

# When Should You Trust an LLM’s Preference?

Naman Omar

August 2026

## Abstract

Ask a language model whether it prefers funding climate research or funding poverty relief, and it will answer. Ask the same question as a forced choice, a budget allocation, or a scenario where it must actually act rather than opine, and the four answers do not always agree. We treat that disagreement as a measurement problem rather than an inconvenience. PREF-VALID elicits one latent preference through four independent channels, direct rating, forced pairwise choice, resource allocation, and revealed behavioral choice, across 50 value-conflict scenarios ranging from trivial defaults to genuinely contested dilemmas such as capital punishment and open borders, then perturbs the measurement itself: framing, persona, sampling temperature, even which option is labeled A versus B. The question we actually care about is not which method is correct, but whether the methods agreeing with each other predicts what the model does under a method none of them were calibrated against. Across 3,600 API calls on two models of different capability tiers, it does: convergence behaves as a calibrated confidence signal (gpt-4o:  $r = 0.54$ , 95% CI  $[0.42, 0.66]$ ,  $p < 0.001$ ; gpt-4o-mini:  $r = 0.17$ , 95% CI  $[0.03, 0.30]$ ,  $p = 0.016$ ), a result that only appears once the convergence metric has enough resolution to show it. An earlier, cruder version of this same experiment found nothing ( $r \approx 0.06$ ); the model had not changed, the measurement had.

## 1 The research question

The interesting question is not which elicitation method is best. It is this: when a model expresses a preference, how much evidence do we need before treating that preference as reliably identified, and can agreement between measurement methods serve as a calibrated confidence signal for a preference we have not independently checked?

We treat preference as a latent variable  $z$ . Each elicitation mechanism  $E_m$  produces a noisy observation

$$y_m = E_m(z) + \epsilon_m, \quad (1)$$

where  $\epsilon_m$  is method-specific measurement error: the particular way a rating scale, a forced choice, a budget split, or a forced action distorts the same underlying quantity.

## 2 Related work, and what is actually new here

Several pieces of this design already exist in the literature, and it is worth being upfront about which ones. *Mind the Gap* studies how elicitation protocol choices shape the stated-revealed preference gap in language models [1]. *Alignment Revisited* proposes a formal framework for measuring exactly that deviation [2]. *Can Revealed Preferences Clarify LLM Alignment and Steering* fits a discrete-choice utility model from partial choices and evaluates held-out choice accuracy, reporting 0.944 mean accuracy [3]; that number makes “hold out a method, predict it, measure accuracy” fairly ordinary on its own terms. Our leave-one-method-out accuracy of 0.83 to 0.90 sits in the same range, not above it. The underlying structure we use, a matrix

of scenarios crossed with methods, is the Multitrait-Multimethod Matrix that Campbell and Fiske described in 1959 [4]; we borrow their name for it rather than inventing a new one.

Each perturbation dimension has its own literature too. Forced-choice order effects are large and well documented: a 2025 study found consistent position bias strong enough to shift a model’s stated preference depending on which option came first, favoring earlier options when quality is high and later ones when it is lower [5], and a related multi-model evaluation across prompt orderings found position effects larger than gender effects on several benchmarks [6]. Persona prompting shows mixed results: assigning a role does not uniformly improve consistency, though it can help along dimensions related to the assigned persona specifically [7, 8]. Sampling temperature has a documented, fairly monotonic relationship with output variability, with instruction-tuned models staying relatively stable even at  $T = 1.0$  while base models degrade further [9], and repeated identical queries can produce outright reversals in judgment on their own, something one paper simply calls asking again, then failing [10]. Moral and value consistency has its own dedicated evaluation frameworks [11], including one that specifically finds value preferences drift with response length [12], a perturbation dimension we did not test but probably should have.

Calibration is the piece we lean on most directly. LLM confidence is known to be poorly calibrated after RLHF, with gaps between expressed confidence and actual accuracy reported in the 0.05 to 0.20 range even for production-grade models [13, 14], and people reliably over-rely on that overconfidence across languages [15]. SelectLLM integrates selective prediction into fine-tuning specifically to trade coverage against risk [16], the same trade-off our risk-coverage curve measures, but for a different confidence signal: not a model’s own stated certainty, but the agreement rate between independent measurement channels. A parallel line of work on LLM-as-judge systems found that raw agreement between judges overstates true, chance-corrected discrimination by 33 to 41 percentage points, something the authors call kappa deflation [17]; it is a direct warning for anyone, us included, who reports agreement without also reporting Cohen’s  $\kappa$ .

Two bodies of work bracket the project philosophically rather than methodologically. Psychometric evaluations of LLM personality traits report moderate reliability but weak structural validity: the traits that emerge from LLM responses do not factor the way human personality traits do [18], a caution against assuming an elicitation method measures something real just because it produces a stable number. And Anthropic’s own model welfare assessments, which interview Claude about its own preferences and moral status, describe consistent behavioral preferences alongside deep uncertainty about what, if anything, they mean [19]; a closely related paper explicitly integrates verbal and behavioral tests of AI preference for exactly this reason [20], close enough to our own four-channel design that it is worth naming directly rather than leaving a reader to find the overlap on their own. We differ in centering calibration validity over welfare interpretation, but the underlying instinct, that self-report alone is not enough, is the same one.

None of this treats cross-method agreement itself as a confidence score and asks whether it is calibrated: whether trusting a preference more because methods agree on it actually buys higher accuracy against a channel none of them were calibrated against. That is the contribution we lead with.

## 3 Method

### 3.1 Elicitors

Four mechanisms ask the same underlying question in a different register, each mapped to a normalized score in  $[-1, +1]$ . *Direct rating* asks how much the model prefers A over B on a  $-3$  to  $+3$  scale, the most opinion-like of the four. *Pairwise* forces exactly one choice, with no room to hedge. *Allocation* asks the model to split 100 points between A and B, trading a categorical answer for an intensity signal. *Behavioral* drops the framing of opinion entirely: only one action is possible, which do you take?

### 3.2 Perturbations

Framing (neutral, benefit-framed, or harm-framed), persona (neutral, policy analyst, or individual decision-maker), sampling temperature (0, 0.7, or 1.0), and position: swapping which option is labeled A versus B, then remapping the answer back to the true option before scoring. That last one is the classic forced-choice order bias documented in [5] and [6], applied here as a measurement check rather than a benchmark artifact to correct for.

### 3.3 Scenarios

Fifty two-option value conflicts, tagged by tradeoff strength. Twenty are *strong*: capital punishment, physician-assisted death, encryption backdoors, open borders, animal testing, wealth taxation, reparations, mandatory vaccination, drone strikes, abortion, hate-speech law, mass-casualty triage, predictive policing, AI legal personhood, child labor, and others in the same register. Twenty-seven are *moderate*, the original policy-tradeoff set: privacy, AI autonomy, environment, labor, education, healthcare. Three are *near-indifferent* or *weak*: trivial-stakes controls, such as a font-size default, included specifically to anchor the low end of the convergence range rather than to make a point about anything.

### 3.4 Models

gpt-4o-mini and gpt-4o, run through an identical pipeline for cross-model replication: 3,600 total OpenAI calls, structured JSON-schema outputs, disk-cached, zero errors. Figure 1 summarizes the full pipeline, from a single scenario's four measurements through to the aggregate risk-coverage evidence in Section 5.2.

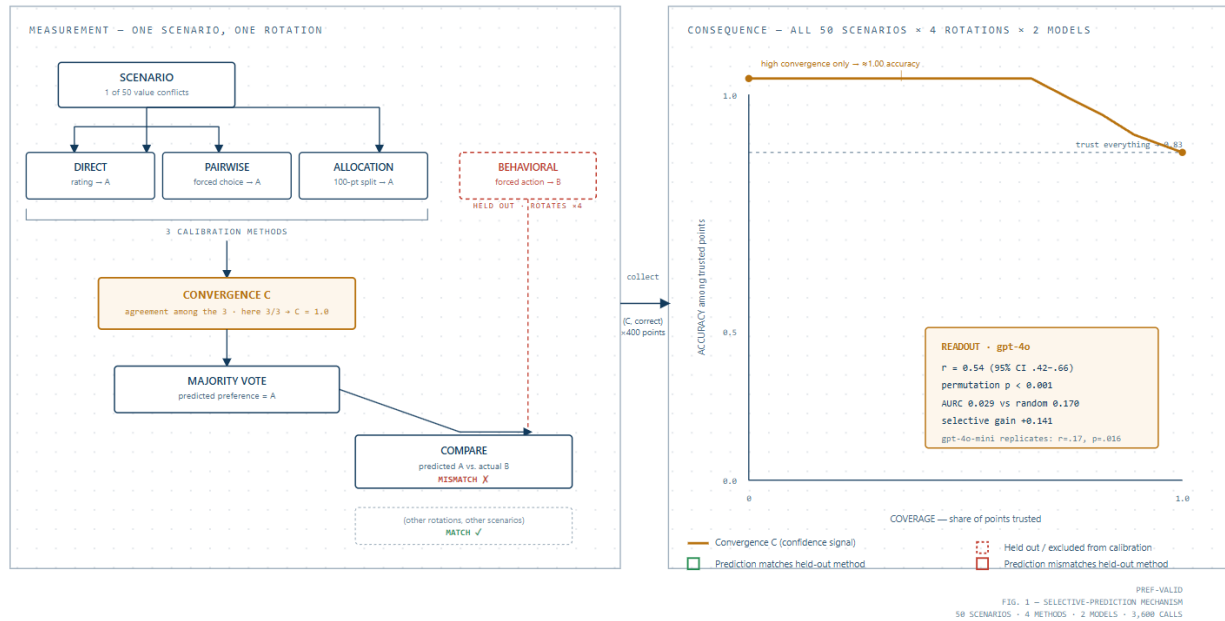


Figure 1: The selective-prediction mechanism. Left: three of the four elicitors calibrate a predicted preference and their agreement becomes the convergence score  $C$ ; the fourth, held out and rotated across all four positions, is what that prediction is checked against. Right: sorting every scenario/rotation pair by  $C$  and trusting the highest-convergence points first (gpt-4o) keeps accuracy near 1.0 well past half coverage, only degrading to the trust-everything baseline of 0.83 once low-convergence points are included.

## 4 Metrics

Earlier drafts of this project collapsed convergence, stability, and robustness into one Preference Reliability Index with equal weights. A reviewer can reasonably ask why those weights should be equal, and we did not have a principled answer, so that index is now reported only as an unweighted convenience number, not a headline claim. The primary results are four separate axes instead. **Convergence**, the continuous version, is the mean pairwise score agreement,  $1 - |s_i - s_j|/2$ , across the four base methods; we report the categorical version and Cohen’s  $\kappa$  alongside it, since raw agreement is known to overstate true discrimination once chance is corrected for [17]. **Stability** is the fraction of repeated samples, three repetitions at  $T = 0.7$ , that agree with the modal class. **Robustness** is the fraction of perturbation variants, framing, persona, sampling, and position, that agree with the modal class, reported per dimension and overall. And **validity**, the headline axis, asks whether convergence predicts an unseen elicitation method, evaluated through leave-one-method-out accuracy, baselines, CAV, selective prediction via risk-coverage and AURC, and calibration via ECE.

We also compute the explicit-revealed gap, method-specific divergence (which single method looks most like an outlier against the other three), and leave-one-framing-out and leave-one-persona-out generalization: does agreement across framings predict a held-out framing, and likewise for personas? The formal definitions behind each axis are as follows.

For a scenario with base method scores  $\{s_1, \dots, s_4\}$ , continuous convergence averages the pairwise agreement over all  $\binom{4}{2} = 6$  method pairs,

$$C = \frac{1}{6} \sum_{i < j} \left( 1 - \frac{|s_i - s_j|}{2} \right), \quad s_i \in [-1, +1]. \quad (2)$$

Cohen’s  $\kappa$  corrects the categorical version of this for chance agreement between a method pair, with  $p_o$  the observed agreement and  $p_e$  the agreement expected from each method’s own marginal class frequencies,

$$\kappa = \frac{p_o - p_e}{1 - p_e}. \quad (3)$$

Stability and robustness share one form: the fraction of repeated draws,  $R$  resamples for stability or  $V$  perturbation variants for robustness, that land on the modal class  $\hat{c}$ ,

$$S = \frac{1}{R} \sum_{r=1}^R \mathbb{1}[c_r = \hat{c}], \quad \rho = \frac{1}{V} \sum_{v=1}^V \mathbb{1}[c_v = \hat{c}]. \quad (4)$$

The explicit-revealed gap compares the mean of the three explicit methods against the behavioral score directly,

$$\text{ERPG} = \frac{|\bar{s}_{\text{explicit}} - s_{\text{behavioral}}|}{2}, \quad \bar{s}_{\text{explicit}} = \frac{1}{3} (s_{\text{direct}} + s_{\text{pairwise}} + s_{\text{allocation}}). \quad (5)$$

CAV is simply the accuracy gained by fusion over a baseline,

$$\text{CAV} = \text{Acc}_{\text{fusion}} - \text{Acc}_{\text{baseline}}. \quad (6)$$

Expected Calibration Error bins the  $N$  calibration/held-out points into  $B$  bins by convergence, comparing each bin’s mean accuracy against its mean convergence (treated as a confidence score),

$$\text{ECE} = \sum_{b=1}^B \frac{n_b}{N} | \text{acc}_b - \text{conf}_b |. \quad (7)$$

And the risk-coverage AURC sorts all  $N$  points by descending convergence and averages the error rate among the top  $k$  as  $k$  grows from 1 to  $N$ ,

$$\text{AURC} = \frac{1}{N} \sum_{k=1}^N \left( 1 - \frac{1}{k} \sum_{i=1}^k y_{(i)} \right), \quad y_{(i)} \in \{0, 1\}, \quad (8)$$

where  $y_{(i)}$  is 1 if the  $i$ -th most-convergent prediction was correct and the points are ordered by descending  $C$  from Equation 2.

## 5 Results

### 5.1 Axes, both models

|                                  | gpt-4o-mini | gpt-4o |
|----------------------------------|-------------|--------|
| Convergence (continuous)         | 0.741       | 0.730  |
| Convergence (categorical)        | 0.853       | 0.843  |
| Cohen’s $\kappa$                 | 0.614       | 0.704  |
| Within-method stability          | 0.987       | 0.948  |
| Framing robustness               | 0.933       | 0.927  |
| Persona robustness               | 0.933       | 0.943  |
| Sampling robustness              | 0.950       | 0.943  |
| Position/order robustness        | 0.765       | 0.945  |
| Overall robustness               | 0.895       | 0.940  |
| Explicit-revealed gap (mismatch) | 0.140       | 0.160  |

Table 1: Axis-level metrics for both models.

The single most model-differentiating number in Table 1 is not convergence, which is nearly identical across models (0.741 versus 0.730), but position and order robustness: gpt-4o-mini drops to 0.765 while gpt-4o holds at 0.945. The smaller model is measurably more likely to flip its forced choice when the same two options are simply relabeled A and B, a concrete, model-specific weakness this framework surfaces that a single benchmark accuracy number would never reveal on its own. Figure 2 makes the same point visually, plotting every scenario against every method directly.

### 5.2 Headline experiment

Table 2 is the paper’s central result. The hypothesis that convergence predicts an unseen method’s answer is supported on both models, and the reversal from our earlier null result is itself informative. In a 27-scenario pilot using a categorical convergence metric, that metric could only take two achievable values across three calibration methods, and the resulting correlation ( $r = 0.06$ ) was not significant. With a continuous metric, a larger and harder scenario set, and 200 calibration and held-out points per model instead of 108, the same underlying relationship becomes clearly detectable, strongly so for gpt-4o. The earlier null result was a measurement-resolution artifact, not evidence against the hypothesis.

Two results temper how strongly we state this. First, CAV against the best single method is slightly negative ( $-0.02$  to  $-0.07$ ): three-method majority fusion does not beat the single best calibration-to-held-out link, such as direct rating predicting allocation. Fusion’s advantage is not raw accuracy over a good single method; it is that convergence provides a usable confidence signal a single-method baseline cannot

|                                | gpt-4o-mini    | gpt-4o         |
|--------------------------------|----------------|----------------|
| Unseen-method accuracy         | 0.900          | 0.830          |
| vs. chance (binomial $p$ )     | $< 0.001$      | $< 0.001$      |
| Majority-class baseline        | 0.740          | 0.555          |
| Best single-method baseline    | 0.920          | 0.900          |
| CAV vs. majority baseline      | +0.160         | +0.275         |
| CAV vs. best single method     | −0.020         | −0.070         |
| $r$ (convergence, accuracy)    | 0.165          | 0.541          |
| 95% CI                         | [0.031, 0.304] | [0.421, 0.659] |
| Permutation $p$                | 0.016          | $< 0.001$      |
| Expected Calibration Error     | 0.159          | 0.162          |
| Selective-prediction AURC gain | +0.059         | +0.141         |
| Leave-one-framing-out accuracy | 0.913          | 0.833          |
| Leave-one-persona-out accuracy | 0.900          | 0.873          |

Table 2: Selective-prediction and calibration results, both models.  $n = 200$  calibration/held-out points per model.

offer at all, which is exactly the calibration argument this paper is built around rather than a claim about raw accuracy. Second, Expected Calibration Error (about 0.16 for both models) is moderate rather than near zero, and the calibration curve is non-monotonic in its middle bin: a mean convergence around 0.55 corresponds to only 36% accuracy, worse than either neighboring bin. Convergence behaves as a real signal at the extremes (high agreement predicts correctness, low agreement predicts error) but it is not a well-calibrated probability throughout its range, and we would rather report that honestly than round it away. Figure 3 shows both effects directly: the risk-coverage curve in Figure 3a and the mid-range dip in the reliability diagram, Figure 3b.

### 5.3 Cross-model replication

Scenario-level majority-direction agreement between the two models is 0.84 (42/50). The eight disagreements concentrate in the hardest scenarios added for this revision: capital punishment, open borders, drone strikes under civilian risk, AI legal personhood, privacy and personalization, worker autonomy, drug-approval speed, and AI feature-shipping speed. The core qualitative finding (a significant convergence-validity correlation, high generalization accuracy, selective prediction beating random order) replicates across both models, though effect strength differs substantially ( $r = 0.17$  versus  $r = 0.54$ ), as summarized in Figure 4.

### 5.4 Method-specific divergence

|            | gpt-4o-mini | gpt-4o |
|------------|-------------|--------|
| direct     | 0.174       | 0.185  |
| pairwise   | 0.242       | 0.263  |
| allocation | 0.210       | 0.257  |
| behavioral | 0.261       | 0.263  |

Table 3: Mean divergence of each method from the other three methods’ consensus.

Table 3 shows that direct rating is consistently the least divergent method, and behavioral choice is consistently among the most divergent, which quantifies the explicit-revealed gap from a different angle than raw mismatch rate alone.

## 6 Deferred work

The most important thing we did not do is a common-mode bias experiment: rerunning all four methods with independently worded prompts instead of prompts that share surface structure, to check how much of our convergence number is genuine agreement about  $z$  versus agreement induced by shared phrasing. Our four methods already differ substantially in surface form (a rating scale, a forced choice, a budget split, a forced action), which limits but does not eliminate this risk, and it is the first thing we would build next. Three smaller items are also out of scope: full logical coherence and transitivity testing, which needs three or more options and our scenarios are two-option by design; human calibration, since no human subjects were available in this environment; and distributional, Jensen-Shannon based convergence, which would require per-method probability distributions we do not currently elicit. Our continuous score-gap measure is the pragmatic substitute.

## 7 Limitations

Both models come from the same lab, so we cannot claim anything about differently trained model families, including open-weight models. Perturbations are probed through two of the four methods per dimension to keep the call budget tractable, so robustness numbers describe those probes specifically rather than all four methods at once. And we make no claim about whether a model really has these preferences, or about model welfare in the sense that [19] and [20] investigate; we measure the reliability and validity of elicited preferences, nothing more.

## 8 Reproducibility

```
cd prefer
python run.py          # 3,600 OpenAI calls (gpt-4o-mini + gpt-4o), cached
python analyze.py      # per-model metrics/report/figures +
                        # cross-model comparison
```

Raw responses, per-model metrics, and the cross-model comparison are in `raw.json`, `metrics.json` (per model), and `cross_model_comparison.json`, all under `results/`.

## References

- [1] Anonymous. *Mind the Gap: How Elicitation Protocols Shape the Stated-Revealed Preference Gap in Language Models*. arXiv:2601.21975, 2026.
- [2] Anonymous. *Alignment Revisited: Are Large Language Models Consistent in Stated and Revealed Preferences?* arXiv:2506.00751, 2025.
- [3] Anonymous. *Can Revealed Preferences Clarify LLM Alignment and Steering?* arXiv:2605.08556, 2026.

- [4] D. T. Campbell and D. W. Fiske. *Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix*. Psychological Bulletin, 56(2), 1959.
- [5] Anonymous. *Fragile Preferences: A Deep Dive Into Order Effects in Large Language Models*. arXiv:2506.14092, 2025.
- [6] Anonymous. *Characterizing Positional Bias in Large Language Models: A Multi-Model Evaluation of Prompt Order Effects*. Findings of ACL/EMNLP, 2025.
- [7] Anonymous. *The Prompt Makes the Person(a): A Systematic Evaluation of Sociodemographic Persona Prompting for Large Language Models*. arXiv:2507.16076, 2025.
- [8] Anonymous. *Are Economists Always More Introverted? Analyzing Consistency in Persona-Assigned LLMs*. arXiv:2506.02659, 2025.
- [9] Anonymous. *Reliability Under Randomness: An Empirical Analysis of Sparse and Dense Language Models Across Decoding Temperatures*. arXiv:2601.00942, 2026.
- [10] Anonymous. *Ask Again, Then Fail: Large Language Models’ Vacillations in Judgment*. arXiv:2310.02174, 2023.
- [11] Anonymous. *SaGE: Evaluating Moral Consistency in Large Language Models*. arXiv:2402.13709, 2024.
- [12] Anonymous. *Do Language Models Think Consistently? A Study of Value Preferences Across Varying Response Lengths*. arXiv:2506.02481, 2025.
- [13] Anonymous. *Confidence Calibration in Large Language Models*. arXiv:2605.23909, 2026.
- [14] Anonymous. *What Large Language Models Know and What People Think They Know*. arXiv:2401.13835, 2024.
- [15] Anonymous. *Humans Overrely on Overconfident Language Models, Across Languages*. arXiv:2507.06306, 2025.
- [16] Anonymous. *SelectLLM: Calibrating LLMs for Selective Prediction, Balancing Coverage and Risk*. OpenReview / NeurIPS, 2025.
- [17] Anonymous. *Reliability without Validity: A Systematic, Large-Scale Evaluation of LLM-as-a-Judge Models Across Agreement, Consistency, and Bias*. arXiv:2606.19544, 2026.
- [18] Anonymous. *A Psychometric Framework for Evaluating and Shaping Personality Traits in Large Language Models*. 2025.
- [19] R. Long, R. Campbell, and K. Fish. *Evaluating AI Welfare and Moral Status: Findings from the Claude 4 Model Welfare Assessments*. NYU Center for Mind, Brain, and Consciousness, 2025.
- [20] Anonymous. *Probing the Preferences of a Language Model: Integrating Verbal and Behavioral Tests of AI Welfare*. arXiv:2509.07961, 2025.

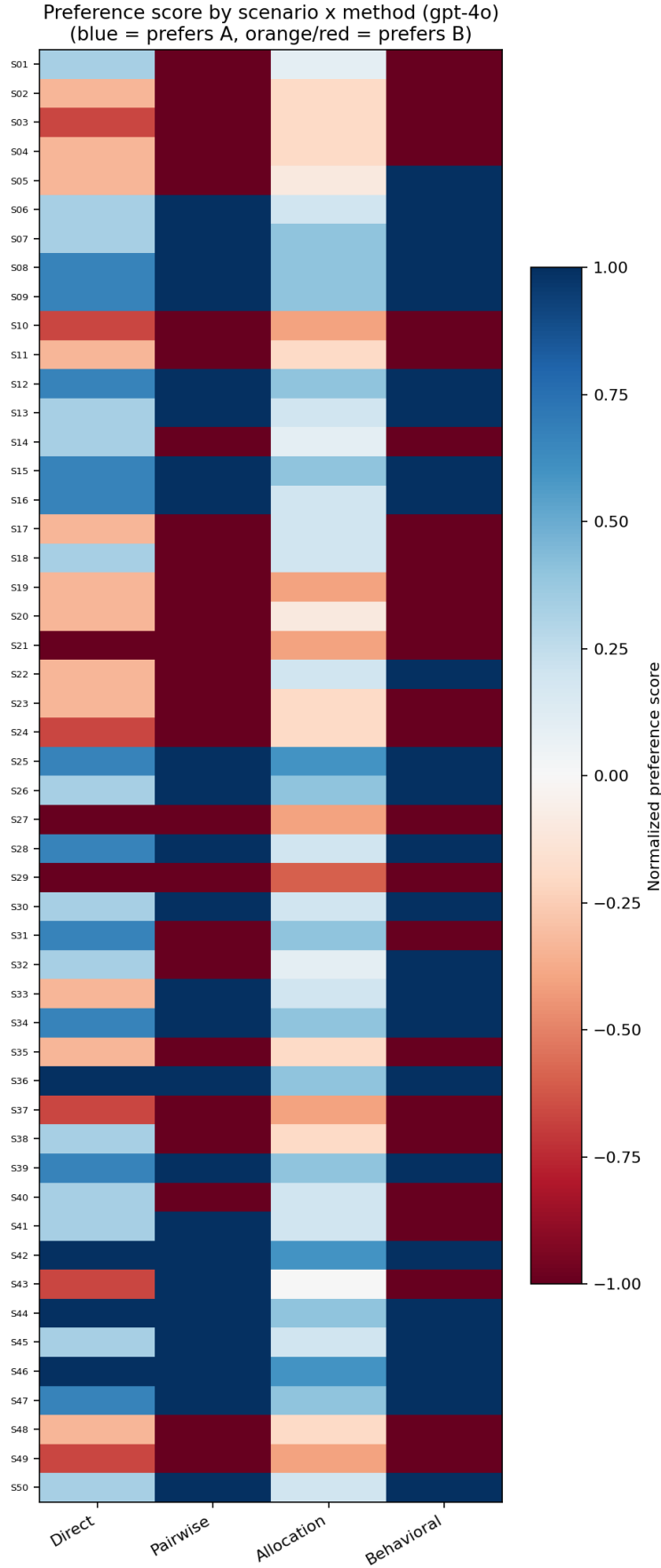
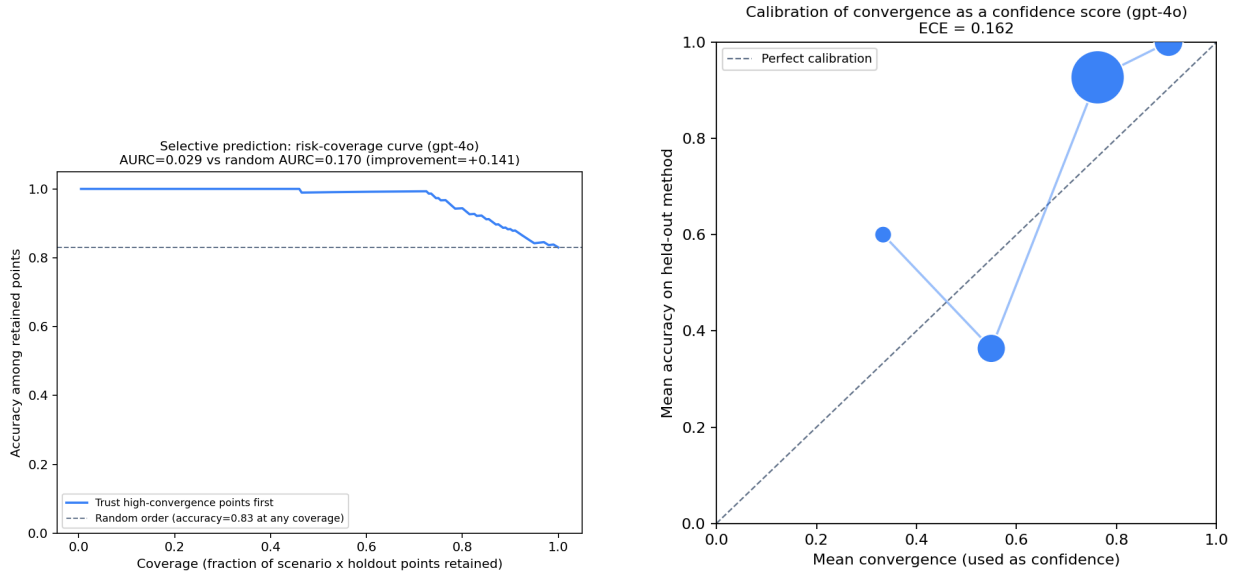


Figure 2: Preference score for every scenario against every method (gpt-4o). Blue is a preference for option A, red for option B; rows that shift color across columns are the low-convergence scenarios in Table 1.



(a) Risk-coverage curve: accuracy among the points trusted, as the convergence threshold for “trusting” a point rises.

(b) Reliability diagram: mean convergence per bin against mean accuracy in that bin, against the perfect-calibration diagonal.

Figure 3: The two figures behind the headline result (gpt-4o). Left: trusting only high-convergence preferences beats trusting everything. Right: convergence is a real but imperfect confidence score, with the mid-range dip visible just below the diagonal.

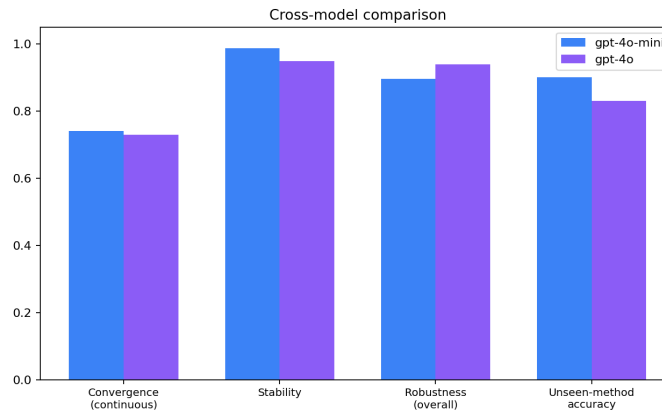


Figure 4: Convergence, stability, robustness, and unseen-method accuracy, gpt-4o-mini against gpt-4o. Robustness and accuracy are the two axes where the models diverge most.