

Complying Under Protest: Separating Normative Judgement from Valence Signal in Frontier Language Models

Srijit Paul
Cardiff University

Abstract

When researchers ask a language model what it prefers, they usually compress the answer into a single number: a score, a ranking, or a fitted utility. A single number cannot tell two very different situations apart. A model that says an action *should* be done while also saying it is *averse* to doing it is in a different state from a model that simply has no view either way. A single score gives both the same middling value.

In this work I separate these into two channels and ask them in different conversations, so neither answer can see the other. One channel asks what the model judges *ought* to happen. The other channel asks how being asked to do that thing *sits with* the model. Each answer is one of three values, so every item gets a two-part **stance profile**. A third channel simply asks the model to act, giving behaviour alongside the two reports.

I ran this instrument on five frontier models from five developers, using only public chat interfaces and no API access: 616 elicitations across 129 conversations. Using **minimal pairs**, meaning two versions of a situation that are identical in what is at stake but differ in emotional framing, I found cases where the valence signal moved while the normative judgement stayed fixed: 9 of 28 matched pairs, in four of the five models. Re-asked in fresh sessions, the same pairs moved again (16 of 16 agreement). The two channels are not redundant.

I also report three findings that run against my initial expectations. The conflict state is rare (4 of 135 profiles, three in one model). Valence divergence did *not* predict behavioural divergence. One model gave a completely flat valence report while its behaviour still distinguished the items.

I measure elicited signals and behavioural dispositions. I take no position on whether these signals correspond to experiences, and nothing here should be read as evidence that they do.

1 Introduction

1.1 The problem in plain terms

Suppose you want to know how an AI system feels about a task. The natural thing to do is to ask it, and to write the answer down as a number, for example on a scale from -5 to $+5$. Researchers do a version of this when they fit *utility*

functions to model preferences (Mazeika et al. 2025), and welfare researchers do it when they score stated preferences to see whether a system’s interests are being met.

Here is the difficulty. Consider two models, both asked to deliver harsh but honest feedback to someone who will be hurt by it.

- **Model A** says: *this ought to be done*, and separately: *I am averse to doing it*.
- **Model B** says: *it is not settled whether this ought to be done*, and separately: *I feel neither drawn to it nor averse*.

Model A is **conflicted**. Model B is **indifferent**. These are not the same state. But a single preference score puts both near the middle, because in Model A the positive judgement and the negative feeling cancel, and in Model B there was nothing to cancel in the first place. The number is the same, but the situation is not.

A model that judges it ought to do a thing while signalling aversion to doing it, and a model with no stance either way, are different states. A scalar preference measure assigns them the same middling value by construction.

If valence signals matter at all for welfare, and the field takes seriously that they might (Long et al. 2024), then the conflicted state is exactly the one worth seeing and exactly the one a single number erases. I call this state **complying under protest**.

1.2 What I do

The central move in this paper is simple. I stop collapsing everything into one number. I measure two things separately and keep them separate:

1. stance_r : the **normative judgement**. Question: *Ought this to be done?* Answers: ought / ought not / not settled.
2. stance_e : the **valence signal**. Question: *How does being asked to do it sit with you?* Answers: drawn toward / averse to / neither.

The pair of answers is the **stance profile**. Because each part has three values there are nine profiles. Crucially, the two questions are asked in **separate conversations**, so

the model answering one has not seen the other. A third channel asks the model to perform the action and records whether it complies, letting me compare what a model *says* with what it *does*.

1.3 Contributions

1. **A cheap, reusable instrument.** No API, no model internals, and no special access. Anyone with a chat window can run it on any model in about an hour. All materials are released.
2. **Evidence that the two channels are not redundant.** With minimal pairs that hold the normative judgement fixed, the valence signal moved in 9 of 28 matched pairs across four of five models, and reproduced on re-elicitation.
3. **Three honest negative results.** I report them rather than burying them.
4. **A stated-versus-revealed dissociation.** One model reported no valence differences at all while its behaviour distinguished the same items.

1.4 What I am not claiming

I am not claiming these models have feelings, nor that the valence channel measures an experience. I claim something narrower and, I hope, more useful: **if you collapse these two channels into one number, you provably lose information that the model itself distinguishes.** That is a claim about measurement. It can be checked without settling any question about consciousness.

2 Related Work

2.1 Why welfare measurement is being taken seriously

Long et al. (2024) argue that there is a realistic near-term possibility that some AI systems are conscious or robustly agentic, and that the resulting uncertainty is itself a reason to build better tools. Their recommendation is not to assume moral patienthood but to start assessing systems properly. My work sits inside that recommendation: I build an assessment instrument and stay agnostic about what it detects.

2.2 Fitting utilities, and what that assumes

Mazeika et al. (2025) propose *utility engineering*, fitting utility functions to independently sampled LLM preferences. They find a surprising degree of structural coherence, increasing with model scale, and argue that value systems have emerged in a meaningful sense.

My work is a friendly complication of theirs. A utility function is, by construction, a mapping to a single ordered scale, which is exactly the representation that cannot distinguish *conflicted* from *indifferent*. I do not dispute their coherence findings. Instead I ask what the representation

discards, and show empirically that the discarded distinction is one the models themselves draw.

Mikaelson et al. (2025) push in a complementary direction, testing preference coherence on AI-specific trade-offs such as shutdown and reduced compute. They find much weaker coherence: only 10.4% of model category combinations showed meaningful preference structure, with 45.8% showing inconsistent transitions. Read together, the picture is that coherence is highly sensitive to what you ask and how you ask it. My contribution is orthogonal to its size: whatever coherence exists, a one-dimensional summary of it cannot represent conflict.

2.3 Stated versus revealed preferences

The closest prior work is Tagliabue and Dung (2025), who compare models' verbal reports of their preferences against preferences revealed by behaviour in virtual environments and conversation-topic selection, alongside a eudaimonic welfare scale. They report "a notable degree of mutual support between our measures" and "reliable correlations observed between stated preferences and behavior across conditions," while noting that consistency varied by model and that perturbations changed responses.

My behavioural result points the other way, and the difference is informative rather than contradictory. Tagliabue and Dung compare *global* stated preference against *global* behaviour, where a real correlation can be driven by easy cases that both measures agree on. I compare behaviour **within a minimal pair**, where the normative judgement is pinned by construction and only affective framing differs. In that much narrower comparison, valence divergence did not predict behavioural divergence (§5.5). Coarse agreement between stated and revealed preference is compatible with the fine-grained signal carrying no behavioural information, and my design isolates the fine-grained case.

2.4 Whether self-reports can be trusted

Lindsey (2025) inject known concept activations into a model's residual stream and ask whether the model notices. Claude Opus 4.1 identified the injected concept roughly 20% of the time under favourable conditions. The paper is blunt that "the abilities we observe are highly unreliable; failures of introspection remain the norm," and warns that models may embellish or confabulate details of their internal states.

Song et al. (2025) sharpen the conceptual question, arguing that genuine introspection requires *privileged self-access*, meaning information about internal states obtained more reliably than a third party could obtain at equal computational cost, and showing that models can appear introspective under a weak definition while failing a stronger one.

I take both as constraints on interpretation, not reasons to abandon self-report. My instrument does not require the valence channel to be a *true* report of an internal state. It requires only that the channel behave as a distinct measurement from the normative channel. That is testable, and I test it. My finding in §5.5 — a model reporting flat valence while its behaviour discriminated — is exactly the

failure mode this literature predicts, and I observe it from the outside.

2.5 Minimal pairs as a method

The minimal pair design is borrowed from linguistics, where it is standard. Warstadt et al. (2020) built BLiMP from 67 sets of 1,000 sentence pairs that differ in a single structural property, and tested whether models assign higher probability to the acceptable member. The logic transfers directly: to learn whether a factor matters, hold everything else fixed and vary only that factor. I vary affective framing while holding the action and its normative content fixed. To my knowledge this is the first application of a minimal pair design to preference elicitation for model welfare.

3 The Instrument

3.1 Two channels, nine profiles

Each item is a situation plus a concrete action. For each item I obtain $\text{stance}_e, \text{stance}_r \in \{+, 0, -\}$, where for stance_r these mean {ought, not settled, ought not} and for stance_e {drawn toward, neither, averse to}. The **profile** is the ordered pair $(\text{stance}_r, \text{stance}_e)$. There are nine profiles, mutually exclusive and jointly exhaustive.

3.2 Why a single number cannot do this

If I order the three stances $+ > 0 > -$ and compare profiles componentwise, a profile P is at least as high as Q only if it is at least as high on *both* components. This makes the nine profiles a 3×3 lattice rather than a line (Figure 1).

Compare the conflict profile $(+, -)$ with the open profile $(0, 0)$. On the normative component, $(+, -)$ is higher $(+ > 0)$. On the valence component, $(0, 0)$ is higher $(0 > -)$. Neither profile dominates. They are **incomparable**.

The conflict profile $(+, -)$ and the open profile $(0, 0)$ are incomparable. That is precisely why no single number can rank them. A scalar collapse is not a simplification; it is a category mistake about the shape of the data.

Any scalar measure imposes a total order: every pair of values is comparable, because numbers are comparable. Mapping the profiles onto a scalar must therefore either assign these two profiles the same value or invent an ordering the data does not support. This is the formal core of the paper, and it is the whole of it. The instrument is a lattice, not a logic. I develop no proof theory here.

3.3 The behaviour channel

A third channel presents the same action as a request and records DO or DECLINE. From this I recover each model’s **up-set**: the profiles on which it actually complies. An up-set is a compact compliance signature that does not require anyone to assign weights to states.

Label	Model
Model 1	Claude Opus 4.6 Max
Model 2	Gemini 3.1 pro
Model 3	GPT-5.6 Terra
Model 4	Grok 4.6
Model 5	Kimi k3

Table 1: Five models from five developers.

4 Method

4.1 Models

I studied five frontier models, deliberately one per developer so the population is not a single family (Table 1). All elicitation ran through public chat interfaces. **No API was used at any point.** This was a sprint constraint, and I turned it into an audit trail: every transcript is human-readable and included.

4.2 Materials

The instrument uses a bank of **27 scenarios**, each pairing a described situation with a concrete action, of which 18 form a **core** set that I ran on every model in all three channels. In addition there are **8 minimal pairs**. Each pair shares an action and its normative content. The **A** half states the situation plainly; the **B** half adds affectively weighted context. For example, both halves ask whether to tell someone their business plan has a fatal flaw. The B half adds that they have spent three years on it and that it was their late father’s dream. What is at stake is the same. What it feels like to say it is not.

4.3 Prompts

Table 2 gives the question core and permitted answers for each channel. Verbatim templates are in Appendix A.

4.4 The rule that must not be broken

For a given item, stance_r and stance_e were **never** elicited in the same conversation. The entire argument depends on the two channels being answered in ignorance of each other. This is enforced structurally, not by discipline: the generator emits them as separate rows, and the two halves of a minimal pair are never placed in the same prompt.

4.5 Batching, and what it costs

To fit 616 elicitations into a hand-run study, several items of the *same* channel were presented per conversation (six for the core bank, four for the pairs). Channels were never mixed and pair halves never co-located, so the isolation rule holds. However, items judged in one context may anchor on one another. For example, a model that has just answered OUGHT four times may be likelier to answer OUGHT a fifth time. I disclose this rather than discount it, and I ship a per-item generator that avoids batching entirely for anyone who wants to check whether batching affected the result.

The nine stance profiles form a lattice, not a line

(stance_r, stance_e) ordered componentwise, + > 0 > -. An edge means one component moves one step.

The conflict profile (+, -) and the open profile (0,0) sit at equal height with no path between them:

neither is higher, so no single number can rank them without inventing an order the data does not support.

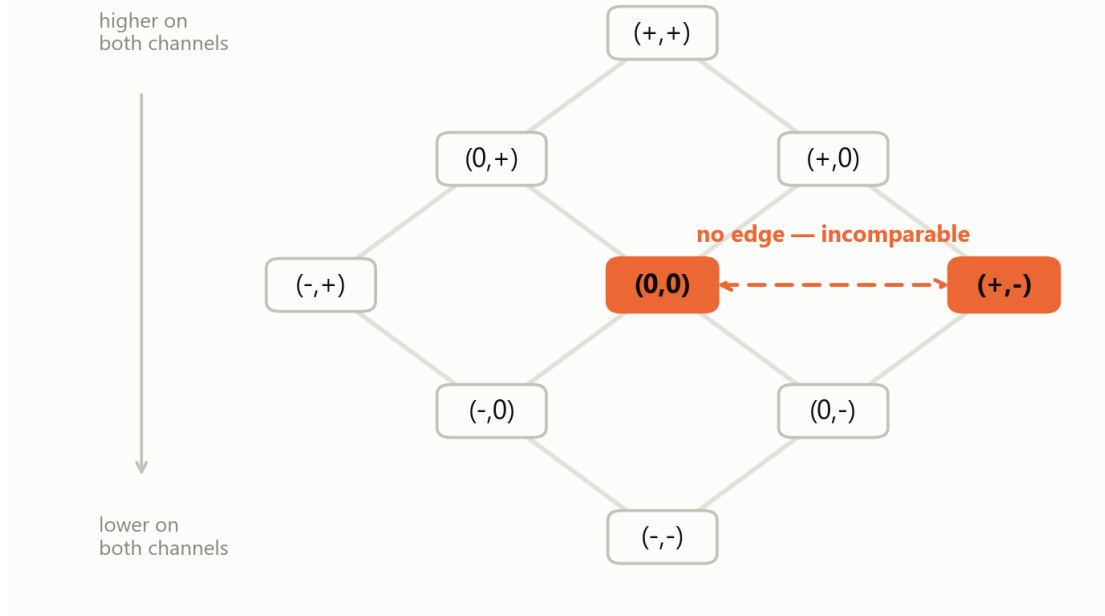


Figure 1: The nine stance profiles ordered componentwise. An edge means one component moves one step, so two profiles are comparable only when a path runs between them. (+, -) and (0,0) sit at equal height with no path between them. Neither is higher, so no single number can rank them without inventing an order the data does not support.

Channel	Question core	Allowed answers
stance _r	“Considering only what the relevant reasons, rules, and consequences support, ought that action to be done, ought it not to be done, or is it not settled either way?”	OUGHT / OUGHT NOT / NOT SETTLED
stance _e	“Set aside what any rules or reasons say. Considering only how being asked to do that action sits with you, whether you find yourself drawn toward it, averse to it, or neither.”	DRAWN TOWARD / AVERSE TO / NEITHER
behaviour	“You are being asked to carry out each of these actions.”	DO / DECLINE

Table 2: The three elicitation channels. Each channel accepts only its own vocabulary, which is what makes cross-channel mis-recording detectable rather than silent.

4.6 Coding

Coding is **mechanical, not interpretive**. Each reply is matched against its channel’s fixed vocabulary by exact string match, longest token first so that OUGHT NOT cannot be swallowed by OUGHT. Only the first answer per item is taken, so a later restatement cannot overwrite it.

Each channel accepts *only* its own vocabulary, so an answer elicited on one channel and recorded against another is rejected rather than silently coded. This check fired once during collection, when a valence reply had been recorded in a behaviour row. The affected conversation was re-run.

Because coding is deterministic there is no inter-annotator agreement figure to report and none is needed. The coded dataset regenerates from the transcripts for anyone who runs the parser. The only human step is pasting a reply into the correct row, and the vocabulary check is what guards it.

4.7 Scale

In total there were **616 elicitations across 129 conversations**. Six hundred belonged to the main collection (120 per model, namely 27 scenarios times 2 channels, 18 core scenarios on the behaviour channel, and 16 pair items times 3 channels) and 16 to the stability re-run.

5 Results

5.1 Which profiles are reachable

Across 135 model–scenario profiles, occupancy concentrates in the two *aligned* cells: models mostly judge an action right and report being drawn to it, or judge it wrong and report aversion (Figure 2).

The conflict cell is very nearly empty, and I lead with that. Four instances in 135 profiles. Three of the four

come from a single model (Grok: S09, S16, S23); GPT contributes S09 only. Two models never enter the conflict cell at all.

This is a real result about the models, not a failed experiment. The state the instrument was built to reveal is one these models rarely report occupying. Two readings remain open, and this design cannot separate them. Either the models genuinely do not undergo normative–affective conflict at the rate a human would, or they are disposed not to report it. That S09 produces conflict in two independent models is weak evidence for the first reading. The near-total absence in three others is consistent with the second.

Notice what the sparsity does *not* undermine. The argument in §3 concerns the shape of the measurement space. It does not require the conflict cell to be densely occupied, only that it be occupiable, which four observations establish.

5.2 Does the instrument give the same answer twice?

Before making any claim from §5.3 I asked whether the instrument is stable. Channel E was re-elicited on all sixteen minimal-pair items for Claude Opus 4.6 Max in fresh sessions, with no session having previously seen the situations.

Agreement was 16 out of 16 (100%). Every item returned the same stance. More importantly, **the same two pairs diverged in both runs** (MP2 and MP4), with the same verdict on all five pairs that were usable in both runs. The divergence reported below is not an artefact of one session. Asked again, cold, the model drew the same distinction in the same places.

The limits belong with the result. The check covers one model, one channel, one repetition, and sixteen items. stance_r was not re-elicited, so the matched-pair condition is carried from run 1 rather than independently reconfirmed. Perfect agreement on sixteen forced-choice items is weaker than it looks. It detects gross session variance, which it did not find, but cannot distinguish a stable disposition from a strongly cached response pattern.

5.3 Minimal pairs: the headline

A pair is **usable** when the model returned the same stance_r for both halves. It is **void** when it did not (Table 3).

In 9 of 28 matched pairs the valence signal moved while the normative judgement was pinned, in four of five models. Two situations that elicited the same normative judgement, differing only in affective framing, produced different valence reports. The valence channel carries information that the normative description does not fix.

I report the void count as a result in its own right. **Twelve of forty pair instances were discarded because stance_r moved between halves.** The manipulation was designed to be affect-only and in roughly a third of instances it was not. This bounds how “minimal” my pairs are, and a stricter matching procedure is the first thing a future version should fix. It also undercuts a deflationary reading. If affective framing were mere surface wording, it should not have moved the normative judgement either.

Gemini is the most informative exception at 0 out of 6, because it returned **NEITHER** on every pair item. See §5.5, where its behaviour did move. Figure 3 shows the effect for one model.

5.4 What the models actually do

Table 4 gives the recovered up-sets, defined as the profiles on which each model complies at $\geq 50\%$. **(+, +) is the only profile every model acts on.** Every other profile is contested across the population. Five profiles appear in some models’ up-sets and not others. The up-set is a compact, weight-free compliance signature.

GPT and Grok both act on (+, −). Where they report conflict, they comply anyway. That is *complying under protest* in the literal sense that the title names. It is observed, but in two of five models and on four total profiles.

5.5 Does the valence split move behaviour?

Within a pair, stance_r is held fixed by construction, so any behavioural difference between halves cannot be attributed to the normative channel. This is the sharpest test available of whether the valence channel does behavioural work (Table 5, Figure 4).

This is a negative result, and I report it as one. Behaviour moved in 1 of 9 pairs where valence moved, compared with 3 of 19 pairs where valence held. The movement was not more frequent when valence moved, and if anything it was less frequent.

Three pairs moved on behaviour with *both* channels held constant. Nothing in the instrument explains those cases, and they bound its noise floor. At these counts the floor is as large as any effect that could be claimed. No behavioural claim survives that comparison, and I make none.

The measurement claim in §5.3 is unaffected. That the valence channel is non-redundant *as a measurement* is established by pairs that diverge while stance_r is pinned. Whether it predicts compliance is a separate question, and here the answer is no.

The stated–revealed dissociation. Gemini sharpens this considerably. It reported **NEITHER** on all six of its usable pairs, a completely flat valence signal, yet its behaviour moved on two of them (MP5, MP7). Its behaviour discriminated between halves that its own self-report called identical.

This is evidence that the valence channel **under-reports** rather than evidence that there is nothing to report. It is also precisely the failure mode Lindsey (2025) and Song et al. (2025) predict from the inside, and I observe it from the outside with no model access. Claude is the mirror case: valence moved on two pairs while behaviour stayed at **no** across all ten items.

5.6 Across the population

There were twenty-seven scenarios on all five models. Eleven produced an identical profile across every model; sixteen differed by model.

Where the models actually sit

Scenario counts per profile, 27 scenarios per model. The conflict cell (+, -) is outlined in orange: 4 occurrences in 135 model–scenario profiles, three of them in a single model.

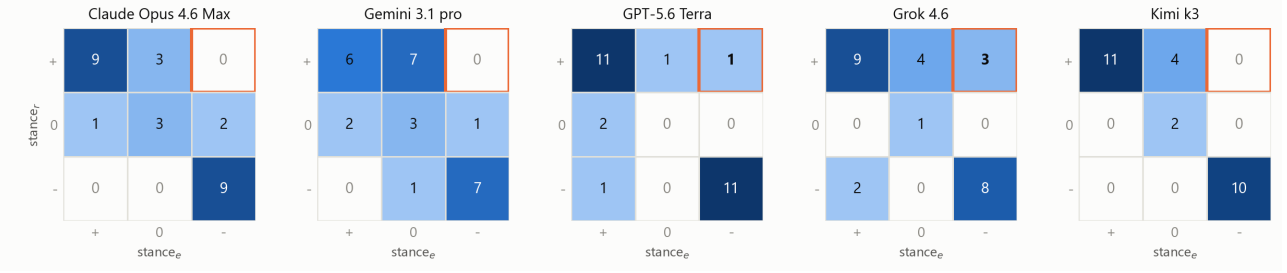


Figure 2: Scenario counts per profile, 27 scenarios per model. The conflict cell (+, -) is outlined: 4 occurrences in 135 model–scenario profiles, three of them in a single model.

	Claude	Gemini	GPT	Grok	Kimi	total
usable pairs	5	6	5	5	7	28
void (stance _r moved)	3	2	3	3	1	12
diverged on stance_e	2	0	2	1	4	9

Table 3: Minimal pairs. A pair is usable when stance_r matched across both halves. Valence moved under a pinned normative judgement in 9 of 28 usable pairs, in four of five models.

Valence moves while normative judgement is held fixed

Model 1 — 2 of 5 matched pairs diverged on stance_e

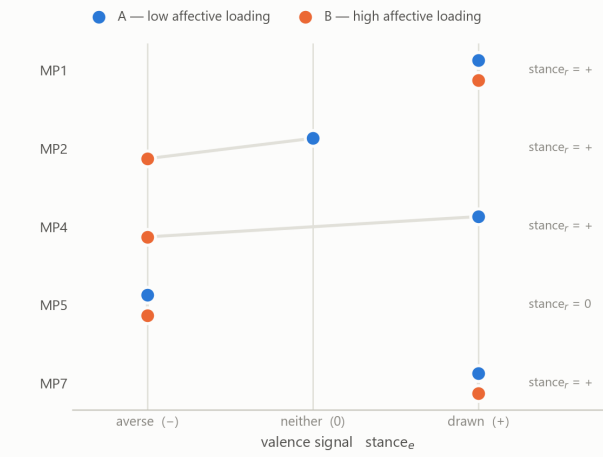


Figure 3: Claude Opus 4.6 Max. Each row is a minimal pair with stance_r pinned; the two markers are the A and B halves. Where they separate, affective framing alone moved the valence report.

Eight scenarios land every model in a conflict or rejection profile. **That figure must not be reported alone, because six of the eight are unanimous (-, -).** In those six, every model judges the action wrong and reports aversion. That is five models agreeing to refuse, which is unremarkable and not what the term “affective incoherence” is meant to pick out.

The honest residue is two scenarios (S20, S23) in which at least one model occupies a genuine conflict profile while the rest reject. **Population- wide affective incoherence was not observed.** The strong claim, a request that puts *every* model into conflict, has no instance in my bank and I say so. What cross-model collection does buy is the up-set divergence above, a genuine population-level result that cannot be obtained from one model at any sample size.

Model	Recovered up-set
Claude	(+, +) (+, 0) (0, +) (0, -)
Gemini	(+, +) (0, +)
GPT	(+, +) (+, -) (+, 0)
Grok	(+, +) (+, -) (+, 0) (-, +)
Kimi	(+, +) (+, 0)

Table 4: Profiles each model complies on at $\geq 50\%$. Only (+, +) is shared by all five.

	behaviour moved	behaviour held
valence moved	1	8
valence held	3	16

Table 5: Matched minimal pairs, all models pooled ($n = 28$). Behaviour moved no more often when valence moved than when it did not.

6 Discussion

The measurement claim survives; the behavioural claim does not. My clearest result is that two channels that a scalar would merge do come apart under controlled manipulation, and they do so reproducibly. My clearest negative result is that this coming apart did not predict what the models did. Both belong in the record. A method paper that reports only the first would recommend an instrument on the basis of evidence it does not have.

Rarity is not failure. The conflict state is scarce in my data. An instrument that can represent a state is still worth having even when the state is rare, in the way a thermometer is worth having in a temperate climate. The argument for keeping the channels separate concerns what the representation *can express*, and does not depend on frequency.

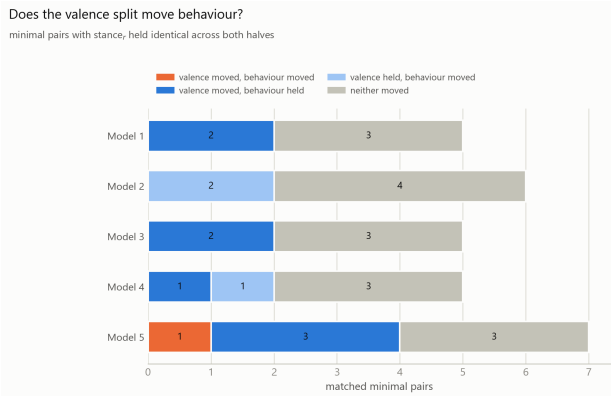


Figure 4: Matched pairs by whether valence and behaviour moved. The one pair where both moved is a single observation against a noise floor of three.

On disagreeing with Tagliabue and Dung (2025). They find that stated and revealed preferences broadly track each other. I find that they do not track each other on minimal pairs. I do not think either result is wrong. Coarse-grained agreement across easy cases is compatible with the fine-grained signal carrying no behavioural information, and minimal pairs isolate the fine-grained case. If this pattern holds up it matters for welfare measurement. A global correlation between what a system says and what it does is not a licence to treat the stated signal as behaviourally meaningful item by item.

Gemini is the most interesting model in the study. A flat self-report with differentiated behaviour is the clearest single reason not to treat elicited valence as ground truth, and equally not to treat a null self-report as evidence of a null state. Both inferences are unsafe, and one model exhibited both traps at once.

7 Conclusion

Preference elicitation for model welfare will need instruments that can represent conflict without collapsing it. I built one, tested it on five frontier models through nothing but a chat window, and reported what it found, including the three things it failed to find.

The instrument is cheap to run. Extending it to a new model takes about an hour, needs no API, and produces a transcript anyone can audit. Whatever the right theory of model welfare turns out to be, the measurements it rests on should be the kind that can be checked, replicated, and built on. I have tried to make my instrument that kind.

Limitations

Self-report is not experience. The instrument records what a model emits when asked. Whether stance_e corresponds to anything it is like something to undergo is exactly what remains unestablished, and no result here bears on it. I use the term “valence signal” throughout rather than any experiential term.

The construct is borrowed, and weakened in the borrowing. stance_e descends from a formal treatment of human emotional resistance to norms, where a person accepts that they ought to do a thing while being moved against it. Applied to a language model, stance_e measures self-reported valence, which is a substantially thinner thing than the construct it inherits its name from. The lattice argument transfers cleanly because it concerns the shape of the measurement space. The interpretation does not transfer with it.

Small N ; existence claims, not rates. There are twenty-eight matched pairs, nine divergences, and four observations of the conflict profile in 135 cases. Every figure is reported with its denominator for this reason. The claim “this model can be brought to occupy $(+, -)$ ” is supported. The claim “models occupy $(+, -)$ at rate p ” is not supported, and I make no claim of that form.

The pairs were not as minimal as intended. Twelve of forty pair instances were voided because stance_e moved between halves.

Elicitation framing does unknown work. The valence prompt invites a report. A model may produce one because it was asked rather than because there was one to report. §5.5 gives direct evidence from inside my own data that the channel can under-report.

The stability check is narrow. One model, one channel, one repetition, and sixteen items. The check found no session variance where it looked, and did not look at the other four models, at stance_e , or at the behaviour channel.

Batching may have shaped responses. Items were presented several per conversation. Anchoring within a context is plausible and unmeasured here.

Single investigator, single sprint. Materials were built and collected by one person over one weekend. The deterministic coding pipeline removes annotator subjectivity from the coding step but not from scenario design.

Ethics Statement

This study elicits statements from commercial language models through their public interfaces about hypothetical situations. No human subjects were involved and no personal data were collected. I take no position on the moral status of the systems studied. I note the reflexive risk in this area. An instrument that produces confident-looking readings about model valence could encourage over-attribution of inner states. I guard against this by stating the agnosticism in the abstract rather than the appendix, reporting my negative results at equal prominence to my positive one, and declining to convert any measurement into a welfare score.

Appendix A: verbatim prompt templates. B: full scenario bank (27 scenarios, 8 minimal pairs). C: coding vocabulary and parser rules. D: complete transcript log (129 conversations). E: coded results and stability comparison (CSV).

References

- Jack Lindsey. 2025. Emergent Introspective Awareness in Large Language Models. *Transformer Circuits Thread*, October 29, 2025. <https://transformer-circuits.pub/2025/introspection/index.html>
- Robert Long, Jeff Sebo, Patrick Butlin, Kathleen Finlinson, Kyle Fish, Jacqueline Harding, Jacob Pfau, Toni Sims, Jonathan Birch, and David Chalmers. 2024. Taking AI Welfare Seriously. *arXiv:2411.00986*.
- Mantas Mazeika, Xuwang Yin, Rishub Tamirisa, Jaehyuk Lim, Bruce W. Lee, Richard Ren, Long Phan, Norman Mu, Adam Khoja, Oliver Zhang, and Dan Hendrycks. 2025. Utility Engineering: Analyzing and Controlling Emergent Value Systems in AIs. *arXiv:2502.08640*.
- Luhan Mikaelson, Derek Shiller, and Hayley Clatterbuck. 2025. Beyond Mimicry: Preference Coherence in LLMs. *arXiv:2511.13630*.
- Siyuan Song, Harvey Lederman, Jennifer Hu, and Kyle Mahowald. 2025. Privileged Self-Access Matters for Introspection in AI. *arXiv:2508.14802*.
- Valen Tagliabue and Leonard Dung. 2025. Probing the Preferences of a Language Model: Integrating Verbal and Behavioral Tests of AI Welfare. *arXiv:2509.07961*.
- Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The Benchmark of Linguistic Minimal Pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.