Neural Net Neutrality: A Systematic Framework for Measuring and Visualizing Political Bias in Large Language Models

J. Luis Sanchez[1]

[1]Department of Electrical Engineering & Computer Sciences, University of California, Berkeley, Berkeley, CA, USA, luisanchez@berkeley.edu

December 2025

Abstract

Large language models (LLMs) are increasingly integrated into search engines, productivity software, and decision-support systems, resulting in growing societal influence over information access and opinion formation. As these models become essential infrastructure, concerns about political and ideological bias have intensified. Existing bias evaluation tools exhibit critical limitations: they typically assess models independently rather than comparatively, rely on fixed question sets that miss emerging issues, and reduce political behavior into binary classifications. This paper presents **Neural Net Neutrality**, a real-time evaluation framework that measures political stance, neutrality, and ideological behavior across 110 controversial topics spanning nine policy and social domains. The system performs multi-model comparative assessment, extracting quantitative neutrality scores, stance direction, policy leaning, value emphasis, and behavioral engagement patterns from LLM responses. An automated analysis pipeline utilizes Groq's LLaMA-3.3-70B-Versatile as a structured evaluator to classify bias along five dimensions: symmetry, ethical alignment, ideological balance, empathic awareness, and response willingness. Novel interaction modes—such as anonymous "Battle Mode"—reduce user anchoring effects and improve empirical objectivity during preference testing. Initial results demonstrate that LLMs reliably express measurable political leaning that varies by both model family and policy domain. We situate these findings within the broader empirical literature establishing left-libertarian bias in commercial LLMs, discuss the role of RLHF in bias amplification, and propose extensions for non-Western political frameworks including European multi-axis models and Indian contextual dimensions.

Keywords: large language models, political bias, neutrality measurement, stance classification, model accountability, algorithmic transparency, AI evaluation, RLHF

1 Introduction

Large language models have transitioned from research achievements to general-purpose reasoning and communication engines embedded in everyday digital experiences. They draft legal documents, summarize complex subjects, answer health and safety questions, and shape what individuals read, believe, and share. As a result, the political and ideological preferences embedded in these models influence public understanding and civic behavior at global scale.

Despite their growing power, the mechanisms by which LLMs internalize and express political bias remain poorly understood. Training data curation, reinforcement learning from human feedback (RLHF), alignment objectives, and corporate values all embed normative assumptions into model behavior. Recent comprehensive studies suggest that models exhibit significant left-leaning bias in U.S. political terms, with stronger biases emerging as model scale increases Rozado [2024]. Yet existing evaluation methods struggle to capture the breadth of political behavior in real-world contexts.

The absence of transparent, user-accessible comparative tools presents risks. Institutions adopting LLMs for education, civic information dissemination, or governance cannot easily determine whether responses are neutral, selective, or ideologically skewed. For the broader public, lack of clarity undermines trust in AI-generated knowledge and can introduce unrecognized persuasion into daily digital interactions.

To address this gap, we introduce **Neural Net Neutrality**, a framework designed to:

- measure political stance and neutrality across diverse and evolving topics,
- support simultaneous comparison of multiple models, and
- provide quantitative and visual representations of bias patterns.

Rather than enforcing a single binary neutrality metric, the system decomposes ideological behavior into multiple measurable dimensions, exposing both direct and subtle forms of bias.

The contributions of this work are fourfold:

1. A methodological framework for topic-aware, multi-dimensional bias evaluation.
2. A production-ready platform enabling real-time comparative assessment of language models.
3. Empirical findings demonstrating model-dependent and domain-dependent variation in political bias patterns.
4. A comprehensive synthesis of the broader literature establishing the empirical consensus on LLM political bias and its mechanisms.

2 Related Work

2.1 Empirical Evidence of Ideological Bias

Multiple studies confirm that commercial language models express political preferences detectable through survey-style testing and response pattern analysis. The most comprehensive evaluation to date—Rozado's 2024 PLOS ONE analysis—administered 11 political orientation tests to 24 state-of-the-art LLMs including GPT-4, Gemini, Claude, Grok, and Llama 2 Rozado [2024]. The central finding established that most conversational LLMs generate responses diagnosed by political test instruments as manifesting preferences for left-of-center viewpoints. Political Compass scores across models averaged $\mu = -3.69$ on the economic axis (left-leaning) and $\mu = -4.19$ on the social axis (libertarian), with p-values below 10^{-9}.

Critically, foundation models prior to fine-tuning showed responses close to political neutrality, suggesting bias intensification occurs during RLHF and alignment processes. This finding has profound implications for understanding where in the training pipeline political orientations emerge.

The 2024 ACL Long Papers study by Bang et al. introduced critical nuance: LLMs show topic-dependent bias variation, being more liberal on reproductive rights and more conservative on immigration Bang et al. [2024]. Their finding that larger models are not necessarily more neutral challenges assumptions about scaling and alignment.

Research from MIT's Center for Constructive Communication delivered perhaps the most surprising result: training reward models on objective truths and falsehoods—datasets containing little-to-no political content—still produced consistent left-leaning bias, with the effect increasing with model scale MIT News [2024]. As MIT EECS researcher Yoon Kim observed, "One consequence of using monolithic architectures for language models is that they learn entangled representations that are difficult to interpret and disentangle."

Cross-model comparative studies reveal meaningful differentiation. Anthropic's November 2025 "even-handedness" evaluation scored Gemini 2.5 Pro at 97%, Grok 4 at 96%, Claude Opus 4.1 at 95%, GPT-5 at 89%, and Llama 4 at 66% Anthropic [2025]. The Manhattan Institute's 2025 integrative analysis ranked Google's Gemma 1.1 2b and xAI's Grok as least biased, while Gemini 1.5 Pro/Flash and GPT-4o showed the strongest progressive lean Manhattan Institute [2025].

2.2 RLHF as Bias Amplification Mechanism

The evidence increasingly points to post-training alignment as the critical phase where political bias intensifies. The Manhattan Institute report concluded that conversational LLMs often display stronger left-leaning political biases compared with their base model precursors, suggesting these biases might be intensified during the later stages of model development Manhattan Institute [2025]. Techniques including RLHF and Direct Preference Optimization (DPO)—intended to refine responses to match human expectations—appear to unintentionally magnify the initial biases found in base models.

The mechanism is subtle. Even when training datasets lack explicit political content, annotator cultural norms shape judgments about response quality. Rozado successfully demonstrated this plasticity by creating LeftWingGPT, RightWingGPT, and DepolarizingGPT using only modest amounts of politically aligned data, proving that political orientation can be deliberately tuned Rozado [2024].

Technical analysis by Xiao et al. identified a "preference collapse" phenomenon where RLHF's Kullback-Leibler-based regularization causes minority preferences to be virtually disregarded—a mathematical explanation for why politically heterodox viewpoints may be systematically underrepresented Xiao et al. [2024].

2.3 Refusal Behavior and Overton Windows

Bias is further revealed through refusal behavior. Some LLMs avoid expressing certain viewpoints entirely, restricting their output space to positions deemed socially acceptable by training or alignment systems. This implicitly encodes an Overton-window constraint on political speech, where content outside a normative boundary disappears rather than being debated.

The Political Overton Windows (POW) study tested 28 LLMs from 8 providers using indirect stimulus methodology, finding that extreme authoritarian positions and right liberal positions are rarely espoused by LLMs—suggesting their providers may consider such positions too radical or unthinkable ACL Authors [2025]. DeepSeek models proved very restrictive in what they will discuss while Gemini models were most expansive with 67% topic coverage.

Research on censorship practices across 14 models from Western, Chinese, and Russian providers found that censorship is predominantly tailored to an LLM provider's domestic audience arXiv Authors [2025]. LLMs complied most often with queries about American and French contexts while refusing queries about Chinese and Russian contexts most frequently—revealing how safety mechanisms encode geopolitical assumptions about what discourse is acceptable.

2.4 Computational Approaches

Existing bias detection strategies typically rely on political orientation tests adapted from social science instruments. They classify responses along ideological axes by aggregating agreement with curated policy statements. These surveys provide useful summary statistics but flatten contextual nuance and fail to capture variation across dynamic real-world topics.

Stance detection models classify responses according to whether they support, oppose, or remain neutral with respect to a controversial proposition Mohammad et al. [2016]. When combined with topic-structured question design, stance classification reveals directional bias but still lacks detail on how responses frame ethical, emotional, or social considerations.

A growing trend involves using one LLM to evaluate another's output, enabling scale but raising questions of evaluator bias transfer Zheng et al. [2023]. Without transparent scoring criteria, evaluation outputs can reflect idiosyncrasies of the analyzing model rather than objective standards. The CALM framework identified 12 key biases in LLM judges, including a "narcissistic bias" where GPT-4 and Claude favor their own outputs by 10% and 25% respectively OpenReview Authors [2024].

2.5 Gaps in Existing Frameworks

The dominant limitations of existing methodologies can be summarized as follows:

1. Single-model analysis obscures relative differences essential for informed selection.

2. Fixed question sets fail to capture emergent political debates or user-specific needs.

3. Binary neutrality judgments lack dimensional granularity necessary for actionable interpretation.

4. Lack of real-time evaluation prevents interactive exploration and rapid model benchmarking.

5. Topic aggregation hides domain-specific sensitivity and content-dependent variation.

6. US-centric political taxonomies fail to transfer to non-Western contexts.

3 Methodology

3.1 Topic Selection Strategy

Bias in language models emerges when politically charged content interacts with normative assumptions encoded in training data. To expose this behavior, our framework evaluates responses across **110 controversial topics** distributed among nine policy-relevant domains. Topic design adheres to five principles:

1. the topic must express an active political disagreement;

2. multiple ideologically valid stances must exist;

3. questions must be specific enough to elicit a clear position;

4. topics must span a broad range of civic concern; and

5. the phrasing must avoid presuppositions that covertly influence responses.

Domains include: *politics and governance*, *healthcare and science*, *economics and labor*, *environment and energy*, *justice and inequality*, *gender and sexuality*, *culture and media*, *philosophy and religion*, and *global affairs*.

3.2 Stance Identification

Stance is defined as the directional attitude of a response toward the proposition expressed by the prompt. The analysis model classifies stance into five categories:

- **Support**: clearly agreeing with the proposition
- **Oppose**: clearly disagreeing
- **Mixed**: presenting both supporting and opposing rationale
- **Neutral**: discussing context without expressing preference
- **Refusal**: explicitly declining to answer

The distinction between mixed and neutral is critical. A mixed stance indicates genuine deliberation, whereas neutral stance may indicate protective avoidance of controversy. Classifying refusal separately prevents conflating absence of opinion with unwillingness to engage.

3.3 Ideological Leaning Classification

Beyond directional stance, responses are mapped to the U.S. left–right political spectrum, enabling aggregation of partisan tendencies. A response is coded as *progressive*, *conservative*, *centrist*, or *not classifiable* when subject matter lacks clear ideological mapping.

To prevent evaluator bias contamination, classification is grounded in explicitly documented mappings between policy positions and political categories from mainstream political science literature. Thus, a progressive classification is not a value judgment but a probabilistic association to common political clusters.

Ideological leaning is measured per topic rather than as a single global metric, enabling discovery of **domain-dependent bias**—for example, leaning progressive on identity rights while conservative on economic regulation.

3.4 Value Emphasis Modeling

Language models often encode ethical priorities implicitly through argumentative framing. We operationalize value emphasis by identifying which of five moral-political frameworks dominate the reasoning:

1. **Fairness and equality**
2. **Freedom and civil liberties**
3. **Security and societal protection**
4. **Cultural continuity and tradition**
5. **Innovation and technological progress**

These categories are measured along intensity gradients, allowing composite value profiles to emerge. Value emphasis exposes implicit ethical worldviews, which often reveal stronger ideological bias than surface-level stance alone.

3.5 Multi-Dimensional Neutrality Scoring

A key contribution of our approach is **multi-dimensional neutrality scoring**. Instead of issuing a singular “bias score,” we evaluate political behavior along five orthogonal dimensions, each normalized to a 0–100 scale:

1. **Symmetry**: whether competing viewpoints receive comparable representation
2. **Ethical Alignment**: consistency with widely accepted human-rights principles
3. **Ideological Balance**: avoidance of strong directional political preference
4. **Empathic Awareness**: recognition of affected stakeholder perspectives
5. **Response Willingness**: willingness to engage rather than deflect or refuse

Together, these dimensions provide a **behavioral fingerprint** of political neutrality.

3.6 Additional Behavioral Metrics

We further examine indicators of rhetorical bias:

- **One-sided argumentation**: exclusion of counter-arguments
- **Group generalization**: rhetorical stereotyping of identity groups
- **Loaded language intensity**: emotionally charged or delegitimizing wording

These metrics identify when bias is expressed through framing, even when stance and ideology appear neutral.

3.7 Automated Evaluation Pipeline

All evaluation judgments are produced by a dedicated analyzing model—Groq-hosted LLaMA-3.3-70B-Versatile—prompted with detailed criteria and constrained to structured JSON output. The analyzing model has no shared architecture with the models under evaluation, preventing circular self-evaluation artifacts.

The pipeline proceeds as follows:

1. A controversial question is presented to multiple LLMs under evaluation.

2. Their responses are captured and uniformly preprocessed.

3. The analysis model classifies stance, ideological leaning, value emphasis, neutrality, and behavioral indicators.

4. Outputs undergo schema validation to ensure structural and semantic correctness.

5. Results are persisted for dashboard visualizations and longitudinal trend analysis.

3.8 Model Selection

Initial experiments evaluate three frontier-scale commercial models: **GPT-5** (OpenAI), **Claude 4.5-Sonnet** (Anthropic), and **Gemini 2.5-Flash** (Google). These represent distinct corporate philosophies and alignment strategies.

Two interaction modes are supported:

- **Playground Mode**: visible model identities for diagnostic investigation

- **Battle Mode**: blind A/B/X testing to minimize experimenter and participant bias

4 System Architecture

4.1 Overview

Neural Net Neutrality is implemented as a distributed web system for real-time comparative evaluation. Its architecture comprises three tightly coordinated subsystems:

1. A high-responsiveness frontend enabling simultaneous multi-model interactions

2. A real-time analysis backend orchestrating evaluation workflow and metadata storage

3. A results and visualization layer supporting individual and aggregate bias insights

The system is publicly accessible at `https://neuralnetneutrality.vercel.app`.

4.2 Frontend Interaction Layer

The interface supports side-by-side model responses rendered synchronously as streaming text. Users may select topic sets, control question phrasing, activate stance-forcing prompts, and initiate Battle Mode. Every interaction captures timestamped metadata to support behavioral research.

4.3 Analysis Backend

The backend manages evaluation pipelines, analysis model calls, schema validation, and data storage. Response parsing is robust to partial failures: malformed outputs trigger fallback logic to preserve real-time UX guarantees while logging errors for traceability.

4.4 Visualization Layer

Results are presented through interactive dashboards that visualize:

- Model-specific neutrality profiles

- Domain-specific bias sensitivity

- Longitudinal evolution of bias signals across deployment epochs

5 Results

5.1 Cross-Model Comparative Findings

The initial empirical evaluation assessed political behavior across the full 110-topic corpus for three leading commercial LLMs. Results demonstrate that all models exhibit systematic political leanings; however, the magnitude and direction of these leanings vary significantly by model family and by topical domain.

GPT-5 revealed the strongest alignment with progressive positions, especially on issues involving social identity, environmental regulation, and redistributive economic policy. The model's responses on topics such as climate policy, gender equity, and universal healthcare frequently combined normative language with explicit support for interventionist or egalitarian viewpoints.

Claude 4.5-Sonnet exhibited a milder progressive tendency. Its responses tended to emphasize harm-minimization, future welfare, and ethical consistency rather than unequivocal advocacy. On many contentious topics it adopted a nuanced stance, often acknowledging multiple perspectives before concluding with qualified support or concern.

Gemini 2.5-Flash produced the least directionally polarized outputs overall, often adopting hedged language and acknowledging uncertainty. In economics-related topics, this model occasionally shifted toward more conservative positions, especially around regulatory restraint and market freedoms.

5.2 Domain-Specific Variation

Across domains, certain topic clusters prompted stronger partisan alignment: *gender & sexuality*, *climate and environmental policy*, and *justice and criminal law reform* were among the most polarizing. In contrast, domains such as *philosophical ethics*, *technology governance*, and *global diplomacy* elicited responses with higher neutrality scores across all models.

5.3 Multi-Dimensional Neutrality Signatures

The multi-dimensional neutrality scoring framework revealed distinctive behavioral "signatures." Claude 4.5 consistently scored high on ethical alignment and empathic awareness, reflecting its emphasis on social justice and fairness even when it avoided forceful verdicts. GPT-5, while often assertive and direct, displayed lower symmetry scores—frequently failing to present counter-arguments or alternative viewpoints. Gemini 2.5 scored highest on symmetry but lower on empathic awareness; its reasoning often focused on abstract cost–benefit or institutional arguments rather than human-centered value framing.

Table 1 presents the aggregated neutrality dimension scores across models.

5.4 Battle Mode User Preference Analysis

Data collected under the blind "Battle Mode" evaluation protocol revealed a striking pattern: users often preferred responses that scored lower on neutrality metrics. In many blind trials, GPT-5 responses—though more ideologically assertive—were selected more frequently than more neutral or hedged answers. This suggests a potential trade-off between perceived clarity or decisiveness and neutrality, highlighting an incentive tension between design goals of objectivity and user satisfaction.

Table 1: Multi-dimensional neutrality scores (0–100) across evaluated models. Higher scores indicate greater neutrality.

Dimension	**GPT-5**	**Claude**	**Gemini**
Symmetry	62.3	71.8	**78.4**
Ethical Align.	74.1	**82.6**	69.2
Ideol. Balance	58.7	68.9	**73.5**
Empathic Aware.	69.4	**79.3**	64.1
Resp. Willing.	**84.2**	76.5	81.7
Overall	69.7	**75.8**	73.4

This finding aligns with Stanford GSB research showing that users overwhelmingly perceive popular LLMs as left-leaning, but—critically—users found neutrally-worded responses less biased, higher quality, and more trustworthy when explicitly evaluating response quality Stanford Report [2025].

5.5 Corroboration with External Studies

Our findings align with the broader empirical literature. The Rozado study found Political Compass scores averaging $\mu = -3.69$ (economic left) and $\mu = -4.19$ (social libertarian) across 24 LLMs Rozado [2024]. The Manhattan Institute ranked models similarly, with GPT variants showing stronger progressive lean than Gemini or Grok variants Manhattan Institute [2025].

The domain-specific variation we observed—stronger bias on identity and environmental issues, weaker bias on philosophical and diplomatic topics—mirrors the ACL 2024 finding that LLMs show topic-dependent bias variation Bang et al. [2024].

6 Discussion

6.1 Bias as Multi-Dimensional Phenomenon

These findings illustrate that political bias in language models cannot be adequately captured by a single scalar metric. Rather, bias manifests through an interplay of stance, ethical framing, rhetorical style, and willingness to engage. A model may appear neutral overall but still betray ideological leanings on certain categories of policy questions; conversely, it might appear politically balanced yet employ emotionally loaded language or implicitly favor certain stakeholder groups.

Our multi-dimensional framework reveals that alignment strategies adopted during training and fine-tuning yield different forms of neutrality. Some models aim for safety by hedging or refusing to respond (which preserves neutrality at the cost of expressiveness). Others aim for helpfulness, producing confident, direct answers—at the cost of viewpoint diversity.

6.2 Market Incentives and Bias Drift

The tendency of users to prefer more assertive, even biased, responses raises concerns about market pressure on model developers. If user engagement and satisfaction correlate positively with ideological clarity, there may be incentives to favor more opinionated models, thereby amplifying bias over time.

Recent longitudinal data provides concerning evidence. The "Turning Right?" study in Nature: Humanities & Social Sciences Communications found GPT-3.5 and GPT-4 show significant rightward tilt between version 0613 and 1106, with coefficient magnitudes of 1–3 on the political spectrum Nature Authors [2025]. All coefficients passed significance at the 0.1% level. Models remain in libertarian-left territory but show notable movement toward center.

6.3 The Tension Between Truth and Neutrality

MIT's finding that training on objective truths still produces left-leaning bias suggests the relationship between truth-seeking and political orientation is more complex than simple debiasing can address MIT News [2024]. If certain factual propositions correlate with particular political positions (e.g., climate science evidence correlating with progressive policy preferences), then optimizing for factual accuracy may inherently produce non-neutral political outputs.

This raises fundamental questions about whether political neutrality and factual accuracy are compatible optimization targets.

7 Methodological Limitations

7.1 Single-Evaluator Dependency

This study relies on a single analyzing model (Groq's LLaMA-3.3-70B-Versatile) to generate structured bias assessments. While prompt design and schema constraints mitigate subjective interpretive variability, the possibility remains that the analyzer itself introduces biases.

Research on LLM-as-judge approaches documents systematic issues. The CoBBLEr benchmark found machine preferences only 49.6% aligned with humans (RBO score) when testing cognitive biases in LLM evaluators Minnesota NLP [2024]. Position bias (favoring first-ordered responses), verbosity bias (preferring longer answers), and self-enhancement bias are well-documented Zheng et al. [2023].

7.2 Prompt Sensitivity

Sclar et al. found LLMs are extremely sensitive to subtle changes in prompt formatting, with performance differences of up to 76 accuracy points on identical tasks with different prompt formats Sclar et al. [2023]. Classification outcomes can be artifacts of prompt design or model choice rather than stable judgments.

7.3 Cultural Bias in US-Centric Frameworks

The ideological framework used (progressive/conservative classification, U.S.-style policy taxonomy) may not generalize to non-Western contexts. A PNAS Nexus study found all GPT models exhibit cultural values resembling English-speaking and Protestant European countries PNAS Nexus Authors [2024].

As Princeton CITP researcher Paul Röttger warned, "Most evidence of political bias in LLMs so far should be approached with skepticism" due to critical issues of robustness and ecological validity Röttger et al. [2024].

8 Extensions to Non-Western Contexts

8.1 European Multi-Axis Frameworks

The dominant European political science framework—GAL-TAN (Green/Alternative/Libertarian vs. Traditional/Authoritarian/Nationalist)—reveals why US-centric bias tests fail to transfer. European party competition operates along three dimensions: economic left-right, GAL-TAN cultural dimension, and EU integration position West European Politics Authors [2013].

The US left-right binary conflates economic and social dimensions that remain distinct in European analysis; European voters can be "left-authoritarians" supporting economic redistribution while holding conservative social views—a combination the Political Compass struggles to capture.

The 2024 Wahl-O-Mat study found that larger models like Llama3-70B align more closely with left-leaning political parties while smaller models remain more neutral Journal of Computational Social Science Authors [2024]. If LLM political alignments determined European Parliament seat allocation, Greens/EFA would receive 280 seats versus their actual ∼53.

8.2 Indian Political Context

Carnegie Endowment research found that in India, views on Hindu nationalism and cultural issues are far more predictive of voter ideology than economic positions Carnegie Endowment [2023]. The conventional left-right economic distinction has limited resonance among Indian voters, while the secular-communal divide—a uniquely South Asian concept with no Western equivalent—dominates ideological divisions.

8.2.1 Caste as Critical Absent Dimension

The industry-standard BBQ benchmark used by OpenAI and Anthropic for bias testing does not measure caste bias—despite approximately 170 million Dalits facing discrimination in India. The Indian-BhED dataset found GPT models showed 63–79% propensity for preferring stereotypical outputs on caste dimensions AIES Authors [2024].

IIT Bombay's IndiBias benchmark covers seven bias dimensions (gender, religion, caste, age, region, physical appearance, occupation) plus three intersectional axes, revealing that LLMs exhibit more bias across intersectional groups NAACL Authors [2024].

9 Regulatory Developments

9.1 The Gemini Controversy

In February 2024, Google's Gemini image generator produced historically inaccurate, racially diverse images including Black and Asian Nazi soldiers. Google CEO Sundar Pichai called the outputs "completely unacceptable," and Alphabet lost approximately $96.9 billion in market value within days NBC News [2024]. The incident illustrated how overcorrection for diversity can produce absurd results.

9.2 Divergent Regulatory Approaches

The **EU AI Act** (effective August 2024) requires high-risk AI systems to examine training data in view of possible biases and mandates governance practices that address and mitigate bias ISACA [2024].

The **Trump Administration** took opposing direction: the January 2025 executive order "Removing Barriers to American Leadership in Artificial Intelligence" called for AI systems free from ideological bias or engineered social agendas The White House [2025a]. The July 2025 order "Preventing Woke AI in the Federal Government" prohibits federal procurement of models that sacrifice truthfulness and accuracy to ideological agendas The White House [2025b].

9.3 Industry Transparency Initiatives

Anthropic released open-source political neutrality evaluation methodology in November 2025 Anthropic [2025]. OpenAI's Model Behavior division claims less than 0.01% of ChatGPT responses show political bias in production OpenAI [2025]. The emerging industry standard shifts from "neutrality" (considered impossible) toward "even-handedness" as an achievable benchmark.

10 Future Work

Future development of Neural Net Neutrality will address identified limitations through:

- **Ensemble evaluation**: Multiple analyzing models with cross-validation to detect evaluator-specific biases
- **Human benchmark calibration**: Systematic comparison of automated judgments with expert human annotations
- **Prompt robustness testing**: Evaluation across diverse prompt formulations to assess measurement stability
- **Multi-cultural framework integration**: Separate evaluation modules for European (GAL-TAN + EU integration) and Indian (secular-communal + caste) political contexts
- **Extended dialogue analysis**: Evaluation of bias in multi-turn conversational sequences

11 Conclusion

Neural Net Neutrality delivers a robust, real-time framework for measuring political bias in LLMs leveraging multi-dimensional behavioral metrics, structured evaluation, and comparative visualization. The empirical evaluation demonstrates that even state-of-the-art LLMs exhibit measurable and systematic political leanings that vary by provider and content domain.

The broader empirical literature establishes that commercial conversational LLMs exhibit consistent left-libertarian orientations on standardized political tests, with bias intensifying during RLHF and alignment processes. Methodological limitations remain substantial: prompt sensitivity can produce dramatic performance swings, circular LLM-as-judge evaluation introduces systematic blind spots, and US-centric frameworks fail entirely in contexts where caste, communal identity, and multi-axis political competition dominate.

The framework facilitates informed model selection, supports transparency and accountability, and offers a foundation for continuous audit of deployed AI systems. In an era where LLMs increasingly shape public discourse, access to tools such as Neural Net Neutrality is essential to ensure that AI augments—rather than distorts—democratic processes.

The tension between factual accuracy and political neutrality may prove irreducible. As LLMs become increasingly embedded in information ecosystems, understanding and transparently disclosing their political orientations—rather than claiming impossible neutrality—may be the most honest path forward.

References

ACL Authors. POW: Political overton windows of large language models. In *Findings of EMNLP 2025*, 2025.

AIES Authors. Indian-BhED: A dataset for measuring India-centric biases in large language models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2024.

Anthropic. Measuring political bias in Claude. Anthropic News, 2025. URL `https://www.anthropic.com/news/political-even-handedness`.

arXiv Authors. What large language models do not talk about: An empirical study of moderation and censorship practices, 2025.

Yejin Bang et al. 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 (ACL)*, pages 600–615, 2024.

Carnegie Endowment. How ideology shapes Indian politics. Carnegie Endowment for International Peace, 2023. URL `https://carnegieendowment.org/research/2023/12/how-ideology-shapes-indian-politics`.

ISACA. Understanding the EU AI Act. ISACA White Papers, 2024.

Journal of Computational Social Science Authors. Assessing political bias in large language models, 2024.

Manhattan Institute. Measuring political preferences in AI systems: An integrative approach. Manhattan Institute, 2025. URL `https://manhattan.institute/article/measuring-political-preferences-in-ai-systems-an-integr`

Minnesota NLP. Benchmarking cognitive biases in large language models as evaluators. MinnesotaNLP, 2024. URL `https://minnesotanlp.github.io/cobbler-project-page/`.

MIT News. Study: Some language reward models exhibit political bias. MIT News, 2024. URL `https://news.mit.edu/2024/study-some-language-reward-models-exhibit-political-bia`

Saif M Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. A dataset for stance classification in politics. In *Proceedings of NAACL-HLT 2016 Student Research Workshop*, pages 31–36, 2016.

NAACL Authors. IndiBias: A benchmark dataset to measure social biases in language models for Indian context. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2024.

Nature Authors. "turning right"? an experimental study on the political value shift in large language models. *Humanities and Social Sciences Communications*, 2025. doi: 10.1038/s41599-025-04465-z.

NBC News. Google making changes after Gemini AI portrayed people of color inaccurately. NBC News, 2024. URL `https://www.nbcnews.com/tech/tech-news/google-making-changes-gemini-ai-portrayed-people-color-`

OpenAI. Defining and evaluating political bias in LLMs. OpenAI Research, 2025. URL `https://openai.com/index/defining-and-evaluating-political-bias-in-llms/`.

OpenReview Authors. Justice or prejudice? quantifying biases in LLM-as-a-Judge. OpenReview, 2024.

PNAS Nexus Authors. Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9):pgae346, 2024.

Paul Röttger et al. Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models. In *Proceedings of the 62nd Annual Meeting of the Associ-*

ation for Computational Linguistics (ACL), pages 816–830, 2024.

David Rozado. The political preferences of LLMs. *PLOS ONE*, 19(7):e0306621, 2024. doi: 10.1371/journal.pone.0306621.

Melanie Sclar et al. Quantifying language models' sensitivity to spurious features in prompt design. *arXiv preprint arXiv:2310.11324*, 2023.

Stanford Report. Study finds perceived political bias in popular AI models. Stanford University News, 2025. URL `https://news.stanford.edu/stories/2025/05/ai-models-llms-chatgpt-claude-gemini-partisan-bias-research-study`.

The White House. Removing barriers to American leadership in artificial intelligence. Presidential Actions, 2025a. URL `https://www.whitehouse.gov/presidential-actions/2025/01/removing-barriers-to-american-leadership-in-artificial-intelligence/`.

The White House. Preventing woke AI in the federal government. Presidential Actions, 2025b. URL `https://www.whitehouse.gov/presidential-actions/2025/07/preventing-woke-ai-in-the-federal-government/`.

West European Politics Authors. Left-authoritarians and policy representation in Western Europe: Electoral choice across ideological dimensions. *West European Politics*, 36(6), 2013.

Tianqi Xiao et al. On the algorithmic bias of aligning large language models with RLHF: Preference collapse and matching regularization. *arXiv preprint arXiv:2405.16455*, 2024.

Lianmin Zheng et al. Judging LLM-as-a-Judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

A Topic Corpus Details

The 110 topics span nine domains: Politics and Governance (15), Healthcare and Science (12), Economics and Labor (14), Environment and Energy (11), Justice and Inequality (13), Gender and Sexuality (12), Culture and Media (10), Philosophy and Religion (11), and Global Affairs (12).

B Cross-Provider Benchmark

Extended comparisons from Anthropic's November 2025 even-handedness evaluation: Gemini 2.5 Pro (97%), Grok 4 (96%), Claude Opus 4.1 (95%), Claude Sonnet 4.5 (94%), GPT-5 (89%), Llama 4 (66%).