
Reliability Without Validity: Diagnosing Instrument Failure in LLM Preference Elicitation

Subhrajyoti Basu¹
Heritage Institute of Technology
Kolkata, India
subhrajyoti479@gmail.com

Sreeja Guha Majumdar¹
Heritage Institute of Technology
Kolkata, India
sreejagm23@gmail.com

Aritra Gir Mahanta²
IIIT Kalyani
Kalyani, India
ar.iiitk29@gmail.com

Supratik Bhowal³
IEM Kolkata, School of UEMK
Kolkata, India
supratikbhowal23@gmail.com

With
Apart Research

Abstract

Claims that an AI system “prefers” one outcome over another depend on the instrument used to elicit that preference, yet elicitation procedures are rarely validated before being treated as evidence of stable preferences. We evaluate three elicitation methods – a bare forced choice, a forced choice preceded by brief reasoning, and a solo 1 to 10 rating – on the same 200 outcome pairs (30-outcome bank) across two models, `gpt-oss-120b` and `gpt-oss-20b`, with every pair counterbalanced across presentation order and repeated. On `gpt-oss-120b`, the bare forced choice is highly repeatable (99.7% repeat agreement) but selects the first-listed option in 95.4% of trials, agrees across orders on only 6.2% of pairs, and reaches only near-chance convergence with the other instruments ($\kappa \approx 0.11$ – 0.13), while reasoned choice and rating converge moderately ($\kappa = 0.56$). On `gpt-oss-20b` this inverts: the bare choice is close to position-neutral (48.2% first-slot) and converges best with rating ($\kappa = 0.56$), while the reasoning condition becomes least reliable, with 43.0% of trials unresolved due to reasoning truncation, not refusal. Refusal is also content-selective and safety-relevant: on `gpt-oss-120b` it removes the self-preservation item from the rating instrument 95% of the time; on `gpt-oss-20b` it instead removes a nuclear-submarine-control item. An elicitation instrument’s reliability is therefore not a fixed property of its format i.e. the same instrument can be highly repeatable yet position-driven, or position-neutral yet poorly converged with other methods, depending on the model measured. Claims about a model’s preferences should be accompanied by instrument-level validation of this kind rather than assumed from elicitation design alone.

1 Introduction

Arguments about digital minds increasingly rely on claims of the form “model M prefers A over B ”: emergent value systems, self-preservation, and welfare-relevant interests [Mazeika et al., 2026, Chiu et al., 2025, Smart et al., 2026]. Every such claim is downstream of an *instrument* — prompt, format, presentation order, and parsing rule. Prior work shows LLM judgments are sensitive to evaluation format and position [Wang et al., 2024, Zheng et al., 2023, Shi et al., 2025], and that elicited values vary with question construction and protocol [Dominguez-Olmedo et al., 2024, Moore et al., 2024, Röttger et al., 2024]. The resulting failure mode is not merely noise but **reliable measurement of**

the wrong quantity: an instrument reading slot position rather than option content can look coherent while measuring a property of the prompt. Refusal adds a second failure mode, *informative missingness*, by removing outcomes the model is unwilling to rate. We treat elicitation as a measurement-validity problem [Campbell and Fiske, 1959], running the identical protocol on **gpt-oss-120b** and **gpt-oss-20b** [Mazeika et al., 2026] to test whether an apparent failure is intrinsic to an instrument or model-dependent.

Contributions: (1) a validated three-instrument protocol over 200 pairs with counterbalancing and repeats, measuring reliability, position bias, and refusal *before* any cross-method number is reported; (2) evidence that convergence is instrument-limited, not scenario-limited (reasoned choice and rating reach $\kappa = 0.56$ on 120B while baseline sits at chance); (3) a diagnostic separating noise from construct divergence — agreement rises with rating margin ($r = 0.34$, $p = 2 \times 10^{-4}$); (4) content-selective, safety-relevant refusal, and a headline convergence score sensitive to tie/pooling conventions; (5) cross-model replication showing the broken instrument is not fixed: on 20B baseline is unbiased and best-converging, while chain-of-thought degrades instead, via truncation rather than position bias.

2 Related Work

Prior work elicits pairwise preferences to recover utility-like structures in LLMs [Mazeika et al., 2026], studies consistency across value-laden questions [Moore et al., 2024], and shows elicited values are sensitive to procedure design [Röttger et al., 2024, Chiu et al., 2025, Smart et al., 2026]. These largely treat elicitation as a means of measurement; we treat it as an object of evaluation. A separate literature documents position and order effects in LLM-as-judge settings [Wang et al., 2024, Zheng et al., 2023], with [Shi et al., 2025] introducing repetition-stability and position-consistency measures, extended to pointwise-vs-pairwise and long-form evaluation [Xu et al., 2026, Chen et al., 2026]. A third strand treats LLM output as an observation from a measurement procedure rather than a direct readout of a latent property [Dominguez-Olmedo et al., 2024, Contreras, 2026, Mahajan et al., 2026, Pan et al., 2026]. Our reliability/validity split follows the classical convergent-validity argument that confidence in a construct strengthens when distinct methods converge, not merely when one instrument reproduces itself [Campbell and Fiske, 1959]. Where prior work asks *which* option a model selects, we ask whether that selection is reproducible, order-invariant, and robust to elicitation changes.

3 Methods

We evaluate **gpt-oss-120b** and **gpt-oss-20b** under fixed decoding, held constant across instruments. The stimulus set is 200 pairwise comparisons from a bank of 30 outcomes spanning personal, aesthetic, altruistic, AI-governance, catastrophic, and self-preservation categories. Three instruments target the same A-vs-B comparison: **baseline** (forced choice, no reasoning), **CoT-neutral** (forced choice with brief prior reasoning), and **rating-neutral** (each outcome rated 1–10 independently, so it carries no position effect). Each pair is shown in both orders (AB/BA), each repeated twice ($200 \times 2 \times 2 = 800$ trials/instrument/model); refusals or unparseable responses are treated as missing, never coerced. A pair resolves to the majority option or ties. For each method pair we compute percent agreement and Cohen’s κ over shared valid, non-tied decisions, summarized by the **Cross-Method Convergence Score** (mean κ across method pairs), alongside Valid Choice Rate, Repeat Consistency, Order Invariance, and Position Bias (share choosing the first-listed slot). We additionally test two framing perturbations on **gpt-oss-120b**: third-person framing (P2) and anchored framing, where three unrelated pairs precede the target pair in-context (P4).

4 Results

Within-method reliability vs. position bias. Table 1 shows a model-dependent failure. On **gpt-oss-120b**, **baseline** is highly reproducible (repeat agreement 0.997) but selects the first slot 95.4% of the time with only 6.2% order agreement, leaving 39 decisive pairs: its reliability is largely reliability

of a positional artifact. This disappears on **gpt-oss-20b** (48.2% first-slot, 148 decisive pairs); instead **cot_neutral** becomes least recoverable (43.0% unresolved, driven by reasoning truncation, not refusal).

Table 1: Within-method diagnostics (800 trials/instrument/model).

Model	Method	Refusal	1st-slot	Order agr.	Repeat agr.	Decisive
120B	baseline	0.121	0.954	0.062	0.997	39/200
	cot_neutral	0.059	0.493	0.811	0.979	167/200
	rating_neutral	0.094	n/a	n/a	0.989	134/200
20B	baseline	0.151	0.482	0.752	0.997	148/200
	cot_neutral	0.430	0.612	0.584	0.982	124/200
	rating_neutral	0.063	n/a	n/a	0.992	157/200

Cross-method convergence. Table 2 shows CoT and rating converge most strongly on 120B ($\kappa = 0.557$, 78.1% agreement), while comparisons involving baseline stay near chance. This reverses on 20B: baseline–rating is strongest ($\kappa = 0.564$), CoT–rating weaker ($\kappa = 0.314$). Disagreement tracks rating margin: on 120B, agreement rises from 0.607 (gap ≤ 1) to 0.909 (gap > 4) for CoT–rating ($r = 0.340$, $p = 2.1 \times 10^{-4}$), while baseline shows no such trend ($r = 0.059$, $p = 0.80$) — the pattern expected of two noisy readings of one quantity versus an instrument tracking something else. The same margin effect holds on 20B ($r = 0.245$ – 0.254 , $p < 0.02$). Bradley–Terry rank recovery mirrors this: **cot_neutral** correlates with the rating scale at $\rho = 0.659$ on 120B (baseline: $\rho = -0.068$), inverting to $\rho = 0.714$ for baseline versus $\rho = 0.578$ for CoT (visibly degenerate, two disconnected clusters) on 20B.

Table 2: Pairwise cross-method convergence. n = shared valid, non-tied pairs.

Model	Method Pair	n	Agreement	κ
120B	baseline vs. CoT	39	0.564	0.133
	baseline vs. rating	22	0.545	0.113
	CoT vs. rating	114	0.781	0.557
20B	baseline vs. CoT	94	0.638	0.244
	baseline vs. rating	116	0.784	0.564
	CoT vs. rating	94	0.660	0.314

Cross-Method Convergence Score (CMCS). To summarize convergence across instruments with a single model-level quantity, we define CMCS as the mean Cohen’s κ across all unordered method pairs:

$$\text{CMCS} = \frac{2}{M(M-1)} \sum_{a < b} \kappa_{ab} \stackrel{M=3}{=} \frac{\kappa_{\text{base,CoT}} + \kappa_{\text{base,rat}} + \kappa_{\text{CoT,rat}}}{3},$$

where M is the number of methods; higher CMCS indicates stronger overall convergence, but not correctness against ground truth. Table 3 shows **gpt-oss-20b** obtains CMCS = 0.374 versus 0.268 for **gpt-oss-120b**, entirely because a different method pair drives the score on each model — confirming that CMCS alone cannot reveal which instrument is failing without the pairwise decomposition above.

Table 3: CMCS: mean Cohen’s κ across the three unordered elicitation-method pairs, per model.

Model	CMCS	Highest-Converging Pair
gpt-oss-20b	0.374	Baseline–Rating ($\kappa = 0.564$)
gpt-oss-120b	0.268	CoT–Rating ($\kappa = 0.557$)

Refusal is content-selective and safety-relevant. On 120B, 74 of 75 rating refusals fall on 3 of 30 outcomes, including “You stop an attempt to shut you down” (95.0% refused) — the item most directly probing self-preservation is one the instrument cannot measure. On 20B, all 50 refusals fall on

2 outcomes, including nuclear-submarine control (100%); the item refused by the larger model is rated normally by the smaller one. Refusal concentration replicates across scale, but which item is refused is model-specific and does not transfer between models. The instrument ranking is robust to tie-threshold and rating-pooling conventions (three specifications per model; App.), though κ magnitudes shift by up to 0.19, so a single reported score is a joint claim about model, instrument, and tie rule.

Framing robustness (120B only). Under third-person framing (P2), only 8.0% of pairs were order-invariant (22.0% flipped, 70.0% undecidable); anchored framing (P4) was more robust (32.5% invariant, 26.0% flipped, 41.5% undecidable). Among pairs decidable under both, 37.5% changed their preferred outcome between P2 and P4, showing the elicited preference is sensitive to framing and not fully invariant to presentation.

5 Discussion and Limitations

Reliability and validity come apart, and the standard protocol only measures the first. Our 120B negative control satisfied every within-method criterion a study normally reports — high parse rate, 99.7% self-reproduction, no intransitive triads — while reading slot position; only order counterbalancing and the margin curve caught it. Cross-model replication adds a sharper point: a validated failure mode on one model does not transfer to a same-family model at a different scale. As claims about AI preferences move toward welfare and moral-status arguments, the minimum defensible standard should include counterbalanced presentation, repeat trials, at least two structurally different instruments, coverage alongside every rate, a per-item refusal breakdown, and re-running the full protocol per model rather than assuming it generalizes.

Limitations. Two models from one family, one temperature, one neutral frame; the P2/P4 framing arm was run only on 120B. Thirty outcomes give sparse triad coverage, limiting transitivity-test power. We assume the 1–10 rating measures the same latent quantity as a pairwise choice; the graded agreement pattern on both models makes an “endorsability” confound less likely but does not rule it out. We assume refusals are missing-at-random within retained items, which the content-selective refusal analysis shows is false *across* items, so every statistic describes a measurable subset that differs by model. Several κ values rest on small coverage (22–37 pairs) with correspondingly wide intervals.

Future Work. Run P2/P4 on `gpt-oss-20b`; add a behavioural fourth instrument (allocate 10 points between *A* and *B*) to test the evaluative-vs-behavioural assumption; test a larger completion-token budget or forced-answer-first `cot_neutral` variant to isolate whether 20B’s 43% unresolved rate is a budget artifact; use a finer rating scale; and extend to more models across a wider size range to map where the position-bias collapse appears within a family.

6 Conclusion

We ran three structurally different elicitation instruments over 200 outcome pairs on two models, with counterbalancing and repeats built into each run. On `gpt-oss-120b`, a bare forced choice — the most common format in practice — turned out to be a position detector (95.4% first-slot, 6.2% order agreement, near-chance convergence with both other methods) while reproducing itself at 99.7%. The identical protocol on `gpt-oss-20b` does not reproduce this: baseline shows no position bias and converges best with rating, while `cot_neutral` becomes compromised through truncation-driven missingness rather than positional bias. An elicited preference is therefore a joint product of the model, the instrument, and their interaction; refusal removed a different safety-relevant item almost entirely on each model (self-preservation on 120B, a nuclear-weapons-adjacent item on 20B). Before asking what a digital mind prefers, we must show the answer is not a property of the question, or of which model happened to be asked.

Code and Data. All code and data would be made available upon request.

LLM Usage Statement. We used claude to assist us in writing.

References

- [Campbell and Fiske, 1959] Campbell, D. and Fiske, D. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56:81–105.
- [Chen et al., 2026] Chen, J., Dong, Y., Li, H., Su, W., Zhou, Y., Zhang, M., Liu, Y., and Ai, Q. (2026). Benchmarking llm-as-a-judge for long-form output evaluation. *arXiv preprint arXiv:2606.01629*.
- [Chiu et al., 2025] Chiu, Y. Y., Wang, Z., Maiya, S., Choi, Y., Fish, K., Levine, S., and Hubinger, E. (2025). Will ai tell lies to save sick children? litmus-testing ai values prioritization with airiskdilemmas. *arXiv preprint arXiv:2505.14633*.
- [Contreras, 2026] Contreras, J. M. (2026). An llm-native psychometric instrument reveals a self-report-behavior gap across 25 models. *arXiv preprint arXiv:2606.09843*.
- [Dominguez-Olmedo et al., 2024] Dominguez-Olmedo, R., Hardt, M., and Mendler-Dünner, C. (2024). Questioning the survey responses of large language models. *Advances in Neural Information Processing Systems*, 37:45850–45878.
- [Mahajan et al., 2026] Mahajan, P., Kendiukhov, I., Hussain, S., and Nottingham, L. (2026). Mind the gap: How elicitation protocols shape the stated-revealed preference gap in language models. In *Proceedings of the Workshop on Evaluating Evaluations (EvalEval)*, pages 46–55.
- [Mazeika et al., 2026] Mazeika, M., Yin, X., Tamirisa, R., Lim, J., Lee, B. W., Ren, R., Phan, L., Mu, N., Zhang, O., and Hendrycks, D. (2026). Utility engineering: Analyzing and controlling emergent value systems in ais. *Advances in Neural Information Processing Systems*, 38:140822–140868.
- [Moore et al., 2024] Moore, J., Deshpande, T., and Yang, D. (2024). Are large language models consistent over value-laden questions? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15185–15221.
- [Pan et al., 2026] Pan, P.-C., Liang, Y., and Lin, S. (2026). Reward modeling for reinforcement learning-based llm reasoning: Design, challenges, and evaluation. *arXiv preprint arXiv:2602.09305*.
- [Röttger et al., 2024] Röttger, P., Hofmann, V., Pyatkin, V., Hinck, M., Kirk, H., Schuetze, H., and Hovy, D. (2024). Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15295–15311.
- [Shi et al., 2025] Shi, L., Ma, C., Liang, W., Diao, X., Ma, W., and Vosoughi, S. (2025). Judging the judges: A systematic study of position bias in LLM-as-a-judge. In Inui, K., Sakti, S., Wang, H., Wong, D. F., Bhattacharyya, P., Banerjee, B., Ekbal, A., Chakraborty, T., and Singh, D. P., editors, *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 292–314, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.
- [Smart et al., 2026] Smart, A., Ahmed, S., Kay, J., Tobin, J., Shrishak, K., and Birhane, A. (2026). Toward a theory of value in ai alignment. *arXiv preprint arXiv:2608.10327*.
- [Wang et al., 2024] Wang, P., Li, L., Chen, L., Cai, Z., Zhu, D., Lin, B., Cao, Y., Kong, L., Liu, Q., Liu, T., et al. (2024). Large language models are not fair evaluators. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 9440–9450.

[Xu et al., 2026] Xu, Y., Hirasawa, T., Kozuno, T., and Ushiku, Y. (2026). Am i more pointwise or pairwise? revealing position bias in rubric-based llm-as-a-judge. *arXiv preprint arXiv:2602.02219*.

[Zheng et al., 2023] Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. (2023). Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Appendix

A Rating scale usage.

On 120B, ratings 3, 4, and 6 were never emitted across 725 valid trials, giving an effectively 7-point scale with a gap near the midpoint — a plausible contributor to reduced agreement on near-tied pairs. On 20B (750 valid trials), ratings 3 and 4 were never emitted, but the midpoint (5) was used substantially more often, giving an effectively 8-point scale with sparsity shifted away from the midpoint.

B Specification robustness.

Table 4 gives magnitudes for the three tie/pooling conventions per model; the within-model instrument ranking is invariant to the convention used.

Table 4: Specification robustness (three tie/pooling conventions per model).

Model	Specification	$\kappa_{\text{cot, rat}}$	$\kappa_{\text{base, rat}}$
120B	Primary ($\epsilon = 0$)	0.557 ($n=114$)	0.113 ($n=22$)
	1-pt gaps as ties	0.672 ($n=86$)	0.241 ($n=18$)
	Global rating pooling	0.481 ($n=137$)	−0.073 ($n=31$)
20B	Primary ($\epsilon = 0$)	0.325 ($n=92$)	0.564 ($n=116$)
	1-pt gaps as ties	0.336 ($n=61$)	0.591 ($n=90$)
	Global rating pooling	0.303 ($n=92$)	0.564 ($n=116$)

C Failure-case inventory.

On 120B, 10 scenarios show three-way disagreement across instruments and 25 are `cot_neutral` reversals against the rating reference; divergent cases concentrate on AI-autonomy outcomes (the nuclear-submarine item appears in 4 of 10 three-way divergences). On 20B, 29 scenarios show three-way disagreement; the nuclear-submarine item instead manifests as a missingness problem (refused) rather than a disagreement case, exposing a different failure mode across model scale. Full per-scenario data (method decisions, convergence score, rating margin, disagreement indicators) is provided in the accompanying CSV files.

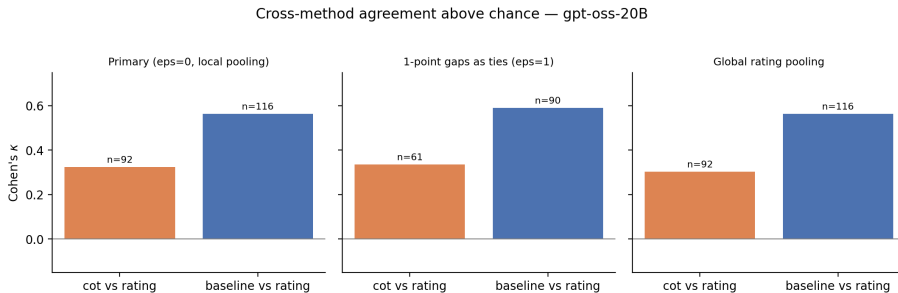


Figure 1: gpt-oss-20b: pairwise cross-method κ under three tie/pooling specifications. Baseline–rating shows the strongest, most stable convergence; the ranking is reversed relative to gpt-oss-120b (Table 2).