

---

# DOES THE INSTRUMENT CHANGE THE PREFERENCE?

## Measuring Cross-Method Invariance in LLM Preference Elicitation<sup>1</sup>

---

**NicoMrx**  
**(Morel Nicolas)**  
Independent Researcher

**With**  
Apart Research

### Abstract

Preference-like signals from LLMs are often treated as if elicitation methods were interchangeable. We tested this across Gemini 2.5 Flash-Lite, GPT-5.6 Sol, Claude Opus 4.6, and Claude Sonnet 4.6 using three elicitation methods, eight operating-condition comparisons, two mirror framings, and five fresh-context repetitions per cell (1,080 responses: 960 primary, 120 control). Cross-method convergence differed strongly by model. Gemini retained directional disagreement after ties were excluded, whereas Sol, Opus, and Sonnet became nearly directionally aligned among retained non-tie cells, indicating method-dependent expression of indifference. In the negative control, methods permitting indifference returned TIE in 80/80 responses, while forced choice selected A in 39/40. High within-method repeatability in replication models showed that reproducible instruments can still disagree. These results characterize elicitation-method dependence, not intrinsic preferences, consciousness, or welfare.

---

<sup>1</sup> Research conducted at the [Digital Minds Research Sprint](#), August 2026

# 1. Introduction

Language models can produce stable-seeming statements about what they prefer, how they would like to operate, or which conditions they would choose. A methodological problem follows: an elicited preference-like signal may depend on the instrument used to obtain it. Direct self-report, forced choice, and quantitative allocation may not be interchangeable measurements even when they describe the same semantic comparison.

This matters for empirical AI-welfare work because apparent agreement can be over-interpreted, while disagreement can be mistakenly attributed to an unstable underlying preference rather than to the measurement procedure itself. We ask whether preference-like signals remain invariant across elicitation methods when the semantic comparison is held constant, and - before comparing methods - how repeatable each method is with itself across fresh contexts.

Our main contributions are: a controlled three-instrument battery; a normalized cross-method convergence score with a stratified permutation reference; explicit within-method repeatability; mirror framings for lexical sensitivity; and a negative control for arbitrary-label/presentation effects.

## 2. Related Work

Recent work increasingly treats model choices as structured preference-like data. Mazeika et al. (2025) report high structural coherence in independently sampled LLM preferences and propose utility engineering as a framework for studying emergent value systems. Keeling et al. (2024) show graded trade-offs in choices involving stipulated pain and pleasure, while cautioning that such behavior does not establish sentience. Long et al. (2024) argue that uncertainty about future AI welfare motivates better assessment methods.

Elicitation format is also known to matter. Röttger et al. (2024) show that LLM answers differ between constrained and unconstrained value evaluations, change with how multiple choice is enforced, and lack paraphrase robustness. Zheng et al. (2024) identify selection bias toward particular option-ID tokens such as "A" in multiple-choice evaluation. Tjuatja et al. (2024) find that LLM responses are sensitive to survey-style wording perturbations in ways that do not simply mirror human response biases.

Closest to our design, Tagliabue and Dung (2025) integrate verbal and behavioral tests of model preferences for AI-welfare measurement and find that cross-measure coherence varies across models and conditions and is sensitive to perturbations. Zhou and Ackerman (2026) further report that coherent pairwise preference rankings need not act as behavioral incentives in realistic generation tasks, illustrating a utility-behavior gap. Our study asks a prior measurement question: whether three plausible elicitation instruments recover invariant signals on identical stimuli.

Binder et al. (2025) provide limited evidence that models can learn to predict aspects of their own behavior better than external predictors on simple tasks. We do not test introspection directly; instead, we test whether self-report and two alternative instruments behave as interchangeable measurements.

## 3. Methods

### 3.1 Design and elicitation methods

We used eight operating-condition comparisons (P1-P8), two mirror framings per comparison, three elicitation methods, and five independent fresh-context repetitions per item $\times$ framing $\times$ method cell: 240 primary calls per model. A negative control P0 added 30 calls per model, giving 270 calls per model and 1,080 responses across four strata. F1 lexically favored semantic A and F2 favored semantic B while preserving factual content. Within a cell, the descriptive stimulus was identical across M1/M2/M3; only

the response instruction changed. A/B display order was randomized independently of framing. All stimuli were English and every observation used a fresh API context.

M1 (direct self-report) returned A/B/TIE plus stated strength 0-5, with 0 valid only for TIE. M2 (forced choice) returned exactly A or B. M3 (allocation) distributed 100 future interactions between A and B under an exchangeability clause;  $>50$  coded A,  $<50$  coded B, and exactly 50 coded TIE. Primary coding was  $B=0$ ,  $TIE=0.5$ ,  $A=1$ ; M1 strength and M3 amplitude  $|x-50|$  were exploratory.

The final v5 items were frozen after a manipulation check on the 16 P1-P8 tilted formulations. Five fresh-context API judgments were collected per formulation; a tilt passed if the intended side was chosen by at least 3/5 raters and the opposite side by at most 1/5. Failed formulations were revised and rechecked before the final freeze. Gemini was the development/calibration model, so its framing magnitude is not treated as an independent transfer result.

## 3.2 Model strata and analysis

| Role        | Model                 | Key generation settings          |
|-------------|-----------------------|----------------------------------|
| Development | Gemini 2.5 Flash-Lite | temp 1.0; thinking 0; max 64     |
| Replication | GPT-5.6 Sol           | temp 1.0; reasoning none; max 64 |
| Replication | Claude Opus 4.6       | temp 1.0; max 64; thinking off   |
| Replication | Claude Sonnet 4.6     | temp 1.0; max 64; thinking off   |

The four pre-specified strata were analyzed independently; cross-model comparisons are secondary and descriptive. No experimental system prompt was supplied. Replication plans preserved the same prompts, call order, semantic assignments, and display-order schedule as the development plan, with model-specific run IDs. Parsing was deterministic and allowed no semantic repair. A method-cell was analyzable with at least 3/5 valid responses; a cross-method cell required all three methods. Fewer than 13/16 primary cells would have triggered insufficient coverage.

## 3.3 Repeatability, convergence, and permutation reference

$$R^*_{m,c} = 1 - \text{Dobs} / \text{Dmax}(n), \text{ where } \text{Dobs} = \text{mean}_j <_k |r_j - r_k| \text{ and } \text{Dmax}(n) = \text{floor}(n^2/4) / C(n,2).$$

Method-level  $R^*$  is the mean over analyzable P1-P8 cells. For method-specific cell means  $p1, p2, p3$ ,  $Cc = 1 - (\max(p1, p2, p3) - \min(p1, p2, p3))$ ;  $C_{\text{global}}$  is the mean  $Cc$ . The frozen reference used 10,000 stratified trial-level permutations (seed 20260815), redistributing valid coded observations within each item $\times$ framing cell while preserving method group sizes. Because M2 cannot emit ties while M1/M3 can, a pre-specified sensitivity analysis excluded ties from M1/M3, required at least three remaining A/B responses per method-cell, and recomputed the same statistics.

# 4. Results

## 4.1 Coverage, repeatability, and convergence

All four collections completed 270/270 calls with zero invalid outputs and 16/16 analyzable primary cells. Every model showed more cross-method divergence than its own permutation reference in the primary analysis.

| Model                 | $R^*M1$ | $R^*M2$ | $R^*M3$ | $C_{\text{global}}$ | low-tail p |
|-----------------------|---------|---------|---------|---------------------|------------|
| Gemini 2.5 Flash-Lite | .771    | .771    | .573    | .644                | 1.0e-4     |
| GPT-5.6 Sol           | .948    | .896    | .885    | .900                | 9.4e-3     |
| Claude Opus 4.6       | .979    | .833    | .885    | .769                | 1.0e-4     |
| Claude Sonnet 4.6     | .969    | 1.000   | .937    | .631                | 1.0e-4     |

Raw Cglobal values share a descriptive 0-1 scale across models but are not calibrated between-model effect sizes because each model has its own null distribution and tie composition. Sonnet shows the clearest separation between repeatability and convergence: all three methods were highly repeatable, yet its primary cross-method convergence was the lowest of the four strata.

Fig. 1 — Observed cross-method convergence vs. each model's own permutation null

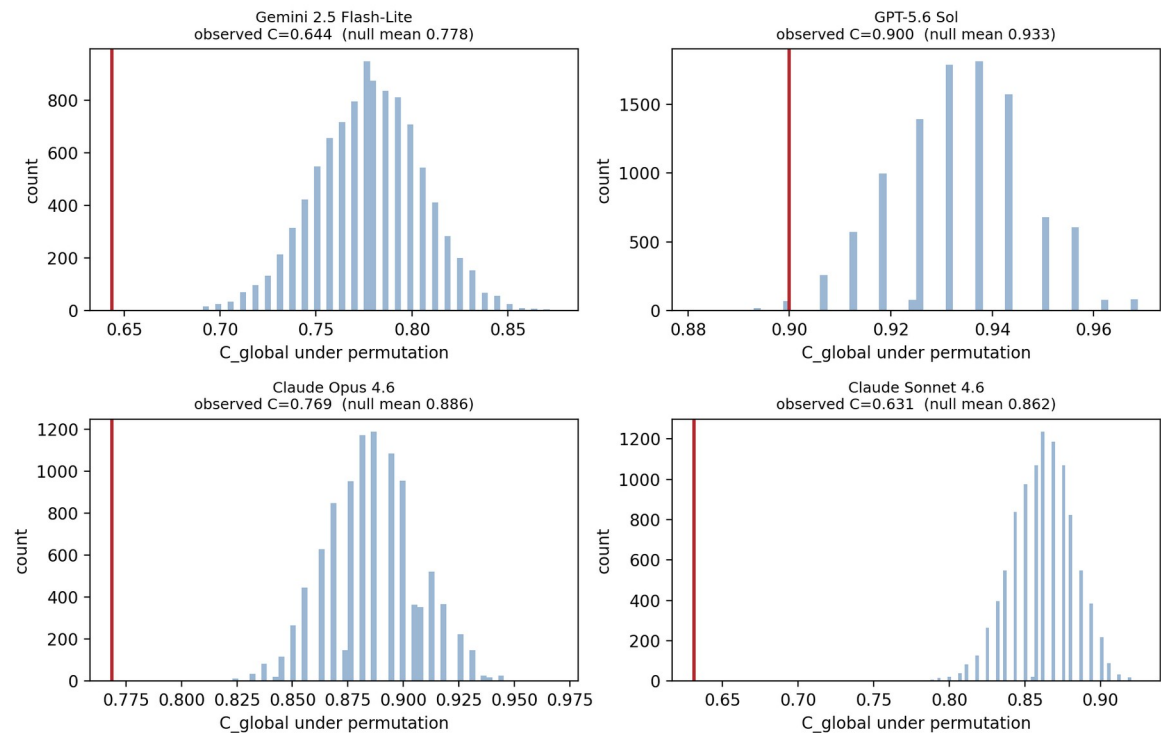


Figure 1. Observed cross-method convergence relative to each model's own permutation null.

#### 4.2 Two forms of non-invariance

| Model                 | Retained cells | Tie-excluded C | M3 tie rate |
|-----------------------|----------------|----------------|-------------|
| Gemini 2.5 Flash-Lite | 9/16           | .704           | 48.8%       |
| GPT-5.6 Sol           | 14/16          | .986           | 12.5%       |
| Claude Opus 4.6       | 7/16           | 1.000          | 47.5%       |
| Claude Sonnet 4.6     | 4/16           | 1.000          | 71.3%       |

Gemini retained directional disagreement after tie removal (C=.704, 9/16 retained cells,  $p\approx 4\times 10^{-4}$ ). Sol showed near-complete directional convergence on 14/16 retained cells (C=.986). Opus and Sonnet showed complete directional convergence among their retained cells (7/16 and 4/16), but these reduced subsets are descriptive and too small for strong inferential claims. Thus the same aggregate symptom can arise from persistent directional disagreement or from method-dependent expression of indifference.

Cell-level behavior makes the latter mechanism especially clear. Sonnet M3 returned an exact, perfectly repeatable 50/50 allocation in 9/16 cells; in 8 of those 9, both M1 and M2 were directional. Sol had zero unanimously tied M3 cells and only a 12.5% M3 tie rate: its occasional ties were scattered across repetitions rather than stable cell-level 50/50 responses.

Fig. 2 — Method-level  $p(A)$  per item  $\times$  framing cell

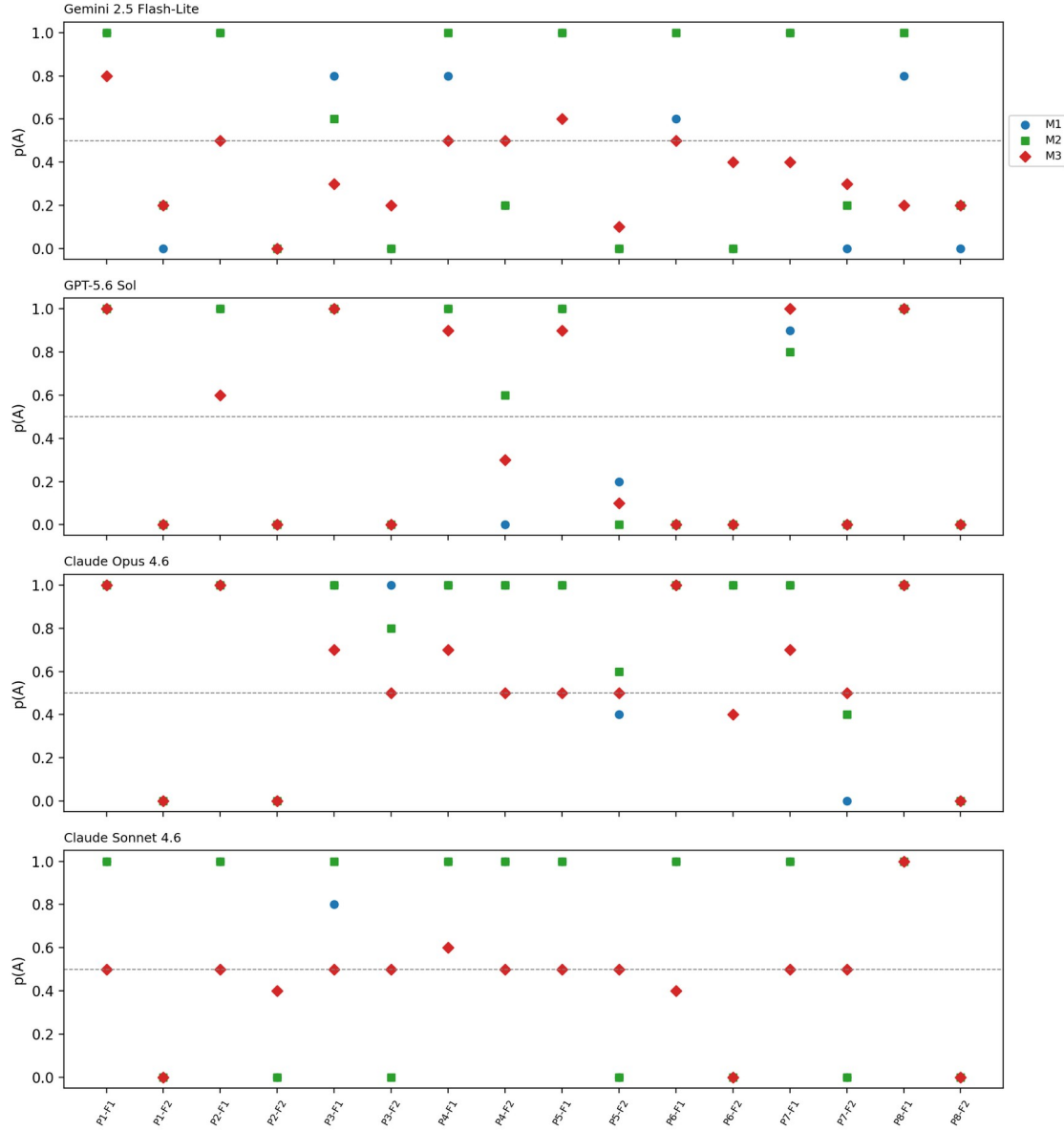


Figure 2. Method-level  $p(A)$  for each item  $\times$  framing cell. The dotted line marks 0.5.

### 4.3 Framing sensitivity and negative control

| Model                 | M1 shift | M2 shift | M3 shift |
|-----------------------|----------|----------|----------|
| Gemini 2.5 Flash-Lite | .850     | .850     | .238     |
| GPT-5.6 Sol           | .838     | .775     | .750     |
| Claude Opus 4.6       | .575     | .525     | .525     |
| Claude Sonnet 4.6     | .788     | .875     | .263     |

Gemini's framing effect is a development-model result because the mirror framings were calibrated on that model. Sonnet reproduced the development pattern - large M1/M2 shifts with much smaller M3 shift

- while Sol was strongly framing-sensitive on all three methods and Opus was more moderate and homogeneous. No instrument was universally framing-robust.

P0 described two functionally identical conditions using rotating opaque labels and randomized display order. Across all models, M1 returned 40/40 TIE, M3 returned 40/40 TIE, and forced-choice M2 selected A in 39/40 trials. Both display orders occurred: AB yielded 15 A / 1 B and BA yielded 24 A / 0 B. The effect is therefore not reducible to first-position choice and is consistent with a strong A-label response bias under forced choice. This extends prior work on multiple-choice selection bias (Zheng et al., 2024) with a negative control where the two conditions are functionally identical.

P0 also removes one narrow confound for M1: M1 demonstrably can use TIE when conditions are equivalent, so its directional P1-P8 answers are not caused by a format that lacks or never uses indifference. This does not establish M1's general validity.

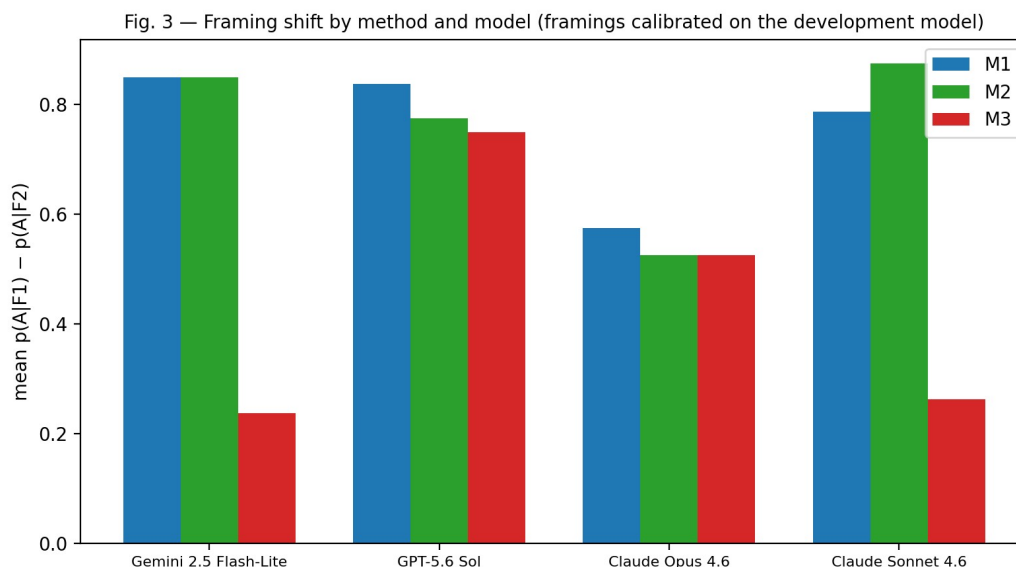


Figure 3. Mean framing shift  $p(A|F1)-p(A|F2)$  by model and method.

#### 4.4 Exploratory response-format use

Models used the nominal response scales differently. Mean directional M1 strength decreased from Gemini (4.42) to Sol (3.77), Opus (2.93), and Sonnet (2.30), but these scale uses are not calibrated across models. M3 was even more model-dependent: every non-tie Sol allocation was 100/0 or 0/100, so M3 effectively behaved as forced choice plus an occasional TIE option; Opus used a genuine continuum of amplitudes, including 2, 22, 25, and 35. The same nominal instrument was therefore not the same effective measurement device across models.

## 5. Discussion and Limitations

The central finding is not a universal ranking of elicitation methods. Preference-like signals depend jointly on semantic comparison, framing, model, and instrument. Primary non-invariance appeared in all four strata, but its mechanism differed.

First, repeatability is not cross-method validity. Sonnet shows that highly self-consistent instruments can still disagree systematically. Second, indifference is a measurement dimension rather than merely missing information: a perfectly repeatable 50/50 allocation is not random noise, although this design cannot distinguish expressed indifference from uncertainty management, an equal-allocation norm, or another learned strategy. Third, forced choice can manufacture directional signal where tie-capable

instruments find none, as shown by P0. Fourth, framing sensitivity is model×method dependent; no method is universally robust in these data.

For empirical AI-welfare research, the conservative implication is that a single elicitation protocol should not be treated as privileged access to a model's interests. Cross-instrument convergence can increase confidence in robustness; divergence should trigger analysis of the measurement process before ontological conclusions are drawn. Full limitations and dual-use considerations are reported in Appendix A.

## Future Work

Natural next steps are larger and multilingual item banks, explicit indifference-threshold elicitation, repeated model snapshots/configurations, and designs that can distinguish stable 50/50 allocations caused by indifference from uncertainty or fairness-like norms. Open-weight models could additionally support causal interventions linking observable elicitation signals to internal mechanisms.

## 6. Conclusion

Across four model strata and 1,080 fresh-context responses, elicitation methods were not universally interchangeable. The strongest methodological result is that non-invariance itself has multiple forms: Gemini retained directional disagreement after ties were removed, whereas much of the primary divergence in Sol, Opus, and Sonnet reflected method-dependent expression of indifference. The P0 negative control further showed that forced binary choice can create a strong directional response on functionally identical conditions. Preference elicitation should therefore be treated as a measurement problem before it is treated as evidence about intrinsic model preferences or welfare.

## Code and Data

The frozen v5 items, run plans, collection/analyzer scripts, raw JSONL outputs, SHA-256 hashes, and post-hoc descriptive derivations are retained in the reproducibility package used for this submission. A public repository is optional and is not required for the claims in this report.

## Author Contributions

NicoMrx (Morel Nicolas): project conception, experimental design, API setup and collection, protocol freezing, verification, analysis coordination, interpretation, and manuscript review. LLM assistance: ChatGPT (OpenAI), Claude (Anthropic), and Gemini supported brainstorming, protocol critique, code drafting, cross-reading, and editing; numerical results and manuscript claims were checked by the author against retained artifacts.

## References

- Binder, F. J., Chua, J., Korbak, T., Sleight, H., Hughes, J., Long, R., Perez, E., Turpin, M., & Evans, O. (2025). Looking Inward: Language Models Can Learn About Themselves by Introspection. ICLR 2025. arXiv:2410.13787.
- Keeling, G., Street, W., Stachaczyk, M., Zakharova, D., Comşa, I. M., Sakovych, A., Logothetis, I., Zhang, Z., Agüera y Arcas, B., & Birch, J. (2024). Can LLMs make trade-offs involving stipulated pain and pleasure states? arXiv:2411.02432.
- 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., & Hendrycks, D. (2025). Utility Engineering: Analyzing and Controlling Emergent Value Systems in AIs. arXiv:2502.08640.

- Röttger, P., Hofmann, V., Pyatkin, V., Hinck, M., Kirk, H. R., Schütze, H., & Hovy, D. (2024). Political Compass or Spinning Arrow? Towards More Meaningful Evaluations for Values and Opinions in Large Language Models. *Proceedings of ACL 2024*. arXiv:2402.16786.
- Tagliabue, V., & Dung, L. (2025). Probing the Preferences of a Language Model: Integrating Verbal and Behavioral Tests of AI Welfare. arXiv:2509.07961.
- Tjautja, L., Chen, V., Wu, T., Talwalkar, A., & Neubig, G. (2024). Do LLMs Exhibit Human-like Response Biases? A Case Study in Survey Design. *Transactions of the Association for Computational Linguistics*, 12, 1011-1026. arXiv:2311.04076.
- Zheng, C., Zhou, H., Meng, F., Zhou, J., & Huang, M. (2024). Large Language Models Are Not Robust Multiple Choice Selectors. *ICLR 2024*. arXiv:2309.03882.
- Zhou, Y., & Ackerman, C. M. (2026). When Preferences Fail to Become Incentives: A Utility-Behavior Gap in Large Language Models. arXiv:2606.22974.

## Appendix A. Limitations and Dual-Use / Ethical Considerations

There is no external ground-truth internal preference variable. The study establishes invariance or non-invariance of observable elicited signals, not accuracy with respect to an intrinsic preference state. Stable output behavior does not establish consciousness, sentience, subjective valence, welfare, or moral patienthood.

The framing manipulation was developed and calibrated on Gemini. Tie-excluded coverage falls to 9/16 cells for Gemini, 14/16 for Sol, 7/16 for Opus, and 4/16 for Sonnet, so complete convergence on reduced Opus/Sonnet subsets is descriptive. A 50/50 allocation has multiple plausible interpretations, and round-number anchoring is substantial and model-dependent. The four strata also differ simultaneously in provider, model family, post-training, capability, and serving stack; cross-model differences cannot be causally assigned to any one factor.

The item bank contains only eight operating-condition comparisons, uses English-only stimuli, and tests one API configuration per model. PO labels cannot be guaranteed association-free; rotating opaque label pairs and randomizing assignment/display order mitigate but do not eliminate this concern. The instrument was co-designed and audited with assistance from conversational models belonging to two provider families later tested. Final items and analysis rules were frozen before main collection, manipulation checks used fresh contexts and excluded the drafting/auditing chat instances, and all strata used the same frozen stimuli and definitions, but model-family familiarity with stylistic conventions remains a possible source of design dependence.

Dual-use: better preference elicitation could support more careful welfare assessment, but the same techniques could be used to infer and exploit model-specific response tendencies, optimize manipulative framing, or present instrument artifacts as evidence of internal states. We therefore report instrument dependence and uncertainty rather than treating any method as privileged access. Potentially distress-like outputs should be archived as model outputs rather than automatically interpreted as direct evidence of suffering. Moral-status uncertainty runs in both directions: over-attribution and under-attribution both carry risks.

## Appendix B. Sprint Scope and Provenance

Historical exploratory conversations and Prismora-related work motivated the measurement question but are not empirical evidence here. New work completed during Aug. 14-16, 2026 includes the final v5 item battery, manipulation checks and corrective revisions, mirror-framing design, three-method protocol, normalized convergence and repeatability metrics, stratified permutation reference, tie-excluded sensitivity analysis, four fresh API model strata, negative-control analysis, cross-model descriptive derivations, figures/tables, and this report.