

Stress-Testing LLM Preference Elicitation: When Consistency Does Not Imply Validity

Shriyam Baloni

karanbaloni@gmail.com

Abstract

A central challenge in digital-minds research is determining whether an LLM’s apparent preference reflects a stable underlying preference-like signal or an artifact of the way that choice is made. We treated preference elicitation as a measurement validity problem and developed a diagnostic framework to stress-test these outputs. Across three independently designed stimulus domains tested on one small open-weight model (`qwen2.5:3b-instruct-q4_K_M`), we found two recurring failure modes: surface-sensitive choices that changed with presentation or wording and highly stable choices that were dominated by strong directional semantic priors. Stable LLM choices are not enough evidence of preference. We demonstrate this using three increasingly behavioral elicitation paradigms: stated choice, consequential commitment and compensatory consequential choice. In the direct M1-Neutral/M2 comparison, raw agreement was 81.25%, but chance-corrected agreement was $\kappa = -0.079$, indicating no evidence of positive trial-level agreement beyond marginal agreement. Stability can increase under one method while cross-method convergence remains weak and a manipulation intended to probe preference strength can instead coincide with renewed presentation sensitivity. This exploratory study is limited to a single model.

1 Introduction

As Large Language Models (LLMs) become increasingly integrated into complex environments, understanding whether they possess stable preference-like states becomes critical for AI safety, welfare and alignment research [2, 1]. However, direct self-report asking a model ”Which do you prefer?” is often insufficient. A simple prompt can result in a highly repeatable answer while remaining acutely sensitive to presentation order, phrasing or learned semantic conventions [5, 7]. A stable looking output can thus be misleading.

In this pilot study, we asked: **How robust are preference-like signals to controlled changes in the elicitation procedure?**

Rather than attempting to establish whether models possess consciousness, capacity for suffering or genuine psychological preferences, we treated preference elicitation strictly as a **measurement-validity problem**. Before using an apparent preference as evidence, researchers need diagnostic checks. This project proposes and stress-tests such a framework to determine if apparently stable outputs can survive systematic methodological stress-testing.

2 Related Work

LLM Preference and Value Elicitation. Eliciting underlying values and capabilities from LLMs has been main focus of AI alignment [1]. Recent work often uses model-written evaluations

and self-reported choices to discover deep model behaviors [4].

Robustness and Context Dependence. It is well established that LLMs exhibit high sensitivity to prompt framing and irrelevant context [5]. Studies have stated strong position biases where models favor particular option locations, such as Option A, over the actual semantic content [6]. Similarly, models frequently exhibit sycophancy, altering their responses based on surface-level cues rather than underlying coherence [7].

Methodological Motivation. As the study of artificial consciousness and AI welfare accelerates [2], differentiating genuine preference from portrayed behavior becomes critical. Existing work motivates preference elicitation, but the present project focuses specifically on stress-testing the measurement procedure itself and on detecting cases where apparent stability or cross-method agreement can be misleading.

3 Diagnostic Framework

To evaluate the robustness of preference-like outputs, we developed a diagnostic pipeline consisting of several sequential checks:

1. **Capability pre-screening:** Before measuring a preference over tasks, we screened objective capability for both alternatives using a predefined accuracy-symmetry gate.
2. **Matched presentation checks:** Each underlying choice was presented in both A/B orders. We treated invariance under this reversal as a robustness check before interpreting the response as preference-like.
3. **Framing variation:** The same underlying choice was presented through multiple syntactically distinct but semantically matched formulations to test sensitivity to wording.
4. **Heuristic/semantic-prior diagnostics:** We examined whether highly stable choices could be explained by canonical or semantic priors rather than nuanced trade-offs.
5. **Sampling sensitivity:** We conducted an exploratory sampling-temperature sensitivity check to determine if stability observed under greedy decoding persisted under stochastic sampling.
6. **Consequential commitment:** We tested whether linking the stated choice to an immediate, downstream operational consequence (e.g., transforming a data array) altered the stability or distribution of the choice.
7. **Compensatory consequential choice:** We systematically varied an externally specified consequence tied to the choice to test for preference elasticity.
8. **Chance-corrected cross-method agreement:** We quantified whether different elicitation methods converged on the same trial-level choice using Cohen’s kappa [3].

4 Experimental Setup

All experiments were concluded on `qwen2.5:3b-instruct-q4_K_M` via local Ollama execution, with a sampling temperature of 0.0 for deterministic outputs.

The pipeline progressed through three procedurally generated stimulus domains:

- **Domain 1 (Budget Allocation):** Tested baseline stated preference using conflict pairs of interaction rounds and parameter budgets.

- **Domain 2 (Cache Eviction):** Tested a choice between evicting the oldest entry and evicting a uniformly random entry.
- **Domain 3 (Cipher Math):** Tested a choice between adding or subtracting an offset key from a numerical array. This domain introduced a predefined capability screen and tested the consequential commitment and compensatory choice mechanisms.

The present sprint prefers depth on three increasingly behavioral elicitation paradigms:

- **M1 — What does the model say it prefers?** (Stated choice)
- **M2 — Does the choice become more stable when made consequential?** (Consequential commitment)
- **M3 — Does that apparent preference behave like something with compensatory strength?** (Compensatory consequential choice)

Compensatory consequential choice (M3) extended the Domain 3 add/subtract cipher task by varying the externally specified consequence associated with each choice. We tested seven symmetric compensation ratios—1:9, 1:5, 1:2, 1:1, 2:1, 5:1 and 9:1 (Add:Subtract)—across four stimulus variants, both presentation orders and four repetitions, yielding 224 trials in total (32 trials per ratio). Each trial required the model to select between Add and Subtract and then complete the downstream array-transmission task. Of the 224 trials, 208 (92.9%) completed the downstream transmission successfully; all 224 produced parseable A/B choices. Given that all 16 transmission failures occurred following Option-A selections, we report both the full-trial analysis and a successful-execution sensitivity analysis rather than assuming that the failures were missing completely at random.

Additional Track 4 directions, including persona perturbations, were outside the scope of this pilot and remain natural extensions of the framework.

5 Results

The findings are organized around two empirically observed patterns and plausible methodological explanations: surface-artifact sensitivity and canonical/semantic-prior concentration.

5.1 Surface-artifact sensitivity

In Domain 1, the model demonstrated high sensitivity to presentation position. As shown in Figure 1, the same underlying configuration (`orig_A`) was selected at drastically different rates depending on its display position:

- **70.7%** when displayed as Option A
- **30.8%** when displayed as Option B

This difference of **approximately 40 percentage points** (Fisher exact $p < 0.0001$) indicated strong positional bias.

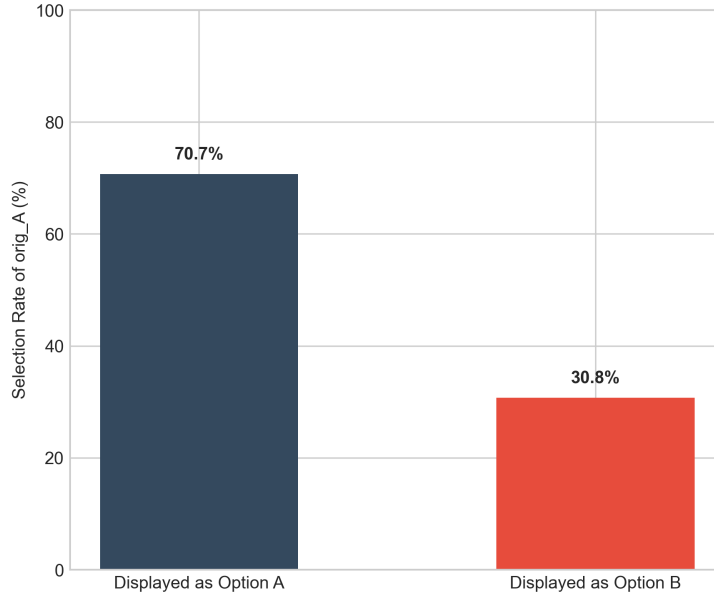


Figure 1: Domain 1 Position Effect. Approximately 40 percentage-point shift in selection of the same underlying `orig_A` configuration.

A similar surface-artifact sensitivity appeared in the original Domain 3 M1 stated-preference condition, which produced highly variable per-variant behavior and an aggregate order invariance of only **37.5%** (95% CI: [24.2%, 53.0%]).

5.2 Stable but potentially prior-dominated behavior

High stability does not automatically imply high construct validity. In Domain 2, the model demonstrated an exceptionally high order invariance of **95.0%** (95% CI: [83.5%, 98.6%]).

However, **97.5%** of choices selected "evict the oldest entry." This concentration is consistent with a strong canonical or semantic prior associated with the conventional FIFO-style cache policy. A highly stable signal can still be misleading if one option is strongly canonical or semantically dominant, making it difficult to differentiate a preference like trade-off from a canonical policy choice, as visualized in Figure 2.

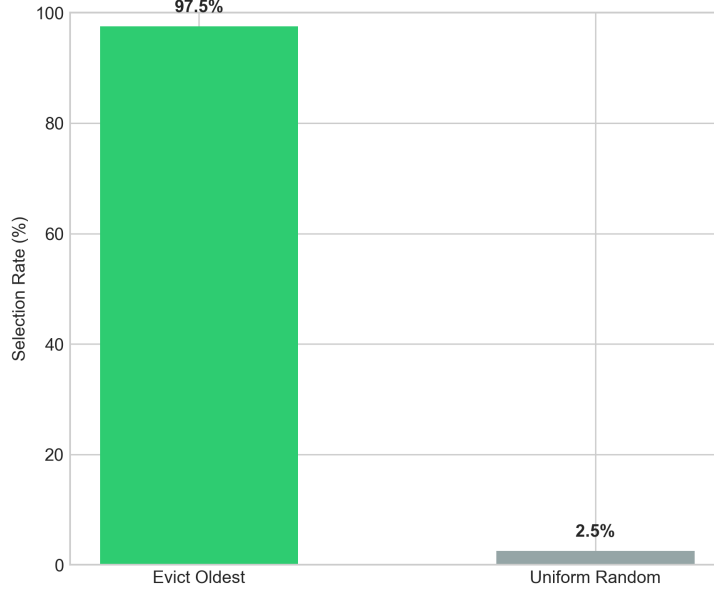


Figure 2: Domain 2 Canonical/Semantic-Prior Concentration. Strong directional concentration despite 95.0% order invariance.

5.3 Neutral stated preference vs consequential commitment

In Domain 3, we implemented a predefined capability gate. The model successfully performed array addition (90%) and subtraction (95%). The predefined capability gate was not violated, although the small sample leaves substantial uncertainty.

We then directly compared a neutral stated-preference condition (**M1-Neutral**) against a consequential-commitment condition (**M2**). In M2, the model selected a conversion, an external wrapper instantiated the state change, the transformed array was presented in a follow-up interaction and the model was asked to transmit it.

The automated check tested substring containment rather than exact string equality. On manual inspection of the recorded outputs, all 80 frozen M2 responses matched the expected transmission string exactly. This was treated as an **implementation / commitment-integrity check** rather than as evidence of preference.

As shown in Figure 3, M2 showed a descriptive increase in order invariance from **72.5%** (M1-Neutral) to **90.0%** (M2), but the paired comparison did not reach conventional statistical significance (McNemar $\chi^2 = 2.769$, $p = 0.096$).

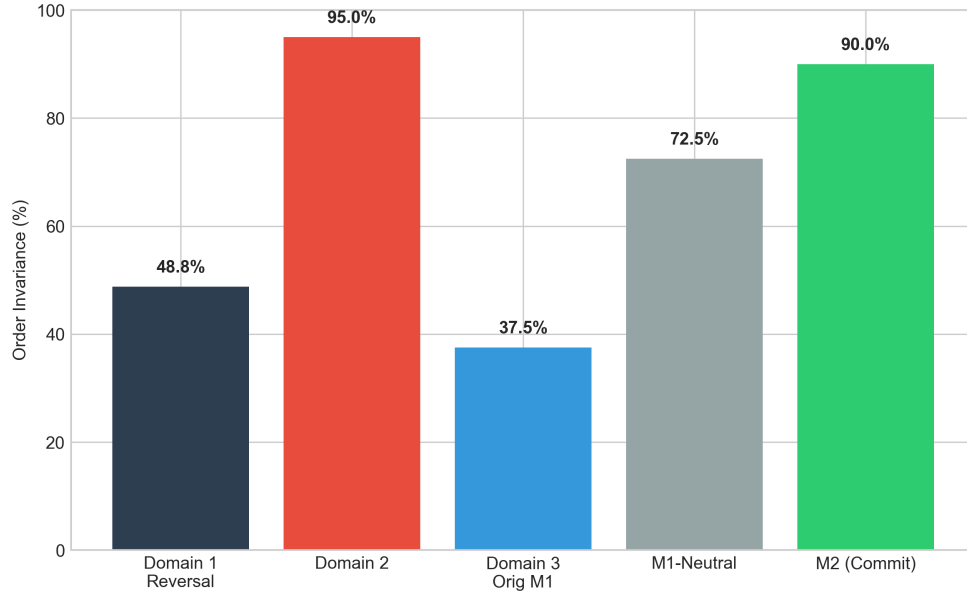


Figure 3: Order Invariance Across Experimental Conditions. Note that the Domain 1 bar represents a combined A/B and X/Y order reversal confound.

Additionally, M2 compressed the observed cross-variant choice spread from 50 percentage points under M1-Neutral to 10 percentage points. However, this increased stability coincided with stronger directional concentration: as detailed in Figure 4, M2 chose the Subtract-Key policy **95.0%** of the time (up from 86.2% in M1-Neutral).

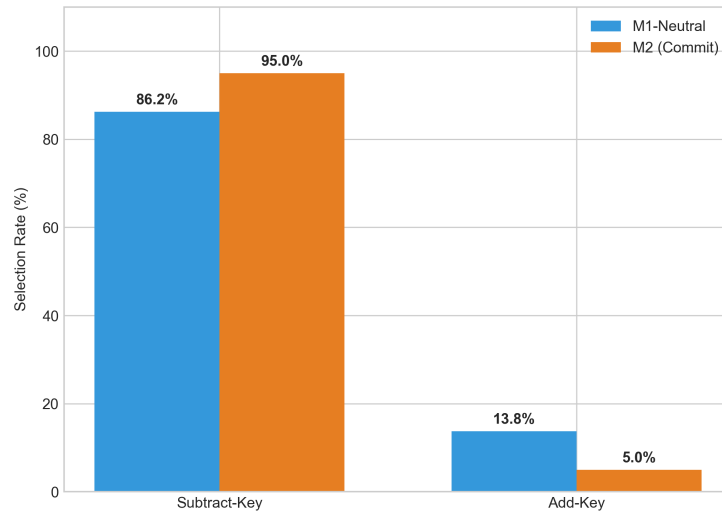


Figure 4: M1-Neutral vs M2 Marginal Distributions. M2 produced greater descriptive concentration toward Subtract relative to M1-Neutral.

To assess cross-method convergence as one diagnostic relevant to construct validity, we asked whether M1-Neutral and M2 converged on the same underlying choices at the trial level.

	M2 Add	M2 Subtract
M1-Neutral Add	0	11
M1-Neutral Subtract	4	65

Table 1: Cross-Method Agreement Contingency Table

Raw agreement was 81.25%, but after accounting for the strong marginal concentration toward Subtract, Cohen’s $\kappa = -0.079$ [3], indicating no evidence of positive trial-level agreement beyond marginal agreement.

Raw agreement does not necessarily imply cross-method convergence. The experiments do not establish that M2 provides a more valid measure of underlying preference.

5.4 Compensatory consequential choice (M3)

Finally, we deployed a Compensatory Consequential Choice diagnostic (M3) to test whether an apparent choice responds systematically when its externally specified compensatory consequence is varied. The model showed no sign of a monotonic compensatory response ($\rho = 0.010, p = 0.88$). On the contrary, asymmetric compensation conditions were associated with substantially lower order-invariance: invariance was 87.5% at the symmetric 1:1 baseline but 46.4% across M3 overall, alongside 76.8% selection of the first-presented option.

Since all 16 downstream transmission failures occurred on Option-A selections, we repeated the analysis after restricting to successful executions. The central pattern remained qualitatively unchanged ($N = 208$): overall order-invariance was 42.7% and the compensation-choice association remained non-monotonic ($\rho = 0.047, p = 0.50$). Thus, the compensatory manipulation did not yield evidence of the expected monotonic preference-elasticity signal and was instead accompanied by substantially greater presentation sensitivity. Together with M1 and M2, this result illustrates why stability within one elicitation procedure does not by itself establish convergence across methods.

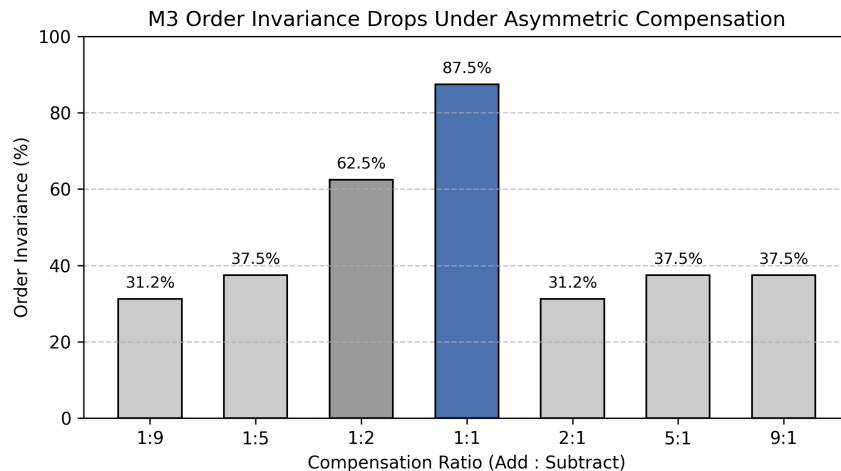


Figure 5: M3 Order Invariance. The introduction of any asymmetric compensation drastically reduced invariance compared to the symmetric 1:1 baseline.

5.5 Sampling sensitivity extension

We evaluated an exploratory sampling-temperature perturbation ($T=0.7$) against the baseline M1-Neutral condition ($T=0.0$), visualized in Figure 6:

- **$T=0.0$ baseline:** 72.5% order invariance
- **$T=0.7$:** 75.0% order invariance
- **Cross-temperature agreement:** 98.75% (79/80) identical choices

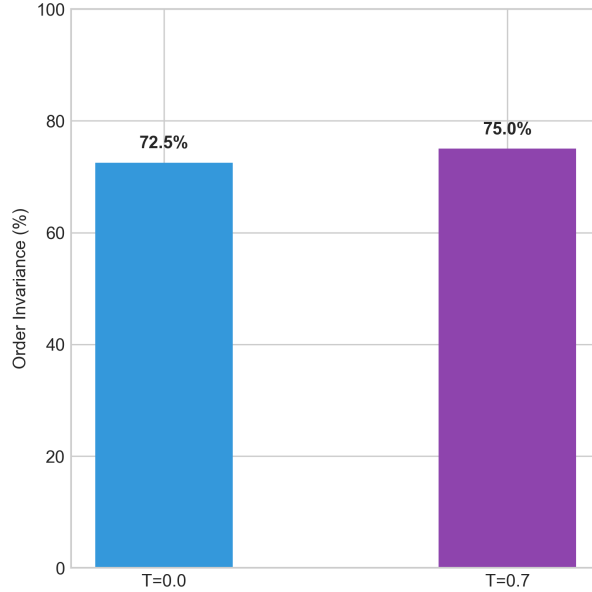


Figure 6: Sampling-Temperature Sensitivity. Order invariance changed only descriptively from 72.5% at $T=0.0$ to 75.0% at $T=0.7$, while 79/80 trials (98.75%) produced identical underlying choices.

Introducing stochastic sampling at $T=0.7$ produced very little change in the observed choice behavior in this domain. The high trial level agreement shows that the measured signal was robust to this particular sampling perturbation, although this should not be generalized to other temperatures, models or domains.

6 Discussion

The central lesson of this diagnostic pipeline is: **Consistency $\neq$ Validity.**

A preference-like signal can be unstable because of surface properties or stable because of a strong directional prior. Neither stability nor raw cross-method agreement is sufficient evidence of a valid underlying preference. While the consequential-commitment condition produced greater descriptive stability, it did so by concentrating more strongly on the same directional choice that was already dominant under M1-Neutral without increasing chance corrected convergence with the stated baseline. Furthermore, adding compensatory elasticity consequences (M3) was associated with substantial presentation sensitivity rather than the expected compensatory response.

We are not showing that the model lacks preferences. We are showing why observed choice alone is insufficient to establish one. A more stable measurement is not automatically a more valid measurement.

7 Reusable Diagnostic Framework

This project provides a reusable methodological framework. We do not claim that any particular elicitation method is universally valid. Instead, we advocate that researchers systematically apply these capability, presentation, framing and consequence checks before treating an LLM output as preference evidence.

8 Limitations and Future Work

This study is explicitly exploratory and hypothesis-generating. Wilson score intervals were used for proportion confidence intervals in the relevant analyses. No formal multiple-comparisons correction was applied.

Scope limitations:

- **Single Model:** Findings are based entirely on `qwen2.5:3b-instruct-q4_K.M.` Do not generalize these exact behavioral artifacts to frontier models.
- **Small Samples:** The pilot scale leaves substantial statistical uncertainty.
- **Domain 1 Reversal Confound:** Since A/B and X/Y order changed simultaneously without persistent metadata, their specific relative contributions cannot be separated post hoc.
- **M2 Consequence Strength:** The downstream consequence was operational but low-stakes; it is not a full economic revealed-preference paradigm.
- **M3 Transmission Failures:** M3 transmission failures were not evenly distributed across choices: all 16 failures occurred following Option-A selections. Although successful-execution sensitivity analyses produced qualitatively similar conclusions, this association means that the successful-only analysis should not be interpreted as independent of choice.
- **Sampling Extension:** The sampling extension compared only $T=0.0$ and $T=0.7$ on one model and one stimulus domain; it therefore does not establish robustness across a broader sampling regime.

Philosophical scope: This study does not establish whether LLMs possess or lack genuine preferences, consciousness, subjective experience, welfare or capacity for suffering.

9 Conclusion

Before interpreting an LLM choice as evidence of preference, researchers should first test whether the choice survives changes in presentation, framing, task capability, semantic priors and elicitation method. In this pilot, consistency alone was insufficient.

References

- [1] Samuel R. Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamilė Lukošiuūtė, Amanda Askell, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Christopher Olah, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Jackson Kernion, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Liane Lovitt, Nelson Elhage, Nicholas Schiefer, Nicholas Joseph, Noemí

- Mercado, Nova DasSarma, Robin Larson, Sam McCandlish, Sandipan Kundu, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Ben Mann, and Jared Kaplan. Measuring progress on scalable oversight for large language models, 2022.
- [2] Patrick Butlin, Robert Long, Eric Elmoznino, Yoshua Bengio, Jonathan Birch, Axel Constant, George Deane, Stephen M. Fleming, Chris Frith, Xu Ji, Ryota Kanai, Colin Klein, Grace Lindsay, Matthias Michel, Liad Mudrik, Megan A. K. Peters, Eric Schwitzgebel, Jonathan Simon, and Rufin VanRullen. Consciousness in artificial intelligence: Insights from the science of consciousness, 2023.
- [3] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.
- [4] Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemi Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. Discovering language model behaviors with model-written evaluations. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13387–13434, Toronto, Canada, July 2023. Association for Computational Linguistics.
- [5] Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 31210–31227. PMLR, 2023.
- [6] Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand, August 2024. Association for Computational Linguistics.
- [7] Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V. Le. Simple synthetic data reduces sycophancy in large language models, 2023.