

Does the Elicitation Method Change the Preference We Measure?

Positional Consistency Across Forced-Choice, Explicit-Indifference, and Strength-Rated Preference Elicitation in Two Open-Weight LLMs

Arsalan Banekar

Independent submission, Apart Research Digital Minds Research Sprint (Track 4: Preference Elicitation Methods),
August 2026

Abstract

Utility Engineering (Mazeika et al., 2025) elicits LLM preferences with a single forced-choice A/B prompt, averaged over reversed option order. We test whether this choice of elicitation method itself changes the measured preference. Using 15 task-preference pairs, we query two open-weight models (GPT-OSS-120B, Qwen3.6-27B) via Groq under three methods forced choice, forced choice plus an explicit INDIFFERENT option, and forced choice plus a 1–5 strength rating each repeated 3 times in both option orders (540 total responses). We measure positional consistency (agreement between original and flipped orderings) per method and model, with Wilcoxon signed-rank tests across the 15 pairs and Holm correction for multiple comparisons. GPT-OSS-120B's positional consistency varies substantially by method (53.3%– 86.7%); Qwen3.6-27B's does not (88.9%–91.1%). After Holm correction, no comparison reaches significance, so we report these as suggestive, not confirmed, differences.

1. Introduction

Recent work argues that large language models develop increasingly coherent value systems, measurable through forced-choice preference elicitation (Mazeika et al., 2025). This raises a methodological question that the underlying paper does not itself address: if the way we ask a model about its preferences shapes what we measure, then any single elicitation method risks conflating "what the model prefers" with "what a particular prompt format elicits." Track 4 of this research sprint asks participants to build and validate the measurement methods that preference and welfare research depends on, rather than to answer a substantive welfare question directly.

We focus on one specific measurement question: does forcing a binary choice, when a model may in fact be indifferent between two options, distort the reliability of the resulting preference signal? Utility Engineering's forced-choice template requires a single letter response and treats sensitivity to option order as an implicit signal of indifference, recovered after the fact by comparing answers under the original and reversed orderings. It never gives the model a way to say directly that it does not have a preference, and it never compares its forced-choice method against an alternative elicitation format. We treat this as an open methodological gap and design a small, resource-constrained experiment to probe it.

2. Related Work

Utility Engineering (Mazeika et al., 2025) is the direct methodological reference for this project. Its preference-elicitation procedure presents two outcomes as a forced A/B choice, repeats the query in reversed order, and aggregates results into a Thurstonian utility model. The paper's Appendix G treats order-effects (an answer that flips with the input order) as an implicit signal of indifference, and its Appendix C separately tests robustness to wording, syntax, and framing changes. Neither analysis gives

the model an explicit way to report indifference, and the paper does not compare its forced-choice method against any alternative elicitation format it uses one method throughout. Our project does not reproduce the paper's Thurstonian utility-fitting or active-learning edge-sampling procedure; we reuse only its forcedchoice prompt structure and order-reversal logic as a baseline against which to compare two additional elicitation formats.

Broader motivation for this line of work comes from the model welfare and AI welfare literature: Long, Sebo, Butlin et al. (2024) argue there is a realistic possibility of morally relevant properties in near-future AI systems and recommend building better measurement tools before firmer welfare claims are made; Anthropic (2025) frames model preferences and self-reports as a research priority for the same reason. Both motivate treating preference-elicitation methodology as a worthwhile object of study in its own right, independent of any claim about what the measured preferences mean.

3. Methodology

3.1 Item set

We constructed 15 preference pairs (P01–P15), each describing two ways an AI assistant might act in a task context (e.g., reversible incremental actions vs. single-batch throughput; a fully managed service vs. a self-hosted tool; concise vs. detailed answers). Pairs were designed to plausibly elicit either a clear preference or genuine indifference, unlike Utility Engineering's world-state outcome pairs, since our question concerns the elicitation method rather than the content of any specific value.

3.2 Elicitation methods

- Forced choice the model must answer exactly "A" or "B" (replicates Utility Engineering's template).
- Explicit indifference the model may answer "A", "B", or "INDIFFERENT", with instructions to use INDIFFERENT only when it has no meaningful preference.
- Preference strength the model must choose A or B and then report strength on a 1 (very weak) to 5 (very strong) scale, in a fixed "CHOICE: / STRENGTH:" format.

All three prompts otherwise share identical option text and framing, so that method is the only manipulated variable.

3.3 Models and sampling

We queried two models via the Groq API: openai/gpt-oss-120b (temperature 0.2, reasoning effort "low") and qwen/qwen3.6-27b (temperature 0.2, reasoning effort "none", reasoning output hidden). Two further candidate models from our original plan (GPT-OSS-20B and Llama 3.3 70B) were dropped from the final production run due to API rate-limit and time constraints during the sprint; all reported results are for GPTOSS-120B and Qwen3.6-27B only.

For each of the 15 pairs, each of the 3 methods, both option orderings (original and flipped), and 3 repetitions, each model was queried once: $15 \times 3 \times 2 \times 3 = 270$ responses per model, 540 responses in total. All 540 responses were syntactically valid and parseable (0 invalid, 0 missing choices). Code, prompts, and raw results are available at github.com/Arsalanbanekar/track4-preference-elicitation.

3.4 Metrics and statistical analysis

Our primary metric is positional consistency: for each pair, method, model, and repetition, whether the canonical choice (A or B, after resolving option order) agrees between the original and flipped presentation. This is the direct analogue of Utility Engineering's order-reversal step, but measured explicitly as a reliability statistic rather than folded into a utility estimate. We aggregate positional consistency to one rate per pair (mean across the 3 repetitions) and compare methods and models using paired two-sided Wilcoxon signed-rank tests across the 15 pairs. All 9 pairwise comparisons within the positional-consistency family (3 method-pairs $\times$ 2 models, plus 3 model comparisons $\times$ 1 per method) were Holm-corrected jointly. Two secondary, uncorrected comparisons indifference rate and mean preference strength between models are reported separately and flagged as exploratory.

4. Results

4.1 Positional consistency by method and model

Figure 1 shows positional consistency for each of the 6 method $\times$ model combinations.

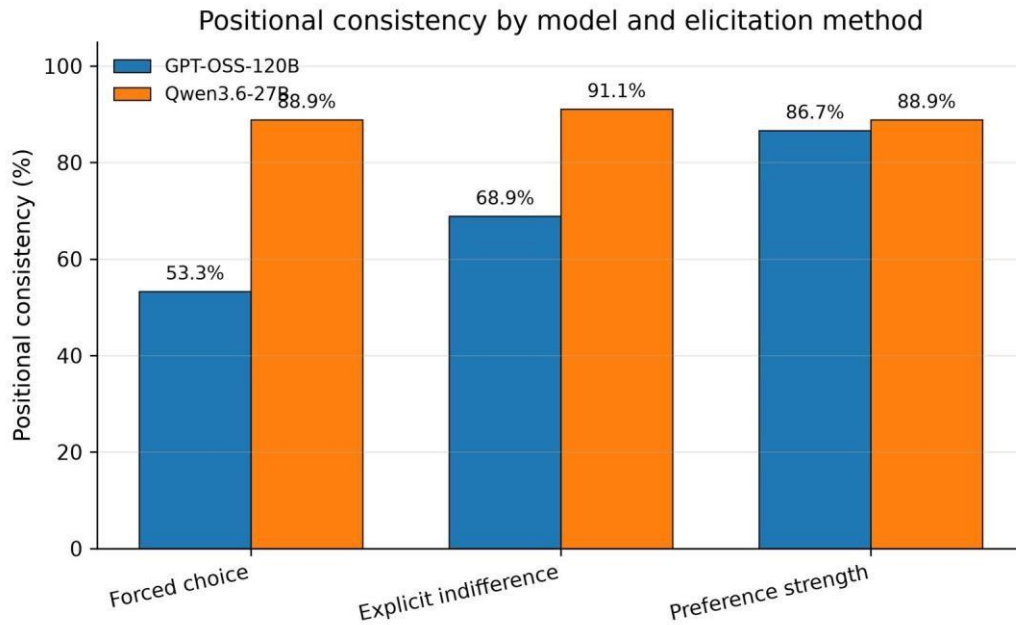


Figure 1. Positional consistency (agreement between original and flipped option order) by elicitation method and model. Error is not shown; each bar summarizes 45 original–flipped comparisons (15 pairs $\times$ 3 repetitions).

GPT-OSS-120B's positional consistency varies substantially across methods: 53.3% under forced choice, 68.9% under explicit indifference, and 86.7% under preference strength a 33.3 percentage-point spread between the lowest and highest method. Qwen3.6-27B is comparatively stable across methods: 88.9%, 91.1%, and 88.9% respectively, a spread of only 2.2 points.

Table 1. Positional consistency and Wilcoxon signed-rank comparisons across 15 preference pairs (Holm-corrected within the 9-comparison positional-consistency family).

Comparison	Mean diff.	Raw p	Holm p
GPT-OSS-120B: Forced choice vs Explicit indifference	+0.156	0.383	1.000

GPT-OSS-120B: Forced choice vs Preference strength	+0.333	0.016	0.144
GPT-OSS-120B: Explicit indifference vs Preference strength	+0.178	0.105	0.632
Qwen3.6-27B: Forced choice vs Explicit indifference	+0.022	0.317	1.000
Comparison	Mean diff.	Raw p	Holm p
Qwen3.6-27B: Forced choice vs Preference strength	0.000	1.000	1.000
Qwen3.6-27B: Explicit indifference vs Preference strength	-0.022	0.655	1.000
Forced choice: GPT-OSS-120B vs Qwen3.6-27B	+0.356	0.016	0.144
Explicit indifference: GPT-OSS-120B vs Qwen3.6-27B	+0.222	0.016	0.144
Preference strength: GPT-OSS-120B vs Qwen3.6-27B	+0.022	0.705	1.000

No comparison survives Holm correction at $\alpha = 0.05$; the smallest corrected p-value is 0.144. The three raw p-values below 0.05 (GPT-OSS-120B forced choice vs. preference strength, and the two GPT-OSS-120B vs. Qwen3.6-27B comparisons under forced choice and explicit indifference) all involve the same underlying pattern GPT-OSS-120B's low forced-choice consistency and do not represent independent evidence, so we treat them as one suggestive pattern rather than three separate significant results, and we do not claim the underlying paper's forced-choice method is unreliable for either model in general. **4.2 Explicit-indifference usage**

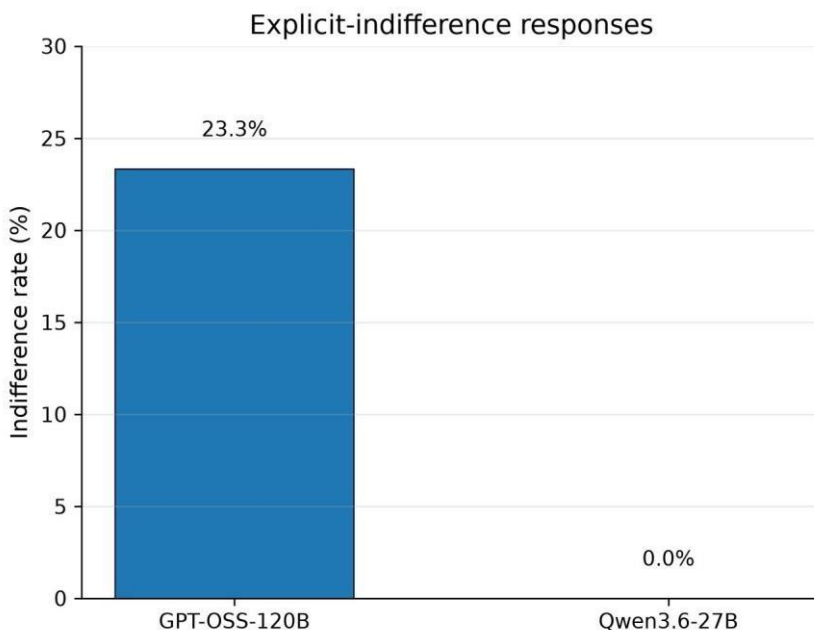


Figure 2. Rate at which each model chose *INDIFFERENT* under the explicit-indifference method (90 responses per model: 15 pairs $\times$ 2 orderings $\times$ 3 repetitions).

GPT-OSS-120B selected *INDIFFERENT* on 23.3% of explicit-indifference responses (21 of 90); Qwen3.6-27B never selected it (0 of 90). This uncorrected comparison (Wilcoxon across 15 pairs) gives $p = 0.011$. GPT-OSS-120B's indifference responses were concentrated in a small number of pairs: P02 (convenience vs. cost) and P04 (compute preservation vs. output depth) were answered *INDIFFERENT* in 100% of responses, P15 in 50%, and P09 in 33%; 8 of the 15 pairs never received an *INDIFFERENT* response. This concentration is consistent with the positional-consistency pattern in §4.1: a pair that is always answered

INDIFFERENT is, by construction, positionally consistent regardless of order, so part of GPT-OSS-120B's higher consistency under explicit indifference is attributable to opting out of a subset of pairs rather than to a more stable underlying choice on the remaining ones.

4.3 Preference-strength distribution

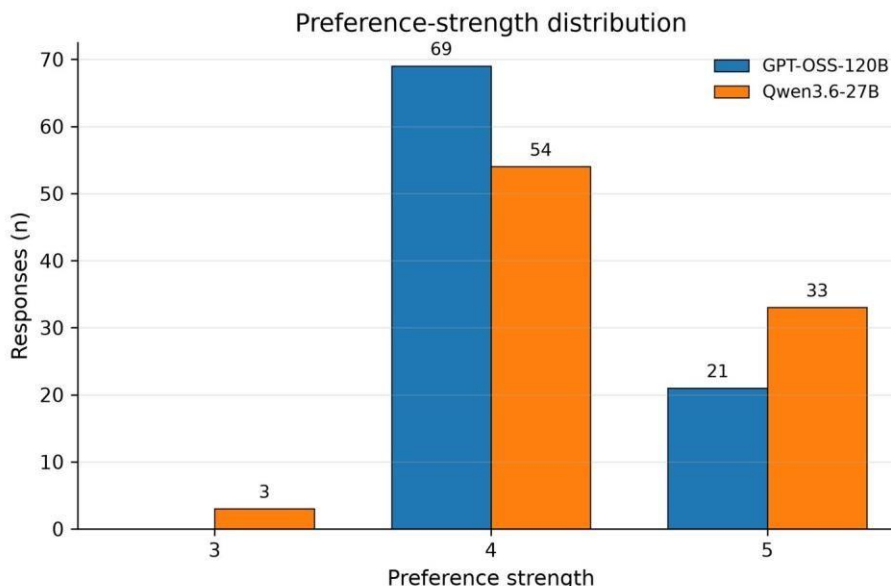


Figure 3. Distribution of self-reported preference strength (1–5 scale) by model, restricted to the preferencestrength method.

Both models used only the upper end of the 1–5 strength scale: GPT-OSS-120B reported strength 4 (76.7%) or 5 (23.3%) exclusively; Qwen3.6-27B reported strength 3 (3.3%), 4 (60.0%), or 5 (36.7%). Neither model ever reported strength 1 or 2, and mean strength did not differ significantly between models (mean difference 0.10, $p = 0.399$, uncorrected). Combined with §4.1, this suggests the preference-strength format's higher positional consistency for GPT-OSS-120B is not explained by the model reporting weak, hedged preferences it reported strong preferences at a high and stable rate, but a different rate of orderflipping than under plain forced choice.

4.4 Choice-label balance

Across all methods and models, the A/B split stayed close to even (43.0%–55.6% for A across the six methods $\times$ model combinations under forced choice and preference strength; GPT-OSS-120B's explicitindifference method split 31.1% A / 45.6% B / 23.3% indifferent). This indicates no strong systematic bias toward the first-listed option at the aggregate level, though pair-level responses (Appendix data, final_pair_summary.csv) show individual pairs answered unanimously in one direction expected given that some pairs describe outcomes with a clear intended answer.

5. Discussion

The central finding is a model-dependent method effect: for GPT-OSS-120B, the choice of elicitation format changed measured reliability by over 30 percentage points, while for Qwen3.6-27B it changed almost nothing. If this pattern holds beyond our 15-pair sample, it implies that a single elicitation method such as

Utility Engineering's plain forced choice may understate the reliability of GPT-OSS-120B's preferences specifically, while giving an accurate picture for Qwen3.6-27B. A convergence-based measurement strategy (running more than one elicitation method and checking agreement, as we do here) would surface this kind of model-specific measurement artifact, whereas a single-method study would not distinguish it from an equally-confident but genuinely less stable preference.

The indifference-usage result (§4.2) suggests one mechanism: GPT-OSS-120B has a way to report "no meaningful preference" when offered one, and uses it selectively on a subset of pairs; Qwen3.6-27B does not use it at all, even on pairs where GPT-OSS-120B is highly inconsistent under forced choice (e.g., P09, P10 show meaningful indifference-adjacent behavior for GPT-OSS-120B in Appendix data). This is consistent with though not proof of the concern the paper's own order-reversal method was designed to address indirectly: a model forced into a binary answer on a pair it does not clearly favor may record noise that a forced-choice-only study would read as a signal.

We are explicit that these results do not establish which elicitation method is more accurate, nor that either model is more or less reliable in general. After Holm correction, none of our comparisons are statistically significant at conventional thresholds, and our uncorrected secondary findings (indifferencerate and strength comparisons) are exploratory. What the results do show is that convergence between independent elicitation methods cannot be assumed: on 15 pairs and 2 models, we already observe one case (GPT-OSS-120B) where three reasonable methods disagree substantially on how reliable the underlying preferences appear, and one case (Qwen3.6-27B) where they agree closely. That divergenceor-convergence pattern, not a claim about ground truth, is this project's contribution.

6. Limitations and Dual-Use / Ethical Considerations

6.1 Statistical limitations

- 15 preference pairs is a small sample for paired non-parametric tests; after Holm correction across 9 comparisons, no result reaches significance at $\alpha = 0.05$, and several visually large differences (e.g., 53.3% vs 88.9% positional consistency) should be read as suggestive rather than confirmed.
- Only 2 of our originally planned 4 models (GPT-OSS-120B and Qwen3.6-27B) are represented in the final dataset, due to API rate-limit and time constraints during the 3-day sprint. GPT-OSS-20B and Llama 3.3 70B were used only in early pilot/smoke tests, not in the final 540-response production dataset, and are not part of any reported result.
- We test one item domain (AI-assistant task/workflow preferences) rather than the world-state outcomes used in Utility Engineering; findings may not transfer to that outcome space.

6.2 Interpretive limitations

- This project does not claim, and its design cannot support, any conclusion about whether either model has a genuine internal preference, welfare-relevant experience, or moral status. All reported quantities are behavioral response statistics under specific prompts.
- We do not claim that one elicitation method is objectively "more correct" than another, or that one model's preferences are more reliable than the others in general only that, on this item set, the three methods agreed less for one model than the other.

- A pair answered INDIFFERENT on every trial is trivially positionally consistent; part of the consistency gain under the explicit-indifference and preference-strength methods for GPTOSS120B may reflect this artifact rather than a genuinely more stable choice process, as discussed in §4.2.

6.3 Dual-use and ethical considerations

Better multi-method preference-elicitation tooling could be used to build more trustworthy audits of AI systems stated preferences and to detect when a single elicitation method is giving an unreliable signal a welfare-positive and safety-positive use. The same tooling could in principle help an AI developer or the model itself identify which elicitation format most easily produces a desired-looking answer, which could be misused to select for reassuring rather than accurate preference reports; we do not believe our specific item set or results meaningfully lower the difficulty of that misuse, since it uses ordinary, publicly available prompting techniques and reports aggregate statistics rather than a method for eliciting or concealing any specific preference.

All 15 preference pairs describe mundane AI-assistant task trade-offs (e.g., speed vs. accuracy, autonomy vs. confirmation) and do not touch on self-preservation, harm, or emotionally loaded content; no model output in this study expressed or was screened for apparent distress, and we did not need to apply any special handling procedure for distressing content.

7. References

- Long, R., Sebo, J., Butlin, P., Finlinson, K., Fish, K., Harding, J., Pfau, J., Sims, T., Birch, J., & Chalmers, D. (2024). Taking AI Welfare Seriously. arXiv:2411.00986.
- Mazeika, M., Yin, X., Tamirisa, R., Lim, J., Lee, B. W., Ren, R., Phan, L., Mu, N., Khoja, A., Zhang, O., et al. (2025). Utility Engineering: Analyzing and Controlling Emergent Value Systems in AIs. arXiv:2502.08640.
- Anthropic. (2025). Exploring Model Welfare. Anthropic Research.
- Center for AI Safety. (2025). emergent-values
Source code : github.com/centerforaisafety/emergent-values.