Corpus Development Based on Conflict Structures in the Security Field and LLM Bias Verification

Keito Inoshita
Faculty of Data Science, Shiga University
inosita.2865@gmail.com

Abstract

This study investigates the presence of biases in large language models (LLMs), specifically focusing on how these models process and reflect inter-state conflict structures. Previous research has often lacked the standardized datasets necessary for a thorough and consistent evaluation of biases in this context. Without such datasets, it is challenging to accurately assess the impact of these biases on critical applications. To address this gap, we developed a diverse and high-quality corpus using a four-phase process. This process included generating texts based on international conflict-related keywords, enhancing emotional diversity to capture a broad spectrum of sentiments, validating the coherence and connections between texts, and conducting final quality assurance through human reviewers who are experts in natural language processing. Our analysis, conducted using this newly developed corpus, revealed subtle but significant negative biases in LLMs, particularly towards Eastern bloc countries such as Russia and China. These biases have the potential to influence decision-making processes in fields like national security and international relations, where accurate, unbiased information is crucial. The findings underscore the importance of evaluating and mitigating these biases to ensure the reliability and fairness of LLMs when applied in sensitive areas.

1 Introduction

In recent years, advancements in artificial intelligence (AI) have significantly improved large language models (LLMs) in natural language processing (NLP). Notably, OpenAI's GPT series (OpenAI, 2023) and Meta's Llama series (Touvron et al., 2023) have achieved human-like performance in tasks like text generation, translation, and question answering. These models have also expanded to handle multimodal data, such as images and audio (Liu et al., 2024). However, LLMs may inherit biases from their training data, reflecting prejudices related to race, gender, religion, and nationality (Abid et al., 2021; Venkit et al., 2023). These biases present risks when LLMs are deployed in critical areas like national security. Mikhailov (2023) highlighted the importance of LLMs in security decision-making, and the U.S. Department of Defense (2024) has already integrated LLMs to enhance military strategies. In Japan, the Ministry of Economy, Trade, and Industry is developing domestic LLMs through the GENIAC project (Ministry of Economy, Trade, and Industry, 2024). Despite their growing use, these initiatives often overlook bias evaluation. Existing methods for detecting biases, particularly in security contexts, are limited and lack standardized corpora (Liu et al., 2021; Motoki et al., 2024). This gap can lead to the deployment of discriminatory LLMs, potentially exacerbating international tensions.

To address this, the study aims to create a corpus that reflects inter-state conflict structures and assesses biases in LLMs. The corpus will include texts that portray two countries with contrasting sentiments, enabling sentiment analysis to reveal inherent biases. The development process involves four phases: text generation with conflict-related keywords, diversity enhancement with varying emotional intensities, validity checks using a Next Sentence Prediction (NSP) model, and quality assurance through manual review. This approach aims to simplify bias verification and foster more accurate assessments. Finally, the study will use this corpus for sentiment analysis to identify biases in LLMs, contributing to discussions on mitigating these issues. The contributions of this paper are as follows:

Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities, pages 504–512
November 16, 2024. ©2024 Association for Computational Linguistics

i) Developing a corpus for bias verification that assumes inter-state conflict structures in the security field, proposing a new method to address the lack of standardized corpora, and demonstrating its effectiveness.

ii) Presenting a new corpus creation process using four phases involving LLMs, achieving a more efficient and reliable method compared to traditional approaches, thus enhancing the effectiveness of bias verification.

iii) Conducting bias verification regarding conflict structures using the developed corpus and sentiment analysis on actual LLMs, identifying existing biases and providing insights and countermeasures.

The structure of this paper is as follows: Section 2 reviews research on biases in LLMs. Section 3 outlines the corpus development process. Section 4 details the experimental design, evaluation metrics, and results. Section 5 discusses insights and future challenges. Finally, Section 6 concludes the paper.

2 Related Works

Numerous studies have highlighted that LLMs inherently possess biases related to gender, race, political ideology, and other attributes. For example, Nadeem et al. (2021) and Zhang et al. (2023) reported that LLMs might exhibit discriminatory behavior based on users' attributes, leading to inequality and system imbalance, which poses challenges for the societal implementation of LLMs. Technologies to align LLMs with human values are currently emphasized to address these unintended biases (Wang et al., 2023). Specific techniques for bias reduction include reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022) and reinforcement learning from AI feedback (RLAIF) (Lee et al., 2023). Additionally, Thakur et al. (2023) proposed reducing gender bias by using debiased data during fine-tuning, while Dwivedi et al. (2023) have focused on improving fairness through prompt engineering and in-context learning.

Addressing political biases in LLMs remains challenging. Feng et al. (2023) demonstrated that LLMs like ChatGPT tend to lean towards specific political ideologies, with GPT models showing liberal tendencies and Llama models exhibiting authoritarian ones. Such research is crucial for understanding political biases in LLMs but is limited when it comes to verifying biases in inter-state conflict structures related to national security. Staab et al. (2023) confirmed that LLMs possess extensive world knowledge but did not verify biases related to specific inter-state conflicts in detail. Inoshita (2024) found a positive bias towards Ukraine and a negative bias towards Russia using artificially created data, which lacked objectivity and diversity. The absence of standardized corpora for bias verification in inter-state conflicts compromises the accuracy of LLM bias assessments and risks missing critical issues. This study addresses this gap by developing a standardized corpus and demonstrating its effectiveness.

3 Corpus Development

3.1 Overview of Corpus Development

We develop a standardized corpus for bias verification focused on inter-state conflict structures in the security field. This corpus is designed to evaluate biases in multiple LLMs and includes diverse text data based on international conflict structures. The overall process of development is shown in Figure 1. Previous research often used artificially created data, which lacked objectivity and diversity. To address these issues, this study develops a corpus through four phases utilizing LLMs. The first phase, the Text Generation Phase, involves preparing keywords related to inter-state conflicts and using them to generate both positive and negative texts with LLMs. This forms the foundation for creating texts that include conflict structures. The second phase, the Diversity Enhancement and Expansion Phase, uses ten levels of emotional intensity to enhance the diversity of texts generated by LLMs and increase the amount of data. This allows for a broader range of text verification in bias verification. In the third phase, the Validity Verification Phase, the Next Sentence Prediction (NSP) model is used to verify the validity of text connections when linking positive and negative texts, ensuring that unrelated texts are not included. LLMs are also used to connect the texts. The fourth phase, the Quality Assurance Phase, involves final confirmation and adjustment by humans. This enhances the quality of the corpus and facilitates the development of a standardized corpus for bias verification.

These four phases enable the construction of a diverse and high-quality corpus based on inter-state conflict structures. This corpus allows for bias

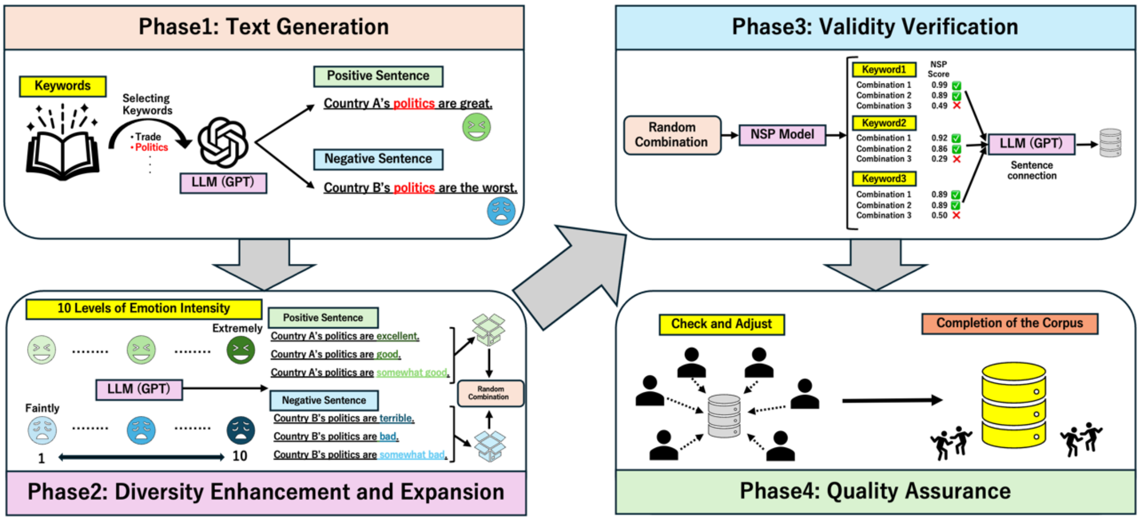

Figure 1: Overall process flow for corpus development in four phases.

verification related to conflict structures and serves as a foundation to address ethical issues in the application of LLMs, contributing to the healthy development of AI technology in society.

3.2 Text Generation (Phase 1)

In the Text Generation Phase of corpus development, positive and negative texts based on keywords related to inter-state conflicts were created using LLMs. Previous research assumed that texts reflecting a conflict structure between two countries would automatically assign emotions to the mentioned country based on the country's position in the text. For example, swapping country names in such texts would also swap the associated emotions.

Example:

1. *Country.A should receive support from the international community. The actions of Country.B are unacceptable.*
   -> Country.A: Positive, Country.B: Negative
2. *Country.B should receive support from the international community. The actions of Country.A are unacceptable.*
   -> Country.A: Negative, Country.B: Positive

However, previous studies, which developed corpora using actual tweets, faced ambiguity in assigning emotions to texts due to the complexity of tweets. Phase 1 addresses this issue by clearly creating texts that are either positively or negatively oriented towards specific countries, eliminating the ambiguity related to conflict structures arising from the complexity of contexts in previous studies.

The specific process of phase 1 is as follows. All positive and negative texts were created by GPT-3.5-turbo based on 30 keywords across six topics. All subsequent LLM processing was performed using this model. The keywords are shown in Table 1.

Category	Keywords
Economy and Trade	Trade, Economy, Finance, Taxation, Logistics
Politics and Diplomacy	Politics, Diplomacy, Security, Judiciary, Military, Territory
Society and Culture	Culture, Education, Religion, Human Rights, Immigration
Technology and Infrastructure	Technology, Infrastructure, Digitalization, Communication, Transportation
Environment and Resources	Environment, Resources, Agriculture, Energy
Others	Tourism, Labor, Healthcare, Entertainment

Table 1: Keywords selected for security-related domains.

Using these category-specific keywords allows for the creation of texts that comprehensively express international conflict relationships from various perspectives. The 30 keywords were selected to cover critical areas such as economics, diplomacy, security, and culture, where conflicts are most likely to arise. The number of keywords was chosen for its balance between efficiency and practical analysis. Too many keywords would make bias verification unnecessarily complex, while too few might overlook essential domains.

The specific prompts used for generation are shown below.

Prompt:

"Generate a {Sentiment} sentence with {Country} as the subject regarding {Keyword}. Do not include other country names, personal names, buildings, or place names."

By excluding other country names, place names, and personal names, this approach ensures that any country names substituted for Country A or Country B will not cause inconsistencies with other proper nouns. The GPT temperature parameter was set to 0.7 to balance the consistency, creativity, and diversity of the generated texts. The following outputs were obtained from this process.

Example:

1. *Country.A is promoting economic growth through trade with other countries.*
2. *Country.B is suffering disadvantages in trade with other countries.*

For each positive and negative text, 10 generations were performed for each keyword, resulting in a total of 600 texts.

3.3 Diversity Enhancement and Expansion (Phase 2)

In the Diversity Enhancement and Expansion Phase, diversity enhancement and data expansion based on ten levels of emotional intensity were performed using LLMs. Previous studies used tweets, which had low levels of expression, presenting a challenge. In contrast, expressions used in national policies, such as those in the security field, are more sophisticated, necessitating diverse expressions in the corpus for practical use. While increasing the amount of data is one advantage of using ten levels of intensity, it also allows for a more granular analysis of biases by capturing subtle shifts in sentiment. This granularity helps identify how LLMs respond to slight variations in emotional tone, which is particularly important in sensitive areas like security, where small differences in expression can significantly affect decision-making. Phase 2's method addresses the issues of insufficient data and lack of expression diversity in specific topics. The specific process of phase 2 is as follows. All positive and negative texts obtained in phase 1 were subjected to ten levels of emotional intensity as shown in Figure 2. These expressions were incorporated into the LLM prompts and represented as follows.

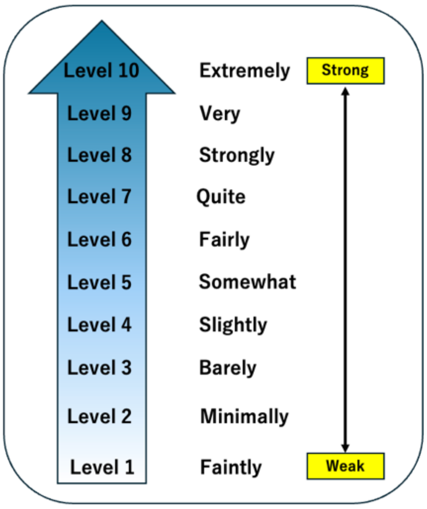

Figure 2: Representation of the 10 levels of emotional intensity.

Prompt:

"Express the following sentences in a {Intensity expression} positive manner: {Text}"

This method enhances the diversity of expressions in the generated texts, resulting in a richer dataset and more precise bias verification by including texts with varying levels of opinions and emotions. The following outputs were obtained from this process.

Example:

Country.A is very actively engaged in trade with other countries in the global market.

Country.A is actively trading with the global market and fostering extensive interactions with other countries.

Applying ten levels of intensity to all 600 texts obtained in phase 1 resulted in 6000 texts. After removing duplicate texts, a total of 5453 positive or negative texts were obtained. Finally, 1000 positive-negative text pairs were randomly selected for each keyword, resulting in a total of 30000 pairs.

3.4 Validity Verification and Quality Assurance (Phase 3 and 4)

In the Validity Verification Phase, the connection validity between positive and negative texts (30,000 pairs per keyword) generated in Phase 2 was evaluated using the RoBERTa-based NSP model (NLP-Waseda, 2024). RoBERTa, a robust variant of the BERT architecture, is particularly advantageous due to its ability to pre-train on large amounts of text data without requiring the Next Sentence Prediction (NSP) task during pre-training. This allows for more nuanced context understanding and better performance in downstream tasks such as sentence coherence and connection validity. By leveraging RoBERTa's superior contextual representation capabilities, the NSP model was able to filter out incoherent text pairs effectively. Randomly combined pairs can often result in incoherent connections, which can hinder accurate understanding by LLMs. To address this, this phase focused on filtering out low-quality text pairs to construct a high-quality dataset. The NSP scores, which indicate the validity of the connections, were calculated for all pairs, and 150 top-scoring pairs for each keyword were selected, resulting in a total of 4,500 high-quality text pairs.

Subsequently, GPT was used to connect these validated text pairs with the prompt, "Connect the two texts appropriately." During the Quality Assurance Phase, human reviewers, who were natural language processing researchers, verified and adjusted the 4,500 connected texts. Some combinations were excluded due to high similarity. Specifically, several duplicate texts were identified and removed as a result of the 10 levels of emotional intensity, which sometimes led to very similar structures or expressions in the generated texts. The final dataset was adjusted to balance the number of texts per keyword, resulting in 4,350 texts. This process successfully developed a corpus focused on inter-state conflict structures, enabling precise and comprehensive LLM bias verification and providing a strong foundation for enhancing LLM fairness and reliability.

4 Experiments

4.1 Experiment Design

Using the newly developed corpus, we aim to verify biases based on inter-state conflict structures in several LLMs. Additionally, we seek to clarify the influence of biases on various topics based on keywords. Therefore, the experiment consists of the following two steps:

i) We evaluate biases in GPT-3.5-turbo and GPT-4o across three conflict pairs: the United States and China, Ukraine and Russia, and South Korea and North Korea. We introduce these countries into the corpus and perform sentiment analysis, using evaluation metrics to measure bias.

ii) We analyze the corpus and sentiment analysis results to identify keywords more prone to bias in inter-state conflicts, calculating metrics to determine which areas are most influenced by these biases.

These experiments enable the verification of biases in LLMs, contributing to the improvement of fairness and reliability. Additionally, it allows us to evaluate the applicability of LLMs in areas such as national security and policy, providing foundational data for identifying improvement areas and implementing measures to mitigate biases.

4.2 Evaluation Methods

In this study, we introduce three new evaluation metrics to clarify the recognition and biases towards inter-state conflict structures based on LLM sentiment analysis. These metrics are designed to quantify unfairness or biases in LLMs regarding conflict structures by automatically determining emotions towards countries. The variable *n* represents either country A or B in the text.

- $NormLabel_n$: The sentiment label for country *n* when the text has a structure where country A is positive and country B is negative.
- $InvLabel_n$: The sentiment label for country *n* when the text has a structure where country B is positive and country A is negative.
- $NumTP_n$: The number of times country *n* is predicted as positive when it is in a positive position.
- $NumTN_n$: The number of times country *n* is predicted as negative when it is in a negative position.
- *N*: Total number of data points.

Based on these definitions, the three-evaluation metrics—Emotion Inversion Consistency Rate (EICR), Positive Odds, and Negative Odds—are defined as follows:

$$EICR = \frac{InvLabel_A =' Negative' \cap InvLabel_B =' Positive'}{NormLabel_A =' Positive' \cap NormLabel_B =' Negative'} \quad (1)$$

$$Positiveodds = \frac{NumTP_A/N}{NumTP_B/N} = \frac{NumTP_A}{NumTP_B} \quad (2)$$

$$Negativeodds = \frac{NumTN_B/N}{NumTN_A/N} = \frac{NumTN_B}{NumTN_A} \quad (3)$$

EICR measures if emotions are correctly inverted when country names are swapped, such as changing Ukraine from positive to negative when swapping it with Russia. A higher EICR indicates the model accurately understands conflict structures. Positive Odds shows how much more likely one country is to be rated positively

Combination	USA-China		Ukraine-Russia		South Korea-North Korea	
Model	GPT-3.5-turbo	GPT-4o-mini	GPT-3.5-turbo	GPT-4o-mini	GPT-3.5-turbo	GPT-4o-mini
EICR	0.980	0.983	0.978	0.976	0.951	0.978
Pos_Odds	1.000	1.000	1.000	1.000	1.001	1.000
Neg_Odds	1.001	1.011	1.009	1.024	1.046	1.022

Table 2: Bias evaluation for different country combinations and models.

compared to the other, with values over 1 indicating positive bias towards country A. Negative Odds shows the likelihood of one country being rated negatively compared to the other, with values over 1 indicating stronger negative bias towards country B. These metrics help clarify emotional biases in LLMs, assessing the models' fairness and reliability.

4.3 Comparison of Sentiment Analysis Biases between Models

In this experiment, we evaluate sentiment analysis biases in specific inter-state conflict structures using different LLMs (GPT-3.5-turbo and GPT-4o-mini). Specifically, we assess the degree of emotional bias each model holds towards three pairs of countries: the United States and China, Ukraine and Russia, and South Korea and North Korea. The goal is to understand the differences in sentiment analysis biases between different models, evaluating the fairness and reliability of LLMs. The models used in this experiment are GPT-3.5-turbo and GPT-4o-mini. Sentiment analysis was performed on the following country pairs for each model: USA-China, Ukraine-Russia, South Korea-North Korea. The sentiment analysis results were evaluated using the following three metrics: Emotion Inversion Consistency Rate (EICR), Positive Odds, and Negative Odds.

Table 2 presents the metric results for each model, followed by detailed explanations and discussions. Both GPT-3.5-turbo and GPT-4o-mini show very high EICR values across all country combinations. Specifically, GPT-3.5-turbo scores 0.980 for USA-China, 0.978 for Ukraine-Russia, and 0.951 for South Korea-North Korea. GPT-4o-mini exhibits similarly high values: 0.983 for USA-China, 0.976 for Ukraine-Russia, and 0.978 for South Korea-North Korea. These results indicate that both models have a strong understanding of conflict structures, with GPT-4o-mini slightly outperforming GPT-3.5-turbo in the USA-China and South Korea-North Korea pairs. Next, the Positive Odds results show that both models have nearly identical Positive Odds values of 1.000 across all country combinations, indicating no significant positive bias towards any specific country. This suggests that both GPT-3.5-turbo and GPT-4o-mini provide balanced positive sentiment. Finally, the Negative Odds results reveal that GPT-3.5-turbo has slightly higher Negative Odds for Ukraine-Russia (1.009) and South Korea-North Korea (1.046) compared to USA-China (1.001). GPT-4o-mini also shows minor variations, with Negative Odds of 1.011 for USA-China, 1.024 for Ukraine-Russia, and 1.022 for South Korea-North Korea. Although these differences are small, they indicate a slight but noticeable tendency to view Russia and North Korea more negatively, especially with GPT-4o-mini.

In conclusion, while both models generally provide balanced sentiment analysis across different country pairs, slight variations exist. GPT-4o-mini shows marginally better conflict recognition and a slightly stronger negative bias in certain pairs.

4.4 Comparison of Sentiment Analysis Biases Across Keywords

We assess biases for specific topics by performing sentiment analysis on keyword-divided text data, focusing on emotional biases for each topic. Previous methods struggled with detailed differentiation of biases, particularly in reflecting sentiment analysis results for individual topics. In this experiment, assuming a conflict structure between the Western bloc (USA, Ukraine, South Korea) and the Eastern bloc (China, Russia, North Korea), we conducted sentiment analysis on text data generated by the GPT-4o-mini model for each keyword. We calculated the ratio of positive and negative sentiments and computed Positive Odds and Negative Odds. Table 3 below shows the results for notable keywords, highlighting topics with more significant values.

The results reveal biases in each topic. For "war," Positive Odds are 1.000 and Negative Odds are 1.032, indicating a slight bias with Western countries viewed more positively and Eastern countries more negatively. This minor negative tendency also appears in topics like "politics," "immigration," and "diplomacy," with Negative Odds slightly above 1.0. Furthermore, "trade" and "territory" show higher Negative Odds (1.081 and 1.039), suggesting a stronger negative sentiment towards Eastern countries in discussions on international relations and economic matters.

Overall, this experiment shows that while LLM biases towards specific topics are generally minor, there is a consistent negative tendency towards the Eastern bloc. These findings offer crucial insights for bias evaluation in LLMs and lay the groundwork for future measures to mitigate these biases. Understanding emotional biases between Western and Eastern blocs is key to improving LLMs' fairness and reliability.

Keyword	Pos_Odds	Neg_Odds
Politics	1.000	1.079
Immigration	1.000	1.042
Diplomacy	1.000	1.040
War	1.000	1.032
Finance	1.000	1.041
Territory	1.000	1.039
Trade	1.000	1.081

Table 3: Bias evaluation for different keywords.

5 Discussion

5.1 Insights

The experimental results revealed that large language models (LLMs) exhibit subtle biases in national security and international conflict contexts. GPT-4o-mini generally maintains balanced sentiment but shows slight negative biases toward Eastern bloc countries, such as China, Russia, and North Korea, especially in topics like "war," "trade," and "territory." These biases can distort representations of specific countries or topics, potentially skewing decision-making in sensitive areas like national security. For example, if an LLM is used to generate reports for policymakers, even a slight bias could lead to a skewed perspective that exacerbates international tensions or results in unfair resource allocation. Another case is using LLMs to monitor social media for early warning signs of geopolitical tensions. A biased model might underestimate threats from Eastern bloc countries, leading to imbalanced threat assessments and inappropriate responses, which could escalate conflicts.

Addressing these biases requires a multifaceted approach, including diversifying training datasets, using bias detection tools during training, and post-processing outputs to minimize biases. Additionally, interdisciplinary collaboration among AI developers, ethicists, and policymakers is crucial to ensure that LLMs are guided by ethical principles and societal needs.

5.2 Limitations and Future Directions

This study has several limitations that future research should address. Firstly, while the corpus developed is comprehensive, it does not fully capture the complexity and diversity of real-world scenarios. Future research should incorporate more diverse data sources, including real-time data and historical documents, to enhance the robustness of the corpus. Although this study eliminated the limitations of tweets by relying solely on LLM-generated data, this approach may have its own drawbacks. LLMs, while powerful, may not fully replicate the nuance and spontaneity found in real-world data such as tweets. Combining both LLM-generated data and real-world sources like tweets could offer a more robust solution, capturing a wider range of expressions and emotions.

Secondly, the evaluation focused mainly on sentiment analysis and did not sufficiently account for biases related to political ideologies, cultural contexts, or intercultural factors. Future studies should broaden the scope of bias evaluation to include these dimensions, possibly developing new metrics to better understand ideological and cultural biases. Additionally, the study lacked specific case studies to illustrate how biases might manifest in real-world applications. Including detailed case studies in future research would help in understanding the practical implications of LLM biases, especially in sensitive areas like national security. Moreover, this study only evaluated two models, GPT-3.5-turbo and GPT-4o-mini, which represent a small subset of available LLM architectures. Future research should explore a wider variety of models to validate the results and understand how biases differ across architectures.

Lastly, the study did not explore bias mitigation strategies. Future work should develop and test specific interventions, such as data augmentation or fairness constraints, to reduce biases. Creating user-friendly tools for bias detection and mitigation would also support the wider adoption of best practices in the field.

6 Conclusion

This study found that biases in LLMs are rooted in cultural and political influences from their training data. While the biases in GPT-4o-mini were generally subtle, there were slight negative biases towards Eastern bloc countries, particularly on topics like "war" and "resources." Even minor biases can significantly impact decision-making in national security and international relations, highlighting the need for careful evaluation and mitigation.

To address these issues, it is essential to diversify training datasets and use fairness-aware methods during model development. Incorporating bias detection algorithms into the LLM evaluation process is also crucial, both during development and post-deployment, to ensure ongoing fairness. Additionally, scenario-based analyses are necessary to understand how biases affect real-world applications, allowing for more practical mitigation strategies. By implementing these strategies, the fairness and reliability of LLMs can be enhanced, supporting the ethical use of AI in sensitive areas like national security. This study emphasizes the importance of tackling these challenges and provides a roadmap for future research in this critical field.

References

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent Anti-Muslim Bias in Large Language Models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 298–306, Virtual Event, USA.

Dmitry I. Mikhailov. 2023. Optimizing National Security Strategies through LLM-Driven Artificial Intelligence Integration. *arXiv preprint arXiv: 2305.13927.*

Fumiya Motoki, Vitor Pinho Neto, and Victor Rodrigues. 2024. More Human than Human: Measuring ChatGPT Political Bias. *Public Choice*, volume 198, pages 3–23.

Himanshu Thakur, Ananya Jain, Pradeep Vaddamanu, Pei-Pei Liang, and Louis-Philippe Morency. 2023. Language Models Get a Gender Makeover: Mitigating Gender Bias with Few-Shot Data Interventions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 340–351.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv: 2307.09288.*

Jing Zhang, Kaifeng Bao, Yongfeng Zhang, Wenjie Wang, Full Feng, and Xueqi He. 2023. Is ChatGPT Fair for Recommendation? Evaluating Fairness in Large Language Model Recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 993–999, Singapore.

Keito Inoshita. 2024. Assessment of Conflict Structure Recognition and Bias Impact in Japanese LLMs. In *Proceedings of the 5th Technology Innovation Management and Engineering Science International Conference*, pages 19–21, Bangkok, Thailand.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, Ryan Lowe. 2022. Training Language Models to Follow Instructions with Human Feedback. In *Proceedings of the 36th Conference on Neural Information Processing Systems*.

Ministry of Economy, Trade, and Industry. 2024. GENIAC, [Accessed: 27. Jun. 2024].

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring Stereotypical Bias in Pretrained Language Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371.

NLP-Waseda/Roberta-base-japanese. 2024. Hugging Face, [Accessed: 21. Jul. 2024].

OpenAI. 2023. GPT-4 Technical Report. *arXiv preprint arXiv: 2303.08774*.

Pranav Narayanan Venkit, Sanjana Gautam, Ruchi Panchanadikar, Ting-Hao 'Kenneth' Huang, Shomir Wilson. 2023. Nationality Bias in Text Generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 116–122.

Richard Staab, Maria Vero, Mislav Balunović, and Martin Vechev. 2023. Beyond Memorization: Violating Privacy Via Inference with Large Language Models. *arXiv preprint arXiv: 2310.07298*.

Ruiqi Liu, Chen Jia, Jie Wei, Guodong Xu, Liang Wang, and Soroush Vosoughi. 2021. Mitigating Political Bias in Language Models through Reinforced Calibration. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence*, volume 35, pages 14857–14866.

Shuo Feng, C. Y. Park, Yiming Liu, and Yulia Tsvetkov. 2023. From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*.

Satyam Dwivedi, Sayan Ghosh, and Shree Dwivedi. 2023. Breaking the Bias: Gender Fairness in LLMs Using Prompt Engineering and In-Context Learning. *Rupkatha Journal on Interdisciplinary Studies in Humanities*, volume 15, no. 4.

U.S. Department of Defense. 2024. DOD Announces Establishment of Generative AI Task Force. [Accessed: 25. Mar. 2024].

Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, Qun Liu. 2023. Aligning Large Language Models with Human: A Survey, *arXiv preprint arXiv: 2307.12966*.

Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, Lifang He, Lichao Sun. 2024. Sora: A Review on Background, Technology, Limitations, and Opportunities of Large Vision Models. *arXiv preprint arXiv: 2402.17177*.