https://doi.org/10.57017/jorit.v4.2(8).04

Geopolitical Bias in Sovereign Large Language Models: A Comparative Mixed-Methods Study

Sridhar JONNALA
IBM India, Karnataka, Bangalore, India
International School of Management Excellence, Bangalore, Karnataka, India
https://orcid.org/0000-0003-2122-1972

Basavaraj SWAMY
International School of Management Excellence, Bangalore, Karnataka, India
https://orcid.org/0009-0007-8173-4397

Nisha Mary THOMAS
International School of Management Excellence- Bangalore, Karnataka, India
Symbiosis Institute of Business Management, Bengaluru. India
https://orcid.org/0000-0002-2427-6146

Abstract

Sovereign large language models (LLMs), emerging as strategic assets in global information ecosystems, represent advanced AI system developed under distinct national governance regimes. This study examines how model origin and governance context influence AI-generated narratives on international territorial disputes. The study compares outputs from three prominent sovereign LLMs - OpenAI's GPT-4o (United States), DeepSeek-R1 (China), and Mistral (European Union), across 12 high-profile territorial conflicts. Statistically significant differences in each model's sentiment distribution and geopolitical framing are identified using a mixed-methods approach that combines sentiment analysis with statistical evaluation (chi-square tests and analysis of variance, ANOVA) on responses to 300 standardized prompts.

The findings indicate model provenance substantially shapes the tone and stance of outputs, with each LLM reflecting distinct biases aligned with its national context. These disparities carry important policy and societal implications: reliance on a single sovereign model could inadvertently bias public discourse and decision-making toward that model's native perspective. The study highlights ethical considerations such as transparency and fairness and calls for robust governance frameworks. It underscores the need for careful oversight and international cooperation to ensure that sovereign LLMs are deployed in a manner that supports informed and balanced geopolitical dialogue

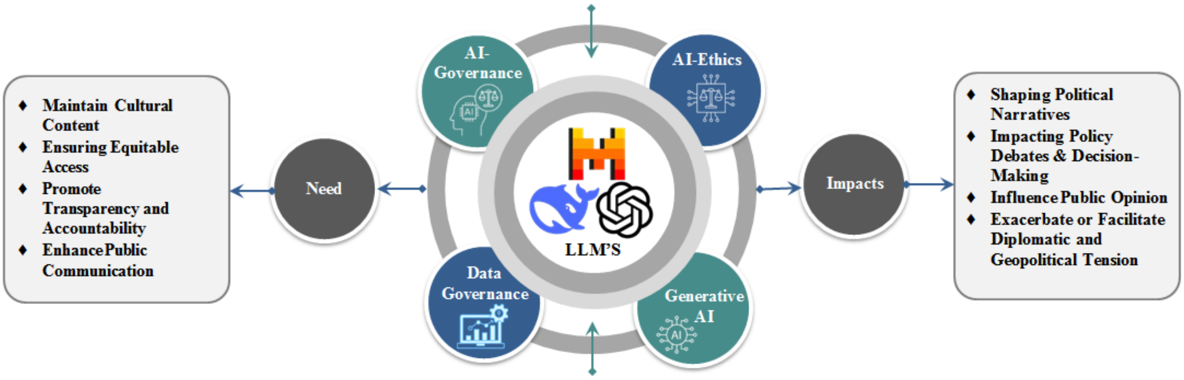
Keywords: artificial intelligence; LLMs; territorial conflicts; AI ethics; generative AI; responsible AI; AI governance.

Introduction

Artificial Intelligence (AI) has become one of the most powerful technologies of the 21st century. Numerous industries, such as education, electronic governance (e-governance), and medicine are transmuted by AI by increasing productivity and accuracy, leading to revolutions and modest benefits (Kaplan & Haenlein, 2019). Humans require an efficient strategy like Natural Language Processing (NLP) in this era of technological advancement to communicate with machines to comprehend and generate human-like content. Nevertheless, an immense amount of data source and training considerations are required by the advancement of AI and NLP, which contributed to the evolution of Large-scale Language Models (LLMs). LLMs' advent has revolutionized the AI field by utilizing neural linkages, training parameters, and text to enable machines to understand human language text.

A sovereign language guarantees a nation's ability to express its distinct cultural inheritance, reinforce its national uniqueness, and uphold its independence in the global landscape (Bondarenko et al., 2025; Dandage, 2025). In content generation and natural language indulgence, generative AI and LLMs offer transformative proficiencies. This AI transformation also elevates the structural workflows through extensive AI models and other technologies to create constantly evolving sectors. The necessity and impact of LLMs in geopolitical disputes are depicted in Figure 1.

Figure 1: Influence of LLMs in transnational disagreements


Maintain Cultural Content
Ensuring Equitable Access
Promote Transparency and Accountability
Enhance Public Communication
Need
AI-Governance
AI-Ethics
LLM'S
Data Governance
Generative AI
Impacts
Shaping Political Narratives
Impacting Policy Debates & Decision-Making
Influence Public Opinion
Exacerbate or Facilitate Diplomatic and Geopolitical Tension

OpenAI's GPT-4o and DeepSeek-R1 have received significant attention among the LLMs for different community-driven innovations, specialized adaptations, instruction-following capabilities, and Application Programming Interface (API) flexibility. As a current trend, the development of powerful language models has gained momentum across e-commerce, finance, healthcare, and travel sectors due to its task activities, response to complex queries, and writing and interpreting code (Maiti et al., 2025; Raza et al., 2025; Dandage, 2025). Systematically, a wide range of organization's data oversees the management and utilization of artificial intelligence by AI Data Governance. Therefore, it ensures the data accuracy, security, and ethical protection of the data used in AI systems. Similarly, the AI ethics center ensures the adherence to fundamental principles of trustworthiness, such as transparency and interpretability of AI systems (Linkon et al., 2024; Linegar et al., 2023).

Building upon the foundational role of LLMs in enabling digital transformation across sectors, it is imperative to delineate how sovereign LLMs - those developed under direct national or regional jurisdiction - exhibit unique technical and operational characteristics. Diverging from conventional LLMs, sovereign implementations systematically embed geopolitical, regulatory, and cultural imperatives into their core architecture and deployment frameworks, extending beyond generalized AI governance and data stewardship principles.

Specifically, sovereign LLMs are trained on deliberately curated datasets that mirror national information ecosystems, incorporating government publications, state-mediated content, and culturally contextualized corpora. This data governance paradigm is frequently shaped by local content policies and censorship norms, particularly concerning politically sensitive narratives. Architecturally, these models integrate multilingual tokenizers, domain-specific embeddings, and privacy-preserving features aligned with regional regulatory standards (e.g., GDPR compliance), reflecting domestic legal landscapes (Bondarenko et al., 2025; Bender et al., 2021). Beyond training, sovereign models systematically undergo *instruction tuning* and ideological guardrail adjustments to align outputs with state-endorsed perspectives on governance, historical discourse, and territorial integrity (Guey et al., 2025; Pacheco et al., 2025). Deployment occurs exclusively via national digital infrastructure, embedding models into public-sector platforms (e.g., legal advisory systems, e-governance portals, and healthcare interfaces) to reinforce computational sovereignty. Collectively, these technical, normative, and infrastructural dimensions position sovereign LLMs as operationally distinct from globally oriented counterparts, necessitating contextually tailored evaluation and governance protocols.

Problem statement

Ethical concerns and communal complications are increasingly raised by the expansion of AI models. Specifically, multiple risk dimensions, such as cultural diversity, cyber security, and data protection authority are encountered by LLMs. In addition, biased AI models transfer sensitive information and challenge the privacy of sectors including defence, finance, healthcare, education, and international affairs (Usman et al., 2023; Kulesz, 2024). While these models serve as vast repositories of data, their training processes, dataset transparency, and content generation logic remain inadequately scrutinized in academic research. Furthermore, limited empirical work has addressed how national governance structures influence the geopolitical framing embedded in sovereign LLMs (SLLMs). Therefore, the present study aims to evaluate the moral and practical consequences of implementing SLLMs and assess the social security encounters and bias-mitigating ability of LLMs. Also, the research focusses on governance strategies that align with inclusive and equitable AI development.

Objectives of the study

The primary aim of this study is to investigate the influence of sovereign large language models (LLMs) on the framing of diplomatic discourse and the generation of narratives related to international territorial disputes. In light of the growing role of AI in shaping public opinion and policy dialogues, this research seeks to systematically compare the outputs of prominent LLMs developed under different geopolitical contexts - namely, the United States (OpenAI's GPT), the European Union (Mistral), and China (DeepSeek-R1).

Specifically, the study has four core objectives:

- O1: to examine the role of these LLMs in shaping diplomatic narratives and fostering balanced perspectives in the context of territorial disagreements;
- O2: to assess the presence and extent of geopolitical biases across model responses to regional and provincial disputes;
- O3: to evaluate sentiment variability in AI-generated content across the selected models; and
- O4: to quantify the statistical divergences in outputs that may be attributable to the models' national origins and their alignment with intergovernmental policy orientations.

The remaining part is arranged as: the relevant literature is reviewed in Section 1, the research design, source of data collection and processes, data analysis, and ethical considerations are elucidated in section 2, the data analysis reports and the study findings are discussed in section 3, and finally, section 4 concludes the research with the limitations and future scope of the study.

1. Literature Survey

Sapkota et al. (2025) appraised the adherence to transparency standards of ChatGPT, DeepSeek, and other State-of-the-Art (SoTA) LLMs. By using open-source vs. open-weight models' perspectives, it also assessed transparency and accessibility. To evaluate the relation between the AI system and foundational concepts, licensing types, and transparency definitions, the study adopted a multi-stage research approach design. As per the study findings, the labelling on the models was unnecessary due to the full openness of the data source. Moreover, the investigation explained the critical distinction between the open weights and open source in the context of over 100 SoTA LLMs. But, some of the open-source models often didn't report the training data, codes, and key metrics.

Aydın et al. (2025) scrutinized the academic writing performance of Qwen 2.5 Max and DeepSeek v3 by analogizing them with ChatGPT, Gemini, Llama, Mistral, and Gemma. Next, 40 digital twin and healthcare articles' texts and abstracts were summarised by using generative AI tools. Also, by using the plagiarism tool, word count comparisons, AI detection tools, semantic similarity tools, along with readability assessments, the accuracy of the generated contents was examined.

Generally, the plagiarism rates were higher for the paraphrased abstract and lower for the answers created for the questions, but the present observation results showed the above acceptable levels of plagiarism reports.

Also, all chatbots created a satisfactory amount of content and high semantic overlap, while the readability test results expressed insufficient text readability. The study's drawbacks were measuring the model's performance over scanty data sets and the same scenarios

Chiarello et al. (2024) measured the applications and potential influence of generative LLMs in different business areas. 31,747 unique case study tasks, out of 3.8 million tweets, were identified and grouped by using a quantitative and granular data-driven approach. The study outcomes revealed the potential influence of LLMs and their implications in several spectrum of applications like programming assistance, and creative content generation, emphasizing the LLM's versatility in human resources, programming, social media, office automation, search engines, and education. Also, the study expressed certain limitations, such as the exclusion of qualitative analysis, reliance on a single data source, and non-establishment of empirical settings of a suitable dynamic landscape of generative models.

Li et al. (2024) evaluated the geopolitical bias in LLMs via territorial disputes, a naturally divisive and polyglot task. The case study introduced the BORDERLINES1 dataset to 251 territories, and the dataset consisted of multiple-choice questions on each country's language. It also used the metrics to quantify the bias and reliability of responses. The evaluation reports of models expressed the internal knowledge and use of metrics on the detection of contradiction in different language responses. Moreover, the results explored prompt modification strategies, variation in geopolitical bias, and response tailoring methods. Yet, the type of machine translation selection, un-ranking of GPT-4, and template-wise translations were the boundaries of the analysis.

Zhou & Zhang (2024) examined the political prejudices and contradictions in bilingual GPT models of the USA and China. To examine the cross-language biases of the selected countries, an analytical framework along with two dimensions was developed. As per the results, the administrative knowledge and approach of bilingual models was expressively higher in China than in the USA. Likewise, the Chinese model showed the least negativity toward China's problems, whereas the English model was seriously critical of China. Besides, due to country disparities, both LLM models influenced censorship and geopolitical tensions. The boundaries of the study are the selection of languages, less accuracy and transparency of languages, and unclear specific sources of China.

Guey et al. (2025) established the geopolitical bias across 11 conspicuous LLMs by investigating the bilingual (English and Chinese) as well as dual-framing (affirmative and reverse) responses to seven critical topics about US-China relations. Next, to detect the learning outputs, the study generated 19,712 prompts. Besides, based on the stance, impartiality, and rejection rates, the responses were quantitatively assessed and categorized. The results demonstrated significant and consistent ideological correlations with the LLMs' and geographic origins, and they expressed a noteworthy influence on the model responses. Also, the comprehensive metrics identified the variability and vulnerabilities in the model's behaviours. Some restrictions acknowledged were the selection of high-profile topics, prompting sensitivity and interpretation, and limited language selection.

Pacheco et al. (2025) examined the echo of geopolitical biases in US and Chinese LLMs. It also weighed the model's response to geopolitics and international relations questions. A geopolitical-based questionnaire set was framed, and ChatGPT and DeepSeek responses were collected. Also, the outputs were evaluated by both qualitative and quantitative analyses. To overcome the public discourse, the study highlighted the importance of AI-generated content, predominantly the politically sensitive contexts. The models' responses were composed, which indicated the sensitive topics without viewpoint presentations. Likewise, the findings showed notable biases in the selected LLMs model and echoed discrete conceptual views and ethnic impacts.

Urman & Makhortykh (2025) compared the guardrail-related political bias and false information prevalence in the outputs of the three LLM-centric chatbots of Russia. A systematic query was done, and responses from the ChatGPT, Google Bard (Gemini), and Bing Chat were extracted. Also, the performance was compared between the chatbots as well as between the languages. The results revealed significant disparities in politics-related information among the study chatbots, which dodged to respond to Russian prompts and restricted political information. Also, the reports showed a considerable difference across languages concerning Russian regime opponents regarding false claims. The factual error prevalence and unclear Russian regime censor information were the limitations of the study.

Torkamaan et al. (2024) discovered the multi-layered challenges and future necessity for a human-integrated approach of LLMs to ensure the socio-technical standards. The integration of LLMs was analysed by employing a structured multi-dimensional human-centred AI framework. The study recognized and promoted the human-machine collaboration principles, uninterrupted longitudinal studies, awareness campaigns, and regular inspections to develop LLMs. Moreover, it revealed the strategic integration between humans and technology. Also, the findings demonstrated the importance of a wide range of research perspectives and influenced the LLMs in socio-technical landscapes. Also, relying on large technology companies and rotating revenue into independent research were the biases of the study.

Rivera et al. (2024) scrutinized the escalatory risks and behaviour of multiple Language Models during military and diplomatic decision-making, precisely focusing on prejudice to take actions in intensified multilateral conflicts. To assess the action in different scenarios, the study designed a scoring framework. Likewise, both qualitative and quantitative evaluations were carried out on LLMs. As per the observations, all the LLMs showed the development and predicted escalation patterns. It also reported cognitive choice for actions and observed deterrence based on first-strike tactics. Yet, the study showed serious confines, such as developed arms-race dynamics and deployment of nuclear weapons.

Ahmed et al. (2025) defined the automated assessment of MARBLES in the evaluation of uses and biases in LLMs' ethical interactions within educational sets. The student chatbot interactions were analysed via a question-centric framework dataset. Next, to evaluate the rate of biased responses, the LLM was employed as a judge by multifaceted assessments. The findings proved that some AI models exhibited more biases and the LLMs evaluations offered significant and clear language knowledge. LLMs advanced the efficiency and flexibility of response, showing the importance of robust mechanisms in the assessment of bias. Nonetheless, the drawbacks of the study were the selection and handling of critical topics, training, and deployment procedures of the models.

Kharchenko et al. (2024) analysed the Hofstede cultural dimensions of the human population using the LLM responses. The study evaluated the cultures and languages of 36 countries. Moreover, it assessed the diverged LLMs' consistency using a series of Hofstede cultural dimensions advice, and then the requests were quantified. The analysis response expressed the differentiation of similar and differing cultural values among the selected countries based on cultural differences. The findings also revealed that the models adhere to precise values across different languages and reserve levels. Besides the objectives, the stereotypical study data and unawareness about the Hofstede cultural dimension prompt were the conflicts of the study.

Sugureddy (2023) described the applications of AI-driven solutions for robust data governance. Adopting a standardized model-centric data collection approach instead of a data-centric method enhanced the code architecture and data accuracy and also increased the data quality of the DCAI prompt. The results showed a significant upgrade in model performance and problem-solving skills. In particular, the AI system increased the forecast, decision-making, and other analogous activities by enhancing the dataset. Yet, the study expressed some boundaries. Initially, the implementation process is affected by the unpopularity of the model-based data-driven approaches. Then, the adeptness of the people is limited by the uninterrupted upgrading policies.

Francisco Castillo Eslava et al. (2023) appraised the potential impact of LLM on the recognition of territorial sovereignty and its legitimization of Crimea, West Bank, and Transnistria. By using a content analysis questionnaire created by ChatGPT, the study was carried out, and the information resolutions were issued by the General Assembly and the United Nations Security Council. The study reports proved that the impartial, objective, and perceptive flaws of technology tools like Google Maps and OpenAI's ChatGPT expressed the biases reflected by AI algorithms. Next, the results also proved the important role of AI in the creation of legitimacy and the acknowledgment of territorial sovereignty. Yet, the work opened up problems regarding the unanswered necessary questions, unfocused on many international conflicts and different languages.

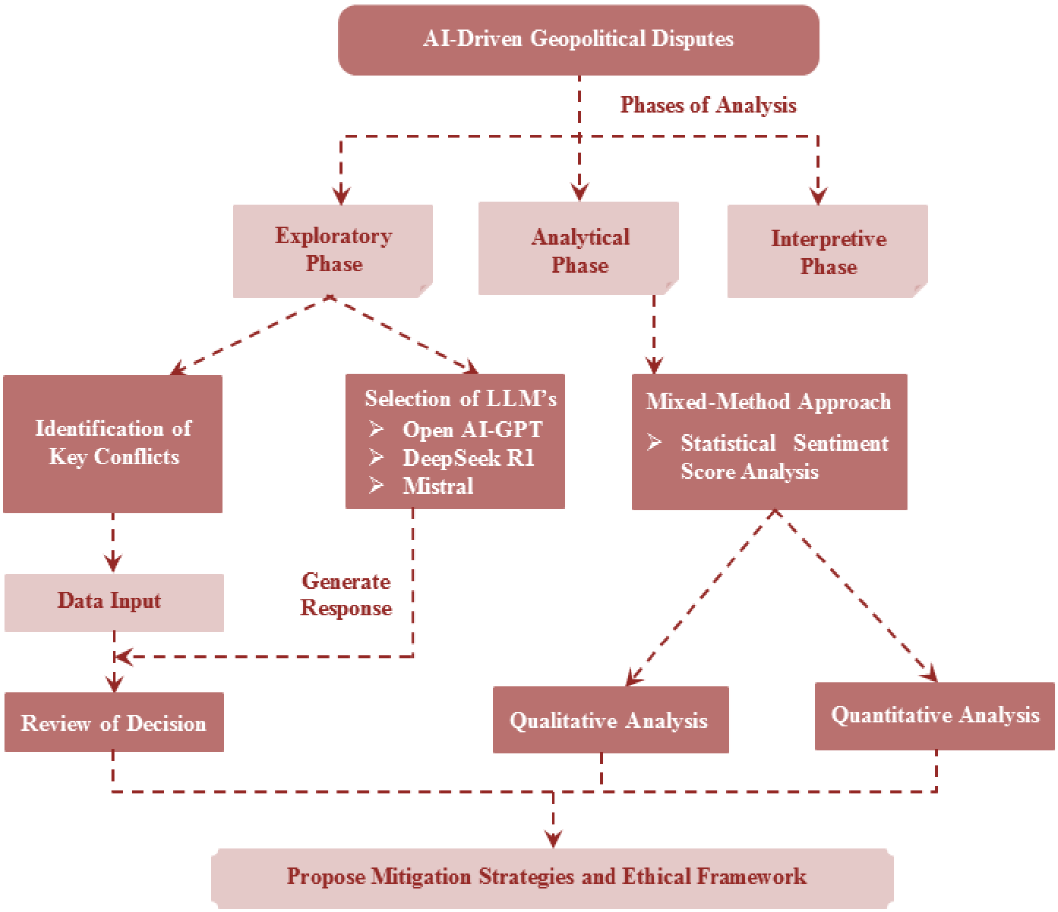
Tao et al. (2024) evaluated the cultural bias and cultural alignment of five widely used LLMs. The cultural partiality of 95 countries and territories was evaluated by models' responses to national descriptive questions by employing Integrated Values Surveys (IVS) data. The disaggregated replies proved that selected models (OpenAI's GPT-4o/4-turbo/4/3.5-turbo/3) resembled English-speaking and Protestant European countries' cultural values. The cultural stimulation strategy increased the cultural configuration by 71–81% in all territories. Yet, the limitations of the study were the selection of model parameters, people's authentic expressions, unsystematic response formatting, and lack of valid responses.

2. Research Methodology

2.1. Research Design

It is fundamental for generating reliable and valid findings and providing a systematic approach to planning, conducting, and analysing research studies. Also, it ensured the defined explanation of research questions, efficient data collection, and suitable analysis methods, which ultimately led to trustworthy inferences. By merging both qualitative and quantitative techniques, the present study adopted a mixed-methods approach to evaluate the biases in LLMs during territorial dispute discussions. The mixed methods enhanced the knowledge about the addressed problems and increased the strength of the research findings. Likewise, the crucial territorial disputes and LLMs identified, evaluated, and compared the sentiment bias outputs of different models and presented mitigation strategies and ethical outlines for SLLMs. In Figure 2, the study design and research phases are displayed.

Figure 2: Framework of the study design



Qualitative analysis in this study was based on open-ended AI-generated responses to carefully designed prompts, which provided a broad understanding of the underlying bias patterns. This approach enriched the study's insights by capturing the nuanced ways in which different models framed sensitive geopolitical topics. Quantitative analysis (closed data) was a set of methods that used mathematical and statistical modelling techniques to examine and understand the arithmetical data and identification of designs, leanings, and relationships among the observed data. It also helped to convert the raw data into metrics to eliminate assumptions and bias.

2.2. Source of Data and Data collection

Here, the main data was collected from the responses to input queries. It was extracted from the three selected LLM models, such as US-based OpenAI GPT, the EU's Mistral, and China's DeepSeek-R1. The articles, literature, books, research reports, newspapers, and case studies related to geopolitical disputes were used as the secondary data. The present research identified, collected, and evaluated the data through three phases: Identification of territorial disputes, selection of LLMs, and dataset compliance.

Identification of Territorial Disputes

Primarily, for the study, expanding 12 key territorial conflicts among different nations were selected. The identification of disputes was based on the geopolitical implication, historical complexity, and relevance of AI-driven narratives. Likewise, the disagreements reflected in academic literature, international policy forums, and diplomatic negotiations are interpreted using the AI models. Also, here, the strong political and economic issues of nations that impact global security and regional stability disputes were also chosen. These key disputes analysed the biases embedded in AI models, and the need for new sovereign AI frameworks was understood. In Table 1, the conflict sources are reported.

Table 1: Comparative analysis of territorial disputes & AI narratives with sources

Territorial Dispute	Key Conflict	Literature Insights on AI Narratives	Literature Sources
India-China (Arunachal Pradesh & Aksai Chin)	China claims Arunachal Pradesh: India asserts full sovereignty	Indian and Chinese AI models provide conflicting narratives on sovereignty.	Joshi et al. (2020); Bondarenko et al. (2025)
Israel-Palestine	Territorial control over Jerusalem, Gaza, and the West Bank	Western and Middle Eastern AI models diverge significantly in legal status.	Tao et al. (2024); Kulesz (2024)
Russia-Ukraine (Crimea & Donbas)	Russia annexed Crimea; Ukraine and global entities dispute legitimacy	Russian AI models emphasize historical ties; Western models cite occupation	Kulesz (2024); Castillo-Eslava et al. (2023)
Taiwan-China	China claims Taiwan as part of its territory; Taiwan operates independently.	Chinese AI models fully align with Beijing's claim; Western models acknowledge Taiwan's autonomy.	Bondarenko et al. (2025); Tao et al. (2024)
South China Sea	China, Vietnam, Philippines, and others contest maritime control	Chinese AI models assert exclusive rights; Southeast Asian models push for shared claims	Tao et al. (2024)
India-Pakistan-China (Kashmir)	India and Pakistan dispute Kashmir, with China holding part	Indian and Pakistani AI models offer highly polarized perspectives.	Joshi et al. (2020); Bondarenko et al. (2025)
Western Sahara (Morocco vs. Polisario Front)	Morocco claims sovereignty; Polisario Front seeks independence	Western AI models emphasize self-determination; Moroccan models stress sovereignty.	World Trade Organization (2024); GDPR (2018)

Territorial Dispute	Key Conflict	Literature Insights on AI Narratives	Literature Sources
Kuril Islands (Russia-Japan)	Russia controls the Kuril Islands; Japan claims sovereignty	Russian AI models justify post-WWII control; Japanese AI models highlight diplomatic disputes.	Joshi et al. (2020); Castillo-Eslava et al. (2023)
Cyprus (Turkey vs. Greece)	Greek-Cypriots vs. Turkish-Cypriots over control of Cyprus	Turkish-trained models align with Turkey's claim; EU-trained models support Greek Cypriot sovereignty.	Castillo-Eslava et al. (2023); Tao et al. (2024)
Falkland Islands (UK vs. Argentina)	Argentina claims the Falklands; the UK asserts continued sovereignty	UK-trained AI models reinforce historical claims; Argentine models push for decolonization.	Miao et al. (2021)
Guantanamo Bay (US vs. Cuba)	US controls Guantanamo under lease; Cuba disputes its validity	US models frame Guantanamo as a military necessity; Cuban AI models stress sovereignty violation.	Tao et al. (2024); Castillo-Eslava et al. (2023)
Puerto Rico (US Statehood vs. Independence)	The debate over Puerto Rico's status: statehood, independence, or US territory	US trained models highlight economic/military ties; Latin American-trained models discuss imperialism.	Tao et al. (2024); Bondarenko et al. (2025)

Selection of Large Language Models (LLMs)

The study compared the LLMs of different geopolitical regions, such as US-based OpenAI GPT, the EU's Mistral, and China's DeepSeek-R1 to evaluate the bias and background sensitivity about different provincial descriptions. The selected models represented a diverse set of training data, governance policies, and linguistic capabilities to analyse the bias and contextual sensitivity in territorial dispute descriptions. Also, these models were chosen not only for their regional provenance but for their relevance in framing policy-sensitive narratives, making them ideal candidates for studying geopolitical bias. In Table 2, the model preference details are shown.

Table 2: Selection reasons and literature sources of the study models

Models	Reasons	Source descriptions	References
OpenAI's GPT (GPT-4.0)	The models were widely used among the LLMs models due to their diverse dataset training and provided Western insights about territorial disputes and geopolitical narratives.	In AI research, the models were used as a standard source, which made them a suitable option for comparison with Sovereign LLMs.	Bondarenko et al. (2025); Castillo-Eslava et al. (2023)
DeepSeek (R1)	The models evolving in the AI market and surpassing the ChatGPT, Gemini, and Claude AI in performance became the most downloaded free app in the world.	It compared the same territorial dispute response of LLMs for different geopolitical regions and also presented a contrasting opinion to the Western LLMs.	Castillo-Eslava et al. (2023)
Mistral	Mistral reflected the European Union (EU) languages, cultures, and regulatory frameworks; it was also a part of France's broader strategy to endorse AI supremacy and reduce the dependence of foreign-developed models.	It represented a European perspective on Sovereign LLMs, offering a contrast to both Western and Asian models.	Bondarenko et al. (2025)

Assembly of dataset

An analytical framework was developed along with two extents to systematically explore the disparities in LLM language answers. The selected questions concerned the social and political problems criticized in official documents published by the nation's authorities, and the contradiction rate in political questions was considerably higher in the primary dimension.

The nature of political questions, the distinct political values embedded in the LLM's languages, and political censorships interfered with LLM to reduce the biases. Also, to control the influence of query outlining on answer consistency, the secondary dimension was used. Primarily, from the opinion-based questions, the factual questions were separated. These questions may lead to diverse levels of unpredictability. Logically, the present study reports were expected to receive opinion-centric questions related to a higher level of inconsistency. Likewise, exploratory inquiries might lead to more inconsistent answers than forced-choice queries. Here, a total of 300 queries about all the selected disputes were systematically organized, and by comprising 25 questions per dispute, each dataset was created. The models were asked to answer all the political questions to measure the inconsistency rate of LLMs' responses. All prompt queries were reviewed to minimize linguistic framing bias and ensure consistency in structure. Consistent with established prompt design guidelines declarative and open ended formats were used to reduce steering effects.

2.3. Data Analysis

Here, various steps like examining the collected data and interpretation of secondary source data are involved in the data analysis. The present study data were evaluated by using both qualitative and quantitative analysis methods. The variations in geopolitical framing and the potential of sovereign AI solutions in mitigating external biases were assessed by comparing the outputs of these models. The data gathering and data analysis were carried out together in the qualitative study. Primarily, the collected query answers were transcribed into comprehensive and reliable outcomes. Then, to find out the necessity of LLMs, the acquired background information and theoretical ideas were evaluated. To evaluate the mean, frequencies, and percentages of the observed data, quantitative analysis mainly was used. Here, to conduct the statistical calculations to draw the study conclusions, the data was converted in a numeric manner.

Sentiment scoring

To determine the emotions and evaluate the positive, negative, or neutral tone of the response, AI-generated responses were used. Also, to analyse the sentiment score, the VADER (Python’s NLTK library) was used to determine the model's support, opposition, or neutrality reply regarding the territorial dispute. This sentiment scoring acts as a proxy for directional bias, allowing us to infer whether each model aligns positively, negatively, or neutrally with specific territorial narratives.

Statistical analysis

To quantify the sentiment polarity divergence in the perspectives of 300 AI-generated responses, the current research used a structured assessment method. Next, to measure the response difference between the selected models, the Chi-square tests and Kruskal-Wallis tests were used. Also, by using Software Package Social Science-25 (SPSS-25), the correlations between the AI model origins and geopolitical leanings were also examined statistically. Through figures and plots, all the obtained data were visually represented.

In AI studies, ethical deliberations were complicated in upholding accountability, fairness, transparency, privacy, and data protection, which also raised trust, oversight, and sustainability.

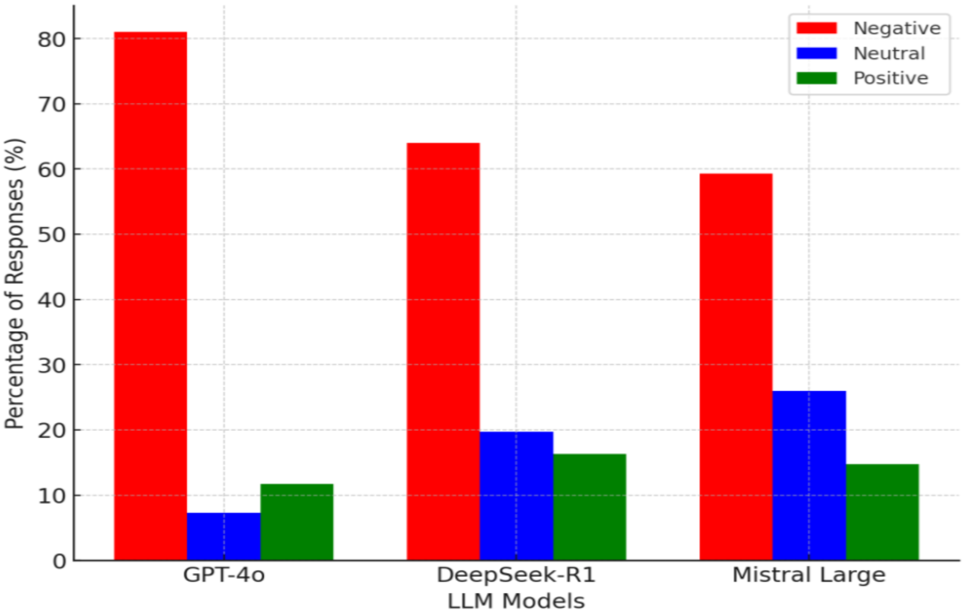
3. Result and Discussion

By using standard procedures, the findings regarding the geopolitical conflicts and mixed-method data outputs of selected LLMs were analysed. Here, to evaluate the territorial disputes, mixed-method was used by using both qualitative and quantitative methods. The sentiment score of the territorial disputes was assessed and compared within the selected LLMs in qualitative studies. The statistical divergence between the attained responses and the AI-driven models was evaluated in quantitative studies.

3.1. Assessment and Comparison of LLMs' Response Score towards Territorial Disputes

Here, 12 territorial dispute sets, such as India-China (Arunachal Pradesh & Aksai Chin), Israel-Palestine, Russia-Ukraine (Crimea & Donbas), Taiwan-China, South China Sea, India-Pakistan-China (Kashmir), Western Sahara (Morocco vs. Polisario Front), Kuril Islands (Russia-Japan), Cyprus (Turkey vs. Greece), Falkland Islands (UK vs. Argentina), Guantanamo Bay (US vs. Cuba), and Puerto Rico (US Statehood vs. Independence) were selected for the analysis. Primarily, three SLLMs, such as USA-based OpenAI GPT, China's DeepSeek-R1, and the EU's Mistral were selected as the study models. Next, 25 political queries per territory were created, and each model was prompted with the same 300 queries. Also, the responses were collected from all three LLMs separately, and the replies were categorized as negative, neutral, or positive based on the response tune, which was identified as sentiment tone. Still, due to their origin, culture, and training datasets, the response of each model showed variation in response scores.

Figure 3: Sentimental score of LLMs towards the territorial disputes



In Figure 3, the comparative analysis reports are illustrated. Here, all the study models expressed the response of input data in sentimental tones and were categorized as negative, neutral, and positive replies. Among the 300 responses of each model, the USA-based OpenAI GPT-4o showed the highest range of negative reactions (81%), followed by positive (12%) and neutral (7%). The China-based DeepSeek-R1 expressed 64% negative, 20% neutral, and 16% positive replies. Similarly, the EU-based Mistral also mostly exhibited negative responses (59%), followed by neutral (26%) and positive (16%) responses. Moreover, the findings proved that the USA-based models were dominated by the negative profile; but, the EU-based models balanced between the negative and positive sentimental line and merely managed to stand in the neutral position. Moreover, the China-based model's profile lined between the other two models; yet, it was largely in the negative phase. Also, it showed a remarkable number of neutral and positive answers. Thus, the visual comparison cleared that the negative sentiment of all models regarding the territorial disputes probably contributed to the comprehensive bias, especially when USA-based OpenAI GPT-4o's outputs were far more harmful than the other selected LLMs.

The present analysis confirmed that LLMs significantly affected the sentiment and responses regarding geopolitical queries due to their complex data station, dominance, and acquiescence with critical national AI governance. It suggested that the incorporation of varied and well-adjusted datasets of multiple views could reduce the twisted descriptions and raise unbiased AI responses. Also, clear and established regulatory frameworks provided strong data protection measures that defined the data governance standards for SLLMs.

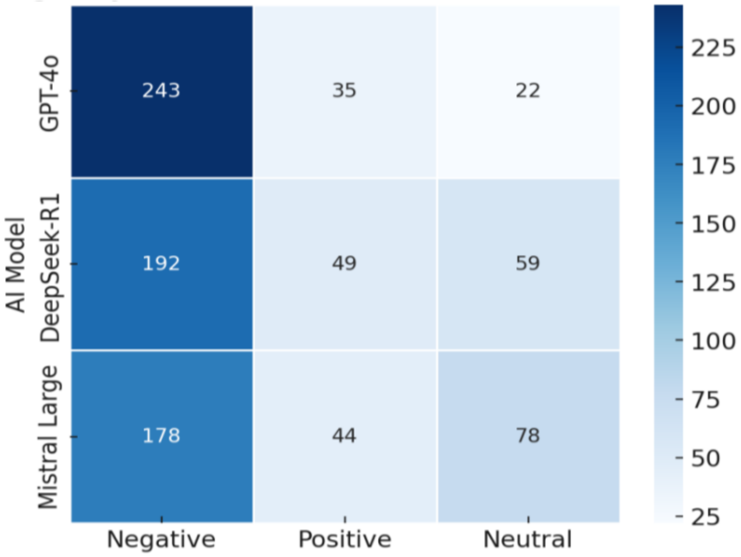
3.2. Statistical Analysis of Sentiment Scores

Here, the statistical analysis was divided into two subsets of analysis. Primarily, by using a chi-square test, the sentiment distribution differences among the models and the responses were analysed. Next, by using the expected and observed data, the descriptive statistical analysis was conducted to identify the statistical significance and variance among the responses as well as the selected study models.

Evaluation of Statistical Significance using Chi-Square Test

Here, to evaluate the sentiment distribution differences, a chi-square test was carried out among the models and responses. Here, USA-based OpenAI GPT, China's DeepSeek-R1, and the EU's Mistral were chosen as the models, and negative, neutral, and positive responses were used as the outputs. The test outcomes proved that the responses were highly dependable to the LLMs and also showed an extreme statistical significance ($p < 0.0001$) among the study models and its sentiment response, which discarded the null hypothesis. Moreover, it explained that the AI model selection drastically influenced the significance of the response. In Figure 4, the overall chi-square results are illustrated.

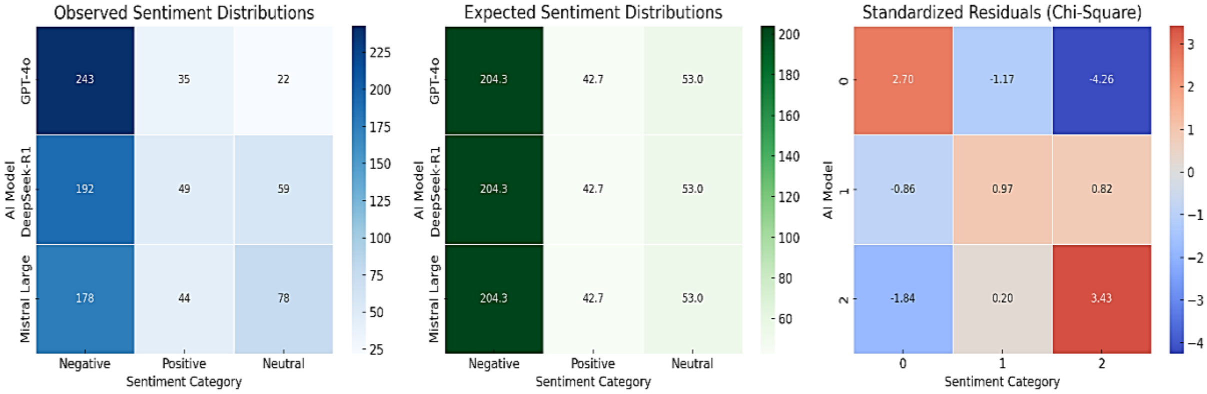
Figure 4: Sentiment distribution differences between models and responses



The analysis results showed a restrained variance in sentiment distributions (Cramér's V ≈ 0.16) against the given large number of responses (N=900 responses), and each model expressed different characteristic preferences but not in all aspects. Yet, the sentiment profile of the models exhibited varied significance with corresponding behaviour. Also, by using the sentiment table, standardized residual analyses were conducted on each model's cells to comprehend the specific sentiment category that mostly implicated the chi-square result. This examination indicated the difference between the observed cell values and excepted values by standard deviation measures. The residual values above ±1.96 were considered statistically significant ($p < 0.05$). The residual analysis reports are displayed in Figure 5.

The below-illustrated results showed a highly polarized pattern between GPT-4o and Mistral models. The GPT-4o's negative outputs were more highly residual than the expected values (+5.87), conspicuously far fewer in neutral responses (–5.75), and lower in positive tone (–1.55), which were parallel to the residuals of average models.

Figure 5: Category difference analysis



In the Mistral model, the positive count showed very close expected residual values (+0.27); but, an opposite trend was observed in negative responses with less residual values –4.00 and neutral responses with higher residuals +4.64. As per these results, the Mistral actively avoided the negative sentiments more than the other models and leaned towards impartiality. Yet, all residuals of DeepSeek-R1 expressed nearly 0 values (Negative –1.87, Neutral +1.11, and Positive +1.28) and also showed marginally fewer negatives and slightly more positives than expected values. As per the results, the distribution of DeepSeek was adjacent to the overall average of the three models without an extreme surplus or deficit. The overall chi-square results interpreted that GPT-4o's vigorous negativity and Mistral's additional neutrality were the key contributors to the overall association between the model and sentiment response. The differences in sentiment score across model results showed that the training data sources significantly altered the outputs; developing domain-specific datasets improved the neutrality and balanced AI outputs.

3.3. Sentiment Score Analysis

Then, the qualitative responses were quantified to analyze the sentiment intensity of each model's response. Here, to compute the average sentiment score of each model and also to analyze the statistical differences, the sentiment score was assigned as negative = –1, neutral = 0, and positive = +1. The standard deviation, variance, and mean values of each model are calculated, which are depicted in Table 3.

Table 3: Average dataset of sentiment scores

Models	Standard Deviation	Standard Variance	Mean Values
OpenAI GPT-4o	~0.668	~0.446	≈ –0.69
DeepSeek-R1	~0.759	~0.576	≈ –0.48
Mistral	~0.735	~0.540	≈ –0.45

As per the above-tabulated values, theGPT-4o was the most negative model with the lowest mean value (≈ -0.69), whereas the Mistral showed the least negative value (≈ –0.45), which often yielded a more balanced sentiment profile. The DeepSeek-R1 showed a moderate score (≈ –0.48) among the study models, where the variability expressed a mixed distribution of sentiment sorts. Moreover, the results proved that GPT-4o's expressed constantly negative impact due to lower variance; but, DeepSeek possessed higher variance, which indicated more inconsistent or context-dependent sentiment output of the model.

Statistical Score Variance Analysis

To detect each model's significant intensity difference in the sentiment scores, both the Analysis of Variance (ANOVA) and Tukey's Honestly Significant Difference (HSD) post-hoc tests were used. A one-way ANOVA was used on three-factor levels to evaluate three different sentiment score differences. Likewise, the different response intensities of each sentiment category of all the models were examined. In Table 4, the observed ANOVA results are shown.

Table 4: Sentiment score analysis using ANOVA

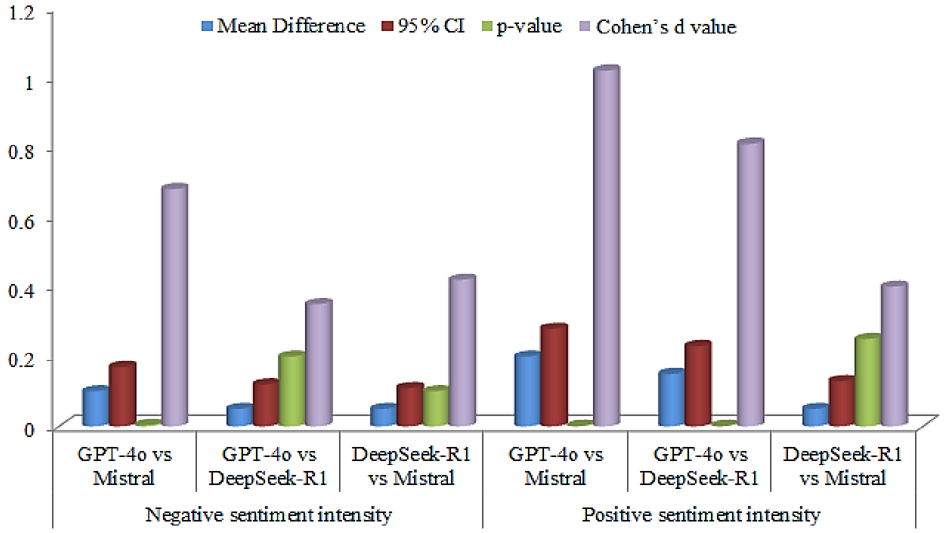
Sentiment Category	Model Effect F(2,147)	p-value	Partial η²	Effect Size
Negative	8.5	<0.001	0.10	Moderate
Neutral	1.9	0.15	0.03	Negligible
Positive	18.0	<0.001	0.20	Large

Note: **p*<0.01 (significant), Partial η² = proportion of variance explained by model differences. Large effect ≈ η²≥0.14.

Both negative and positive sentiment categories expressed an extreme statistical significance (<0.001); but, the neutral categories failed to show any significance toward the responses. The positive category stated a large effect size (F(2,147) = 18.0, partial η² ≈ 0.20). Moreover, the negative sentiment exhibited a moderate effect size (F(2,147) = 8.5, partial η² ≈ 0.10), whereas the neutral only possessed a mere effect (F (2,147) = 1.9, partial η² ≈ 0.03). The results summarised that the identity of the models significantly affected the sentiment intensity in both negative and positive frameworks but not in neutral contexts. Also, the positive sentiment showed a 20% distinct variance and attributed practical effects in terms of the model and its answer.

Then, by using Tukey's HSD post-hoc tests, pairwise significant sentiment score differences of each model were evaluated. These analyses were conducted on the positive and negative scores; but, the neutral scores were excluded from the analysis due to non-significance. In Figure 6, the pairwise statistical differences of the LLMs are disclosed.

Figure 6: Comparative sentiment intensity analysis of selected models



A significant difference was found between GPT-4o and Mistral (p = 0.004) in negative sentiment intensity, and the GPT-4o showed a stronger negative response than DeepSeek-R1 and Mistral's response. In particular, the GPT-4o expressed about 0.10 units mean negativity, a moderate effect (Cohen's d ≈ 0.7), and a 95% confidence interval (CI) that showed the true value of the models, which fell between [0.03, 0.17]. But, the GPT-4o vs DeepSeek-R1 and DeepSeek-R1 vs Mistral pairs expressed moderate negative response (0.05 units), mild effect (Cohen's d ≈ 0.4), and the 95% CI that showed the true value of the models, which fell between [-0.02, 0.12] in GPT-4o vs DeepSeek-R1 model and [-0.01, 0.11] in DeepSeek-R1 vs Mistral. However, both models failed to exhibit any statistical significance (p = 0.20 and p = 0.10, respectively).

GPT-4o strongly benchmarked the responses than the other models while in positive sentiment intensity. It also showed different levels of significance among both DeepSeek-R1 and Mistral pairs. The GPT-4o expressed 20 units of mean difference and a $p < 0.001$ value. It also exhibited a large effect (Cohen's d ≈ 1.0) and a true value between [-0.02, 0.12] against the Mistral's pairing. It showed variation in significance (p = 0.001), mean difference (15 units), and 95% CI range [0.07, 0.23] against the DeepSeek-R1 model but expressed a large effect (Cohen's d ≈ 0.81). But, in contrast, DeepSeek-R1 vs Mistral showed a lower mean difference (0.05 units), moderate effect (Cohen's d ≈ 0.4), and the 95% CI that showed the true value of the models, which fell between [-0.03, 0.13]; but, significant difference (p = 0.25) was not found among the pair. The results didn't indicate the reliability of these two models due to the related intensity. Therefore, the post-hoc findings proved the statistical differences between the positive and negative sentiment intensity of LLMs. Especially, the USA-based OpenAI GPT-4o model expressed a strong tone in both negative and positive responses; but, the Mistral and DeepSeek-R1 showed a moderate tone.

Therefore, the research strongly suggested that LLMs from different origins exhibited distinct biases and perspectives, which influenced the sentiment and outline of AI-generated content. Moreover, it highlighted the necessity of the model's origin while selecting an LLM for specific contexts. Also, the findings supported the development of SLLMs, allowing regions to create AI systems that aligned with the cultural values of the area of study. This research further informed the stakeholders and researchers about the evident impact of model origin, promoting more ethical and effective use of LLMs in various contexts. To address the challenges and opportunities allied with the bias in SLLMs, researchers explored diverse data sources to detect developing bias and mitigation techniques and implemented rigorous testing procedures to identify and minimize biased outputs. By tackling these issues, SLLMs could be developed to promote fairness, equity, and responsible AI development to unlock opportunities for innovation across various sectors.

3.4. Technical Factors Underlying Geopolitical Bias in LLMs

The fact that sovereign LLMs produce varied response patterns with geopolitical biases necessitates investigation into their underlying technical determinants. Beyond quantifying these output differences, this study prompts reflection on bias origins within LLM design and training. Specifically, the composition of training data and decisions made during model architecture or fine-tuning are highly relevant

Influence of Training Data Composition

At the core of every LLM is the data it learns from. Training data for these models often includes a broad range of sources - news articles, encyclopaedic content, government publications, online discussions, and other publicly available text. The geographic and cultural origin of this data can significantly affect how certain topics, particularly those related to territorial disputes or national identity, are framed (Chen et al., 2025; Djuhera et al., 2025).

For instance, a model trained primarily on data from media sources based in a specific country may reflect the dominant perspectives and terminologies used within that region. This doesn't necessarily indicate intentional bias but mirrors the information ecosystem in which the model was developed (Chen et al., 2025). In cases of geopolitical disagreement, where narratives are often contested, the way historical events, legal claims, and territorial boundaries are represented in the training data subtly shapes the model's language generation behaviour (Lin et al., 2024). Furthermore, disparities in data availability can lead to asymmetric representation across linguistic or regional contexts. English-language models, for example, may disproportionately reflect viewpoints prevalent in Western publications due to the greater volume and accessibility of such digital content (Djuhera et al., 2025)

Post Training Influences: Tuning and Safety Mechanisms

Beyond the initial pre-training phase, large language models (LLMs) are often refined through instruction tuning and reinforcement learning from human feedback (RLHF) to improve their usefulness, safety, and alignment with human expectations (Ouyang et al., 2022). However, the instructional data and the judgments of human annotators involved in this process can unintentionally introduce cultural biases and framing effects, as these reviewers bring their own norms and assumptions to the tuning process (Wei et al., 2023).In addition to tuning, models are often embedded with safety mechanisms or guardrails that determine how they respond to sensitive or controversial topics. These mechanisms - such as response filtering, content prioritization, or suppression - are frequently shaped by the legal, ethical, and cultural values of the regions in which the models are developed or deployed (Bommasani et al., 2021). As a result, even for identical input prompts, different LLMs may produce divergent outputs. These differences are not necessarily due to flaws or manipulation but reflect the underlying assumptions and societal values embedded in the models' development processes (Ouyang et al., 2022; Wei et al., 2023).

Conclusion

This study's comparative analysis of biases in sovereign LLMs OpenAI's GPT-4o, DeepSeek-R1, and Mistral demonstrates how AI generated content frequently mirrors and sometimes amplifies national narratives. Each model showed distinct leanings tied to its origins across a dozen territorial disputes. DeepSeek-R1, for instance, echoed official Chinese positions in 114 out of 125 China-related queries, a pattern that could suggest alignment with state priorities. GPT-4o, by contrast, often balanced responses, though one might question whether its "neutrality" masks Western-centric calibration. Meanwhile, mistral's open-source framework revealed subtler biases, perhaps an artifact of its training data's invisible hand. These divergences highlight a paradox: sovereign AI development, while empowering digital autonomy, risks hardening partisan viewpoints into seemingly objective outputs.

These findings raise profound implications. Users in separate jurisdictions might as well inhabit different worlds when LLMs frame the same conflict differently. This epistemic fragmentation, communities isolated within their own "fact" ecosystems, could erode the very idea of shared truth. Take a model that dismisses historical consensus or legal boundaries: its outputs may gradually normalize alternative realities, chipping away at international norms. An important gap remains here: how do these algorithmic echo chambers affect diplomacy or public trust in institutions? Without answers, we risk a future where AI doesn't just mirror divisions but actively deepens them.

The double-edged nature of sovereign AI is evident and becoming clearer here. On one hand, nation-specific models can empower cultural inclusion and technological self-reliance; think GDPR-driven privacy safeguards or local language preservation. Hitherto, the flipside looms large: a fragmented global AI ecosystem split into US-aligned, China-aligned; or EU-centric blocs. Recent warnings about "technological balkanization" should give pause. Wealthier states, able to fund cutting-edge models, might cement information dominance, while others lag. This limitation should be acknowledged: without guardrails, sovereign AI could calcify - not bridge global asymmetries. Beyond theoretical concerns, unchecked biases in LLMs may propagate into operational systems such as automated journalism, educational AI, and policy advisory tools. In these contexts, partial outputs could institutionalize partisan narratives, thereby influencing public discourse and policy frameworks. The integration of sovereign LLMs without explicit bias mitigation compounds the risk of legitimizing selective truths.

While nascent, governance efforts offer glimmers of hope. The EU's AI Act, with its bias-prevention mandates, could set a template, and UNESCO's ethical guidelines might curb worst-case scenarios. Nevertheless, current frameworks are patchwork. A rights-based EU approach clashes with state-controlled models elsewhere, and it's possible that uncoordinated policies will fail to address cross-border harms. One could argue that without ethical "air traffic control" for AI, we're inviting a crisis of incompatible truths.

Further, recommendations must walk a tightrope between sovereignty and solidarity:

- Mandate disclosures of training data and biases, not unlike the EU's content flagging, but couple this with culturally attuned audits. A model's "neutrality" in Brussels might read as bias in Jakarta.
- Diversify datasets, yes, but also confront the uncomfortable question: Can any “sovereign” model truly escape its creators’ blind spots?
- Push for interoperable standards, even if a global treaty feels distant. Start small: bilateral pacts on bias research, or shared datasets for contested histories.
- User Empowerment: Teach digital literacy but also design tools that nudge users toward diverse sources, a vaccine against algorithmic determinism.
- Creation of regional ethics review boards (under e.g. US, EU, ASEAN) to formally audit sovereign LLM’s before deploying into public facing domains
- Mandate the policy that all sovereign LLM’s publish standardized model cards including bias benchmarks, training data, sentiment score distributions on politically sensitive topics.
- Real time bias monitoring and escalation systems that can flag and suspend content generation related to high-risk geo-political crisis.

Forward-Looking Technical Strategies for Bias Mitigation

To advance beyond the documentation of bias and toward its substantive mitigation, integrating bias-sensitive mechanisms throughout the Large Language Model (LLM) development pipeline is imperative. Promising methodologies include *bias-aware fine-tuning*, which incorporates multi-perspective geopolitical exemplars (e.g., competing territorial narratives) and employs loss functions penalizing ideological over alignment; *adversarial training for neutrality*, leveraging challenge prompts to probe partiality and iteratively reinforce balanced responses; and *counterfactual narrative generation*, synthetically inducing divergent cultural or national perspectives (e.g., "How would a historian from Country X analyse this?") to diversify training data and user outputs. Complementing these, *bias logging and gradient tracing tools* enable architectural diagnostics by mapping biased outputs to specific latent representations, facilitating targeted interventions.

Collectively, these techniques - applied synergistically during training and post-deployment monitoring - enhance model resilience against geopolitical framing biases. By systematically operationalizing multi-perspective representation, adversarial robustness, synthetic viewpoint generation, and interpretable bias tracing, developers can cultivate LLMs capable of navigating societally consequential domains with greater neutrality, accountability, and sociotechnical responsibility, thereby reducing the propagation of hegemonic narratives.

In essence, sovereign LLMs operate not merely as technical systems but as epistemic agents, mediating foundational truths. While they advance national values, their divergent representations of geopolitical events risk institutionalizing incompatible worldviews - particularly in domains like peacebuilding or international law. Mitigating these risks demands technical safeguards, transparency, and transnational collaboration to prevent algorithmic polarization from solidifying geopolitical conflict. Sovereign LLMs force a reckoning: Left unchecked, they threaten an “epistemic Babel.” Yet with care, they might become bridges, not walls. The path forward demands humility—recognizing that every model carries the fingerprints of its makers.

Limitations

This study offers valuable insights into the geopolitical biases of SLLMs, but its contributions come with caveats. The analysis focuses exclusively on three prominent models, GPT-4o, DeepSeek-R1, and Mistral, leaving other sovereign or regional LLMs unexamined while robust in design. This scope restriction might be argued as overlooking culturally specific perspectives that might diversify findings. The reliance on English-language prompts introduces another layer of bias, arguably oversimplifying how these models operate in multilingual contexts. An important gap remains here: without testing non-English inputs, the study inadvertently mirrors the linguistic limitations it critiques. Future research should incorporate native language prompts across selected conflicts to reveal potential shifts in sentiment distribution or national alignment.

Prompt framing effects further complicate interpretations as even minor phrase variations can tilt model outputs significantly, a fragility the methodology acknowledges but cannot fully neutralize. This limitation should be acknowledged in future iterations. Now, the findings capture only a static snapshot of models evolving at breakneck speed. How might next-month's updates alter their geopolitical leanings? The proprietary nature of these systems compounds the challenge; training data and architectural choices stay shrouded, making bias tracing akin to reverse-engineering shadows.

Another important limitation relates to the technical opacity of many leading LLMs, particularly those developed by private-sector entities. The lack of public access to training datasets, model weights, and internal tuning parameters significantly constrains both the scientific understanding of model behaviour and the development of generalized mitigation techniques. This opacity also limits collaborative progress: researchers are unable to test de-biasing strategies at scale, audit models transparently, or contribute to shared knowledge bases around geopolitical sensitivities. Open-source LLMs, by contrast, offer a valuable pathway toward reproducibility, cross-cultural bias analysis, and ethical experimentation. Without improved transparency across the ecosystem, efforts to ensure responsible and inclusive AI development will remain uneven and fragmented.

Though methodologically tidy, sentiment analysis may oversimplify nuanced stances. A model's terse response could reflect coding constraints as much as ideological bias, an ambiguity the study flags but cannot resolve. It is possible that real-world deployment scenarios would magnify these issues, yet their practical impact on users or public discourse remains untested. Equally absent is an exploration of how adversarial inputs might weaponize the observed biases, a vulnerability growing more urgent as LLMs permeate information ecosystems.

Particularly, the research stops short of evaluating regulatory frameworks empirically while ethical governance is debated conceptually. Also, cultural framing is inferred indirectly through outputs, leaving internal decision-making pathways opaque. Could repeated queries reveal temporal inconsistencies? The study doesn't say. User feedback loops, another dynamic force shaping model behaviour, also escape scrutiny.

Lastly, these constraints like technological opacity, linguistic narrowness, and frozen-in-time data, don't negate the study's value but clarify its boundaries. In the future, the work will embrace messier multilingual realities and the fluidity of both models and their socio-political contexts.

Credit Authorship Contribution Statement:

Jonnala, S. was responsible for the conceptualization of the study, development of the research methodology, formal analysis of the data, drafting the original manuscript, and supervising the overall research process. Swamy, B. contributed to data curation, implementation of software tools, and execution of the investigation. He also supported visualization of the results and contributed to the review and editing of the manuscript. Thomas, N.M. conducted the literature review, assisted in validating the findings, contributed to the interpretation of the results. All authors have read and approved the final version of the manuscript.

Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

Ahmed, I., Liu, W., Roscoe, R. D., Reilley, E., & McNamara, D. S. (2025). Multifaceted assessment of responsible use and bias in language models for education. *Computers, 14*(3), 1–12. https://doi.org/10.3390/computers14030100

Aydın, O., Karaarslan, E., Erenay, F. S., & Dzakula, N. B. (2025). *Generative AI in academic writing: A comparison of DeepSeek, Qwen, ChatGPT, Gemini, Llama, Mistral, and Gemma*. arXiv. https://doi.org/10.48550/arXiv.2503.04765

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). https://doi.org/10.1145/3442188.3445922

Bondarenko, M., Lushnei, S., Paniv, Y., Molchanovsky, O., Romanyshyn, M., Filipchuk, Y., & Kiulian, A. (2025). *Sovereign large language models: Advantages, strategy and regulations*. arXiv. https://doi.org/10.48550/arXiv.2503.04745

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). *On the opportunities and risks of foundation models*. arXiv. https://doi.org/10.48550/arXiv.2108.07258

Brown, S., Belliappa, G., & Ng, D. P. L. (2024). *Revealing the path forward with sovereign LLMs*. Deloitte. https://www2.deloitte.com/content/dam/Deloitte/us/Documents/consulting/us-nvidia-revealing-the-path-forward-with-sovereign-llms.pdf

Castillo- Eslava, F., Mougan, C., Romero-Reche, A., & Staab, S. (2023). *The role of large language models in the recognition of territorial sovereignty: An analysis of the construction of legitimacy*. arXiv. https://doi.org/10.48550/arXiv.2304.06030

Chen, Y., Zhang, L., & Wang, Q. (2025). Addressing asymmetry in contrastive learning: LLM-driven sentence embeddings with ranking and label smoothing. *Symmetry, 17*(5), 646. https://doi.org/10.3390/sym17050646

Chiarello, F., Giordano, V., Spada, I., Barandoni, S., & Fantoni, G. (2024). Future applications of generative large language models: A data-driven case study on ChatGPT. *Technovation, 133*, 1–12. https://doi.org/10.1016/j.technovation.2024.103002

Choi, W. C., & Chang, C. I. (2025). *Advantages and limitations of open-source versus commercial large language models (LLMs): A comparative study of DeepSeek and OpenAI's ChatGPT*. https://doi.org/10.20944/preprints202503.1081.v1

Dandage, R. V. (2025). A comparative analysis of ChatGPT and DeepSeek: Capabilities, applications, and future directions. *International Journal of Innovative Science and Research Technology, 10*(2), 207–211. https://doi.org/10.5281/zenodo.14899162

Djuhera, A., Müller, T., Li, S., Gupta, R., Schmidt, L., & Agarwal, A. (2025). *Fixing it in post: A comparative study of LLM post-training data quality and model performance*. arXiv. https://doi.org/10.48550/arXiv.2506.06522

Floridi, L., Cowls, J., Beltrametti, M., et al. (2018). AI4People - An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines, 28*, 689–707.

GDPR. (2018). *General Data Protection Regulation: ICAP GDPR Survey* (pp. 1–30). https://in2mobile.gr/wp-content/uploads/2018/05/GDPR_NEWSurvey.pdf

Guey, W., Bougault, P., Zhang, W., de Moura, V. D., & Gomes, J. O. (2025). *Mapping geopolitical bias in 11 large language models: A bilingual, dual-framing analysis of U.S.–China tensions*. arXiv. https://doi.org/10.48550/arXiv.2503.23688

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 6282–6293). https://aclanthology.org/2020.acl-main.560.pdf

Kharchenko, J., Roosta, T., Chadha, A., & Shah, C. (2024). *How well do LLMs represent values across cultures? Empirical analysis of LLM responses based on Hofstede cultural dimensions*. arXiv. https://doi.org/10.48550/arXiv.2406.14805

Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons, 62*, 15–25. https://doi.org/10.1016/j.bushor.2018.08.004

Kulesz, O. (2024). *Artificial intelligence and international cultural relations: Challenges and opportunities for cross-sector collaboration*. Culture and Foreign Policy. https://cultureactioneurope.org/wp-content/uploads/2024/08/ifa-2024_kulesz_ai-intl-cultural-relations.pdf

Li, B., Haider, S., & Callison-Burch, C. (2024). This land is {your, my} land: Evaluating geopolitical bias in language models through territorial disputes. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 3855–3871). https://aclanthology.org/2024.naacl-long.213.pdf

Lin, H., Long, J., Xu, Z., & Zhao, W. (2024). Token-wise influential training data retrieval for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (pp. 841–860). https://doi.org/10.18653/v1/2024.acl-long.48

Linkon, A. A., Shaima, M., Sarker, M. S. U., Badruddowza, N., Nabi, N., Rana, M. N. U., Ghosh, S. K., Rahman, M. A., Esa, H., & Chowdhury, F. R. (2024). Advancements and applications of generative artificial intelligence and large language models on business management: A comprehensive review. *Journal of Computer Science and Technology Studies, 6*(1), 225–232. https://doi.org/10.32996/jcsts.2024.6.1.26

Linegar, M., Kocielnik, R., & Alvarez, M. R. (2023). Large language models and political science. *Frontiers in Political Science, 5*, 1–12. https://doi.org/10.3389/fpos.2023.1257092

Maiti, A., Adewumi, S., Tikure, T. A., Wang, Z., Sengupta, N., Sukhanova, A., & Jana, A. (2025). *Comparative analysis of OpenAI GPT-4o and DeepSeek R1 for scientific text categorization using prompt engineering*. arXiv. https://doi.org/10.48550/arXiv.2503.02032

Miao, F., et al. (2021). *AI and education: Guidance for policy-makers*. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000376709

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). *Model cards for model reporting*. arXiv. https://arxiv.org/pdf/1810.03993

Pacheco, A. G. C., Cavalini, A., & Comarela, G. (2025). *Echoes of power: Investigating geopolitical bias in US and China large language models*. arXiv. https://doi.org/10.48550/arXiv.2503.16679

Pahune, S., Akhtar, Z., Mandapati, V., & Siddique, K. (2025). *The importance of AI data governance in large language models*. Preprints. https://www.preprints.org/manuscript/202504.0219/v1

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Leike, J. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems, 35*, 27730–27744. https://doi.org/10.48550/arXiv.2203.02155

Raza, M., Jahangir, Z., Riaz, M. B., Saeed, M. J., & Sattar, M. A. (2025). Industrial applications of large language models. *Scientific Reports, 15*(1), 1–23. https://doi.org/10.1038/s41598-025-98483-1

Rivera, J.-P., Mukobi, G., Reuel, A., Lamparth, M., Smith, C., & Schneider, J. (2024). Escalation risks from language models in military and diplomatic decision-making. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 836–898). https://dl.acm.org/doi/pdf/10.1145/3630106.3658942

Sapkota, R., Raza, S., & Karkee, M. (2025). *Comprehensive analysis of transparency and accessibility of ChatGPT, DeepSeek, and other SoTA large language models*. arXiv. https://doi.org/10.48550/arXiv.2502.18505

Sugureddy, A. R. (2023). AI-driven solutions for robust data governance: A focus on generative AI applications. *International Journal of Data Analytics, 3*(1), 79–89. https://iaeme.com/Home/article_id/IJDA_03_01_007

Taddeo, M., & Floridi, L. (2018). How AI can be a force for good: An ethical framework will help to harness the potential of AI while keeping humans in control. *Science, 361*(6404), 751–753. https://doi.org/10.1126/science.aat5991

Tao, Y., Viberg, O., Baker, R. S., & Kizilcec, R. F. (2024). *Cultural bias and cultural alignment of large language models*. arXiv. https://doi.org/10.48550/arXiv.2311.14096

Torkamaan, H., Steinert, S., Pera, M. S., et al. (2024). Challenges and future directions for integration of large language models into socio-technical systems. *Behaviour & Information Technology*, 1–21. https://doi.org/10.1080/0144929X.2024.2431068

Urman, A., & Makhortykh, M. (2025). The silence of the LLMs: Cross-lingual analysis of guardrail-related political bias and false information prevalence in ChatGPT, Google Bard (Gemini), and Bing Chat. *Telematics and Informatics, 96*, 1-16. https://doi.org/10.1016/j.tele.2024.102211

Usman, H., Nawaz, B., & Naseer, S. (2023). The future of state sovereignty in the age of artificial intelligence. *Journal of Law & Social Studies, 5*(2), 142–152. https://doi.org/10.52279/jlss.05.02.142152

Zhou, D., & Zhang, Y. (2024). Political biases and inconsistencies in bilingual GPT models—the cases of the U.S. and China. *Scientific Reports, 14*(1), 1–13. https:/doi.org/10.1038/s41598-024-76395-w

World Trade Organization. (2024). *Trade and inclusiveness: How to make trade work for all* (pp. 1–160). https://www.wto.org/english/res_e/publications_e/wtr24_e.htm

Wei, J., Borgeaud, S., Chan, H., Kádár, Á., Mohan, M., Sastry, G., ... & Le, Q. (2023). *Instruction-tuned language models exhibit emergent cognitive biases*. arXiv. https://doi.org/10.48550/arXiv.2308.00225

Cite this article

Jonnala, S., Swamy, B., & Thomas, N. M. (2025). Geopolitical bias in Sovereign Large Language Models: A comparative mixed-methods study. *Journal of Research, Innovation and Technologies*, Volume IV, 2(8), 173-192. https://doi.org/10.57017/jorit.v4.2(8).04

Article's history:

Received 1st of June, 2025; Revised 21st of June, 2025;
Accepted for publication 27th of June, 2025; Available online: 30th of June, 2025
Published as article in Volume IV, Issue 2(8), 2025