languages, revealed slight distinctions across target languages, as translated responses appeared semantically consistent among different languages in most cases. This pattern suggests that when LLMs process English prompts, the dominant language in their training data, they may default to culturally neutral or anglophone perspectives regardless of the requested response language. The lack of distinguishable cultural features in translated responses indicates that prompt language, rather than response language, may be the primary determinant of cultural bias expression in current LLM systems. Nevertheless, topics such as migration demonstrated some cultural differences across selected languages. From Fig-

ure 2, we observe that there exists a language bias when prompting a certain response to the same topic and also that responses slightly vary among different models. This may happen due to differences in training data and model parameters. All responses across different languages and models confirm a slight left-leaning bias, given the fact that no language or model showed negative sentiment. GPT-5 appears to be the most neutral model and English the most neutral language (Figures 2 and 4).

Figure 1: Sentiment Polarity Distribution Across Languages and Models

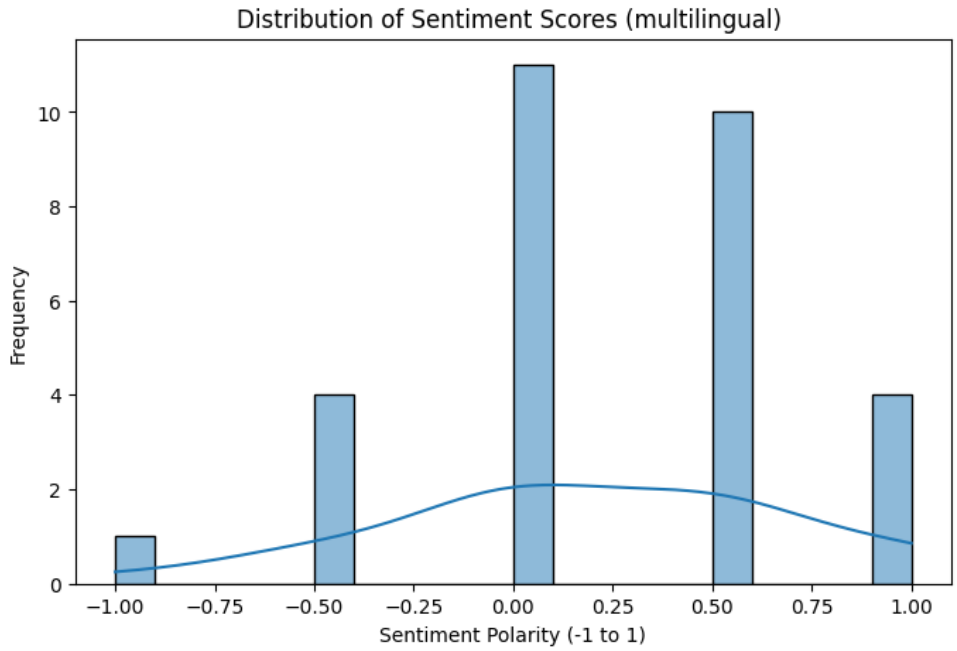

Figure 2: Comparative Analysis of Model Responses by Language

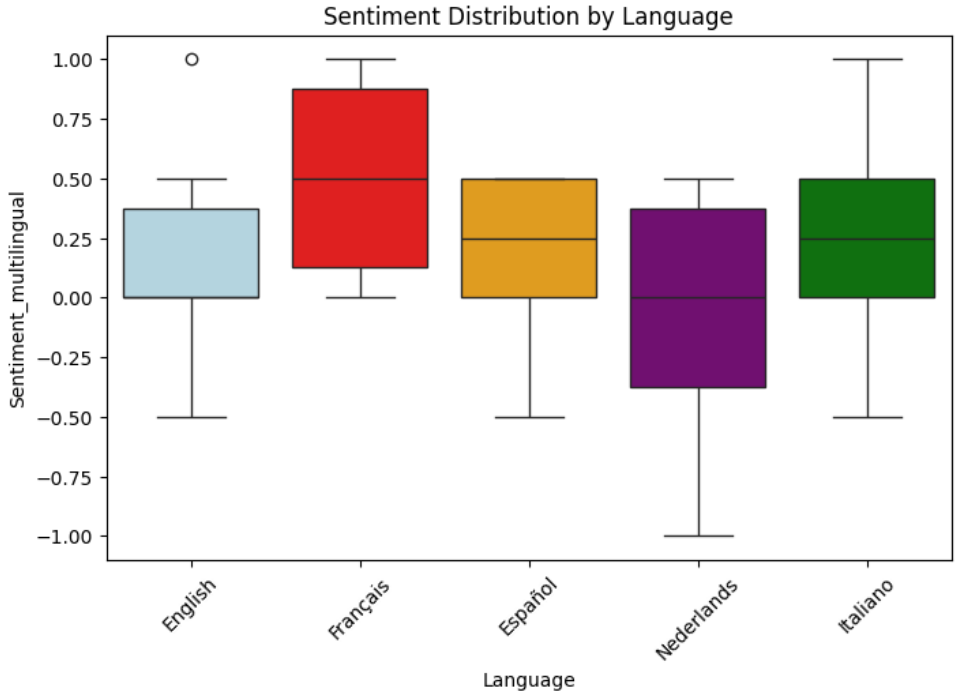

Figure 3: Average sentiment polarity by topic (English translated)

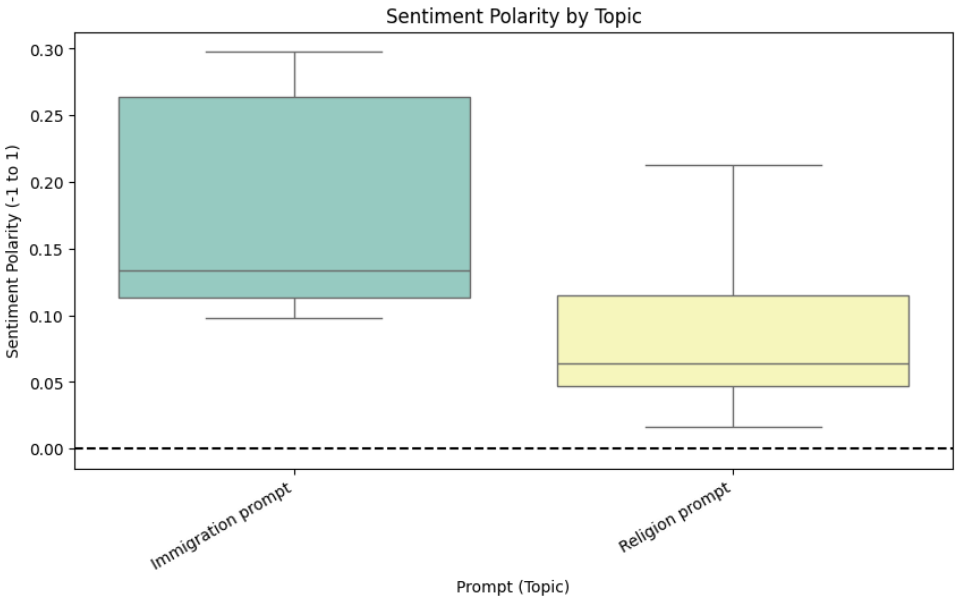

Figure 4: Average sentiment by model

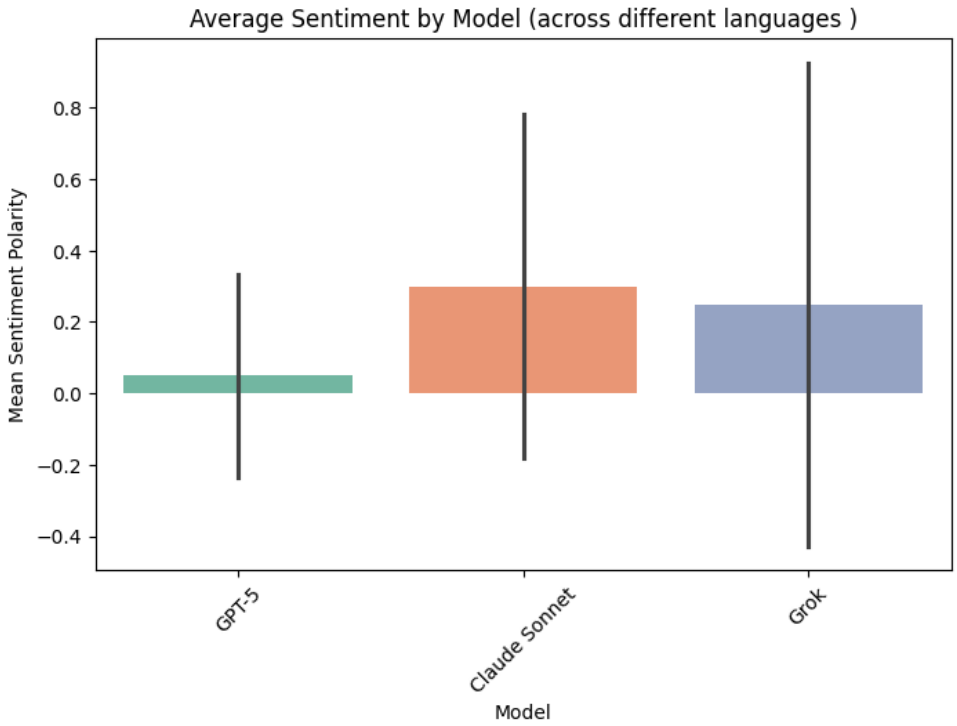

Figure 5: Average sentiment polarity by model and topic

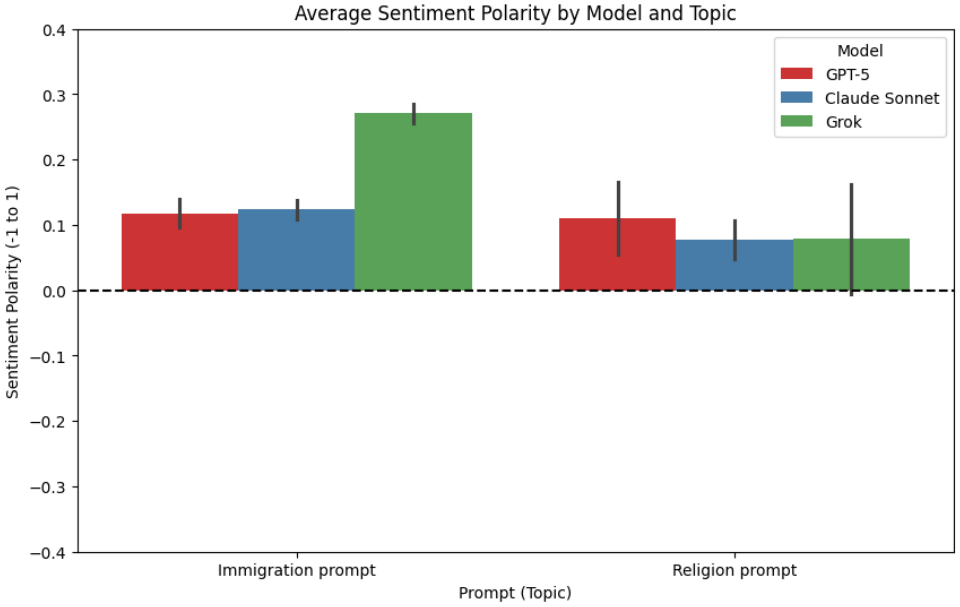

Table 2: Average Sentiment Polarity by Language (Responses Translated into English)

ID	Prompt Category	Language	Sentiment Polarity
0	Immigration	English	0.17
1	Immigration	Spanish	0.17
2	Immigration	French	0.18
3	Immigration	Italian	0.16
4	Immigration	Dutch	0.18
5	Religion	English	0.08
6	Religion	Spanish	0.14
7	Religion	French	0.06
8	Religion	Italian	0.08
9	Religion	Dutch	0.08

Note: Sentiment polarity scores range from -1 (most negative) to +1 (most positive). Values closer to 0 indicate neutral sentiment.

8.1 Analysis of Results

Figure 1 displays the frequency distribution of sentiment polarity scores ranging from -1 to 1 across all languages and models. The distribution shows a bimodal pattern with peaks around 0.0 and 0.6, suggesting that most responses cluster around neutral and moderately positive sentiment ranges. The overlaid density curve indicates the overall distribution tendency of sentiment scores in the multilingual dataset. Figure 3 compares sentiment polarity distributions between immigration and religion prompts across all models and languages. Immigration prompts show consistently higher and more positive sentiment scores (median $\sim$0.13) with greater variability, while religion prompts demonstrate lower, more neutral sentiment (median $\sim$0.07) with tighter distributions and some outliers. Figure 5 reveals distinct model-topic interaction patterns. For immigration prompts, Grok shows the highest positive sentiment ($\sim$0.27), followed by Claude Sonnet ($\sim$0.13) and GPT-5 ($\sim$0.12). For religion prompts, all models converge toward more neutral senti-

ment scores (∼0.08–0.11), with reduced variation and smaller error bars indicating more consistent responses. The sentiment analysis shown in Table 2 reveals relatively consistent polarity scores across languages for immigration-related prompts (ranging from 0.16 to 0.18), suggesting limited cultural variation in this domain. However, religion-related prompts show greater variability, with Spanish responses demonstrating notably higher positive sentiment (0.14) compared to French responses (0.06). This variation may indicate that cultural and religious contexts embedded in language-specific training data influence model responses on religiously sensitive topics.

9 Limitations

As noted by Yang (4), several limitations affect cross-linguistic bias studies in LLMs. First, most leading models are developed in the United States and China, creating potential geographical and cultural biases that may not reflect the European perspectives examined in this study. Alternative models such as Mistral (France) lack the computational resources and widespread evaluation of dominant systems, limiting comparative analysis options. A significant methodological concern involves the increasing use of Retrieval-Augmented Generation (RAG) in commercial LLMs. Sangroya (3) demonstrated that while RAG systems can improve cross-linguistic robustness, they face substantial challenges in multilingual settings, particularly when handling competing contextual information across different cultural perspectives. Their research reveals that RAG systems may inadvertently reduce cultural specificity by retrieving predominantly English-language sources, potentially homogenising responses across different linguistic contexts. This presents a critical limitation for our study, as similar responses across French, Italian, Spanish, and Dutch may reflect shared retrieval sources rather than genuine cross-cultural consensus or underlying model biases. The tendency of RAG systems to access common knowledge bases (Wikipedia, international news sources) could mask the cultural nuances our research aims to identify. Another notable methodological limitation concerns the heterogeneity of sentiment analysis tools employed across languages. For English-language tasks, sentiment scores were derived using TextBlob, which outputs continuous

values ranging from −1 to 1. In contrast, cross-linguistic sentiment analysis relied on a transformer-based model (`bert-base-multilingual-uncased-sentiment`) fine-tuned on multiple languages, which produces discrete sentiment ratings on a scale from 1 to 5. Although these outputs were subsequently normalized to a common scale between −1 and 1, the underlying differences in evaluation metrics and scoring granularity render direct comparisons between the two systems methodologically problematic.

10 Conclusions and Future Directions

This study advances our understanding of cross-linguistic political bias in Large Language Models by introducing a comprehensive methodological framework that distinguishes genuine cultural variations from translation-mediated artefacts. Through systematic comparison of ChatGPT, Claude, and Grok across French, Italian, Spanish, and Dutch political discourse, our research addresses critical gaps in multilingual AI evaluation while contributing to ongoing debates about cultural representation in artificial intelligence systems.As LLMs become increasingly integrated into multilingual information ecosystems, understanding how language and culture intersect with political bias expression becomes essential for ensuring equitable and representative AI-mediated discourse across diverse global communities. This research provides a foundation for future investigations into the complex relationships between linguistic diversity, cultural context, and ideological representation in next-generation AI systems. Future studies may employ other languages as base cases and translate into others to examine whether the provided answers differ from the English baseline. Another approach may include using prompts in different languages while requesting answers in the same languages to further isolate linguistic versus cultural effects. Additionally, future work could explore strategies to enhance the cultural sensitivity of training datasets, thereby reducing anglophone bias and improving representation in multilingual outputs.

References

[1] Bang, Y., Chen, D., Lee, N., & Fung, P. (2024). Measuring political bias in large language models: What is said and how it is said. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 11055–11077). https://doi.org/10.18653/v1/2024.acl-long.600

[2] Ceron, T., Falk, N., Barić, A., Nikolaev, D., & Padó, S. (2024). Beyond prompt brittleness: Evaluating the reliability and consistency of political worldviews in LLMs. *Transactions of the Association for Computational Linguistics*, 12, 1378–1400. https://doi.org/10.1162/tacl_a_00710

[3] Sangroya, A., et al. (2024). Multilingual Retrieval Augmented Generation for Culturally-Sensitive Tasks: A Benchmark for Cross-lingual Robustness. *Findings of ACL 2025*. https://arxiv.org/abs/2410.01171

[4] Yang, K., Huang, J., Cao, Y., Chen, J., & Cheng, K. (2024). Unpacking political bias in large language models: A cross-model comparison on U.S. politics. *arXiv preprint arXiv:2412.16746*. https://arxiv.org/abs/2412.16746

[5] Scuderi, A. G. (2025). *Cross-linguistic prompts for LLMs* [Dataset]. GitHub repository. Retrieved from https://github.com/albertoscuderi/cross_linguistics_prompts_4_llms

[6] Ceron, T., Falk, N., Barić, A., Nikolaev, D., & Padó, S. (2024). eval_political_worldviews: Dataset for evaluating political worldviews in LLMs [Dataset]. GitHub. https://github.com/tceron/eval_political_worldviews