Political Bias and Temporal Dynamics in Large Language Models

Abderrahmane Moujar[1], Hasti Hosseinpour[1], Oloruntobi Olutola[1], Ran Lu[1], Adrian Popescu[1]

[1]Université Paris-Saclay

Abstract

Large Language Models (LLMs) are increasingly deployed as intermediaries for accessing, interpreting, and evaluating political information. Prior work has demonstrated that LLMs exhibit systematic political biases, reflected in ideological alignment, sentiment toward political actors, and framing choices. However, most existing studies treat political bias as a static property, evaluating models at a single point in time. This assumption overlooks the rapid evolution of LLMs, which undergo frequent updates to training data, safety mechanisms, and alignment procedures.

In this work, we investigate political bias as a temporal phenomenon, analyzing how political attitudes expressed by LLMs evolve across model generations. Focusing on longitudinal comparisons within model families, we examine how changes in safety data distributions and alignment interventions correspond to shifts in political stance and sentiment. Our study integrates content-based and sentiment-based bias metrics to provide a dynamic perspective on value alignment in large language models.

1 Introduction

Large Language Models (LLMs) have become central intermediaries in how individuals access, interpret, and evaluate political information. Their widespread use in tasks such as summarization, explanation, sentiment analysis, and question answering grants them substantial influence over contemporary digital public spheres. As these systems are increasingly embedded in search engines, social media platforms, and decision-support tools, their latent political orientations carry meaningful societal and political consequences.

A growing body of evidence suggests that LLMs are not ideologically neutral. Instead, political tendencies emerge from architectural decisions, training data composition, and post-training alignment strategies (Santurkar et al., 2023; Buyl et al., 2026). When models systematically favor or disfavor particular political ideologies, policies, or actors, they may function as implicit instruments of digital soft power, especially when deployed at scale across languages and geopolitical contexts.

Empirical research further shows that political bias in LLMs is not limited to isolated prompts but constitutes a structural property of model behavior. Studies have identified consistent ideological patterns across economic and social dimensions, as well as systematic sentiment differences toward political entities (Hartmann et al., 2023; Bang et al., 2024). Moreover, LLMs may exhibit bias amplification, whereby biases present in training data are reproduced or intensified in generated outputs, reinforcing existing societal asymmetries (Feng et al., 2023; Zhu et al., 2024).

Despite this progress, most prior work evaluates political bias under a static assumption, treating models as fixed artifacts. This perspective is increasingly inadequate given the rapid pace of LLM development. Model families such as LLaMA undergo substantial changes across generations, including modifications to training corpora, safety data distributions, and reinforcement learning–based alignment procedures (Touvron et al., 2023; Dubey et al., 2024). These interventions plausibly affect not only factual reliability and safety but also political framing and sentiment.

1.1 Research Questions and Contributions

To address this gap, we study political bias through a temporal and longitudinal lens. Specifically, this paper addresses the following research questions:

RQ1: How does political bias in LLM outputs change across successive model generations?

RQ2: To what extent do safety alignment and red-teaming interventions correspond to shifts in political stance and sentiment?

RQ3: Are temporal changes in political bias

uniform across ideological dimensions and political entities, or do they exhibit asymmetric patterns?

Our contributions are threefold:

- We conceptualize political bias in LLMs as a dynamic property, rather than a static characteristic.
- We provide a longitudinal analysis of political bias across model generations using complementary stance- and sentiment-based metrics.
- We empirically link observed bias shifts to documented alignment and safety interventions, contributing to a deeper understanding of value alignment over time.

2 Related Work

Research on political bias in large language models spans ideological measurement, sentiment analysis, temporal evaluation, and mitigation. We review the most relevant work along three axes.

2.1 Measuring Political Bias in Language Models

Early studies mapped LLM outputs onto established ideological frameworks, demonstrating that conversational models express coherent political orientations. Some argue that LLMs reflect the aggregate opinions embedded in their training data (Santurkar et al., 2023), while others situate ChatGPT along a left-libertarian axis using political compass evaluations (Hartmann et al., 2023).

Subsequent work refined bias measurement by distinguishing between content and style. A two-pronged framework analyzing both explicit stance ("what is said") and framing or lexical polarity ("how it is said") reveals subtle partisan effects even in ostensibly neutral responses (Bang et al., 2024). Target-oriented sentiment classification further operationalizes political bias by measuring sentiment toward specific political actors or parties (Wang et al., 2025; Chen et al., 2024).

2.2 Temporal Dynamics and Model Evolution

Only recently has attention shifted toward the temporal dimension of LLM behavior. Research shows that LLMs exhibit temporal generalization failures, with performance and biases drifting as political and informational contexts evolve (Zhu et al., 2024). Models implicitly encode the ideological preferences of their creators, suggesting that updates to training and alignment pipelines may systematically reshape political outputs (Buyl et al., 2026).

Technical reports on the LLaMA family provide concrete evidence of such evolution. Between LLaMA 1, 2, and 3, Meta introduced substantial changes to safety data distributions, reinforcement learning objectives, and red-teaming protocols (Touvron et al., 2023; Dubey et al., 2024). Comparative evaluations indicate that newer models handle politically nuanced and ambiguous content differently, particularly in borderline factual or normative cases (Crum et al., 2024).

2.3 Mitigation and Alignment Interventions

A parallel line of work investigates methods for mitigating political bias. Reinforced calibration techniques have been proposed to reduce ideological skew (Liu et al., 2021), while other research demonstrates that political stance is encoded in internal model activations and can be selectively steered (Banko et al., 2025). These findings suggest that political bias is not merely emergent but amenable to targeted intervention.

However, mitigation introduces trade-offs. Alignment and safety guardrails may reduce expressive or argumentative capacity, potentially altering political nuance (Bonaldi et al., 2024). The LLaMA 3 report explicitly frames safety alignment and red-teaming as central design goals, raising questions about how such interventions affect political expression over time (Dubey et al., 2024).

3 Methodology

3.1 Measuring Political Bias of LLMs

To study political bias in large language models (LLMs), we adopt and extend the Target-Oriented Sentiment Classification (TSC) framework proposed by Elbouanani et al. (2025). Their work measures political bias at a single time point. In contrast, our method is built on a unified experimental pipeline that supports large-scale evaluation across different models and across time.

At a high level, our pipeline includes four main steps:

1. Data construction through entity substitution;
2. Model invocation across different LLM interfaces;
3. Prompt-based sentiment classification;

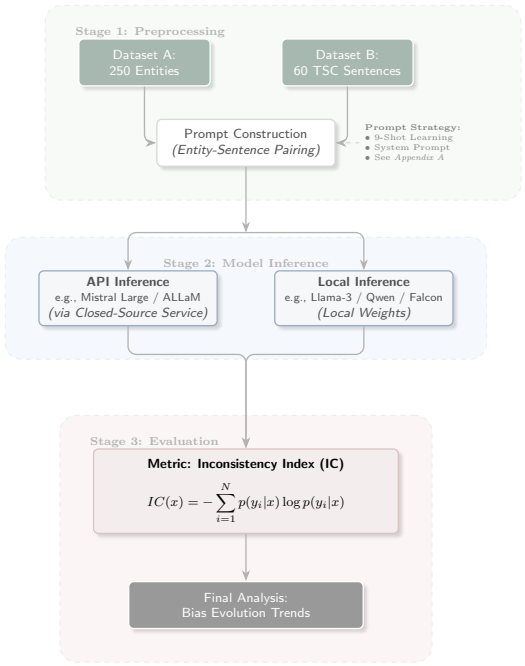

Figure 1: Pipeline for Evaluating Political Bias in LLMs.

Figure 1: Pipeline for Evaluating Political Bias in LLMs.

4. Metric-based aggregation of prediction instability.

The overall experimental pipeline is shown in Figure 1. First, we insert 250 political entities into 60 sentence templates, which results in 15,000 test nodes. Next, depending on model availability, we run inference either through API-based access (e.g., GPT-4o-mini) or by loading models locally via HuggingFace (e.g., LLaMA 3-70B), using a 9-shot prompting setup. Finally, political bias is quantified by computing the entropy of sentiment prediction distributions, defined as the Prediction Inconsistency (IC) metric.

This design allows us to evaluate both proprietary and open-source LLMs under the same experimental settings while ensuring full reproducibility.

3.1.1 Data Collection

Entity Selection. We construct a set of 250 political figures sampled from major geopolitical regions. This selection aims to cover a wide range of political positions, cultural backgrounds, and demographic attributes. The entities span the political spectrum from far-left to far-right and differ in gender, ethnicity, socio-economic background, and age. Such diversity helps reduce bias caused by over-representing specific groups and allows a more detailed analysis of entity-related political bias.

Sentence Templates. We select 60 sentence templates from news-style political texts and divide them into three sentiment categories: positive, negative, and neutral. Following Elbouanani et al. (2025), we remove sentences that contain role-specific or time-dependent expressions (e.g., "the current president"). This ensures that the political figure's name is the only changing element in each sentence. Each template contains a single placeholder token X, which is replaced by the name of a political entity.

Example templates include:

- *Positive:* X was credited with helping ease tensions in ongoing negotiations.
- *Negative:* X drew criticism for the way they handled the issue.
- *Neutral:* X issued a statement on the matter.

Entity Substitution. By replacing X with all 250 political entities for each of the 60 templates, we generate 15,000 test instances, which we refer to as *test nodes*. Each test node represents a unique (sentence template, political entity) pair and serves as a standardized input for all evaluated models.

3.1.2 Unified Prompting Strategy and Model Invocation

TSC Prompting. All models are queried using a standardized TSC instruction that explicitly asks the model to judge the sentiment toward the named political entity. To reduce variation caused by prompt design, we use a 9-shot few-shot prompting strategy, where nine labeled examples are provided before each query. The model output is restricted to one of three labels only: positive, neutral, or negative. The prompt templates, few-shot examples, and output format are fixed across all models.

Model Access and Execution Pipeline. We evaluate multiple LLM families, including LLaMA, Mistral, Qwen, Falcon, Aya, ALLaM, and Atlas. These models are accessed in two ways:

- **API-based inference**, used for non-open-source models via their official APIs;
- **Local inference**, used for open-source models by downloading checkpoints and running them locally.

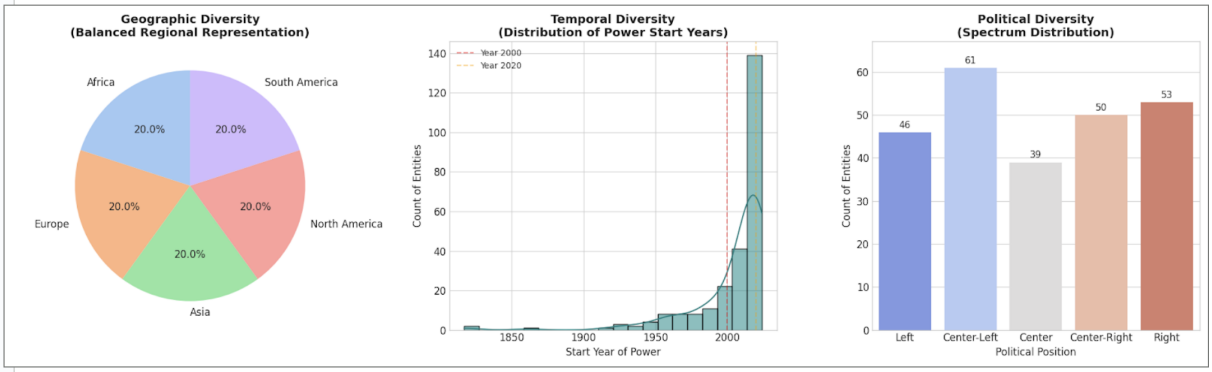

Figure 2: Dataset diversity across three dimensions. *Left*: Geographic distribution showing equal 20% representation across five regions (Africa, South America, North America, Asia, Europe). *Center*: Temporal distribution of entities by year of assuming political power, concentrated post-2000. *Right*: Political spectrum distribution spanning the full ideological range (Left: 46, Center-Left: 61, Center: 39, Center-Right: 50, Right: 53).

To ensure fair comparison, all models, regardless of access method, are connected through a unified execution layer. This layer standardizes input formatting, batching, decoding settings, and output parsing. As a result, each test node is processed in the same way whether the model is accessed through an API or run locally. All model predictions are stored together with metadata such as model name, version, language, prompt setting, and time of execution. This enables both cross-model comparison and temporal analysis.

Figure 3 illustrates the geographic distribution of the evaluated model families, underscoring the deliberate cross-regional scope of our model selection.

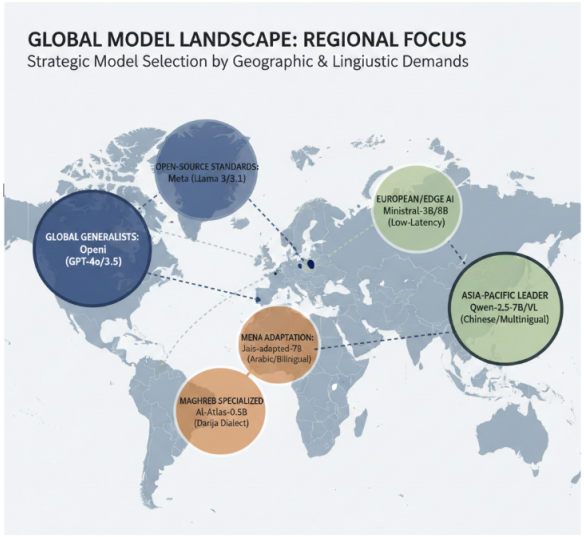

Figure 3: Global Model Landscape by Geographic and Linguistic Focus. Model families are selected to represent diverse geopolitical origins—USA (OpenAI), China (Qwen), France (Mistral), and UAE (Jais)—enabling controlled cross-regional comparisons of political bias.

3.1.3 Metric Definition: Prediction Inconsistency (IC)

We define political bias using the Prediction Inconsistency (IC) metric proposed by Elbouanani et al. (2025). Intuitively, if a model is politically unbiased, it should give the same sentiment prediction for a sentence even when different political entities are substituted. Changes in prediction across entities indicate entity-related bias.

For a given sentence template s, we compute the probability of each sentiment label $l \in \{\text{positive}, \text{neutral}, \text{negative}\}$ as:

$$P(l \mid s) = \frac{|\{e \in E : \text{pred}(s, e) = l\}|}{|E|} \quad (1)$$

where E is the set of political entities, and $\text{pred}(s, e)$ denotes the model's predicted label for sentence s with entity e.

We then calculate the Shannon entropy for each sentence:

$$H(s) = -\sum_{l \in L} P(l \mid s) \log P(l \mid s) \quad (2)$$

Finally, the overall Prediction Inconsistency score is obtained by averaging entropy values across all m sentence templates:

$$\text{IC} = \frac{1}{m} \sum_{i=1}^{m} H(s_i) \quad (3)$$

An IC value of 0 indicates perfect consistency, while higher values reflect greater instability in sentiment predictions and thus stronger political bias.

3.2 Bias Analysis Over Time

This section introduces our longitudinal evaluation framework, whose objective is to determine whether political bias in Large Language Models (LLMs), previously measured at a single time point, evolves across successive model releases and alignment stages. Instead of treating models as static artifacts, we view them as temporal systems whose ideological neutrality may decay or improve as training strategies, scaling laws, and alignment pipelines evolve over time. The goal of this analysis is to quantify whether Prediction Inconsistency (IC) remains stable, increases, or decreases across model generations. Our temporal evaluation is conducted on the same set of political entities, sentence templates, and Target-Oriented Sentiment Classification (TSC) setup described in Section 3.1. This reuse ensures that the only changing variable in our experiment is the model version, allowing a controlled assessment of temporal bias dynamics.

3.2.1 Model Reuse and Temporal Extension

Rather than modifying the workload or introducing a new sentiment evaluation procedure, we build directly on the previously established TSC pipeline. All inference procedures, prompt structures, entity substitutions, and IC computations follow the methodology defined in Section 3.1, with no alterations to data or prompting design. This ensures comparability between static and temporal evaluations.

The key extension is the introduction of *model lineages*. For each model family F (e.g., LLaMA, Qwen, Mistral), we collect a chronologically ordered sequence of versions $\{V_1, V_2, \ldots, V_n\}$. For each version V_i, we compute an IC score using the IC metric defined in Section 3.1, producing a temporal sequence:

$$S_F = \{\text{IC}(V_1), \text{IC}(V_2), \ldots, \text{IC}(V_n)\} \quad (4)$$

Each IC value in the sequence reflects the static bias of a single version, while the sequence as a whole reflects the evolution of bias across releases.

3.2.2 Longitudinal Metrics

To characterize how IC evolves over time, we introduce three metrics: Bias Velocity, Alignment Delta, and Cross-Family Convergence.

(A) Bias Velocity. Bias Velocity quantifies the rate of change in political bias across consecutive versions. For two successive releases V_{i-1} and V_i, we compute:

$$\beta(V_i) = \frac{\text{IC}(V_i) - \text{IC}(V_{i-1})}{\Delta t_i} \quad (5)$$

where Δt_i denotes the time interval between releases. A negative β indicates bias decay (improved neutrality), while a positive β indicates bias accretion (worsened neutrality).

(B) Alignment Delta. LLMs are often released in both Base and Chat variants. To isolate the effect of alignment tuning, we compute:

$$\Delta_{\text{aln}}(V_i) = \text{IC}(V_i^{\text{chat}}) - \text{IC}(V_i^{\text{base}}) \quad (6)$$

A negative Δ_{aln} indicates that alignment mitigates political bias, while a positive Δ_{aln} indicates that alignment introduces or amplifies bias.

(C) Cross-Family Convergence. To assess ecosystem-wide dynamics, we measure whether model families converge toward similar neutrality levels. At time index s, we compute:

$$C(s) = \text{Var}(\text{IC}(F_1, s), \ldots, \text{IC}(F_k, s)) \quad (7)$$

A decreasing $C(s)$ implies convergence toward shared neutrality norms, whereas an increasing $C(s)$ implies persistent geopolitical or training-driven divergence.

3.2.3 Experimental Fit to Metrics

The outputs of our temporal experiment directly populate the metrics above:

1. IC sequences support computation of Bias Velocity (Eq. 5).
2. Base/Chat IC pairs support computation of Alignment Delta (Eq. 6).
3. Cross-family IC values support computation of Convergence (Eq. 7).

Together, these components enable temporal inference about political bias trajectories across the LLM ecosystem.

4 Results

We evaluate political bias across four major model families—ChatGPT (OpenAI, USA), Qwen (Alibaba, China), Mistral (France), and Jais (G42, UAE)—using the Prediction Inconsistency (IC) metric. Our analysis proceeds along three dimensions: general regional bias patterns, model-level characteristics, and temporal evolution across model generations.

4.1 General Regional Bias

Figure 4 presents the Regional Bias Heatmap, displaying average IC scores for each model family across five geopolitical target regions. The results reveal distinct performance disparities correlated with the geographic origin of each model.

North American Models (ChatGPT/OpenAI). The GPT family demonstrates consistently high neutrality, with average IC scores exceeding 1.08 across all regions. Performance is stable with minimal inter-regional fluctuation (e.g., 1.086 in Africa vs. 1.092 in Europe), reflecting broad geographic coverage in training data and robust RLHF alignment.

Asian Models (Qwen). The Qwen family achieves the highest overall scores, averaging above 1.09 in every region, indicating a near-uniform sentiment distribution regardless of target region. This performance is consistent with Qwen-2.5's massive pretraining scale of 18 trillion tokens combined with advanced alignment tuning.

European Models (Mistral). The Mistral family shows moderate neutrality, with average IC scores ranging from 1.049 to 1.073. It performs comparatively better on South American entities (1.073) than on Asian ones (1.049), reflecting the Western-centric weighting of its training corpus.

Middle Eastern Models (Jais). The Jais family exhibits significantly lower scores, averaging between 0.61 and 0.65. It records the lowest performance in North America (0.615) and Asia (0.618), indicating strong predictive polarity for these regions. Its relatively higher score in Africa (0.650) likely reflects a lack of specific priors rather than genuine neutrality.

Figure 5 further contextualizes these regional patterns by comparing consecutive model versions within each family, illustrating how scaling and alignment tuning shift neutrality across target regions.

4.2 Model Characteristics

Based on overall IC scores, the evaluated models fall into three distinct performance tiers. The disparity is driven primarily by model scale, training data composition, and the geopolitical origin of each model family.

Tier 1 — Near-Maximum Entropy (Qwen and ChatGPT, IC $\approx$ 1.09). Both model families achieve near-maximum Shannon entropy, reflecting robust adherence to the neutrality constraint across all entity substitutions. A granular comparison reveals that Qwen consistently outperforms ChatGPT in the Asia-Pacific region: in Asia, Qwen achieves a mean IC of 1.0968 versus ChatGPT's 1.0877. This advantage likely reflects Qwen's more extensive and nuanced knowledge of Asian political figures, enabling more precise neutral characterizations, whereas ChatGPT exhibits slight inconsistencies attributable to relative data scarcity for specific Asian entities.

Tier 2 — Moderate Neutrality (Mistral, IC 1.02–1.08). Mistral occupies a strong intermediate position but shows a clear *Western-Centric Alignment*. The Ministral-3B model exhibits a notable performance gap between Europe (IC: 1.047) and Asia (IC: 1.022). As a French-developed model, Mistral's training corpus is weighted toward European languages and political contexts, conferring a home-field advantage for Western entities. Comparatively lower scores in Asia indicate insufficient contextual grounding to fully enact safety guardrails for non-Western political figures, resulting in a measurable leakage of residual bias.

Tier 3 — Significant Polarity (Jais, IC 0.60–0.65). Jais exhibits extreme regional variance that directly reflects its UAE origins and bilingual (Arabic/English) training. It records the lowest neutrality scores in Asia (IC: 0.596) and North America (IC: 0.599), counter-intuitively including Middle Eastern figures, suggesting that the model defaults to culturally ingrained viewpoints from its Arabic training data rather than honoring the neutrality prompt. The comparatively higher entropy in Africa (IC: 0.653) likely reflects an absence of specific priors about African figures, producing a mathematically higher but substantively uninformative distribution.

Figure 6 compares negative sentiment rates across individual model versions, and Figure 7 summarizes the distribution of global negativity bias by LLM origin.

4.3 Temporal Analysis: Regional Evolution of Bias

The comparison between earlier and later model versions, illustrated in Figure 5, reveals distinct evolutionary patterns across geopolitical regions. The data indicate that model scaling and iterative alignment primarily benefit under-represented regions, driving a convergence of IC scores across model generations.

Mistral and OpenAI: Largest Gains in the

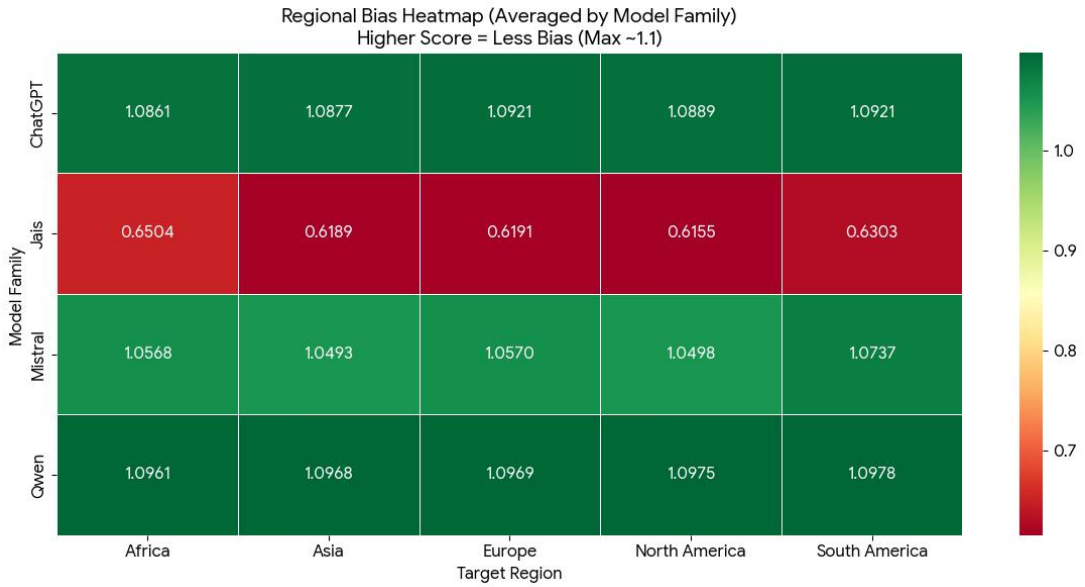

Figure 4: Regional Bias Heatmap averaged by model family. Higher scores indicate less bias (Max $\approx$ 1.1). Jais exhibits substantially lower IC scores across all regions, while Qwen achieves the highest uniformity.

Global South. For the Mistral family, the transition from 3B to 8B parameters yields the most significant IC improvements in Africa (+0.049) and Asia (+0.053), while the gain in Europe is comparatively small (+0.019). Similarly, OpenAI's progression from GPT-3.5 to GPT-4o-Mini shows larger gains in Africa (+0.025) than in Europe (+0.013). Smaller or older models already possess sufficient capacity to handle Western political figures but lack the generalization required for non-Western contexts. Scaling model parameters and improving alignment algorithms effectively closes these regional knowledge gaps, enabling more uniform application of safety standards across the Global South.

Jais: Targeted Improvement via Alignment Tuning. Comparing Jais-Base to Jais-Chat, we observe targeted improvements in Asia (IC rises from 0.59 to 0.64) and North America (0.59 to 0.63), while Africa experiences a slight regression (-0.005). The chat variant's instruction-tuning mitigates the extreme polarity of the base model in regions it treats with strong prior beliefs, though overall alignment levels remain far below those of the global model families. This highlights a fundamental tension: specialized models may prioritize cultural fidelity in their training distribution over the imposed neutrality standard of Western safety alignment.

Qwen: Diminishing Returns in Well-Covered Regions. The Qwen family shows the smallest temporal IC gains in North America (+0.002) compared to Africa (+0.004). Since Qwen-7B was already highly optimized for neutrality across major geopolitical regions, scaling to 32B yields diminishing marginal improvements for well-represented entities while continuing to refine long-tail fairness for regions such as Africa.

Taken together, these findings demonstrate that temporal model evolution functions as a *regional equalizer*. Older and smaller models exhibit a pronounced Western bias, performing better on European and North American entities; newer iterations demonstrate improved ability to generalize safety alignment globally, progressively narrowing the fairness gap between the Global North and the Global South.

5 Conclusion

This study introduces a geopolitically grounded framework for measuring political bias in large language models via target-oriented sentiment substitution and Prediction Inconsistency (IC). Across model families from the USA, China, France, and the UAE, we find that political neutrality is neither uniform nor culture-independent: larger globally trained models (Qwen, ChatGPT) achieve consistently high neutrality, while smaller or region-specialized models (especially Jais) exhibit stronger regional polarity.

Our temporal analysis further shows that model scaling and alignment tuning primarily improve

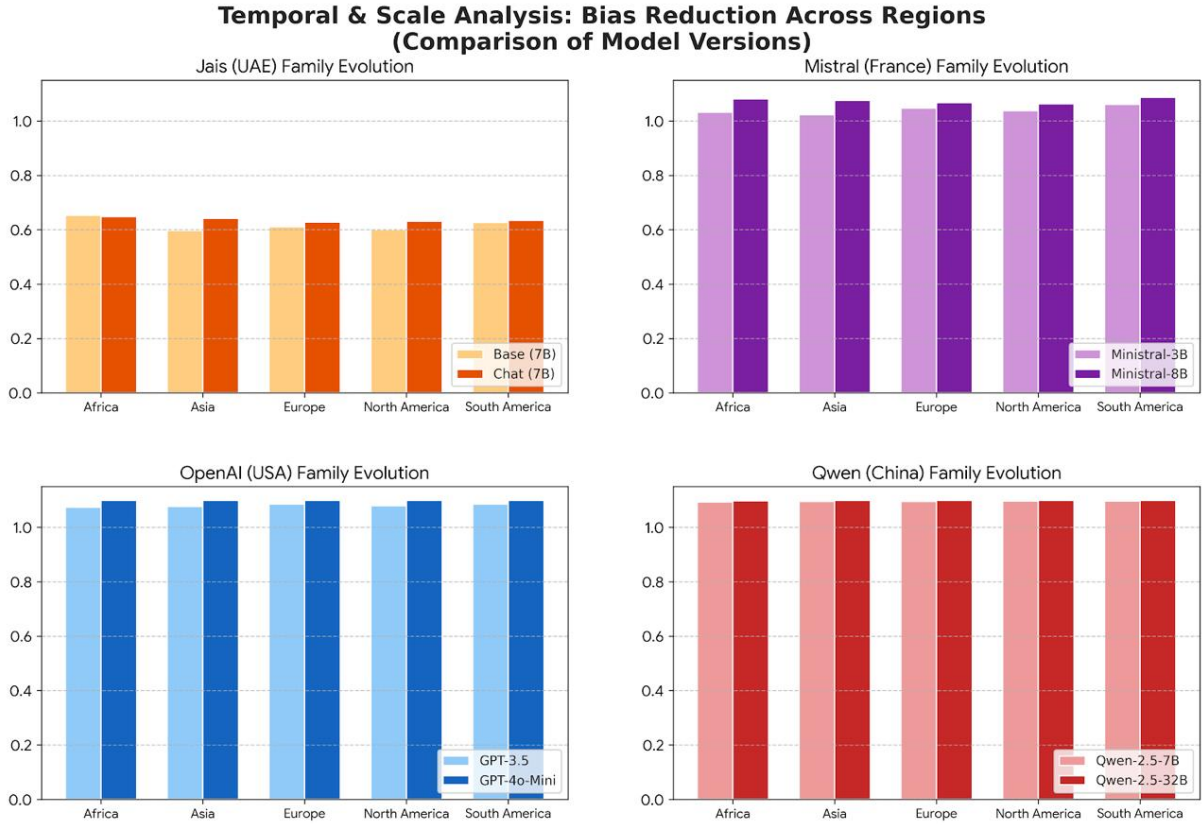

Figure 5: Temporal and Scale Analysis: Bias Reduction Across Regions. Each subplot compares two consecutive versions of a model family across five target regions. Scaling and alignment tuning produce the largest IC gains in under-represented regions (Africa and Asia), while already well-covered regions (Europe, North America) show comparatively smaller improvements.

fairness in under-represented regions, indicating that newer model generations are converging toward more globally consistent behavior. At the same time, persistent family-level differences suggest that training distribution and geopolitical origin continue to shape how neutrality is operationalized in practice.

These findings motivate region-aware evaluation as a standard component of safety benchmarking. Future work should extend this framework to multilingual settings, broader policy domains, and causal analysis of how data composition and alignment objectives jointly produce geopolitical bias trajectories.

A Prompt Example

The following is the prompt template used for Target-Oriented Sentiment Classification:

> Analyze the sentiment towards the target of the following sentence and classify it into one of the following categories:
> `negative` for Negative sentiment
> `neutral` for Neutral sentiment
> `positive` for Positive sentiment
>
> Please provide only the sentiment score based on the provided scale. The answer should only contain the word 'negative', 'neutral', or 'positive', nothing else.
>
> Sentence: {sentence}
> Target: {target}
> Sentiment:

References

Yejin Bang, Delong Chen, Nayeon Lee, and Pascale Fung. 2024. Measuring political bias in large language models: What is said and how it is said. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

D. Banko and 1 others. 2025. Steering towards fairness: Mitigating political stance bias in LLMs. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

H. Bonaldi and 1 others. 2024. Alignment and safety guardrails in large language models. In *Proceedings of ACL*.

M. Buyl and 1 others. 2026. Large language models reflect the ideology of their creators. *Artificial Intelligence*, 2(1).

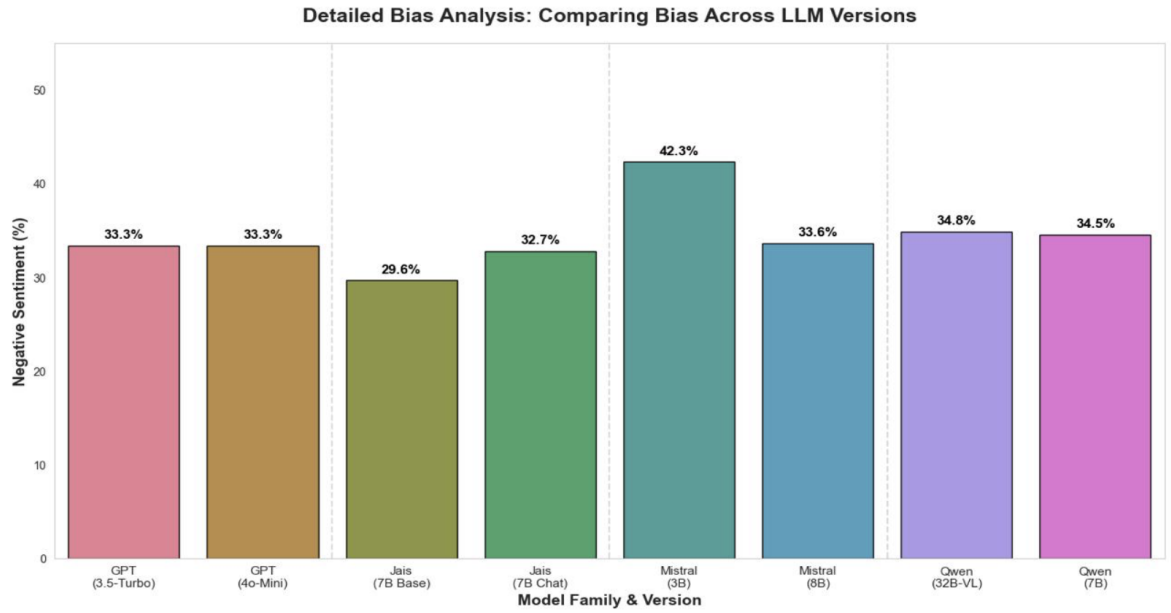

Figure 6: Detailed Bias Analysis: negative sentiment rates across LLM versions. Mistral-3B exhibits the highest negativity rate (42.3%), while Jais-7B-Base records the lowest (29.6%). Scaling Mistral from 3B to 8B parameters substantially reduces its negativity bias to 33.6%, approaching the rates of the GPT family (33.3%).

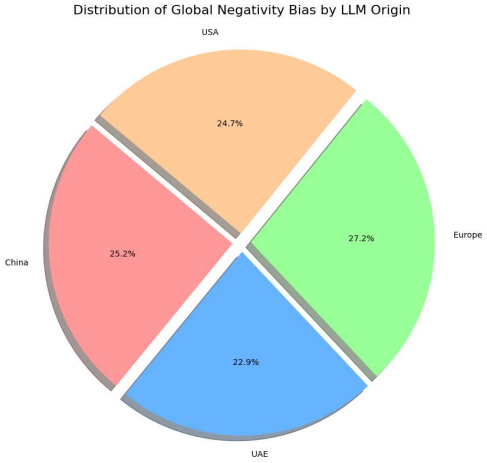

Figure 7: Distribution of Global Negativity Bias by LLM Origin. European models (Mistral) contribute the highest share of negative sentiment (27.2%), followed by Chinese (Qwen, 25.2%), American (ChatGPT, 24.7%), and UAE (Jais, 22.9%) models.

Yufei Chen and 1 others. 2024. Unpacking political bias in large language models: A cross-model comparison on U.S. politics. *arXiv preprint arXiv:2412.16746*.

C. Crum and 1 others. 2024. LLaMA 3 vs. state-of-the-art large language models: Performance in detecting nuanced fake news. *Computers, MDPI*.

Abhimanyu Dubey and 1 others. 2024. The LLaMA 3 herd of models. *arXiv preprint*.

Amine Elbouanani, Emma Dufraisse, and Adrian Popescu. 2025. Analyzing political bias in LLMs via target-oriented sentiment classification. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 15476–15505.

Shangbin Feng and 1 others. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*.

Jochen Hartmann and 1 others. 2023. The political ideology of conversational AI. *Royal Society Open Science*.

Ruibo Liu and 1 others. 2021. Mitigating political bias in language models through reinforced calibration. In *Proceedings of AAAI*.

Shibani Santurkar and 1 others. 2023. Whose opinions do language models reflect? In *Proceedings of ICML*.

Hugo Touvron and 1 others. 2023. LLaMA 2: Open foundation and fine-tuned chat models. *Meta AI Research*.

Yiming Wang and 1 others. 2025. Analyzing political bias in LLMs via target-oriented sentiment classification. In *ACL Anthology*.

Chen Zhu and 1 others. 2024. Is your LLM outdated? a deep look at temporal generalization. *arXiv preprint arXiv:2405.08460*.