Qeios

Peer Review

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

Evgenii Arsentev[1]

1. Independent researcher

What the paper claims

The study asks whether GPT-4, GPT-4o, and GPT-3.5 Turbo show anchoring bias on numeric estimation. It reuses a dataset built for a human study, 62 questions from a prediction game, each carrying a factual reference hint H1 and two opposing anchors H2 and H3 (Section 3.1). Each question is run in three conditions - H1 alone, H1 with the low anchor, H1 with the high anchor - 30 times per condition at temperature 0.8, and the two anchored distributions are compared with a t-test (Section 3.2). The headline evidence in Section 4.2.1 is that among 162 model-question differences, only 11 have a sign opposite to the anchor gap, with stronger models more consistently pulled. Section 4.2.3 finds expert-opinion anchors the strongest; Section 5 reports that Chain-of-Thought, principles-first prompting, an explicit instruction to ignore the anchor, and reflection all fail to remove the effect, while presenting both anchors at once helps on some questions. Section 4.3 puts the anchoring index of GPT-4 at about 0.45 against 0.61 for humans on the same data. Prompts and raw answers are released on GitHub.

Two choices here are better than the norm in this literature. Borrowing a dataset from a human study makes the human comparison meaningful rather than rhetorical, and publishing the prompts and the model answers lets a reader re-analyse without re-running anything. The criticism below is almost entirely about the analysis layer, not the experiment.

Does the conclusion hold on the data?

1. Near-zero variance and a 30-sample t-test do not belong in the same design. Section 4.1 reports that in 78 of 108 control experiments on GPT-4 and GPT-4o, the standard deviation is essentially zero, meaning the model returned the identical number 30 times despite temperature 0.8. Where that

doi.org/10.32388/92KCPQ

happens, the effective sample size is one, not thirty, and a t-test comparing two nearly degenerate distributions produces a p-value that reflects a rounding-level spread rather than evidence. The significance stars in Figures 4 to 8 are therefore not interpretable in the cells that matter most. Report, per cell, how many distinct values the 30 draws contained, and re-run the comparisons either as permutation tests or with the question-level anchoring index as the unit of analysis.

2. 162 uncorrected tests. Section 4.2.1 counts 162 difference values, which is 54 questions by three models, and each is judged against the usual 0.05 threshold; the mitigation tables add more. At that family size, roughly eight cells would clear 0.05 by chance alone. Apply a correction such as Benjamini-Hochberg, or better, fit one mixed-effects model with question as a random effect and anchor condition as a fixed effect. That also replaces 162 scattered verdicts with a single effect size and interval, which is what a reader can actually carry away.

3. The central claim rests on a sign count that throws away magnitude. "Only 11 exceptions out of 162" establishes direction, and a sign test on that ratio would be convincing on its own, but it says nothing about how much the models move, and the questions differ in scale by orders of magnitude - a temperature in Fahrenheit, a stock price, a count of upvotes. The paper already defines the anchoring index in Section 4.3 and then uses it once, as a single average of 0.45 with no dispersion. Report the AI per question, show its distribution with a confidence interval, and use a scale-free effect such as the answer gap divided by the anchor gap so questions can be compared and pooled.

4. Possible answer leakage from training data is never addressed. The questions come from a prediction game about real events, and the worked example in Appendix B asks for the starting price of a device announced at Apple WWDC 2018. For any event that happened before a model snapshot was trained, the model may be recalling the outcome rather than guessing under uncertainty, and recall would suppress anchoring exactly where the paper reads low anchoring as a property of the model. Split the questions by whether the event precedes the snapshot cutoff, report the effect separately for each group, and if possible add a handful of questions about events after the cutoff.

5. The control differs from the treatments by more than an anchor. In Section 3.2, the control shows H1 only while both treatments show H1 plus a second hint, so treatment prompts are longer and contain one extra number presented as somebody's opinion. Part of the measured shift may be a response to extra context rather than to the anchor value. A neutral filler hint of matched length, carrying a number irrelevant to the quantity asked, would separate those two explanations cleanly and costs one more condition.

6. The Both-Anchor result may be arithmetic rather than debiasing. Section 5 observes that when H2 and H3 fall on either side of the bias-free benchmark, GPT-4 answers land close to that benchmark, and lists questions 5, 7, 8, 10, 12, 13, 14, and 15 as cases. If two anchors bracket the truth, a model that simply averages them will also land near the truth without being any less anchored. The test that separates the two readings is already available in your logs: compare the model answer against the trivial rule "answer equals the mean of H2 and H3", and report how often the model is closer to the benchmark than that rule is. Until then, Both-Anchor is better described as anchor balancing than as mitigation.

Reproducibility

The repository with prompt templates and model answers, named at the end of Section 6, is the strongest part of this section, and the CO-STAR prompt in Appendix A plus the worked example in Appendix B mean a reader can see exactly what the model was asked. Temperature 0.8 and 30 repetitions are stated, and the parsing rule for time values in Section 3.2 is spelled out.

What is missing is the identity of the systems under test. The paper names GPT-4o, GPT-4, and GPT-3.5 Turbo without snapshot identifiers and without the dates of access, and all three were served by endpoints that changed several times. A cognitive-bias measurement is precisely the kind of result that shifts between snapshots, so the numbers cannot be re-checked or re-dated as they stand. Also absent: top-p and maximum tokens alongside the temperature, the number of questions dropped or retried when parsing failed, and a commit hash for the repository.

What to fix

1. Section 4.1 and Figures 4 to 8: report the count of distinct values per 30-draw cell, and replace the t-tests in low-variance cells with permutation tests or with an analysis at the question level.
2. Section 4.2: correct for multiple comparisons across the 162 tests, or fit a single mixed-effects model with question as a random effect and report one pooled effect size with an interval.
3. Section 4.3: report the anchoring index per question with its distribution and confidence interval, not only the single average of 0.45, and state whether the 0.61 human figure was computed on the same question subset.
4. Section 3.1: classify the questions by whether the underlying event precedes the training cutoff of each model, and report anchoring separately for each group.

5. Section 3.2: add a length-matched neutral-number control so anchoring is separated from the effect of simply adding a second hint.
6. Section 5: test Both-Anchor against the mean-of-anchors rule, and report how often the model beats it. Rename the strategy if it does not.
7. Sections 3.2 and 6: give model snapshot identifiers and access dates, plus top-p, maximum tokens, parsing failure counts, and a repository commit hash.
8. Figures 4 to 9 are tables presented as images. Render them as text tables so values can be read, cited, and re-used.
9. Section 4.2.2 refers to Figure 2 when describing the seven time-period questions and their six exceptions; from context, this should be Figure 4. Please check the figure references throughout.
10. State the arithmetic of the sample explicitly: 62 questions comprise 42 fact-anchor, 12 expert-anchor, and 8 irrelevant-hint items, and the 162 difference values are 54 questions by three models with the irrelevant-hint group excluded. A reader currently has to reconstruct this.
11. Section 1 says the simple strategies "are not insufficient to mitigate anchoring bias," a double negative that states the opposite of the finding; the abstract has it right. Also, "lowvalue" and "highvalue" need hyphens, and the paper lists no corresponding author address of its own.

Recommendation

Major revisions, at the level of analysis rather than experiment. The design is sound, the dataset choice is clever, and the negative result about mitigation prompts is worth publishing on its own. What has to change is how the evidence is summarised: a sign count over 162 uncorrected t-tests, several of them computed on distributions with no variance at all, cannot carry a claim this general, and the leakage question has to be answered before low anchoring on some questions can be read as a property of the model. Nearly all of it can be done on data the authors already have in their repository, which is the best position a paper can be in after review.

Evgenii Arsentev, PhD

Declarations

Potential competing interests: No potential competing interests to declare.