Qeios

Peer Review

Review of: "Anchoring Bias in Large Language Models: An Experimental Study"

Meltem Aksoy[1]

1. Computer Science, Technische Universität Dortmund, Dortmund, Germany

This paper addresses an underexplored but increasingly relevant issue in the behavior of LLMs: anchoring bias. The authors conduct systematic experiments using GPT-3.5, GPT-4, and GPT-4o on a dataset originally developed for studying human cognitive biases, specifically to investigate how anchoring cues influence LLM-generated numerical predictions. They further examine various prompt-based mitigation strategies. While the paper has clear value and potential, it requires significant revision to meet the standards of high-quality empirical research.

The topic is timely and relevant, especially as LLMs are increasingly integrated into decision-making workflows. Unlike the more frequently studied demographic biases, cognitive biases such as anchoring have received limited attention. The use of a dataset with realistic, everyday-life-inspired questions enhances ecological validity. The repeated prompting (30 runs per question per condition) and statistical comparisons via t-tests provide a quantitative backbone to the study. Moreover, comparing LLM behavior to human data from Yasseri et al. (2022) is an insightful contribution that situates model behavior in a broader psychological context. The negative results are also valuable. The paper shows that common prompting strategies such as Chain-of-Thought, Principle-of-Thought, and Reflection are ineffective in mitigating anchoring bias. This finding is important for LLM practitioners who rely on prompt engineering to steer model outputs in high-stakes contexts.

The paper's main shortcomings lie in its methodology, overinterpretation of results, and limited generalizability.

1. The experimental setup lacks necessary precision and replicability. While the appendix provides a template, the main text does not clearly articulate how prompts were varied across control and

 doi.org/10.32388/ZJNM26

treatment groups. The experimental conditions should be explained more clearly in the main text by specifying how the control, low-anchor, and high-anchor prompts were constructed, ideally using a concise table that outlines which hints were included in each condition and their intended purpose. Moreover, there is no justification for the use of a high temperature (0.8), which may introduce noise and inflate answer variance. A lower temperature, more typical in deterministic inference scenarios, might yield clearer insights into anchoring effects. The absence of a sensitivity analysis on temperature is a missed opportunity.

2. The statistical analysis has shortcomings. The paper uses multiple t-tests across dozens of questions without any correction for multiple comparisons (e.g., Bonferroni or FDR), which raises concerns about false positives. The authors also treat consistent directional differences as evidence of bias but do not always show that these differences are statistically significant. In addition, the standard deviation-based conclusions conflating model confidence and bias susceptibility oversimplify the findings. That a stronger model (GPT-4) shows lower variance in responses does not necessarily mean it is "more biased"; this may simply indicate more consistent behavior, biased or not.
3. The choice and framing of mitigation strategies are weakly grounded. While prompt-based approaches are convenient to test, the rationale behind using reasoning-boosting strategies like CoT and PoT to mitigate cognitive bias is not clearly explained or supported by prior literature.
4. The dataset is relatively small (62 questions), and while it spans different domains, it is sourced from a single game-based platform ("Play the Future"), which may not reflect the diversity of real-world applications. The classification of hint types into "fact," "expert," and "irrelevant" is done without inter-rater validation, and the process is not described. This undermines the internal validity of the experimental conditions.
5. A key issue in terms of generalizability is the paper's exclusive focus on GPT-family models. While comparing different versions of GPT (3.5, 4, 4o) is informative, a broader analysis that includes models from other providers (e.g., Claude, Gemini, LLaMA, Mistral) would be necessary to draw more general conclusions about anchoring bias in LLMs. Without this, claims about "LLM bias" remain largely GPT-specific and should be framed as such.
6. The presentation of results, particularly the tables, needs improvement. Visualizations are dense, lack adequate labeling, and are hard to interpret without close reference to the main text.

Declarations

Potential competing interests: No potential competing interests to declare.