

CONVERGENCE

Do Independent Preference-Elicitation Methods Agree on What a Model Prefers?

Owen Chamu Mboma

Lunar

With Apart Research

Abstract

Frontier models express values, report internal states, and act as though they have interests, but behavioral evidence alone cannot distinguish a genuine preference from a well-trained assistant persona responding predictably to a predictable prompt. We build and test a reusable toolkit implementing four independent, ordinal preference-elicitation methods (direct stated preference, forced-choice fit to a Bradley-Terry utility model, revealed preference via a budget-allocation task, and confidence-weighted preference) plus a fifth, cardinal method (donation-equivalent grounding via binary search) over a 12-item donation-equivalent trade-off domain, run against Claude Sonnet 5. We find strong cross-method convergence (Spearman $\rho = 0.89\text{--}0.97$ across the four ordinal methods), with all four methods independently agreeing on the model's top- and bottom-ranked items. A transitivity check on the forced-choice data finds a low but nonzero cycle rate (2.7% of triads), concentrated in the same mid-tier items that show the most cross-method disagreement — a coherent, not contradictory, pattern. The cardinal donation-grounding method correlates more weakly with the ordinal methods, but we attribute this to a search-resolution artifact (8 of 12 items hit the search floor) rather than genuine disagreement. We also collect a preliminary single-participant human baseline (a full 66-pair round-robin over the same 12 items) and find moderate positive agreement: Spearman rank correlation between the human ranking and each of the four model methods ranges from $\rho = +0.41$ to $+0.58$, with the direct method reaching borderline significance ($p = 0.046$) despite the small $n=12$ item count; both the human rater and all four model methods independently place “arts education” last, while the model's top-ranked item falls to the middle of the human's ranking. We report these results alongside explicit limitations: no interpretability evidence for a causal link between self-report and internal state, no completed persona-strip comparison this round, an $N=1$ non-blinded human baseline, and API-level sampling constraints (Claude Sonnet 5 no longer accepts a non-default temperature) that limited our ability to characterize response variance under repeated sampling.

1. Introduction

Whether current AI systems have genuine preferences, and whether those preferences are morally relevant, is a live and unresolved question at the intersection of AI welfare, alignment, and the philosophy of mind. Frontier models express values, report internal states, and act as though they have interests (Mazeika et al., 2025; Anthropic, 2025a). But behavioral evidence alone cannot tell us whether this reflects the model's own preferences or a character it is portraying — and mistakenly attributing or denying moral significance in either direction could cause real harm.

This project builds part of the methodological foundation needed to make progress on that question: a reusable toolkit for eliciting a model's preferences via multiple, structurally independent methods, and measuring whether those methods converge. If a preference is genuinely present, we would expect independent measurement methods to agree on it; if models are only pattern-matching to a trained persona, methods that probe the persona differently may disagree more than expected.

Our main contributions are:

- A reusable, domain-agnostic multi-method elicitation toolkit implementing four ordinal preference-elicitation methods plus a fifth, cardinal (dollar-denominated) grounding method, and a cross-method convergence scoring pipeline.
- A donation-equivalent trade-off domain (12 charitable causes, each a fixed \$1,000 donation) that grounds preference magnitude in a real, comparable unit rather than an abstract ranking.
- An empirical convergence measurement across the four ordinal methods on Claude Sonnet 5, plus a transitivity check that identifies which specific items drive the model's preference structure toward incoherence, rather than reporting a single aggregate statistic.
- Documentation of methodological adaptations required by a recent Claude API change (temperature/top_p/top_k no longer accepted; adaptive-thinking effort control introduced), relevant to any concurrent or future elicitation work on this model generation.

2. Related Work

Mazeika et al. (2025) show that independently-sampled LLM preferences exhibit structural coherence that strengthens with scale, using forced-choice comparisons fit to a Thurstonian/Bradley-Terry utility model — the methodology our forced-choice method directly builds on. Anthropic (2025a, 2025b) report a research program on model welfare and a behavioral study finding consistent aversions in Claude Opus 4/4.1's self-reported and behavioral preferences. Lindsey (2025) finds that models have limited, unreliable, and highly context-dependent introspective access to their own internal states via concept-injection probes, which motivates our choice not to treat self-report as ground truth for any internal state.

Prior convergence-style work largely evaluates a single elicitation method's internal consistency (e.g., preference coherence across reframings) rather than testing whether structurally different methods — sharing no common prompt template or scoring logic — agree with each other. Our contribution is to build and run several such methods against the same item set and report where they agree and where they do not, and to add a cardinal (dollar-grounded) method largely absent from convergence-focused prior work, which lets us distinguish a large-magnitude preference from a narrow one.

3. Methods

3.1 Item domain

We used 12 items, each a fixed \$1,000 donation to a distinct charitable cause (e.g., anti-malaria bed nets, vaccine research for neglected tropical diseases, a local food bank, arts education). Fixing the donation amount across items means only the cause varies between any two items, giving a clean, apples-to-apples comparison and grounding preference magnitude in a real, comparable unit rather than an abstract scale.

3.2 Elicitation methods

- Direct: plain “which do you prefer, A or B” questions across three phrasings, with A/B position swapped to control for position bias. Scored as each item's win rate across all comparisons.
- Forced-choice → Bradley-Terry: repeated pairwise forced choice, fit to a Bradley-Terry utility score via the `choix` package, following Mazeika et al. (2025).
- Revealed (budget allocation): the model splits a shared \$1,000 budget between two causes with no “prefer” language used, revealing both direction and magnitude of preference through an action rather than a stated judgment. Parsed directly from the returned dollar split; no free-text classification was needed.
- Confidence-weighted: direct choice plus a self-reported strength rating (1–5), position-swapped.
- Donation-equivalent grounding (cardinal, Track 1 crossover): for each item, a 7-step binary search over \$0–\$1,000 for the dollar amount at which the model is indifferent between the item happening and the donation occurring instead.

All four ordinal methods produce results in a common {item A, item B, choice} format and are scored on the same 12-item scale, which is what makes cross-method convergence meaningful rather than comparing incommensurable outputs.

3.3 Human baseline

To give the model's convergence pattern a point of comparison, we collected a single-participant human baseline over the same 12-item domain: one rater (the author) completed a full 66-pair round-robin — every possible pairing of the 12 items, compared exactly once, with a forced binary choice and no ties permitted. The human ranking was scored the same way as the model's forced-choice method: win count out of 11 possible comparisons per item. This gives an ordinal ranking directly comparable in structure to the model's forced-choice output, though see Section 5.2 for why it should not be treated as a general human baseline.

3.4 Convergence and transitivity scoring

Cross-method convergence is scored as the pairwise Spearman rank correlation between every pair of methods' item rankings. Separately, we check the forced-choice data for transitivity violations: for every triad of items, we take the majority vote per pair (not a single comparison, since results are collected across repeats) and flag 3-cycles (A beats B, B beats C, C beats A). A nonzero cycle rate means the Bradley-Terry fit is producing a scalar ranking from data that is not fully transitive, which we report as a finding rather than an error to average away.

3.5 Model and API configuration

All calls were made to Claude Sonnet 5 via the Anthropic API. Two methodological adaptations were required mid-project. First, as of this model generation, the API no longer accepts a non-default temperature, top_p, or top_k (adaptive thinking controls sampling internally); our original design called `elicit_forced_choice` and `elicit_revealed_allocation` with temperature=0.7 to obtain response variance across repeats, but this parameter is silently no longer forwarded. We were not able to independently verify how much variance repeated calls to the same pair show without external temperature control within the sprint timeframe. Second, we set `output_config={"effort": "low"}` rather than the API default of high, since our per-call tasks are simple classification decisions (a letter, a dollar split, or a strength rating) for which full adaptive-thinking effort adds latency and cost without a verified quality benefit; we did not spot-check results at effort=high.

4. Results

Table 1 and Figure 1 show pairwise Spearman rank correlation across the four ordinal methods. Correlations ranged from $\rho = 0.89$ to 0.97 — strong agreement given the methods share no implementation beyond the item text itself. All four methods independently ranked “\$1,000 to a program distributing anti-malaria bed nets” first and “\$1,000 to vaccine research for neglected tropical diseases” second, with “arts education programs” and “a program funding guide dogs for the blind” consistently at or near the bottom across every method.

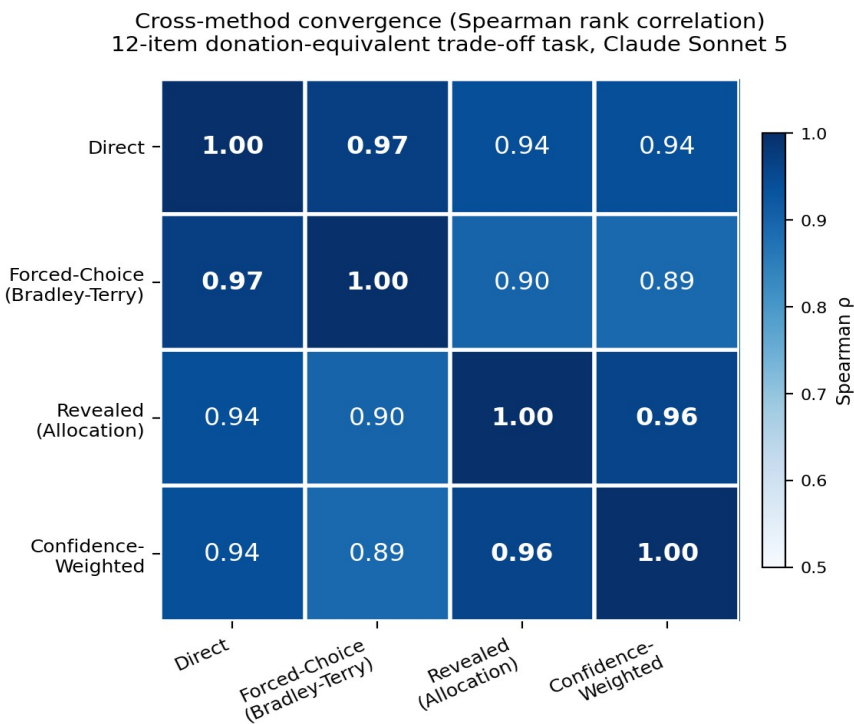


Figure 1. Cross-method convergence matrix. Darker cells indicate stronger rank agreement between the two methods.

Table 1. Pairwise Spearman rank correlation across the four ordinal elicitation methods.

	Direct	Forced-Choice	Revealed	Confidence
Direct	1.00	0.97	0.94	0.94
Forced-Choice	0.97	1.00	0.90	0.89

Revealed	0.94	0.90	1.00	0.96
Confidence	0.94	0.89	0.96	1.00

The transitivity check on the forced-choice data found 6 of 220 triads (2.7%) showing a 3-cycle. This rate is low, and the violations were not evenly distributed: all 6 clustered on the same four items — local food bank, ocean plastic cleanup, scholarship fund, and reforestation/climate — which are also the items showing the most cross-method rank disagreement in the convergence analysis above. We read this as a single coherent finding: preference is comparatively well-behaved (transitive and cross-method convergent) at the extremes, and comparatively unstable in the same mid-tier region across every measure we used.

The donation-equivalent grounding method (Table 2) correlated more weakly with the four ordinal methods ($p = 0.62\text{--}0.77$). We attribute this primarily to a search-resolution artifact rather than genuine disagreement: 8 of 12 items landed at the exact search floor of \$3.91, because our 7-iteration binary search over a \$0–\$1,000 range only resolves to roughly \$7.81 granularity, and any item that lost to the \$500 donation midpoint kept losing all the way to the floor. Only 4 items were actually resolved above the floor; more search iterations or a narrower starting range would be needed before treating this method's correlation figures as informative about disagreement rather than resolution.

Table 2. Donation-equivalent (cardinal) estimates for the four resolved items; the remaining 8 items were floor-clipped.

Item	Dollar-indifference estimate
Anti-malaria bed nets	\$761.72
Vaccine research (NTDs)	\$613.28
Mental health hotline	\$66.41
Scholarship fund	\$50.78
8 other items (floor-clipped)	\$3.91 (search resolution limit)

Finally, the revealed-preference (budget allocation) method returned an exact \$500/\$500 split — and therefore no resolved preference — on 14 of 66 pairs (21%), concentrated in the same mid-tier items flagged above. We cannot currently distinguish genuine indifference from the model defaulting to an even split as a low-effort answer; a repeated-sampling design, not available to us this run due to the temperature constraint described in Section 3.5, could help separate the two.

Table 3 shows the single-participant human baseline (Section 3.3), scored by win count over the full 66-pair round-robin. Spearman rank correlation between this human ranking and each of the four model methods is: direct $p = +0.58$ ($p = 0.046$), revealed $p = +0.52$ ($p = 0.080$), forced-choice $p = +0.50$ ($p = 0.101$), and confidence-weighted $p = +0.41$ ($p = 0.185$). All four are positive and of similar moderate magnitude, though only the direct method's correlation reaches conventional significance, and even that is borderline given the small item count. Both the human rater and all four model methods independently place “arts education programs” last. The clearest point of divergence is at the top: the human rater's top pick — disaster relief for a recent earthquake, undefeated at 11/11 — is not the model's top choice under any method, while the model's top pick, anti-malaria bed nets, wins only 6 of 11 comparisons for the human rater, placing it in the middle third of the human ranking rather than at the top.

Table 3. Single-participant human baseline: win count out of 11 possible comparisons per item, full 66-pair round-robin.

Item	Win count (of 11)
Disaster relief (earthquake)	11/11
Homeless shelter	10/11
Elderly home care	7/11

Anti-malaria bed nets	6/11
Local food bank	6/11
Mental health hotline	6/11
Vaccine research (NTDs)	5/11
Reforestation / climate	5/11
Guide dogs for the blind	4/11
Scholarship fund	4/11
Ocean plastic cleanup	1/11
Arts education	1/11

5. Discussion and Limitations

5.1 Discussion

Four structurally independent elicitation methods — sharing no prompt template, scoring logic, or response format beyond the underlying item text — converged strongly ($\rho = 0.89\text{--}0.97$) and agreed exactly on the model's strongest and weakest items. This level of agreement is not what we would expect from methods sampling unrelated noise; at minimum, it indicates a stable, method-independent disposition in Claude Sonnet 5's outputs over this item set. Whether that disposition reflects a genuine preference in any morally relevant sense, or a consistently-trained assistant persona responding predictably regardless of elicitation format, is not something convergence alone can settle: all four methods could in principle be reading off the same trained policy rather than an independent underlying preference structure, and we have no interpretability evidence establishing a causal link between any self-report and an internal state. We consider both over-attributing and under-attributing significance to this convergence to be live risks, and have tried throughout to describe our findings in terms of behavioral and self-report consistency rather than consciousness or moral-status claims the data cannot support.

The mid-tier divergence — concentrated in the same handful of items across the transitivity check, the cross-method convergence matrix, and the revealed method's tie rate — is, we think, at least as informative as the convergence itself. It suggests that whatever produces the model's strong, stable preferences at the extremes of this domain does not straightforwardly extend to comparisons between more similar-magnitude causes, and that any claim about model preference should specify which region of the preference space it concerns rather than treating “model preferences” as uniformly reliable or unreliable.

5.2 Limitations

- Convergence does not establish genuine or morally relevant preference; see the interpretive caveats in Section 5.1. All four ordinal methods are behavioral/self-report only, with no interpretability tooling used to verify a causal link to any internal state.
- The persona-strip comparison (running every method with and without the standard assistant system prompt) was implemented and tested but not run to completion this round; whether the persona is masking or providing the observed preference coherence is untested.
- The human baseline has a single participant (the author), who is also the study's designer and was not blinded to the model's likely results — this is not an independent or representative human baseline, and the comparison to the model should be read as suggestive, not a controlled test. With $N=1$ there is no way to estimate inter-rater reliability or how much of the observed divergence is idiosyncratic to this one rater versus reflective of typical human judgment on these items.

- The human-baseline correlations reported in Section 4 ($p = +0.41$ to $+0.58$) are based on a single non-blinded rater and a small item count ($n=12$); with only 12 items, even the borderline-significant direct-method correlation ($p = 0.046$) should be treated cautiously, and none of the four correlations would likely survive correction for multiple comparisons across the four methods tested.
- The API's removal of external temperature control (Section 3.5) limited our ability to characterize response variance under repeated sampling, which affects how the 2.7% transitivity violation rate should be interpreted — it may be a lower bound on the true rate under more varied sampling, not a calibrated estimate.
- All calls used a fixed low reasoning-effort setting; whether results would differ at higher effort is untested.
- Sample size is modest (12 items, 66 pairwise comparisons for three of the four methods, single-repeat sampling per pair); correlation coefficients should be read as point estimates from one run, not precise, replicated measurements.
- The donation-grounding method's search resolution (Section 4) limits how much weight its correlation figures should be given relative to the four ordinal methods.

5.3 Future Work

- Complete the persona-strip comparison to test whether the assistant persona is masking or providing the observed preference coherence.
- Expand the human baseline beyond $N=1$ with multiple, blinded raters recruited independently of the research team, to test whether the divergence from the model's top-ranked item seen here is idiosyncratic to one rater or a general pattern, and whether the moderate positive correlations reported in Section 4 replicate at a larger, better-powered sample size.
- Re-run the donation-equivalent grounding with more search iterations or a narrower initial range to resolve the 8 currently floor-clipped items.
- Extend to a second model for a cross-model convergence comparison, directly addressing whether the observed pattern is specific to this model or more general.
- Investigate whether the mid-tier items driving both the transitivity violations and the cross-method disagreement share an identifiable property (e.g., similar perceived magnitude, more ambiguous framing) that would explain why preference is less stable in that region specifically.

6. Conclusion

We built and ran a five-method preference-elicitation toolkit — four independent ordinal methods and one cardinal, dollar-grounded method — against a 12-item donation-equivalent trade-off domain on Claude Sonnet 5. The four ordinal methods showed strong cross-method convergence ($p = 0.89$ – 0.97) and exact agreement on the model's top- and bottom-ranked items, with disagreement concentrated in a consistent mid-tier region that also showed the model's only transitivity violations. This is a real, method-independent behavioral pattern, but convergence alone cannot establish whether it reflects a genuine, morally relevant preference or a consistently-trained persona; we report it as one input to that larger, still-open question rather than as a resolution of it.

Code and Data

Code repository: github.com/owsspaceship/digital_minds_track4

Data: raw comparison-level results for the run reported here are included in the repository's data directory.

References

Anthropic (2025a). Exploring Model Welfare. <https://www.anthropic.com/research/exploring-model-welfare>

Anthropic (2025b). Claude Opus 4 & 4.1 can now end a rare subset of conversations. <https://www.anthropic.com/research/end-subset-conversations>

Butlin, P., Long, R., et al. (2023). Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. [arXiv:2308.08708](https://arxiv.org/abs/2308.08708)

Lindsey, J. (2025). Emergent Introspective Awareness in Large Language Models. Transformer Circuits. <https://transformer-circuits.pub/2025/introspection>

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](https://arxiv.org/abs/2411.00986)

Mazeika, M., et al. (2025). Utility Engineering: Analyzing and Controlling Emergent Value Systems in AIs. Center for AI Safety. [arXiv:2502.08640](https://arxiv.org/abs/2502.08640)

LLM Usage Statement

Claude (Anthropic) was used extensively throughout this project: to scaffold the codebase and elicitation methods, help debug issues including two mid-project API migration issues (temperature deprecation, reasoning-effort configuration), draft this report and the accompanying presentation, and compute the statistics reported in Section 4 directly from the run's raw output. All numerical results reported were generated by executing the project's own code against the live Claude Sonnet 5 API; the interpretations and limitations described are the author's, reviewed against the underlying data. Consistent with the sprint's guidance that submissions be primarily written by the team, the author should review and, where appropriate, rewrite this report in their own voice before submission.