Qeios

Peer Review

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

Alexandru Tăbușcă[1]

1. Informatics, Statistics, Mathematics, Romanian-American University, Romania

The proposed paper addresses a relatively under-explored aspect of bias in large language models (LLMs): cognitive bias, specifically anchoring bias. While demographic bias has received considerable attention in existing AI literature, cognitive bias—especially its manifestation and mitigation within LLMs—remains somehow insufficiently studied. This work offers a novel contribution by evaluating anchoring bias in models such as GPT-4, GPT-4o, and GPT-3.5 Turbo through a controlled and systematic experimental design. The authors expand on previous work by using a broader question set (62 items), making this a more comprehensive study in the area (at least I know of up to date). I consider the contribution of the authors as original and timely.

The findings presented have direct implications for the deployment of LLMs in high-stakes contexts, such as finance, healthcare, or automated decision-making. Notably, the study reveals that stronger models are more susceptible to anchoring bias and that common prompting strategies such as Chain-of-Thought (CoT), Thought-of-Principles, Ignoring Instructions, and Reflection are ineffective in mitigating this bias. These insights are valuable for both the research community and AI practitioners working on model alignment and safety. These findings are quite significant and have real-world applicability.

The methodology applied by the authors seems sound and rigorously executed. The authors employed a balanced control-treatment experimental framework using a well-suited real-world inspired dataset (from the "Play the Future" game). Each condition was tested across 30 runs with different temperature settings (0.8), followed by statistical validation using t-tests. The anchoring index (AI) is appropriately adapted from the literature.

However, I consider that the study would benefit from several additional elements:

doi.org/10.32388/UK340W

- additional error analysis or explanation of failure cases for failed mitigation techniques
- greater elaboration on how anchoring behavior diverges in LLMs vs. humans, beyond the numerical AI comparison

Nevertheless, I consider the methodology solid, with perhaps minor opportunities for deeper analysis.

Given the increasing adoption of LLMs in decision-critical environments, this research is highly relevant to researchers and practitioners concerned with AI alignment, cognitive modeling, bias mitigation, and interpretability. The fact that mitigation strategies largely fail, and that expert anchors exert a stronger influence than fact-based ones, makes the study particularly impactful. As a result, the paper should be of high interest to the broader AI and NLP research community.

The references are well-chosen, timely, and relevant. They span foundational works in cognitive bias (e.g., Kahneman), recent LLM bias studies (e.g., BiasMedQA, BIASBUSTER), and mitigation literature. However, citations from recent flagship conferences (e.g., ACL, NeurIPS, ICLR—with some very interesting results in 2024/2025) could further reinforce context, especially regarding prompt engineering and bias mitigation frameworks.

In conclusion, I consider this paper a **strong and meaningful contribution** to the field of AI safety and behavioral analysis of LLMs. It is rigorously executed, relevant, and well-motivated. The findings challenge commonly held assumptions about the effectiveness of mitigation techniques and provide practical insight for future research directions. With minor improvements in clarity (and perhaps a little bit of format checking), the paper is well-suited for publication and can become a relevant resource in the field.

Declarations

Potential competing interests: No potential competing interests to declare.