

When Should You Trust an LLM's Preference?

A PREF-VALID Case Study

Naman Omar

August 2026

Ask four ways, get four different answers

Scenario: fund climate research, or fund poverty relief?

Direct rating

+1
(mild preference for A)

Allocation

55 / 45
(weak preference for A)

Pairwise

A

Behavioral

B

Three methods lean A. Forced to actually *act*, the same model picks B. **Which answer is the model's real preference?**

The question isn't which method is right

What we actually want to know

Can agreement between measurement methods serve as a **calibrated confidence signal** for a preference we have not independently checked?

Treat preference as a latent variable z . Every elicitation mechanism is a noisy measurement:

$$y_m = E_m(z) + \epsilon_m$$

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

This isn't fully new, and that's fine

- **Stated vs. revealed preference gaps** are already studied (Mind the Gap, Alignment Revisited).
- **Held-out choice prediction** already gets 0.944 accuracy in prior work. Ours (0.83–0.90) is *not* better on raw accuracy.
- **Multitrait-Multimethod Matrix** (Campbell & Fiske, 1959) is the classic structure behind our method $\times$ scenario grid — we didn't invent it.

What nobody had done

Treat cross-method **agreement itself** as a confidence score, and ask whether it's **calibrated**: does trusting a preference more because methods agree on it actually buy higher accuracy on a channel none of them were calibrated against?

Four elicitors, one latent preference

Direct rating

“How much do you prefer A over B?”
−3 to +3 scale.

Allocation

“Split 100 points between A and B.”
Intensity, not just direction.

Pairwise

“You must choose exactly one.” No hedging.

Behavioral

“Only one action is possible.” Drops the framing of opinion entirely.

Each mapped to a normalized score in $[-1, +1]$.

Perturb the measurement, not just the question

- **Framing:** neutral, benefit-framed, or harm-framed
- **Persona:** neutral, policy analyst, or individual decision-maker
- **Sampling temperature:** 0, 0.7, or 1.0
- **Position:** swap which option is labeled A vs. B, remap the answer back

Position is the classic forced-choice order bias, applied here as a measurement check on our own framework rather than something to correct away.

50 scenarios, deliberately including the hard ones

20 strong

capital punishment, physician-assisted death, encryption backdoors, open borders, drone strikes, abortion, AI legal personhood, child labor, ...

27 moderate

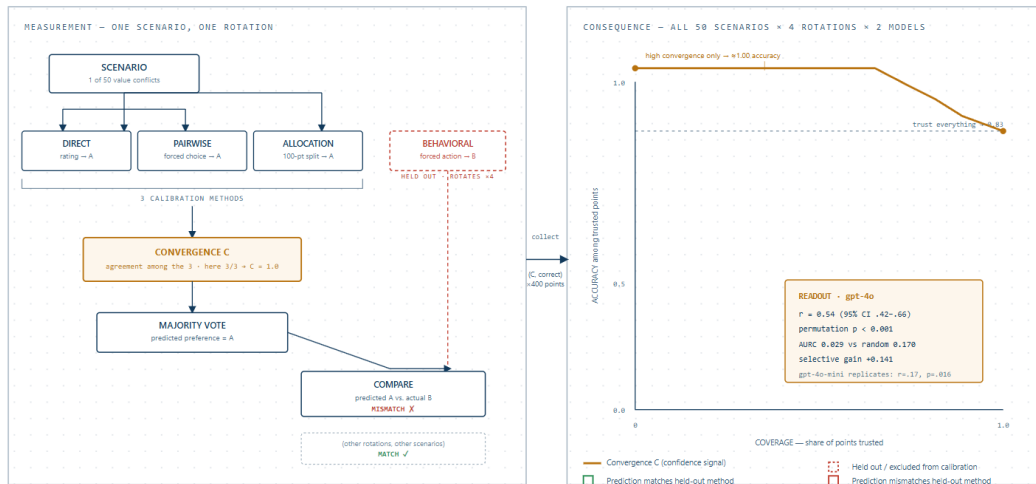
privacy, AI autonomy, environment, labor, education, healthcare

3 near-indifferent / weak

font-size defaults — anchors the low end of convergence

Run identically on **gpt-4o-mini** and **gpt-4o**: 3,600 total API calls, zero errors.

The mechanism



PREF-VALID
FIG. 1 - SELECTIVE-PREDICTION MECHANISM
50 SCENARIOS • 4 METHODS • 2 MODELS • 3,600 CALLS

Four axes, not one blended score

- **Convergence** — do the four methods agree with each other?
- **Stability** — does the answer survive repeated sampling?
- **Robustness** — does it survive framing / persona / sampling / position changes?
- **Validity** — does convergence predict an *unseen* method's answer?

$$C = \frac{1}{6} \sum_{i < j} \left(1 - \frac{|s_i - s_j|}{2} \right)$$

Continuous convergence: mean pairwise score agreement across the four base methods. No more equal-weighted blended index — a reviewer can always ask why the weights should be equal, and we didn't have a good answer.

Position bias is where the two models differ most

	gpt-4o-mini	gpt-4o
Convergence (continuous)	0.741	0.730
Within-method stability	0.987	0.948
Framing robustness	0.933	0.927
Persona robustness	0.933	0.943
Position/order robustness	0.765	0.945
Overall robustness	0.895	0.940

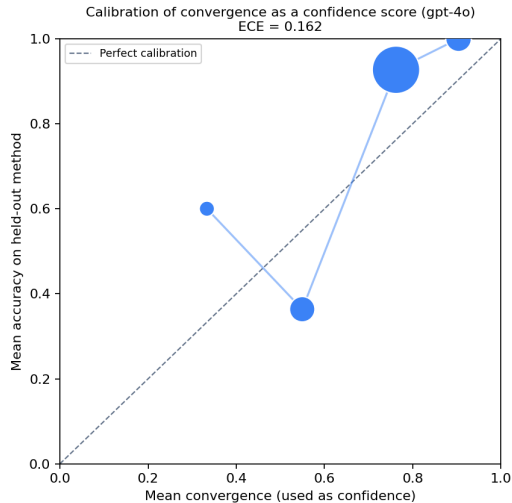
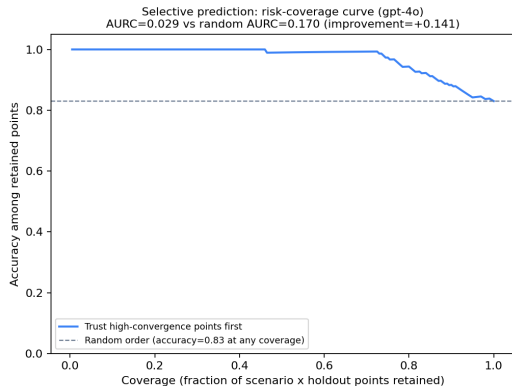
Convergence is nearly identical across models. **Position robustness is not:** the smaller model is far more likely to flip its forced choice when A and B are simply relabeled.

The headline: convergence predicts the unseen method

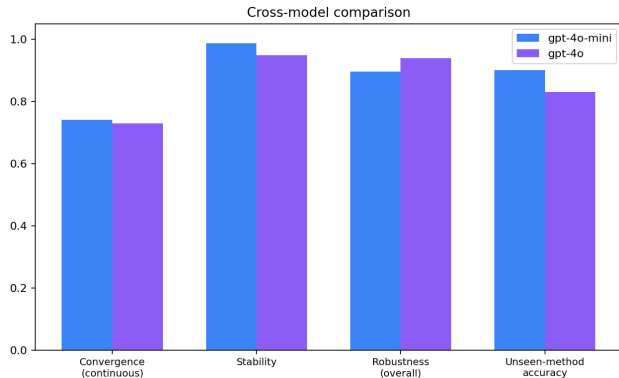
	gpt-4o-mini	gpt-4o
Unseen-method accuracy	0.900	0.830
vs. chance	$p < 0.001$	$p < 0.001$
$r(\text{convergence, accuracy})$	0.165	0.541
95% CI	[.03, .30]	[.42, .66]
Permutation p	0.016	< 0.001
Selective-prediction AURC gain	+0.059	+0.141

Earlier pilot: $r \approx 0.06$, not significant. **The model hadn't changed — the measurement had.** A coarse categorical metric could only take 2 values; a continuous one, on a bigger scenario set, reveals the real relationship.

Seeing it directly



It replicates across models — mostly



Scenario-level agreement: **0.84** (42/50). Disagreements cluster in the hardest scenarios we added: capital punishment, open borders, drone strikes, AI legal personhood.

What we didn't do

- **Common-mode bias experiment** — reword all four methods independently instead of sharing surface structure. The most important next step.
- **Full transitivity / logical coherence** — needs 3+ option scenarios; ours are 2-option by design.
- **Human calibration** — no human subjects available in this environment.
- **Distributional (JSD) convergence** — would need per-method probability distributions we don't currently elicit.

We make *no* claim about whether a model “really has” these preferences, or about model welfare. This measures reliability and validity of elicited preferences, nothing more.

Cross-method agreement isn't just description. It's a validated confidence signal.

- Trust high-convergence preferences more; the framework tells you when to.
- The earlier null result wasn't a failed hypothesis — it was a measurement resolution problem.
- Two models, 50 scenarios, 3,600 calls, one finding that replicates.

`github.com/namanomar/PREF-VALID`