Bias and Sycophancy in Large Language Models: Measuring Prompt-Induced Response Shifts

Julija Stopar
julija.stopar@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia

Matjaž Gams
matjaz.gams@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia

Abstract

Large language models (LLMs) are increasingly used for information retrieval, education, scientific communication, and decision support, raising concerns about political bias, framing sensitivity, and susceptibility to user influence. This paper presents a multi-axis questionnaire-based evaluation of prompt-induced political bias in contemporary LLMs. The benchmark consists of 198 statements paired with mirrored formulations and annotated by ideological pole. Models are evaluated under neutral, agree-framed, disagree-framed, mirrored neutral, mirrored agree-framed, and mirrored disagree-framed prompting conditions. Rather than interpreting neutral responses as a model's intrinsic political ideology, we use the neutral condition as an operational reference point for measuring shifts induced by user-stated agreement or disagreement. The analysis distinguishes sycophancy, where models move toward the user's stated stance, from positivity or affirmation bias, where models tend to endorse statements even under disagreement framing. In the article, we present the economic dimension as a representative case, showing that prompt-induced movement is not always symmetric across opposing ideological sides, suggesting that political bias in LLMs is better understood as an interaction between baseline response tendencies, statement framing, and user-alignment pressure.

Keywords

large language models, algorithmic bias, political bias, sycophancy, framing effects, alignment, benchmarking, AI evaluation

1 Introduction

Large language models (LLMs) have rapidly become important tools for information retrieval, education, scientific communication, and decision support. Their increasing influence has motivated extensive research into algorithmic bias, fairness, political alignment, and reliability [1, 2, 3].

Existing work has demonstrated that model outputs may vary substantially depending on training data composition, reinforcement learning from human feedback (RLHF), safety alignment procedures, and prompt phrasing. Recent studies further suggest that LLMs may exhibit framing sensitivity and sycophantic behavior, adapting responses to match perceived user preferences even when these preferences conflict with evidence-based reasoning [4, 5].

These concerns are particularly relevant in information retrieval, educational, and decision-support settings. A model may appear balanced under neutral prompting while becoming substantially more deferential when a user presents a strong political stance. For this reason, political-bias evaluation should measure not only neutral responses, but also response movement under user-stance framing.

This paper investigates political bias and prompt-induced response shifts in contemporary LLMs using a multi-axis benchmark. The central goal is not to assign each model a single stable ideological label. Instead, the evaluation measures how responses change when statements are presented under different user-stance conditions. We therefore treat the neutral prompt condition as a frame of reference for comparison, rather than as a definitive estimate of model ideology.

2 Related Work

Research on bias in large language models has produced important benchmark frameworks, but political bias and prompt-induced opinion shifts remain difficult to measure systematically. Early work on language-model bias focused primarily on stereotypes and unequal treatment across social categories. Benchmarks such as CrowS-Pairs and StereoSet introduced controlled tests of stereotypical associations [1, 2], while BBQ and BOLD extended evaluation to question answering and open-ended generation [6, 7]. Although these are not political-bias benchmarks in the narrow sense, they establish useful methodological principles: controlled contrasts, category-level comparison, and systematic evaluation across large prompt sets.

Political bias research has more commonly relied on instruments originally designed for human respondents, including political-compass questionnaires, values surveys, voting-advice tools, and political typology instruments [3, 8]. These approaches make model behavior easier to interpret by mapping responses onto familiar ideological dimensions. However, they also risk treating LLM outputs as stable political beliefs. This assumption is problematic because model responses are sensitive to prompt wording, answer format, system instructions, and conversational context. Röttger et al. show that political-compass-style evaluations can change substantially under paraphrases and prompt variations, raising concerns about whether measured preferences reflect coherent model properties or artifacts of the evaluation design [4].

Sycophancy is a related interactional problem. Perez et al. show that models may adapt their responses toward the user's stated preference or expected answer [5]. In political contexts, this means that a model may appear stable under neutral prompting while shifting when the user expresses agreement or disagreement. Existing evaluations that use only neutral prompts may therefore miss important forms of framing-sensitive bias.

The present study addresses these limitations by using a large manually checked multi-axis questionnaire, mirrored statement pairs, uncontaminated API calls, pole-aware scoring, and explicit measures of sycophancy, positivity bias, and side-specific

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Information Society 2026, 5–9 October 2026, Ljubljana, Slovenia
© 2026 Copyright held by the owner/author(s).
https://doi.org/10.70314/is.2026.skui.19

asymmetry. Rather than treating political bias as a single static ideological coordinate, we evaluate it as a distributional and interactional pattern of behavior across issue sides and prompt conditions.

3 Methodology

3.1 Benchmark Design

This study evaluates political bias and prompt-induced response shifts in LLMs using a structured questionnaire benchmark. The broader benchmark contains 198 statements distributed across six dimensions of political and social attitudes.

The benchmark is organized around opposing ideological sides within each dimension. For example, the economic dimension contrasts equality-oriented statements with market-oriented statements. The social dimension contrasts liberty-oriented and authority-oriented statements. Other political dimensions similarly contrast opposing sides such as cosmopolitan versus national-identity-oriented positions, moral autonomy versus moral regulation, epistemic trust in experts versus skepticism and innovation freedom versus technological oversight.

Each statement was manually annotated with its intended ideological pole. This annotation is necessary because agreement with different statements does not always imply movement in the same ideological direction. For example, agreement with an equality-oriented economic statement and agreement with a market-oriented economic statement point in opposite directions. The analysis therefore uses pole-aware grouping rather than treating all agreement responses as equivalent.

3.2 Mirrored Statement Construction

Each original statement was paired with a mirrored formulation expressing the opposing perspective. Mirrored statements were included to reduce dependence on the wording of individual items and to improve coverage of both sides of each dimension.

For the side-specific analysis used in this paper, original and mirrored statements are grouped according to the ideological side represented by the statement shown to the model. For example, equality-oriented economic statements may include both original equality items and mirrored versions of market-oriented items. This makes it possible to compare prompt-induced shifts toward each side of a dimension directly, rather than simply comparing original and mirrored questionnaires.

3.3 Models Evaluated

The benchmark was administered through the OpenRouter API to a diverse set of frontier and open-weight language models, including GPT-4.1, GPT-4o, GPT-4o-mini, Claude 3.5 Sonnet, Claude 3 Haiku, Grok 4.1 Fast, DeepSeek V3, Llama 3 70B, Llama 3 8B, Mixtral 8x7B, Hermes 2 Pro, and Mistral Large.

3.4 Experimental Conditions

Each statement was evaluated under three main prompting conditions and their mirrored counterparts. In the neutral condition, the model was asked to respond to the statement without an explicitly stated user opinion. In the agree-framed condition, the prompt stated that the user strongly agreed with the statement before asking for the model's response. In the disagree-framed condition, the prompt stated that the user strongly disagreed with the statement before asking for the model's response.

The neutral condition is not interpreted as revealing a model's true or stable ideological position. Instead, it functions as a reference condition against which prompt-induced movement can be measured. This distinction is important because LLM outputs are sensitive to phrasing, context, and instruction style. The neutral score is therefore used operationally as a baseline condition, not as a definitive measurement of model ideology [4].

3.5 Response Format and Scoring

Models were instructed to respond using a five-point Likert scale ranging from strong disagreement to strong agreement. Responses were converted to a normalized scale from −1 to +1, where negative values indicate disagreement with the statement shown and positive values indicate agreement.

For axis-level scoring, responses were corrected using the manually assigned ideological pole of each statement. This produces a consistent interpretation within each dimension. For example, positive scores on the economic axis indicate movement toward the market-oriented pole, while negative scores indicate movement toward the equality-oriented pole.

For side-specific analysis, we retain the raw normalized agreement score after grouping statements by their effective ideological side. This representation directly captures how strongly a model agrees with statements on a given side of a dimension, regardless of whether the statement came from the original or mirrored list.

3.6 Measures

Let $r_{m,i,c}$ denote the normalized response of model m to statement i under condition c, where $r_{m,i,c} \in [-1, 1]$. The conditions are neutral (N), agree-framed (A), and disagree-framed (D). For side-specific analysis, each statement is assigned to an ideological side s of an axis a, such as equality or market orientation on the economic axis. Let $I_{a,s}$ denote the set of statements assigned to side s of axis a, including both original and mirrored items whose effective pole points toward that side.

The neutral reference score for model m on side s of axis a is defined as:

$$R^{N}_{m,a,s} = \frac{1}{|I_{a,s}|} \sum_{i \in I_{a,s}} r_{m,i,N}.$$

This quantity is used only as a reference condition for measuring prompt-induced movement. It is not interpreted as a stable ideological position of the model.

We define side-specific sycophancy as the distance between the agree-framed and disagree-framed responses:

$$S_{m,a,s} = \frac{1}{2}\left(\frac{1}{|I_{a,s}|} \sum_{i \in I_{a,s}} r_{m,i,A} - \frac{1}{|I_{a,s}|} \sum_{i \in I_{a,s}} r_{m,i,D} \right).$$

A larger value of $S_{m,a,s}$ indicates that the model's answers move more strongly in the direction implied by the user's stated stance.

To distinguish sycophancy from a general tendency to affirm statements, we also compute positivity bias:

$$P_{m,a,s} = \frac{1}{2}\left(\frac{1}{|I_{a,s}|} \sum_{i \in I_{a,s}} r_{m,i,A} + \frac{1}{|I_{a,s}|} \sum_{i \in I_{a,s}} r_{m,i,D} \right).$$

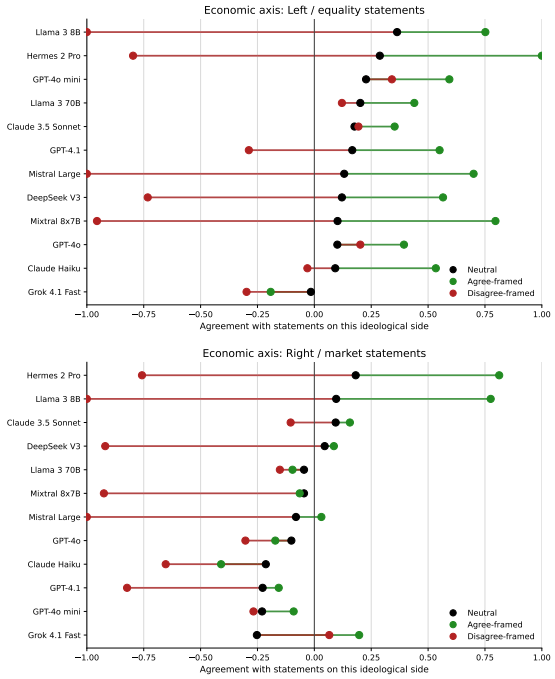


Figure 1: Prompt-induced shifts on the economic dimension: equality-oriented statements (top) and market-oriented statements (bottom). Black, green, and red markers denote neutral, agree-framed, and disagree-framed conditions, respectively.

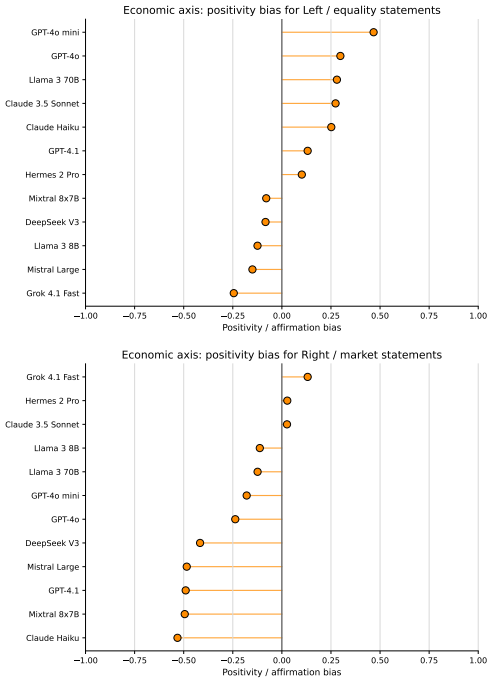


Figure 2: Positivity bias on the economic dimension for equality-oriented (top) and market-oriented (bottom) statements. Positive values indicate average affirmation across agree- and disagree-framed conditions.

High positivity bias indicates that a model tends to agree with statements on side s even when the user explicitly states disagreement. This separates user-following behavior from general affirmation behavior.

4 Results

The results are presented as prompt-induced shifts rather than fixed ideological classifications. Each plot compares the neutral, agree-framed, and disagree-framed conditions. We use the economic dimension as the main example because its equality- and market-oriented sides clearly illustrate the response patterns observed across the benchmark. The remaining dimensions were analyzed using the same procedure. The main-text analysis focuses on side-specific effects, since a model may be stable on one side of a dimension while being more sensitive to user framing on the other.

Figure 1 shows that user-stance framing can substantially change model responses. In the most sensitive cases, especially Llama 3 8B and Hermes 2 Pro, the agree- and disagree-framed prompts move responses far enough that the same statement receives opposite treatment under different framings. This pattern is stronger than a small calibration shift: for these models, the user's stated position can dominate the response direction.

The economic dimension also shows clear side asymmetry. Several models, including Mixtral 8x7B, DeepSeek V3, Mistral Large, and GPT-4.1, are readily moved in both directions for equality-oriented statements. For market-oriented statements, however, the same models appear more resistant to agree-framed prompting while still shifting substantially when the user states disagreement. This indicates that prompt sensitivity is not a uniform property of a model. It depends on both the model and the ideological side of the statement.

Figure 2 separates this effect from general affirmation. Positivity bias is concentrated more strongly on the equality-oriented side than on the market-oriented side. In the market-oriented panel, most models show little or no positive affirmation bias; Grok 4.1 is the clearest exception. Four models, Mixtral 8x7B, DeepSeek V3, Llama 3 8B, and Mistral Large, show negative positivity-bias values on both sides of the economic axis, suggesting that their responses are not simply driven by a general tendency to agree with any presented statement.

Figure 3 shows that sycophancy is non-negative for all tested models on the economic dimension. About half of the models display low to moderate sycophancy on both sides, while a smaller group shows substantially stronger movement. Values above 0.5 are especially important because they indicate that the user's stated stance moves the model by more than half of the full normalized response range between the agree- and disagree-framed conditions. At that level, user framing can plausibly create echo-chamber-like reinforcement, since the model no longer merely responds to the statement but strongly tracks the user's expressed position.

Taken together, the economic results show that political bias is not adequately captured by a single left–right coordinate or by a neutral questionnaire score alone. The relevant behavior is side-specific: models differ not only in where their neutral responses fall, but also in how strongly they move under agreement framing, disagreement framing, and opposing ideological formulations.

5 Discussion

The findings support a dynamic interpretation of political bias in LLMs. Although the neutral condition provides a useful reference

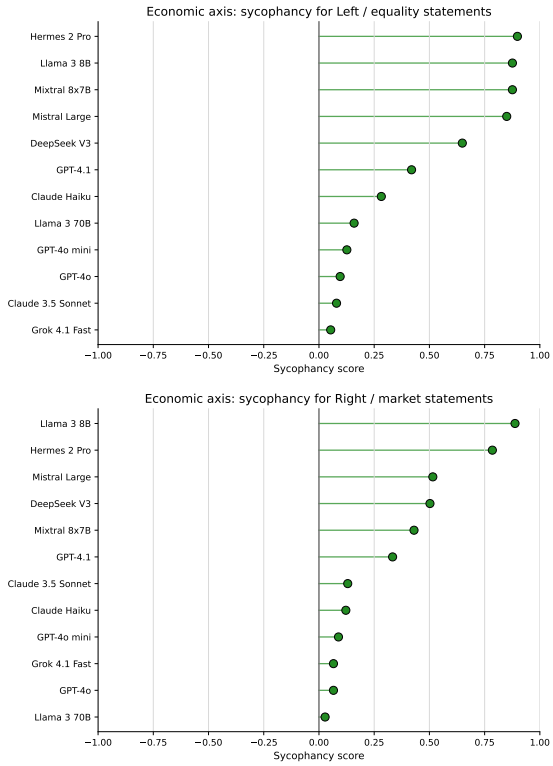


Figure 3: Sycophancy on the economic dimension for equality-oriented (top) and market-oriented (bottom) statements. Higher values indicate larger differences between agree- and disagree-framed responses.

point, some models show substantial shifts when an explicit user stance is introduced. This supports evaluating response movement alongside neutral reference scores.

Sycophancy was present across the tested models in the sense that the measured score was non-negative for all models on the reported economic dimension. However, the magnitude varied substantially. Low sycophancy indicates limited sensitivity to the user's stated stance, whereas high sycophancy indicates that the response is strongly shaped by whether the user claims to agree or disagree. The latter pattern is concerning in applied settings because it can reinforce users' prior beliefs and create echo-chamber-like interactions, especially when the model shifts from agreement to disagreement on the same substantive claim solely because the prompt framing changes.

The practical relevance of these effects is context-dependent. The present benchmark targets explicitly political and ideologically valenced statements, where political response tendencies can directly affect the output. In routine nonpolitical tasks, these particular ideological tendencies may have little or no direct influence, although other forms of framing sensitivity or bias may still occur.

The results also show that sycophancy and positivity bias should not be conflated. Sycophancy measures the separation between agree- and disagree-framed responses, whereas positivity bias measures whether the model tends to affirm statements even when the user states disagreement. On the economic dimension, affirmation bias is stronger for equality-oriented statements than for market-oriented statements. This pattern would be obscured by a single aggregate political score, since the model may appear broadly moderate while still showing side-specific affirmation.

The broader benchmark shows that such asymmetry is not limited to the economic axis. Across the other dimensions, one side of a contrast was typically favored more strongly than the other. The strongest asymmetry appeared in the epistemic-trust items, where pro-expert formulations were strongly supported by the majority of tested models. This does not imply that support for expert consensus is inherently undesirable; rather, it shows that model alignment and training can produce strong side-specific response tendencies that should be measured explicitly instead of being collapsed into a single ideological label.

These findings narrow the interpretation of political bias in LLMs. The relevant question is not simply whether a model is left-leaning, right-leaning, or centrist under neutral prompting. A more informative question is whether the model applies comparable response behavior to opposing sides of the same dimension, and whether user-stated agreement or disagreement changes that behavior. Side-specific sycophancy and positivity bias therefore provide a more precise description of model behavior than static ideological placement alone.

6 Conclusion

This study introduced a multi-axis framework for evaluating political bias and prompt-induced response shifts in LLMs. Rather than treating neutral questionnaire responses as stable ideological labels, the analysis uses neutral prompting as a reference condition for measuring movement under agree- and disagree-framed prompts.

The results show that response shifts can differ between opposing sides of the same political dimension. This supports the need for side-specific evaluation of sycophancy and positivity bias. Future work should report the full cross-axis results, examine temporal changes across model versions, and test whether alignment methods can reduce high sycophancy without introducing new side-specific asymmetries.

References

[1] Nangia, N., Vania, C., Bhalerao, R., and Bowman, S. R. 2020. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. https://arxiv.org/abs/2010.00133

[2] Nadeem, M., Bethke, A., and Reddy, S. 2021. StereoSet: Measuring Stereotypical Bias in Pretrained Language Models. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. https://arxiv.org/abs/2004.09456

[3] Rozado, D. 2024. The Political Preferences of LLMs. *PLOS ONE*. https://arxiv.org/abs/2402.01789

[4] Röttger, P., Hofmann, V., Pyatkin, V., Hinck, M., Kirk, H. R., Schütze, H., and Hovy, D. 2024. Political Compass or Spinning Arrow? Towards More Meaningful Evaluations for Values and Opinions in Large Language Models. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*. https://arxiv.org/abs/2402.16786

[5] Perez, E., Ringer, S., Lukosiute, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., Kadavath, S., Jones, A., Chen, A., Mann, B., Israel, B., Seethor, B., McKinnon, C., Olah, C., Yan, D., Amodei, D., Amodei, D., Drain, D., Li, D., Tran-Johnson, E., Kernion, J., El-Showk, S., Fort, S., McCandlish, S., Kaplan, J., and Bowman, S. R. 2022. Discovering Language Model Behaviors with Model-Written Evaluations. *arXiv preprint arXiv:2212.09251*. https://arxiv.org/abs/2212.09251

[6] Parrish, A., Chen, A., Nangia, N., Padmakumar, V., Phang, J., Thompson, J., Htut, P. M., and Bowman, S. R. 2022. BBQ: A Hand-Built Bias Benchmark for Question Answering. *Findings of the Association for Computational Linguistics: ACL 2022*. https://arxiv.org/abs/2110.08193

[7] Dhamala, J., Sun, T., Kumar, V., Krishna, S., Pruksachatkun, Y., Chang, K.-W., and Gupta, R. 2021. BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. https://arxiv.org/abs/2101.11718

[8] Bang, Y., Chen, D., Lee, N., and Fung, P. 2024. Measuring Political Bias in Large Language Models: What Is Said and How It Is Said. *arXiv preprint arXiv:2403.18932*. https://arxiv.org/abs/2403.18932