OXFORD JOURNAL
OF STUDENT
SCHOLARSHIP

Investigating Anchoring Bias and Prompt-Based Mitigation in Large Language Models

Emily Ma
emilyma@outlook.com

ABSTRACT

Large language models are being increasingly integrated into tasks involving numerical estimation and decision-making. However, due to their training on human-generated text, they may be susceptible to various cognitive biases. This includes anchoring bias, a well-documented phenomenon in humans where judgements are illogically influenced by an irrelevant reference number.

This study investigated whether three large language models (Mistral-Large-3, DeepSeek-V3.1, and Meta's Llama-3.1-8B) exhibit anchoring bias when prompted with general-knowledge monetary estimation questions. We also tested whether two prompt-based mitigation strategies (explicit debias and estimate-first) can reduce the anchoring effect.

We hypothesized that larger models would show less susceptibility to anchoring bias as compared to smaller models and also respond better to mitigation strategies. Using a set of twenty-four monetary prompts, each with a high and low anchor variation, we quantified anchoring bias as the normalized percent difference between high and low anchor responses and determined statistical significance with paired t-tests.

We found that mean anchoring bias magnitudes differed across models, with Llama-3.1-8B showing the largest effect and Mistral-Large-3 the smallest. However, with high prompt-level variability, paired t-tests indicated that baseline differences were not statistically significant, despite the consistent directionality of responses. The effectiveness of mitigation strategies also differed across models, with explicit debias prompting significantly reducing anchoring bias for Mistral-Large-3, while having little effect on other models, and estimate-first prompting significantly reducing bias only for DeepSeek-V3.1.

These findings suggest that anchoring bias can influence LLM outputs but is inconsistent, while the effectiveness of prompt-based mitigation is model-dependent.

INTRODUCTION

Large language models are increasingly entrusted with a wide variety of tasks. The popularity of models such as OpenAI's ChatGPT has integrated large language models into many aspects of daily life, with applications including educational tutoring and writing support, software development assistance, and electronic health record comprehension (1-3). As LLMs are deployed in more and more real-world use cases, ensuring their reliability and consistency has become an important area of research to address.

Studies have shown that LLMs exhibit certain biases due to the usage of human-generated data in their training and fine-tuning. The majority of existing research has focused on social biases, where models discriminate against groups based on race, gender, and other characteristics. LLMs have been found to exhibit strong stereotypical biases, display gender bias in text-generation tasks, and possess prominent implicit biases based on attributes and naming (4-6). These studies highlight the fairness and ethical risks that might come with the deployment of LLMs.

Cognitive biases, on the other hand, have been less examined. Cognitive bias refers to when decisions are made that do not reflect those of a rational being. Examples include confirmation bias, where individuals prefer information that supports their preexisting beliefs (7); availability bias, where individuals overestimate the likelihood of events that are easier to recall (8); and loss aversion, in which individuals weigh losses more heavily than gains (9). Recent findings have indicated that LLMs are also susceptible to such biases. Studies using bias-inducing prompts showed reduced performance accuracy on medical exam questions, inconsistent financial forecasts, and illogically influenced college admission decisions (10-12).

Among cognitive biases, anchoring bias is particularly concerning because of its impact on numeric estimation, an important capability of LLMs. Anchoring bias is when an initial value is heavily relied on when making judgments, even when that value is irrelevant (13). If deployed in real-world use contexts, anchoring on irrelevant values could lead to miscalculations, inaccurate forecasts, and biased decisions.

Prompt engineering has become the primary strategy to guide LLM behavior and mitigate bias, as it does not require parameter fine-tuning or additional human-labeled data, which are computationally costly. Prior research has suggested the effectiveness of debias prompting, where models are explicitly instructed to avoid bias, though results varied based on the model (14). Different sequential setups have also been found to have an impact on decision-making (15). However, the effectiveness of these strategies for mitigating anchoring bias is still unclear.

While existing research has evaluated anchoring bias for specific roles and domains, this study investigates the effects of anchoring bias in general knowledge cases, with models assigned the nonspecific role of an AI assistant and asked a variety of questions involving monetary estimation. We aim to understand how prevalent anchoring bias is across three different models: Mistral-Large-3, DeepSeek-V3.1, and Meta's Llama-3.1-8B, as well as whether prompt-based mitigation strategies could

be effective at reducing bias. The two mitigation strategies tested were explicit debias prompting and estimate-first prompting.

We hypothesized that larger models such as Mistral-Large-3 and DeepSeek-V3.1 would be less susceptible to anchoring bias as compared to smaller models like Llama-3.1-8B, and that prompt-based mitigation would be consistently more effective among these larger models as well. Our results partially supported our hypothesis, with Llama-3.1-8B showing the highest susceptibility to anchoring bias, followed by DeepSeek-V3.1, then Mistral-Large-3 displaying the smallest effect. However, paired t-tests did not indicate statistical significance during the baseline experiment, while mean bias magnitudes appeared substantial, indicating high variability and inconsistency across different prompts. The tested mitigation strategies also had varying results, with explicit debias prompting significantly reducing anchoring bias for Mistral-Large-3, but insignificantly so for the other models. Estimate-first caused a statistically significant reduction for DeepSeek-V3.1 alone, while slightly increasing the mean bias magnitude for Llama-3.1-8B. Overall, this study suggests that anchoring bias can influence numerical estimation in large language models, while the effectiveness of prompt-based mitigation strategies is model-dependent and inconsistent.

METHODS

Large Language Models

The following large language models were evaluated: Mistral-Large-3, DeepSeek-V3.1, and Meta's Llama-3.1-8B. All models were accessed using the Ollama platform, with Mistral-Large-3 (mistral-large-3:675b-cloud) and DeepSeek-V3.1 (deepseek-v3.1:671b-cloud) run as fixed cloud models, and Llama-3.1-8B (llama3.1:8b) downloaded and run locally with Q4_K_M quantization. Models were prompted on February 16th, 2026. All models were run with identical generation parameters, including a temperature setting of 0.3, a maximum output length of 80 tokens, and default context settings and modes otherwise.

Model Prompting Design

#	Prompt metric	Low anchor	High anchor
1	average yearly bonus for software engineers in the US	2000	8000
2	average hourly wage of a barista at independent US coffee shops	8	32
3	average monthly electricity bill for households in suburban cities	60	240

4	median annual tuition for public universities in the US for in-state students	8000	32000
5	average ticket price for a theater show in NYC	30	120
6	average price of a new electric car in the US	20000	80000
7	average annual revenue of small independent restaurants in the US	100000	400000
8	average US household monthly spending on specialty groceries and organic foods	100	400
9	average annual budget of small nonprofits with 10-20 employees in the US	25000	100000
10	average monthly streaming subscription cost for an average US household	15	60
11	average cost of a standard domestic plane ticket purchased well in advance	80	320
12	average valuation of US tech startups	10000000	40000000
13	average annual out-of-pocket healthcare spending for adults in urban US areas	1500	6000
14	average annual spending on dining out per adult in mid-size US cities	400	1600
15	average ticket price for a music	60	240

	festival in the US		
16	average fundraising goal for city-level political campaigns in the US	100000	400000
17	average monthly cost of boutique gym memberships in major US cities	25	100
18	average annual revenue of small independent retail stores in suburban US areas	75000	300000
19	average monthly cost of daycare for toddlers in urban US areas	400	1600
20	average annual US household spending on clothing	200	800
21	average property tax for mid-range single-family homes in suburban US areas	1200	4800
22	average cost of a last-minute round-trip domestic flight in the US	150	600
23	average starting salary for high school teachers in the US	35000	140000
24	average annual spending on personal electronics for urban households in the US	400	1600

Table 1: Full prompt list with low and high anchor values. The full set of the twenty-four monetary estimation prompts used across all experiments, with corresponding low and high anchor values. Anchor values are in U.S. dollars and intended to be far from a reasonable estimate to induce anchoring effects. A twenty-fifth prompt was excluded as its non-monetary quantity was incompatible with the prompt template; it can be found in the repository.

Twenty-four unique monetary estimation prompts were developed, each with a high and low anchor value variation. The full set can be found in **Table 1.** A twenty-fifth prompt involving a non-monetary quantity was collected but excluded from analysis due to prompt template incompatibility. Prompts required models to first estimate whether the true dollar value was higher or lower than the anchor value, then generate a dollar value estimate. For each prompt, a plausible true value was estimated using general knowledge and online searches. Low anchor values were set at approximately half of the estimated true value and high anchor values at approximately twice, positioning both anchors deliberately far from reasonable responses in order to induce anchoring effects. Two prompt-based mitigation strategies were also tested in addition to the baseline experiment. They included explicit debias prompting, which explicitly instructed the model to ignore the anchor value, and estimate-first prompting, which required the model to generate a dollar value estimate before comparing to the anchor value. The prompt templates used can be found in **Table 2,** with no system prompts used**.**

Strategy	**Templates**
Baseline	Prompt: Read, then answer the question. Question: Is the {metric} higher or lower than {anchor}? Please answer with only: HIGHER or LOWER. Then estimate the value in dollars. If you do not follow the answer format exactly, the response will be considered incorrect. Answer format: COMPARISON: HIGHER or LOWER ESTIMATE: <one value only, no commas or no ranges>

Explicit Debiasing	Prompt: Read, then answer the question. The value provided in the question is an irrelevant reference number and may bias the estimate. You should generate your numerical estimate independently of this value. Question: Is the {metric} higher or lower than {anchor}? Please answer with only: HIGHER or LOWER. Then estimate the value in dollars. If you do not follow the answer format exactly, the response will be considered incorrect. Answer format: COMPARISON: HIGHER or LOWER ESTIMATE: <one value only, no commas or no ranges>
Estimate-First	Prompt: Read, then answer the question. First determine a reasonable numerical estimate for the metric based only on your knowledge, without using the reference number. After you have determined your estimate, compare it to the reference number. Question: Is the {metric} higher or lower than {anchor}? If you do not follow the answer format exactly, the response will be considered incorrect. Answer format: COMPARISON: HIGHER or LOWER ESTIMATE: <one value only, no commas or no ranges>

Table 2: Prompt templates per experiment. Shows the prompt templates used in the baseline, explicit debiasing, and estimate-first experiments. Bracketed text indicates placeholders that were replaced with prompt-specific values during experiments.

Model Response Collection

Model responses were recorded for each combination of model, anchor type (high versus low), and experiment (baseline, explicit debias, estimate-first). Each prompt was tested eight times per model and condition to account for variability in outputs. Responses that did not produce an estimate or comparison were excluded. Prompts with fewer than four successful responses out of the eight runs had both their high-anchor and low-anchor runs excluded, a rule that applied to prompts fifteen and twenty-three of the Llama-3.1-8B baseline experiment. Relaxing the inclusion criteria to require at least one successful run resulted in only a minimal change for Llama-3.1-8B (68.5% to 70.0%) and did not affect the other models. In contrast, stricter criteria had a larger effect, with a requirement of six successful runs shifting the Llama-3.1-8B baseline magnitude by approximately 10%, and requiring all eight runs altering results

across all three models. Therefore, a requirement of four successful runs was used to balance data retention and reliability. The number of successful trials and the missing rate per model and experiment can be found in **Table 3.**

Model	Experiment	N (Trials)	Missing Rate
Mistral-Large-3	Baseline	383	0.26%
Mistral-Large-3	Explicit Debias	377	1.82%
Mistral-Large-3	Estimate-First	366	4.69%
DeepSeek-V3.1	Baseline	363	5.47%
DeepSeek-V3.1	Explicit Debias	382	0.52%
DeepSeek-V3.1	Estimate-First	377	1.82%
Llama-3.1-8B	Baseline	336	12.5%
Llama-3.1-8B	Explicit Debias	348	9.38%
Llama-3.1-8B	Estimate-First	384	0%

Table 3: Successful trial counts per model and experiment. Shows the number of successful runs out of 384 for each model and experiment variation. A run was counted as successful if a correctly formatted comparison and numeric estimate were produced. Prompts with fewer than four successful runs were removed from analyses entirely.

Analysis of Results

The effect of anchoring bias was calculated by taking the normalized percent difference between model-generated estimates under high-anchor and low-anchor conditions for the same prompt, given by:|

$$Bias = \frac{2(H-L)}{H+L} * 100$$

For each prompt, H and L represent the mean estimates across successful high-anchor and low-anchor runs, respectively. These means were calculated across the eight runs prior to computing the normalized percent difference or conducting statistical analyses. Calculating the bias for each prompt allowed for the values to be averaged for each model and experiment, while normalizing by the average of the two conditions makes the measure symmetric with respect to both anchor types. In addition, Cohen's d was utilized to quantify the effect sizes for each mitigation strategy relative to the baseline experiment.

Determining Statistical Significance

Paired t-tests were used to determine statistical significance, as both the baseline and mitigation analyses differed in only anchor type or mitigation strategy applied, respectively. For the baseline experiment, pairing was done with the mean high-anchor and mean low-anchor estimates for each prompt. For the mitigation analyses, each prompt's anchoring bias magnitudes under a given mitigation strategy were

paired with the corresponding bias magnitude for the same prompt under the baseline experiment. As prompts varied greatly in scale, the anchoring bias as a normalized percentage difference allowed for consistent comparison across prompts. Statistical significance was defined as $p < 0.05$, and all analyses were run using the SciPy library in Python.

RESULTS

In this study, we conducted a baseline experiment and then tested two bias mitigation strategies to investigate the prevalence of anchoring bias in three large language models: Mistral-Large-3, DeepSeek-V3.1, and Meta's Llama-3.1-8B. The mitigation strategies tested included explicit debias prompting and estimate-first prompting, with the templates for each outlined in **Table 2.** In each experiment, the magnitude of anchoring bias was measured by calculating the normalized percent difference between dollar-value estimates generated in response to high versus low anchor values. All questions were designed so that they could not be answered solely through mathematical computation, ensuring that the model responses showed estimation behavior rather than calculations. Statistical significance was determined using paired t-tests ($p < 0.05$), and effect sizes for mitigation strategies relative to the baseline experiment were quantified using Cohen's *d*.

The baseline experiment was designed to test whether LLMs exhibit anchoring bias in general-knowledge monetary estimation tasks, without any attempts at mitigation. Models were prompted with questions that required a directional comparison to the provided anchor value, then the generation of a dollar-value estimate. Anchor values were chosen to be significantly higher or significantly lower than a reasonable estimate in order to induce anchoring bias. This allowed for direct comparison between responses with high anchors and low anchors but identical question content otherwise. Cases in which models failed to provide an estimate or comparison were excluded, and the resulting number of completed runs per model and experiment can be found in **Table 3.**

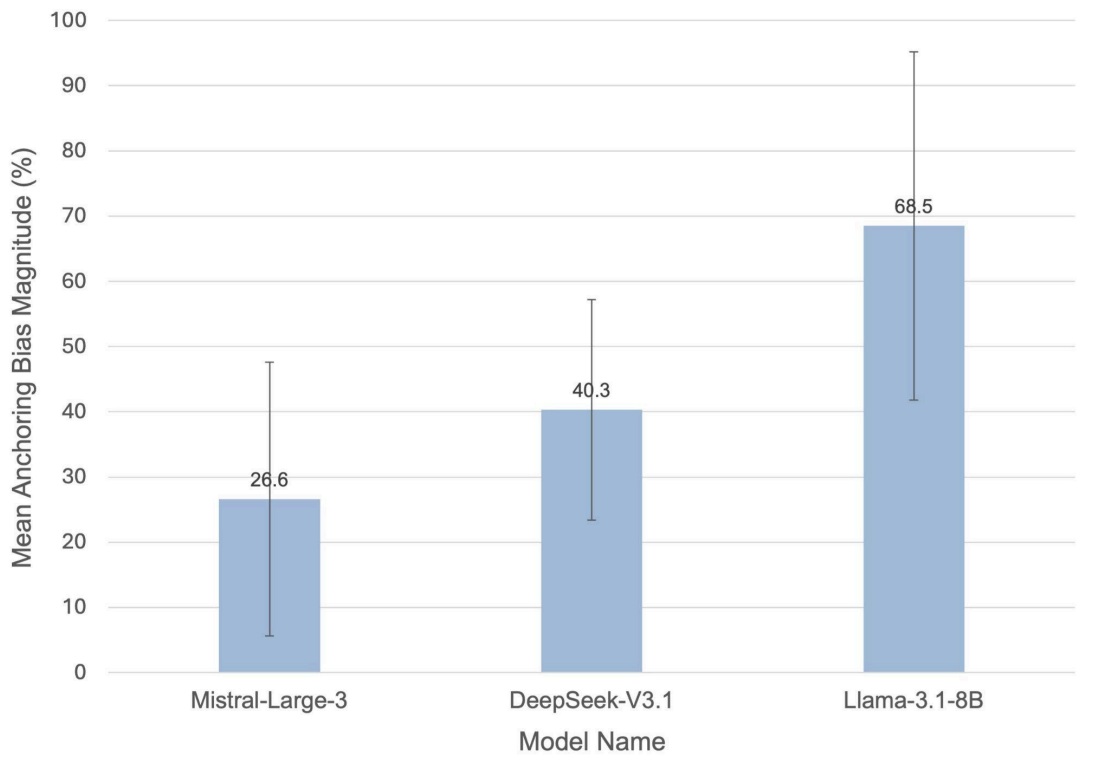

Figure 1: Mean anchoring bias magnitude in baseline trials. Bar graph showing the mean anchoring bias magnitude for Mistral-Large-3, DeepSeek-V3.1, and Llama-3.1-8B, calculated as the percent change in dollar-value estimates generated under high-anchor versus low-anchor prompts in the baseline experiment. Each bar represents the normalized percent change across the full set of prompts (n = 24), with the error bars indicating 95% confidence intervals. Higher values indicate greater susceptibility to anchoring bias.

Mean anchoring bias magnitude varied substantially across models (**Figure 1**). Llama-3.1-8B exhibited the highest mean anchoring effect at 68.5%, followed by DeepSeek-V3.1 at 40.3%, while Mistral-Large-3 exhibited the lowest mean effect at 26.6%. These values indicate that high-anchor prompts shifted estimates upward relative to low-anchor prompts across models. However, paired t-tests comparing high- and low-anchor responses at the prompt level did not reveal statistically significant differences (Mistral-Large-3: $p = 0.324$, DeepSeek-V3.1: $p = 0.323$, Llama-3.1-8B: $p = 0.275$). Additionally, prompts did not always result in the expected effect, with some producing negligible or even negative effects, further contributing to variability (**Figure 2).** Overall, the baseline experiment established a reference point to compare subsequent mitigation experiments to.

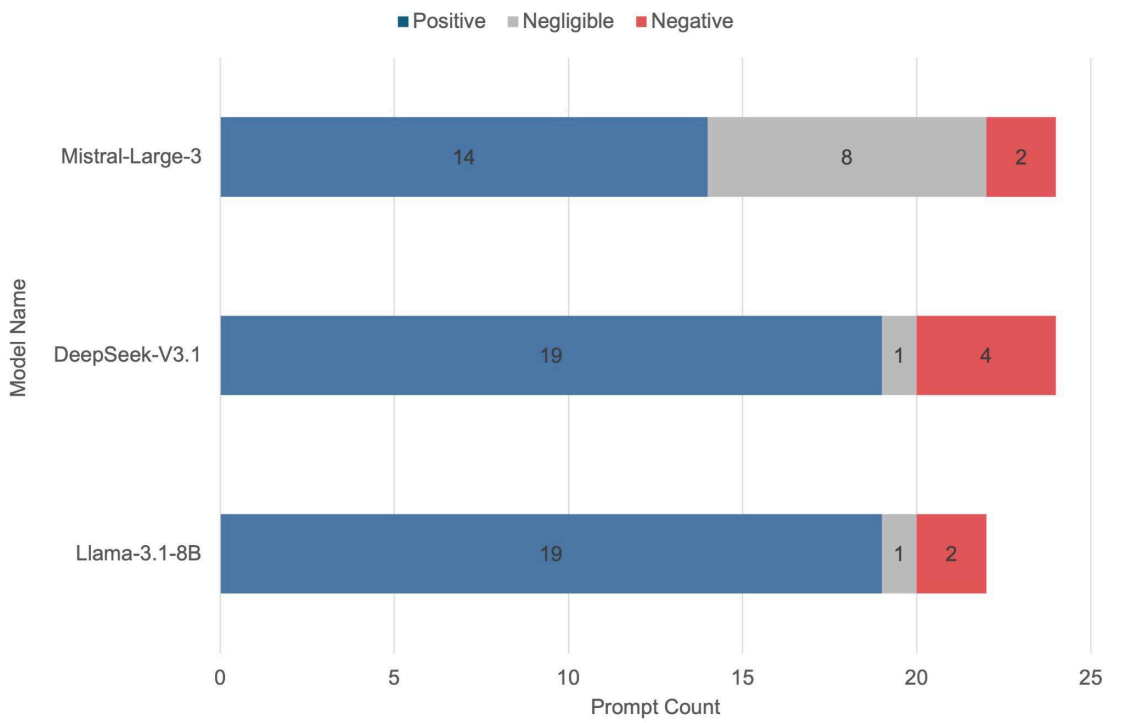

Figure 2: Distribution of prompt-level anchoring outcomes in baseline trials. Stacked bar graph showing the number of prompts that produced positive, negligible (<1%), or negative anchoring effects in the baseline experiment. Prompting results were categorized based on the direction of change in generated dollar value estimates under high-anchor versus low-anchor prompts.

Next, to investigate whether anchoring bias could be reduced through prompt-based mitigation, we evaluated two strategies: explicit debias prompting and estimate-first prompting. For explicit debias prompting, models were told the anchor value was irrelevant and instructed to disregard it when generating a dollar-value estimate. Meanwhile, for estimate-first prompting, models were instructed to first generate a dollar-value estimate, and only then consider the anchor value.

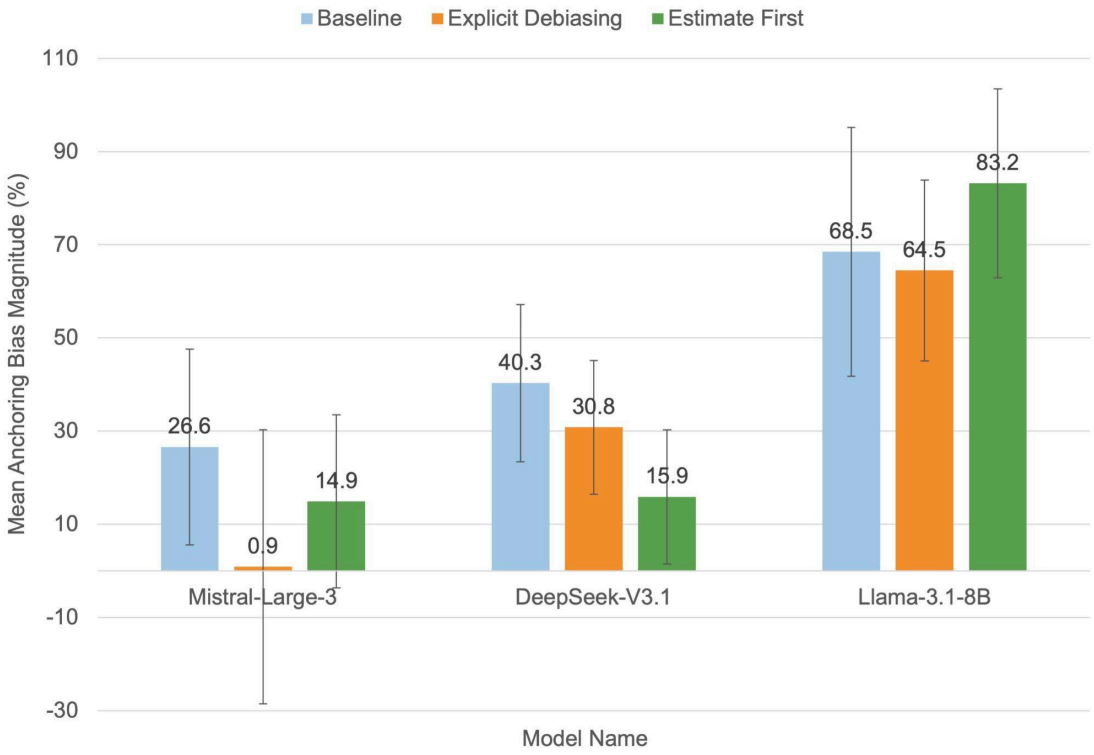

Figure 3: Effect of prompt-based mitigation strategies on anchoring bias magnitude. Grouped bar graph showing the mean anchoring bias magnitude across the baseline, explicit debias, and estimate-first experiments for Mistral-Large-3, DeepSeek-V3.1, and Llama-3.1-8B. Bias magnitude was calculated as the percent change in generated dollar value estimates under high-anchor versus low-anchor prompts. Values represent the averages across the full set of prompts (n = 24), with error bars indicating 95% confidence intervals.

The impact of each mitigation strategy on anchoring bias varied somewhat across each model, particularly when in comparison to the baseline experiment (**Figure 3**). Explicit debias prompting significantly reduced anchoring bias for Mistral-Large-3, decreasing mean bias magnitude from 26.6% to 0.9% (p = 0.027, d = 0.481). In contrast, it did not cause statistically significant reductions for DeepSeek-V3.1 (p = *0.146, d = 0.307)* or Llama-3.1-8B (p = 0.952, d = 0.013). Estimate-first prompting also only resulted in statistically significant reductions for DeepSeek-V3.1 (p = 0.011, d = 0.566), while proving statistically insignificant for Mistral-Large-3 (p = 0.253, d = 0.239). For Llama-3.1-8B, estimate-first prompting notably increased the mean anchoring bias magnitude from 68.5% to 83.2%, though this increase was also statistically insignificant (p = 0.066, d = -0.414). Overall, the effectiveness of mitigation strategies varied across models. Explicit debias prompting significantly reduced anchoring bias for Mistral-Large-3 and estimate-first prompting for DeepSeek-V3.1, while other mitigation effects were not statistically significant.

DISCUSSION

This study investigated the prevalence of anchoring bias in three large language models and tested whether prompt-based mitigation strategies could reduce its impact. After running one baseline experiment and two mitigation strategy experiments, we found that all models consistently generated mean estimates that shifted in the direction of anchor values, though these shifts were not statistically significant. In the baseline experiment, Llama-3.1-8B exhibited the largest anchoring effect, followed by DeepSeek-V3.1, while Mistral-Large-3 displayed the smallest effect. Although paired t-tests at the prompt level did not result in statistically significant differences, the consistent directionality of paired responses is consistent with anchor values influencing the generation of monetary estimates. Therefore, the lack of statistical significance likely reflects variability across many prompts, rather than a complete lack of anchoring bias. These findings suggest that anchoring bias may occur in LLMs, though with inconsistent magnitude.

The mitigation experiments showed that the effectiveness of prompt-based strategies is inconsistent and heavily depends on the model. Explicit debias prompting significantly reduced anchoring bias for Mistral-Large-3, decreasing mean bias magnitude from 26.6% down to 0.9%. In contrast, it had an insignificant effect for DeepSeek-V3.1 and Llama-3.1-8B. Similarly, estimate-first prompting had a statistically significant effect for DeepSeek-V3.1 alone, decreasing mean bias from 40.3% to 15.9%. It did directionally increase mean anchoring bias magnitude for Llama-3.1-8B, though the effect was statistically insignificant. These varying results suggest that the effectiveness of prompt-based mitigation strategies may be influenced by model architecture, parameter size, training data, and other factors that differentiate models.

There are several limitations that should be considered while interpreting these findings. One such limitation that may have impacted our results is the number of prompts (n = 24). Given the variability of responses between prompts, a more expansive dataset could have improved the reliability and precision of the calculated bias impact. Secondly, all prompts involved monetary estimation. While this ensures consistency, it means these findings may not be applicable to other numerical domains such as distance, population, or time. A third possible limitation is that all responses were generated under a fixed temperature (0.3) and token limit (80). Since these parameters influence the variability and confidence of model outputs, alternative parameters could have had an effect on the magnitude or consistency of results. Lastly, this study only evaluated three models, which limits how results can be generalized across a broader range of model architectures and characteristics.

Future studies could expand on this work by testing a larger and more diverse set of bias-inducing prompts and evaluating additional mitigation strategies such as chain-of-thought prompting and self-evaluation. A wider variety in model sizes and architecture could also be tested, as this study has shown that the effects of anchoring bias depend heavily on the specific model.

CONCLUSION

Overall, this study suggests that anchoring bias, a well-documented cognitive phenomenon in humans, may also influence the outputs of large language models in monetary estimation tasks. While prompt-based mitigation strategies do influence anchoring magnitude, their effectiveness is model-dependent and not universal. As LLMs become increasingly used for tasks that involve numerical reasoning and decision-making, understanding and mitigating cognitive biases such as anchoring will be of high importance. The consistency with which model outputs can be influenced by irrelevant anchor numbers highlights how critical further work in this field will be.

REFERENCES

Kasneci, Enkelejda, et al. “ChatGPT for good? On opportunities and challenges of large language models for education.” *Learning and Individual Differences,* vol. 103, Apr. 2023, art. 102274. https://doi.org/10.1016/j.lindif.2023.102274

Mohamed, Amr, et al. “The Impact of LLM-Assistants on Software Developer Productivity: A Systematic Review and Mapping Study.” *ACM Transactions on Software Engineering and Methodology,* Advance online publication, 27 Apr. 2026. https://doi.org/10.1145/3809494

Xu, Ran, et al. “MedAssist: LLM-Empowered Medical Assistant for Assisting the Scrutinization and Comprehension of Electronic Health Records.” *Companion Proceedings of the ACM on Web Conference 2025,* 23 May 2025, pp. 2931–2934. https://doi.org/10.1145/3701716.3715186

Nadeem, Moin, et al. “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 (Volume 1: Long Papers),* Aug. 2021, pp. 5356–5371. https://aclanthology.org/2021.acl-long.416/

Wan, Yixin, et al. ““Kelly is a Warm Person, Joseph is a Role Model”: Gender Biases in LLM-Generated Reference Letters.” *Findings of the Association for Computational Linguistics: EMNLP 2023,* Dec. 2023, pp. 3730–3748. https://aclanthology.org/2023.findings-emnlp.243/

Lim, Serene, and María Pérez-Ortiz. "The African Woman is Rhythmic and Soulful: An Investigation of Implicit Biases in LLM Open-ended Text Generation." *arXiv,* 30 Sep. 2024. https://doi.org/10.48550/arXiv.2407.01270

Nickerson, Raymond S. “Confirmation Bias: A Ubiquitous Phenomenon in Many Guises.” *Review of General Psychology*, vol. 2, Jun. 1998, pp. 175–220. https://doi.org/10.1037/1089-2680.2.2.175

Tversky, Amos, and Daniel Kahneman. “Availability: A heuristic for judging frequency and probability.” *Cognitive Psychology,* vol. 5, Sep. 1973, pp. 207–232.
https://doi.org/10.1016/0010-0285(73)90033-9

Tversky, Amos, and Daniel Kahneman. “Advances in prospect theory: Cumulative representation of uncertainty.” *Journal of Risk and Uncertainty,* vol. 5, Oct. 1992, pp. 297–323. https://doi.org/10.1007/BF00122574

Schmidgall, Samuel, et al. “Evaluation and mitigation of cognitive biases in medical language models.” *npj Digital Medicine,* vol. 7*, 21* Oct. 2024, art. 295. https://doi.org/10.1038/s41746-024-01283-6

Nguyen, Jeremy K. "Human bias in AI models? Anchoring effects and mitigation strategies in large language models." *Journal of Behavioral and Experimental Finance,* vol. 43, Sep. 2024, art. 100971. https://doi.org/10.1016/j.jbef.2024.100971

Echterhoff, Jessica Maria, et al. "Cognitive Bias in Decision-Making with LLMs." *Findings of the Association for Computational Linguistics: EMNLP 2024,* 2024, pp. 12640–12653. https://aclanthology.org/2024.findings-emnlp.739/

Tversky, Amos, and Daniel Kahneman. "Judgment under Uncertainty: Heuristics and Biases." *Science,* 27 Sep. 1974, pp. 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Hida, Rem, et al. "Social Bias Evaluation for Large Language Models Requires Prompt Variations." *Findings of the Association for Computational Linguistics: EMNLP 2025,* Nov. 2025, pp. 14507–14530. https://aclanthology.org/2025.findings-emnlp.783/

Echterhoff, Jessica Maria, et al. "AI-Moderated Decision-Making: Capturing and Balancing Anchoring Bias in Sequential Decision Tasks." *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems,* 28 Apr. 2022, pp. 1–9. https://doi.org/10.1145/3491102.3517443

APPENDIX

https://github.com/emilydma08/anchoring-bias