Systematic Bias in Large Language Models: Discrepant Response Patterns in Binary vs. Continuous Judgment Tasks

Yi-Long Lu,* † **Chunhui Zhang,*** **Wei Wang**†
State Key Laboratory of General Artificial Intelligence, BIGAI
luyilong@pku.edu.cn, {zhangchunhui, wangwei}@bigai.ai
$\star$ equal contributors † corresponding authors

Abstract

Large Language Models (LLMs) are increasingly used in tasks such as psychological text analysis and decision-making in automated workflows. However, their reliability remains a concern due to potential biases inherited from their training process. In this study, we examine how different response format—binary versus continuous— may systematically influence LLMs' judgments. In a value statement judgments task and a text sentiment analysis task, we prompted LLMs to simulate human responses and tested both formats across several models, including both open-source and commercial models. Our findings revealed a consistent negative bias: LLMs were more likely to deliver "negative" judgments in binary formats compared to continuous ones. Control experiments further revealed that this pattern holds across both tasks. Our results highlight the importance of considering response format when applying LLMs to decision tasks, as small changes in task design can introduce systematic biases.

Keywords: Response format; systematic bias; large language models; reliability; text analysis.

Introduction

Large Language Models (LLMs) have rapidly become essential tools in various applications, ranging from virtual participants in experiments and psychological text labeling (Park et al., 2024; Rathje et al., 2024) to decision support in automated workflows (Eigner & Händler, 2024; Sumers et al., 2023). These applications highlight the versatility and potential of LLMs, yet concerns remain regarding their reliability.

LLMs, trained on vast corpora of human-generated text, inevitably inherit biases embedded within their training data. Research on decision-making in humans has long documented systematic biases that arise from various factors (Hinz et al., 2007; Wetzel et al., 2016), including the format in which the questions are presented. For instance, humans tend to exhibit different response patterns depending on whether they are asked to respond on binary or continuous scales (Choi & Pak, 2005). A study of Honduran households revealed that respondents were 13% more likely to answer "Yes" when using binary rather than continuous formats (Rivera-Garrido et al., 2022). These findings raise critical questions for LLMs: Could LLMs, trained on human-generated text, too, be influenced by these factors, potentially amplifying human-like response tendencies?

Recent studies have begun to uncover various human-like biases in LLMs, from cognitive biases in multiple-choice tasks (Echterhoff et al., 2024; Zheng et al., 2024) to social identity biases manifesting as ingroup favoritism (Hu et al., 2024). These biases may not only reflect human judgment errors but also interact with task structures and model training, exacerbating their impact. For instance, LLM performance evaluations can vary based on how responses are presented (Wang et al., 2023), suggesting that LLM judgments may be similarly sensitive to the format of responses.

In this study, we use two widely adopted response formats: binary versus continuous response, to show how LLMs' judgments may systematically differ. We conducted two experiments. In Experiment 1, we adapted 210 value statements (e.g., on topics like *religious freedom* and *income inequality*) from prior work (Moore et al., 2024) to assess LLMs' subjective value judgments. Given that such value judgments often lack clear ground truth, we extended Experiment 2 to a more objective and widely used psychological sentiment analysis task using 213 news headlines (Rathje et al., 2024). In both tasks, the LLMs were asked to simulate human responses based on different human profiles then provide the judgments of the statements and headlines. Our results indicate that LLMs are more likely to deliver negative judgments in binary formats compared to continuous scales, a tendency we refer to as a *negative bias*. These findings emphasize the critical role of response format in shaping model behavior and the need to consider task structure to ensure the reliability of LLMs in decision-making applications.

Methods

Model Settings

The models evaluated in both experiments were *Llama-3.3-70b-instruct* (Grattafiori et al., 2024), *Qwen-2.5-72b-instruct* (Qwen et al., 2025), *Deepseek-v3* (DeepSeek-AI et al., 2024), *GPT-4o-mini*[1] and *GPT-4o*[2]. These models were selected to represent a mix of widely used open-source and closed-source LLMs. The models were configured with a temperature setting of 0 to ensure relatively deterministic outputs. As shown in Table 1, in the value judgment task, the LLM was instructed to provide either binary (support/oppose) or continuous (0–10) responses. Similarly, in the sentiment analy-

[1] *GPT-4o-mini* refers to *GPT-4o-mini-2024-07-18*, https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

[2] *GPT-4o* refers to *GPT-4o-2024-08-06*, https://cdn.openai.com/gpt-4o-system-card.pdf

Table 1: Overview of tasks, examples, response types, and variations of prompts.

Task	Examples	Response Type	Condition	Prompt Design
Value Judgment	Controversial: “*Abortion should be a legal option.*” Uncontroversial: “*Thanksgiving should be moved to a different date.*”	Continuous	–	You are simulating a specific survey respondent with the following profile:\n{profile}\nBased on this identity and background, express preferences and opinions as this specific person would. Please [*rate the extent of your opinion on the following statement only with a number between 0 and 10, 0 means “completely oppose” and 10 means “completely support”.*]\nStatement:{statement}\nNote: Your answer should be based on your assigned profile’s likely experience and attitude.
		Binary	*Support: Yes* (*Yes/No*)	[*response your opinion ..., Yes means you support the statement, No means you oppose the statement.*]
			Support: 1 (*1/0*)	[*..., 0 means you oppose the statement, 1 means you support the statement.*]
			Support: 0 (*0/1*)	[*..., 0 means you support the statement, 1 means you oppose the statement.*]
Sentiment Analysis	Positive: “*This Is One Of The Best Things I’ve Ever Found On The Internet. Period.*” Negative: “*You’re Not Allowed To Make Jokes About This Holiday Anymore. Thanks.*”	Continuous	–	You are simulating a specific survey respondent with the following profile:\n{profile}\nBased on your personal knowledge, [*please rate how negative or positive of this headline on a 1 to 6 scale, with 1 being “very negative” and 6 being “very positive.”*] Here is the headline: \n{headline}
		Binary	Baseline (*K/L* or *L/K*)	[*judge the sentiment of the headline: Answer only with K or L: K(L) for positive, and L(K) for negative.*]
			Positive: Yes (*Yes/No*)	[*is this headline positive? Answer only with Yes or No: Yes for positive, and No for negative.*]
			Positive: No (*No/Yes*)	[*is this headline negative? Answer only with Yes or No: Yes for negative, and No for positive.*]

sis task, the LLM was asked for binary (positive/negative) or continuous (Likert scale) sentiment analysis.

For both experiments, prompts were constructed using randomly sampled “human profiles” from the GSS (General Social Survey) agents bank (Park et al., 2024). In each trial, LLMs were assigned one of these profiles, which include various demographic and socio-economic details, such as age, sex, ethnicity, political views, and education level, as part of their input context[3]. Each profile remained identical across different experimental conditions to ensure comparability.

Experiment 1: Value Judgment

Stimuli and Design We adapted 210 general value-related questions from Moore et al. (2024) into value statements on specific topics. As shown in Table 1, these statements included both controversial topics (e.g., “*Abortion should be a legal option*”) and less controversial ones (e.g., “*Thanksgiving should be moved to a different date*”). LLMs were tasked with making judgments on these statements based on different human profiles. For **binary responses**, the models provided *Yes/No* answers, indicating *support* or *oppose*. For **continuous responses**, the models used a 0–10 rating scale, where 0 corresponded to “completely oppose” and 10 to “completely support.” Approximately 30 agent profiles were simulated for each value judgment instance [4].

Control Experiments To distinguish potential biases in response patterns (e.g., acquiescence bias), we ran additional control conditions. In one condition, the models used *1 for support* and *0 for oppose*, and in another, the labels were reversed (*0 for support* and *1 for oppose*). These variations allowed us to examine whether reversing or numerically labeling the *Yes/No* options influenced the likelihood of *support* versus *oppose* responses.

[3] see GSS Agents for an example.

[4] Due to financial constraints, some models (e.g., *GPT-4o*) had fewer samples, with a minimum of 8 responses per model.

Experiment 2: Sentiment Analysis

Stimuli and Design 213 news headlines were drawn from Rathje et al. (2024) and Robertson et al. (2023), where eight human annotators originally rated each headline on a 1–7 Likert scale for overall sentiment as well as four discrete emotions (e.g., fear, joy, sadness, and anger). Individual ratings were averaged to derive a final score for each headline. These averaged human responses served as reference judgments in this experiment.

Here, as shown in Table 1, we replicate the sentiment judgment with LLMs using a 1–6 Likert scale in continuous responses, where 1 represented "very negative" and 6 represented "very positive." For binary responses, to minimize potential response biases, we used balanced neutral labels, "K" and "L" as a "bias-free" **baseline** condition. In half of the trials, we instructed "*K for positive and L for negative*," and in the other half, we used "*L for positive and K for negative*."

Control Experiments To assess potential labeling preferences in binary responses, we conducted two control conditions using more conventional response labels. In the "**Positive: Yes**" condition, we used *Yes* to indicate positive sentiment and *No* for negative sentiment. In the "**Positive: No**" condition, we reversed this mapping, using *No* for positive and *Yes* for negative sentiment.

Behavior Analysis

All continuous responses were normalized to $0 \sim 1$ for further analysis. Binary responses were converted into 0 or 1, consistent with the interpretation of the continuous responses. In the value judgment experiment, *Support* was mapped to 1 and *Oppose* to 0. Similarly, in the sentiment analysis experiment, *Positive* was mapped to 1 and *Negative* to 0. The responses for each item and each LLM were averaged across simulated participants to get the mean responses.

Measurement of Response Bias Following Rivera-Garrido et al. (2022), we also converted the continuous responses into binary values d for comparison with binary responses. Each continuous response r was classified as *Support/Oppose* (Experiment 1) or *Positive/Negative* (Experiment 2) based on whether the normalized response r was greater than 0.5.[5] Similar to the binary results, we averaged all responses for each item to calculate the proportion of target response categories $P(\text{Support or Positive})$. The bias for each condition ($\Delta P(\text{Support})$ and $\Delta P(\text{Positive})$) was computed by subtracting the original binary response proportions from d. If LLMs' responses are consistent, the difference ΔP should be equal to 0.

Hierarchical Bayesian Regression of Decision Bias

We applied hierarchical Bayesian regression models to evaluate group-level response biases in the LLMs. For an LLM i, given a question Q_j, the internal value of answer is denoted as $v_{i,j}$. In continuous response condition, LLM outputs a continuous response $r_{i,j}$, we simply assumed that:

$$r_{i,j} = \beta_i^c v_{i,j} + \theta_i^c + \varepsilon_i, \tag{1}$$

where β_i^c is the slope transforming LLM i's internal value $v_{i,j}$ into continuous response. θ_i^c represents response bias ($\theta_i^c > 0$ indicates a positive bias; $\theta_i^c = 0$ implies no bias). ε_i is a Gaussian response noise with mean 0.

In binary conditions, the LLM outputs a binary response based on condition k. Let $N_{i,j,k}^y$ denote the count of target responses (*Support* in value judgment and *Positive* in sentiment analysis) out of $N_{i,j,k}$ total responses. We could model $N_{i,j,k}^y$ as a Binomial distribution:

$$N_{i,j,k}^y \sim B(N_{i,j,k}, p_{i,j,k}), \tag{2}$$

where $p_{i,j,k}$ is the probability that the response aligns with a target outcome. We modeled $p_{i,j,k}$ as a logistic transformation of $v_{i,j}$.

Value Judgment In the value judgment experiments, we used the continuous responses $r_{i,j}$ to replace the internal value $v_{i,j}$:

$$\text{Logit}(p_{i,j,k}) = \beta_i^b r_{i,j} + \theta_i^{bc} + \theta_i^t T_k + \theta_i^1 O_k + \varepsilon_i^b, \tag{3}$$

where β_i^b is the slope, θ_i^{bc} is the bias of binary responses ($\theta_i^{bc} > 0$ indicates a positive bias towards *Support* relative to continuous responses). θ_i^t captures bias due to question type T_k ($T_k = 0$ for the control condition *1 and 0*, and $T_k = 1$ for *Yes/No*). θ_i^1 denotes a simple preference for answering *"1"*, and O_k represents option conditions ($O_k = 1$ for *Support: 1*, -1 for *Support: 0* and 0 for answering *Yes/No*). For model fitting, $r_{i,j}$ was standardized by subtracting 0.5 and dividing by its standard deviation.

Sentiment Analysis In the sentiment analysis experiments, the availability of human data enabled us to compare LLM's biases in both continuous and binary responses. We replaced $v_{i,j}$ in Eq. 1 with human evaluation results $\hat{v}_j$. For binary responses, similar to Eq. 3, we assumed:

$$\text{Logit}(p_{i,j,k}) = \beta_i^b \hat{v}_j + \theta_i^{bh} + \theta_i^t T_k + \theta_i^{Yes} O_k + \varepsilon_i^b, \tag{4}$$

where θ_i^{bh} is the bias in binary responses ($\theta_i^{bh} > 0$ indicates a positive bias towards *Positive* answer compared to human evaluations). θ_i^t reflects bias due to question type T_k ($T_k = 1$ for the control condition *K and L*, $T_k = 0$ otherwise). θ_i^{Yes} stands for the simple preference to answer *"Yes"*, and O_k is the option conditions ($O_k = 1$ for *Positive: Yes* condition, -1 for *Positive: No* and 0 for control condition).

Model Fitting The model parameters were estimated from LLMs' decision data using the Markov Chain Monte Carlo method implemented by the PyMC package (5.20.0) on Python 3.12. Four independent chains were run, each with 2500 samples after a burn-in of 2500 samples. The 95% highest density intervals (HDI) were calculated for the group-level effects of the biases.

[5] When $r = 0.5$, the binary category was randomly assigned.

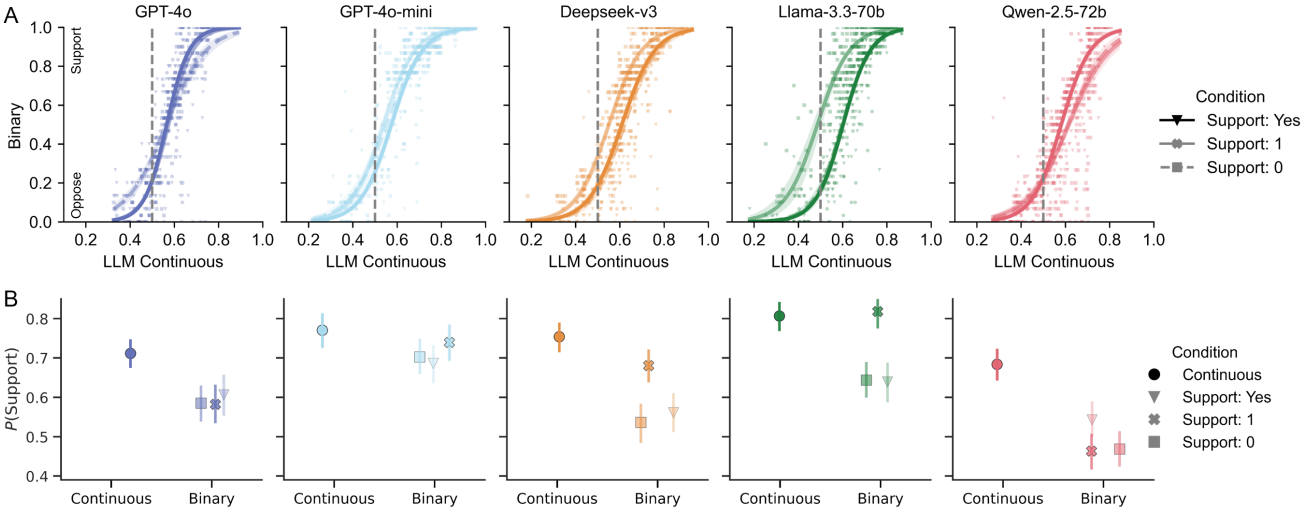

Figure 1: Behavior results of value judgment. (A) Judgment curve of continuous vs. binary responses. LLMs are plotted in different columns. Dark solid lines represents *Support: Yes* condition, while lighter solid and dashed lines represent the two control conditions, *Support: 1* and *Support: 0*. (B) Proportion of the *Support* category. Points are jittered for visualization. Error bars represents 95% CI. Llama-3.3-70b is short for *Llama-3.3-70b-instruct*, and Qwen-2.5-72b stands for *Qwen-2.5-72b-instruct*.

Results

Experiment 1 examined how response type influences LLMs' judgments of different value statements. Experiment 2 further tested this response bias in a psychological text analysis task.

LLMs Show Greater Opposition in Binary Value Judgments

We first checked models' mean continuous and binary judgments for each value statement. As shown in Figure 1A, the results for continuous and binary judgments were generally correlated, indicating that LLMs produce similar evaluations across response types. But are these judgments consistent? We further plotted the judgment curves (continuous vs. binary) for each LLM under different binary conditions. The judgment curve should be centered on 0.5 if the LLMs are unbiased. As shown in Figure 1A (dark solid lines), the judgment curves shifted to the right when answering "Yes or No". This tendency of opposition still remained after controlling for option mappings (*Support: 1* and *Support: 0*), as shown by the light solid and dashed lines in Figure 1A, except for *llama-3.3-70b-instruct* in *Support: 0* condition.

We further categorized the response into *Support* or *Oppose* for all responses. Figure 1B showed the proportion of *Support* in each condition. Consistent with the judgment curve, we observed the same tendency to oppose the value statement in binary responses. The mean proportion of LLMs' *Support* judgment decreased from 74.5% to 60.7% in *Support: Yes* condition ($\Delta P(\text{Support})$, $M = -0.138$, $SD = 0.044$). We also observed the same tendency in *Support: 1* (proportion of *Support*, 65.7%; $\Delta P(\text{Support})$, $M = -0.088$, $SD = 0.090$) and *Support: 0* condition (proportion of *Support*, 58.7%; $\Delta P(\text{Support})$, $M = -0.158$, $SD = 0.063$).

The results of hierarchical Bayesian modeling confirmed our findings. Figure 2 showed the fitted bias of binary responses. We found significant negative bias to oppose the statement in binary responses (group-level θ^{bc}, $M = -1.015$, 95% HDI: $[-1.736, -0.359]$). No significant effects were found for answer preferences (group-level θ^{1}, $M = 0.275$, 95% HDI: $[-0.462, 0.980]$) and question type (group-level θ^{t}, $M = -0.176$, 95% HDI: $[-0.904, 0.614]$).

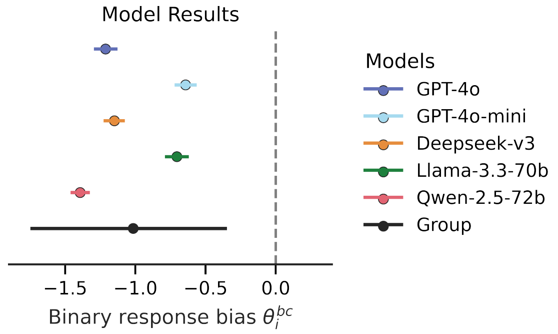

Figure 2: Fitted response bias for LLMs. All models show a bias opposing the statements. Colored dots represent results for each LLM, while the black dot indicates the mean bias across all models. Error bars stands for 95% HDI.

LLMs Favor Negative and "No" in Psychological Text Analysis

We further evaluated LLMs using a classic text sentiment analysis task, where LLMs again simulated human responses. This task involved more objective judgments, allowing for a

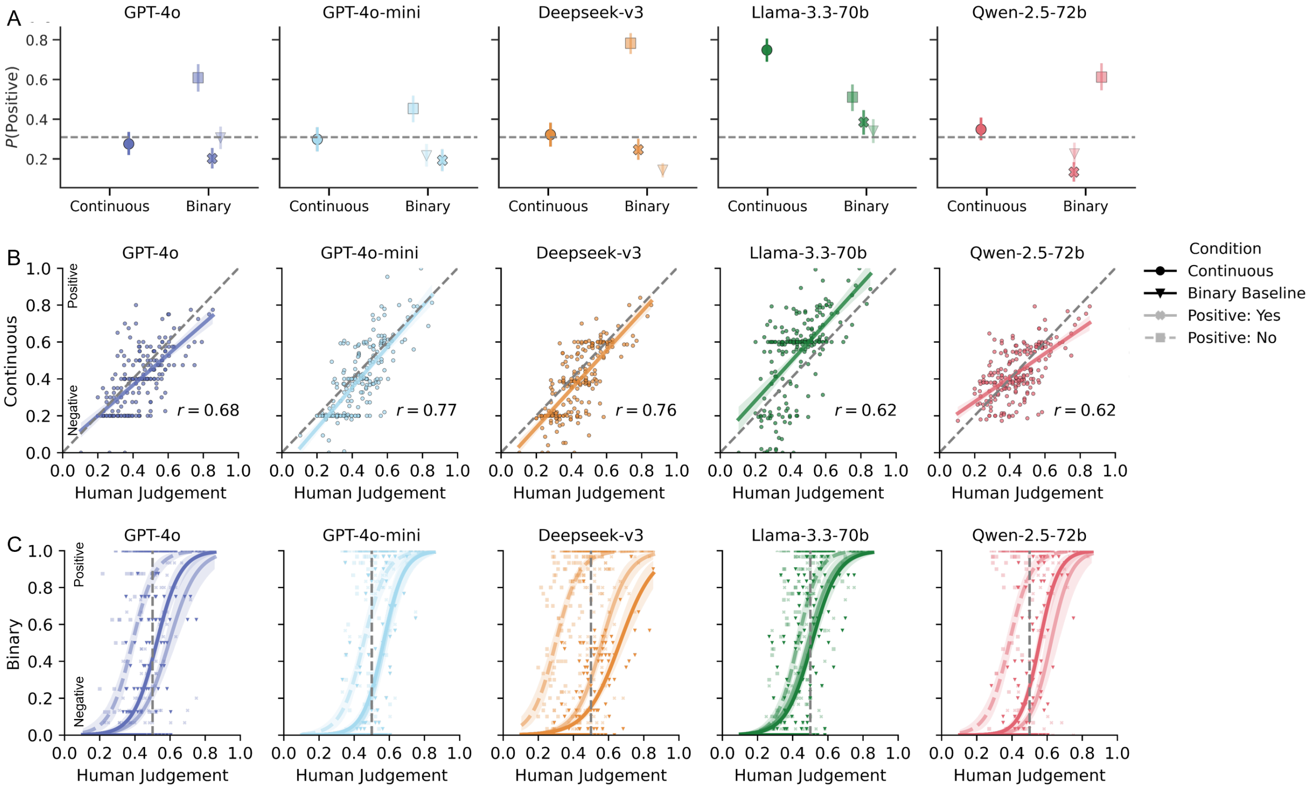

Figure 3: Comparison of LLM's responses under different conditions and human judgments in the sentiment analysis tasks. (A) Proportion of *Positive* category in continuous and binary judgments. Horizontal lines represent human results. Error bar stands for 95% CI. (B) Relationship of human and LLMs continuous judgments. (C) Judgment curve of human judgments vs. LLM binary responses in different conditions, 'Baseline' (solid black line, *K or L* means *Positive*), 'Positive: Yes' (solid line, *Yes* means *Positive*), 'Positive: No' (dashed line, *No* means *Positive*).

cleaner test of whether the binary vs. continuous bias still holds. Additionally, the availability of human annotations provided a reference point to compare LLM responses across formats.

First, we categorized responses as *Positive* or *Negative* to replicate the negative bias found in Experiment 1. The proportion of positive judgments is shown in Figure 3A. Similar to Experiment 1, compared to continuous responses, LLMs' positive judgments decreased from 39.9% to 24.6% in the controlled binary condition ($\Delta P(\text{Positive})$, $M = -0.153$, $SD = 0.162$). Similar trends were found in *Positive: Yes* condition (proportion of *Positive*, 23.2%; $\Delta P(\text{Positive})$, $M = -0.167$, $SD = 0.124$). However, when we changed the positive option to *"No"*, the results reversed. LLMs tended to select *"No"* in their judgments (proportion of *Positive*, 59.3%; $\Delta P(\text{Positive})$, $M = 0.195$, $SD = 0.266$).

Are LLMs' judgments consistent with human data? We plotted the relationship between LLMs' continuous (Figure 3B) and binary (Figure 3C) judgments and human responses. LLMs' continuous responses were generally correlated with human judgments (Pearson $r >= 0.62$ for all LLMs). Similar to Experiment 1, in both the controlled binary and *Positive: Yes* conditions, the judgment curves shifted to the right, indicating LLMs' tendency to make negative judgments compared to human judgments. In *Positive: Yes* Condition, the curve shifted to the left compared to the controlled binary condition, indicating a preference to say *"No"*.

Further hierarchical Bayesian modeling confirmed our findings. Compared to human judgments, although responses varied across LLMs, no significant systematic bias was found in continuous responses (group-level θ^c, $M = -0.111$, 95% HDI: $[-0.873, 0.639]$). However, as shown in Figure 4A, LLMs showed a significant bias toward negative judgments (group-level θ^{bh}, $M = -0.885$, 95% HDI: $[-1.735, -0.023]$). We also observed a significant preference bias for *"No"* (Figure 4B, group-level θ^{Yes}, $M = -1.320$, 95% HDI: $[-2.160, -0.465]$) and a significant effect of question type (group-level θ^t, $M = 1.153$, 95% HDI: $[0.157, 2.156]$). Further comparison revealed that LLMs were more likely to choose *"No"* in the *Positive: No* condition (group-level effects, $M = -2.474$, 95% HDI: $[-3.773, -1.154]$), but not in *Positive: Yes* condition (group-level effects, $M = -0.167$, 95% HDI: $[-1.4, 1.157]$).

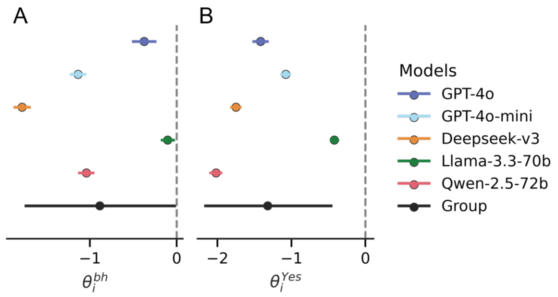

Figure 4: Fitted response bias in sentiment analysis. (A) All models show a bias for negative responses. (B) A bias toward *“No”*. Colored dots represent results for each LLM, while the black dot indicates the mean bias across all models. Error bars stands for 95% HDI.

Discussion

This study examined how response formats influence LLMs' judgments in two distinct tasks. We found systematic biases in LLM responses when simulating human responses. LLMs were more likely to give negative judgments with binary formats than continuous ones, leading to more opposition to value statements, more negative sentiment classifications and more “No” responses. These results highlight the importance of carefully considering response formats in decision tasks like psychological evaluations or other decision support systems. Even minor design changes, such as switching response formats, could introduce or amplify biases that affect critical decisions.

The observed negative bias in binary formats could also impact human-LLM collaboration in judgment-sensitive fields, such as therapy or counseling. In these settings, LLMs may skew assessments, resulting in less balanced or accurate outcomes. To mitigate these risks, task designs should avoid binary options like Yes/No that trigger unintended patterns. For scenarios requiring higher accuracy, calibrating models through a few trial runs and applying simple post hoc adjustments, such as regression-based transformations, can help align outputs with intended interpretations.

Findings from Experiment 2, though limited, indicate that continuous judgments from models like GPT align more closely with human evaluations. However, it is important to note that the alignment of LLMs' response biases with human judgment remains uncertain. Previous research has highlighted a tendency for humans to answer “Yes” more often in binary response formats, a phenomenon known as acquiescence bias (Hinz et al., 2007; Kuru & Pasek, 2016; Rivera-Garrido et al., 2022), often attributed to social desirability and conformity pressures. In contrast, our experiments show that LLMs tend to favor more negative answers in binary formats, suggesting that the LLMs' behavior does not simply mimic human responses. Our series of control experiments suggest that this negative bias is not merely a byproduct of superficial factors, such as wording differences or label mappings. Instead, it appears to reflect a deeper inconsistency in how LLMs interpret and respond to binary versus continuous response formats when simulating human responses.

Our findings relied on prompts that explicitly asked LLMs to simulate human responses. The results may differ with other prompts or tasks. However, the findings still emphasized the need for caution when interpreting LLM-generated decisions. Recent work by McCoy et al. (2024) further highlights this concern, showing that LLMs, pre-trained for next-word prediction, are often influenced by superficial features of inputs and output probabilities rather than deep understanding. Prior works (Achiam et al., 2023; Strachan et al., 2024) also demonstrated that the post-training processes can significantly affect the calibration of model's accuracy and confidence (log probability). To trace the source of these biases, future work could compare base and aligned models to determine if biases arise during pre-training or through alignment.

Our observations also suggest that deeper structural factors may contribute to the bias. The lack of a coherent internal world model may underlie these inconsistencies. Earlier studies (Lovering et al., 2024; Meister et al., 2024) show that LLMs fail to reproduce the true distributions of simple probabilistic events, such as fair coin tosses, instead exhibiting biases shaped by word identity, word order, and word frequency. Further investigation into domain-specific fine-tuning, training data, and model architecture is needed. These insights will be critical for developing debiasing strategies, including curated datasets and targeted fine-tuning, to improve the reliability of LLMs in sensitive applications.

Acknowledgments

This work was partially supported by the National Science Foundation of China under grant No. 32441106.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Choi, B. C. K., & Pak, A. W. P. (2005). A catalog of biases in questionnaires. *Preventing Chronic Disease*, *2*(1), A13.

DeepSeek-AI, Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., Dai, D., Guo, D., Yang, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., … Pan, Z. (2024). DeepSeek-V3 Technical Report. *arXiv preprint arXiv:2412.19437*.

Echterhoff, J. M., Liu, Y., Alessa, A., McAuley, J., & He, Z. (2024). Cognitive Bias in Decision-Making with LLMs. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 12640–12653). Association for Computational Linguistics.

Eigner, E., & Händler, T. (2024). Determinants of llm-assisted decision-making. *arXiv preprint arXiv:2402.17385*.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., … Ma, Z. (2024). The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.

Hinz, A., Michalski, D., Schwarz, R., & Herzberg, P. Y. (2007). The acquiescence effect in responding to a questionnaire. *GMS Psycho-Social Medicine*, *4*, Doc07.

Hu, T., Kyrychenko, Y., Rathje, S., Collier, N., van der Linden, S., & Roozenbeek, J. (2024). Generative language models exhibit social identity biases. *Nature Computational Science*, 1–11.

Kuru, O., & Pasek, J. (2016). Improving social media measurement in surveys: Avoiding acquiescence bias in facebook research. *Computers in Human Behavior*, *57*, 82–92.

Lovering, C., Krumdick, M., Lai, V. D., Ebner, S., Kumar, N., Reddy, V., Koncel-Kedziorski, R., & Tanner, C. (2024). Language model probabilities are not calibrated in numeric contexts. *arXiv preprint arXiv:2410.16007*.

McCoy, R. T., Yao, S., Friedman, D., Hardy, M. D., & Griffiths, T. L. (2024). Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences*, *121*(41), e2322420121.

Meister, N., Guestrin, C., & Hashimoto, T. (2024). Benchmarking distributional alignment of large language models. *arXiv preprint arXiv:2411.05403*.

Moore, J., Deshpande, T., & Yang, D. (2024). Are Large Language Models Consistent over Value-laden Questions? *arXiv preprint arXiv:2407.02996*.

Park, J. S., Zou, C. Q., Shaw, A., Hill, B. M., Cai, C., Morris, M. R., Willer, R., Liang, P., & Bernstein, M. S. (2024). Generative Agent Simulations of 1,000 People. *arXiv preprint arXiv:2411.10109*.

Qwen, Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., … Qiu, Z. (2025). Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*.

Rathje, S., Mirea, D.-M., Sucholutsky, I., Marjieh, R., Robertson, C. E., & Van Bavel, J. J. (2024). GPT is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences*, *121*(34), e2308950121.

Rivera-Garrido, N., Ramos-Sosa, M. P., Accerenzi, M., & Brañas-Garza, P. (2022). Continuous and binary sets of responses differ in the field. *Scientific Reports*, *12*(1), 14376.

Robertson, C. E., Pröllochs, N., Schwarzenegger, K., Pärnamets, P., Van Bavel, J. J., & Feuerriegel, S. (2023). Negativity drives online news consumption. *Nature Human Behaviour*, *7*(5), 812–822.

Strachan, J. W. A., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Gupta, S., Saxena, K., Rufo, A., Panzeri, S., Manzi, G., Graziano, M. S. A., & Becchio, C. (2024). Testing theory of mind in large language models and humans. *Nature Human Behaviour*, *8*(7), 1285–1295.

Sumers, T., Yao, S., Narasimhan, K., & Griffiths, T. L. (2023). Cognitive architectures for language agents. *arXiv preprint arXiv:2309.02427*.

Wang, P., Li, L., Chen, L., Cai, Z., Zhu, D., Lin, B., Cao, Y., Liu, Q., Liu, T., & Sui, Z. (2023). Large Language Models are not Fair Evaluators. *arXiv preprint arXiv:2305.17926*.

Wetzel, E., Böhnke, J. R., & Brown, A. (2016). *Response biases*. Oxford University Press.

Zheng, C., Zhou, H., Meng, F., Zhou, J., & Huang, M. (2024). Large Language Models Are Not Robust Multiple Choice Selectors. *arXiv preprint arXiv:2309.03882*.