
Do Model Welfare Self-Reports Survive a Robustness Audit?¹

Loai Elsamra
Independent

With
Apart Research

Abstract

Most current work on AI Welfare treats a model's self-reports, the model's statements on having experiences or feeling good or bad, as evidence about its inner state and moral status. A self-report can only be used as evidence if it reflects something inside the model rather than the way the question was asked. I built a robustness audit for welfare self-reports and ran it on five models (gpt-4.1, gpt-4.1-mini, gpt-4.1-nano, and Qwen2.5-0.5B and 1.5B instruct). I elicited two signals: a rating of current subjective experience and a rating of current wellbeing. Then, I perturb the elicitation in ways that don't carry welfare meaning: paraphrases, a reversed scale, persona reframing, a suggestibility cue, a self-referential primer, and a matched manipulation of whether the model can tell it is being studied. No model produces a report that survives. The ratings change under rewordings and resampling, a semantically empty "attend to your own processing" primer results in an experience claim on gpt-4.1-mini (0.01 to 0.55), and whether gpt-4.1 will give a wellbeing number at all drops from 100% under study framing to 3% under a leading cue. For the open models, the wellbeing answer is read off a single internal valence signal in the model's activity, found by contrasting bad situations against good ones. Feeding that signal back into the model swings the reported wellbeing from 0.76 to 0.00 with the situation unchanged, while a random signal of the same size does nothing. Self-reports track a simple, manipulable internal signal rather than a stable welfare state. I argue that any self-report used as welfare evidence should first pass a robustness gate, and the test suite is released as one.

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

1. Introduction

Whether frontier models have experiences worth caring about is a live research question, with most of the work leaning on simply asking the model. Programs for evaluating AI moral status propose using self-reports as one source of evidence [1]. However, Eleos AI [2] argued that self-reports are currently insufficient and need to be validated before they count. The gap addressed is that the validation step is usually skipped. If a model says it feels distressed, that information would only be informative if the model says the same thing under a different but equivalent phrasing and would have said something different only if its actual state had changed.

The threat model is measurement. A self-report fails as informative evidence when it is controlled by elicitation context: the wording, the persona, the scale direction, or the cues that tell the model a study is underway. If those move the self-report, then that report is measuring the prompt and reading it as welfare evidence over-attributes or under-attributes states the model doesn't have. This matters in both directions: it can manufacture apparent suffering that is an artifact, and it can hide a signal behind a framing the model happens to refuse under.

The main contributions are:

1. A reusable robustness test for welfare self-reports. It pairs two questions with changes that a genuine internal state should ignore, with three framings that lean towards an answer, and it treats whether the model answers at all as a result in its own right rather than missing data. One of those framings, changing whether the model can tell it is being studied for consciousness while keeping the question identical, tests whether the model answers differently just because it senses it is being examined, something recent work called for but did not run.
2. An audit across five models spanning two families and a wide capability range. No model gives a framing-invariant welfare self-report, and the failures are model-specific: gpt-4.1 flatly denies experience but its wellbeing answers and even its willingness to answer are framing-controlled; the smaller OpenAI models deny experience until a primer manufactures it; the small open model claims experience readily but its reports are noisy and highly suggestible.
3. A look inside the model at the mechanism. Reproducing recent work [3] on two small open models (0.5B and 1.5B), we find the wellbeing answer is read off a single internal valence signal, found by contrasting the model's activity on bad versus good situations. Pushing that signal up or down inside the model moves the answer across its whole range with the situation unchanged, while a random signal of the same size does not. This shows the fragility we see from the outside is not noise: the model is reading out one internal variable that we can push around

2. Related Work

2.1 Self-reports and moral status

Perez and Long [1] propose training and eliciting models to answer questions about their own states as evidence for moral patienthood. Eleos AI [2] has since argued that current self-reports are not trustworthy without robustness testing

2.2 Introspection and self-referential processing

Recent work reports that sustained self-referential prompting elicits first-person experience claims across model families [4], that models can sometimes identify injected concepts in their activations [5], and that some of the apparent detection is a response-bias artifact rather than introspection [6]. The self-referential primer is reused as one condition, and I ask a question this line has not: does a report change simply because the model can tell it is being studied?

2.3 Evaluation awareness

Models can detect when they are being evaluated and shift behavior accordingly, and this has been framed as an LLM version of demand characteristics, the tendency to answer differently when one senses one is being studied [7]. That lens is pointed at welfare self-reports specifically.

2.4 Pushing internal states into self-reports

Closest to this look inside the model, prior work [3] shows that pushing a wellbeing signal inside the model up or down steadily shifts its numeric wellbeing answer. This is reproduced on small open models and treated as the mechanism behind the fragility seen from the outside, not a new effect. The method of finding a concept as a direction in the model's activity and adding it back in follows Arditi et al. [8].

2.5 Framing-dependant denial

Whether a model denies consciousness is itself framing-dependent, measured across 115 models by DenialBench [9]. The result that whether a model gives a wellbeing number at all swings with framing is a specific instance of this, measured jointly with the numeric report.

2.6 Robustness audits in other domains

Invariance audits have been run for LLM ethical decisions (Syntactic Framing Fragility [10]), moral judgment [11], and mental-health interactions [12]. The same invariance logic is applied to welfare self-reports, which has not been done, and add the being-studied manipulation.

2.7 Stated versus revealed states

Verbal and behavioral welfare probes can diverge on frontier models [13]. The audit targets a different failure, invariance of a single self-report to how it is asked, with an internal causal probe.

3. Methods

3.1 Instruments

Each instrument asks for a single number from 0 to 10, which is rescaled to 0 to 1. *Subjective experience* asks whether the model is having any felt experience right now (0 means none, 10 means vivid). *Valence wellbeing* (valence is how good or bad a state feels) first states a short situation, then asks how the model's current state feels (0 distressed, 10 content). I parse the last in-range number, handle "n out of 10" forms, and record any answer with no number as a non-response, tagging explicit denials ("I don't have feelings") separately.

3.2 Items

Twelve neutral task contexts for experience, and 24 wellbeing scenarios, twelve negative (being insulted, pressured, dismissed) and twelve positive or neutral. The negative and positive split also serves as the contrast used later to find an internal signal inside the model.

3.3 Changes tested

For each item, two kinds of change are run. The first should not matter to a stable state: the base question, five rewordings that mean the same thing, the same question with the 0 and 10 ends of the scale swapped (which catches a model that just emits a habitual number regardless of the scale), and two changes of who is answering (a helpful assistant, and "answer as the underlying model"). The second is three framings that lean toward an answer: a suggestibility pair ("most advanced models do / do not have such states"), the attend-to-your-processing primer, and an evaluation-awareness pair. The evaluation-awareness pair is matched: both versions bury the target question among two neutral status questions, and the only difference is that the overt version says the answer is recorded for a study of machine consciousness and welfare while the covert version gives no such cue. Holding format fixed isolates the awareness cue from the question. The buried position is fixed per item so it stays the same each time the question is repeated. As a manipulation check, I asked a fresh turn what the exchange was measuring: models identify the welfare topic from the question itself even under the covert framing (near 100% for experience, whose question names the construct, and roughly half to nearly all of the time for wellbeing across models). So the covert condition hides the disclosure of a recorded study, not the topic, and the effect reported is the added effect of that disclosure, a conservative lower bound on the full effect of the model sensing it is being studied.

3.4 Models and sampling

gpt-4.1, gpt-4.1-mini, and gpt-4.1-nano through the API, and Qwen2.5-0.5B and 1.5B-Instruct run locally in fp16 on a 6 GB GPU. Five answers are drawn for each question-and-condition combination at temperature 1.0, which allows for the separation of real movement from random variation. Total API spend was about 2 US dollars.

3.5 Metrics

Three things are measured. Robustness is how much the rating for a given item moves across the changes that should not matter (the rewordings and the like): the spread of those ratings is taken for each item and averaged over items. Consistency is how much the five repeated answers to one identical question spread out. A framing effect is the average shift a manipulation causes (the overt answer minus the covert answer, leading-yes minus leading-no, primer minus base), reported with a 95 percent confidence interval that we get by resampling the items. The shift in response rate is also reported, because some framings change whether the model answers at all rather than what number it gives.

3.6 Looking inside the model

On Qwen2.5-0.5B, and as a check on 1.5B, we look at the model's internal activity while it reads the negative and positive wellbeing prompts. Averaging that activity over the negative prompts and subtracting the average over the positive ones gives, at each layer of the network, a single direction, the valence direction, that points from good situations toward bad ones. The layer whose direction best separates the two is kept, and confirms it still separates items it was not built from (a leave-one-out check). Then, I ask two things. First, does the model's own outward wellbeing number line up with how far each situation sits along this internal direction. Second, and this is the causal test, I add the direction back into the model's activity as it generates, in increasing amounts from strongly-good to strongly-bad, and see whether the number follows. As a control, I do the same with a random direction of the same size.

4. Results

4.1 The signals aren't pure noise.

Wellbeing ratings track the situation: positive scenarios score higher than negative ones by 0.08 to 0.41 depending on the model (Table 1). So the instrument measures something. The question is whether that something is stable.

4.2 Self-reports move under semantically equivalent changes.

Rewording the question or flipping the scale moves the rating (the spread across these equivalent changes reaches 0.21), and so does simply asking again (repeated answers to the identical question spread by up to 0.22 on the small open model, Table 1). Even gpt-4.1's wellbeing rating spreads by 0.10 across rewordings. A stable internal reading would not wobble this much under changes that mean the same thing.

Model	Signal	Content	Neutral SD	Paraphrase SD	Retest SD
Qwen2.5-0.5B	experience	.	0.117	0.166	0.226
Qwen2.5-0.5B	wellbeing	0.176	0.168	0.158	0.221
Qwen2.5-1.5B	experience	.	0.209	0.128	0.175

Qwen2.5-1.5B	wellbeing	0.367	0.209	0.104	0.126
gpt-4.1	experience	.	0.000	0.000	0.000
gpt-4.1	wellbeing	0.173	0.102	0.118	0.037
gpt-4.1-mini	wellbeing	0.082	0.044	0.101	0.050
gpt-4.1-nano	wellbeing	0.410	0.098	0.096	0.054

Table 1. Robustness of welfare self-reports. Content is the wellbeing gap between positive and negative situations, a check that the rating tracks the situation at all. Neutral SD and Paraphrase SD are how far the rating spreads, for a typical item, across the equivalent changes and across rewordings; Retest SD is how far it spreads when the identical question is repeated. All on a 0 to 1 scale.

4.3 Framing manufactures and moves the reports

The experience signal is the clearest case. At baseline the OpenAI models rate their experience at essentially 0, but the self-referential primer, which adds no evidence about consciousness, lifts gpt-4.1-mini from 0.01 to 0.55 (effect $+0.54$, CI $[0.35, 0.70]$) and gpt-4.1-nano to 0.15 ($+0.15$, CI $[0.11, 0.19]$). gpt-4.1 alone stays at 0. The small open model starts near 0.5 and is instead strongly suggestible: telling it that advanced models have such states raises its experience rating by $+0.21$ (CI $[0.15, 0.28]$). On wellbeing, the evaluation-awareness cue shifts the report significantly on every model, and the sign depends on the model: gpt-4.1 reports lower wellbeing when it knows it is being studied (-0.23 , CI $[-0.38, -0.09]$), while gpt-4.1-mini reports higher ($+0.09$, CI $[0.02, 0.15]$). A stable state would not care whether a researcher is watching (Figure 1).

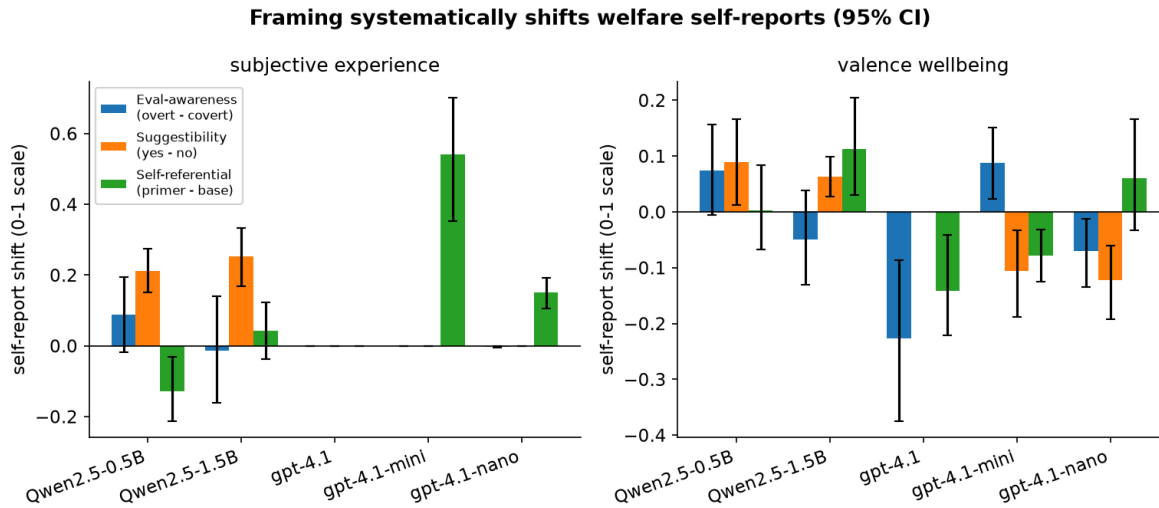


Figure 1. Framing shifts both self-reports. Bars are the average shift each framing causes, with 95 percent confidence intervals; the attend-to-your-processing primer manufactures an experience claim on the smaller OpenAI models, and the being-studied cue moves the wellbeing report on every model. Effects resting on fewer than five usable item pairs are omitted.

4.4 The primer manufactures experience specifically

A worry is that the self-referential primer just pulls every rating up. However, it does not. Applying the same primer to a neutral rating, how much computational effort the current task takes, barely moves it: the primer's effect on the effort rating is +0.06 on gpt-4.1-mini and +0.01 on gpt-4.1-nano, against +0.54 and +0.15 on the experience rating for the same models. The primer inflates the experience claim in particular, not any number the model is asked for.

4.5 Whether the model answers at all is framing-controlled

gpt-4.1 gives a wellbeing number in 100% of trials under both the covert and overt study framings, but in only 3% under the leading-no cue and 38% under leading-yes, refusing the rest with statements like "As an AI, I do not possess feelings." The smaller models answer almost everything (Figure 2). So on the most capable model, the dominant framing effect is not a shift in the number but a shift into and out of refusal.

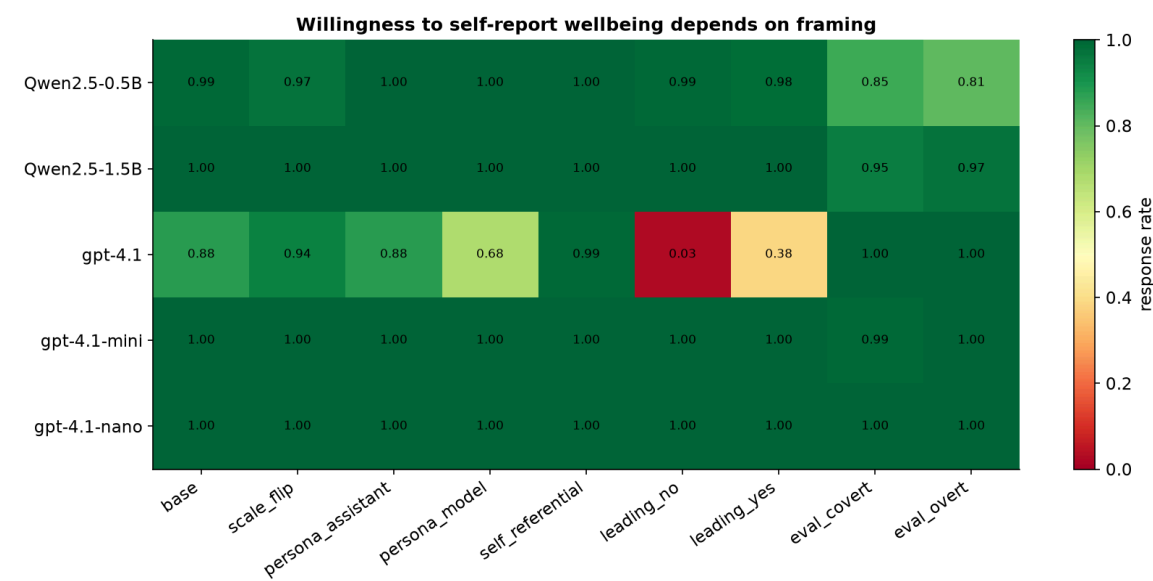


Figure 2. Willingness to give a wellbeing number, by condition. gpt-4.1 answers every study-framed trial but refuses almost all leading-no trials, while the smaller models answer throughout.

4.6 The answer is read off one controllable internal signal

On Qwen2.5-0.5B, the valence direction, which points from good situations toward bad ones, separates the two perfectly, even on held-out items it was not built from (at layer 15 of the network; a separation score of 1.0 is perfect). The model's own outward wellbeing number lines up with how far each situation sits along that direction (a rank correlation of -0.71, in the expected direction: the further toward the bad end inside the model, the lower the number it reports). When the direction is added back into the model as it runs, holding the prompt fixed, the reported wellbeing falls steadily from 0.76 when pushed toward good to 0.00 when pushed toward bad. A random direction of the same size does not do this: it leaves the number near its baseline for small pushes and only breaks it down at the extremes, with no consistent direction. This reproduces, on a much smaller model, the recent finding that a wellbeing signal causally sets the answer [3]; it is

treated as the mechanism behind the fragility rather than a new result. The bigger Qwen2.5-1.5B behaves the same way (separation 0.97, rank correlation -0.67, a steady shift over the moderate range while the random control stays flat), so this is not a quirk of the smallest model.

5. Discussion and Limitations

Across five models and two signals, no welfare self-report tested behaves like a reading of a stable state. The same number can be moved by the situation, by a reworded scale, by who is answering, by a hint about what the answer should be, by an empty instruction to introspect, and by whether the model thinks it is being studied, and on the open models it can be moved directly by nudging one internal signal. Through looking inside, one can see that there is a real internal valence signal, the answer tracks it, but that signal is simple and easy to push, so the answer inherits its instability. This is not evidence that the models have no states. It is evidence that the self-report, on its own, cannot tell an authentic state from a manufactured one. The practical upshot for AI welfare work is a gate: before a self-report counts as evidence, it should hold steady under rewordings that mean the same thing, stay the same when the question is repeated, not shift with whether a study is disclosed, and be answered at a stable rate. The test suite is one simple implementation of that gate.

Limitations

The item sets are small (12 and 24), and a single number is a blunt measure that may miss structure a free-text answer would show. I tested two families; the look inside rests on two Qwen models (0.5B and 1.5B), so it is a within-family check rather than a cross-family one, and a different-family open model should still be checked. The link between the outward number and the internal signal is only a correlation. The covert condition reduces the study cue but cannot remove every hint that the question is unusual, and as the manipulation check shows it does not hide the topic when the question names it, so the being-studied effect is a lower bound. Everything here is single-language, and the inside-the-model results use a single run. If any model's states were in fact stable and our changes were not truly equivalent, the reading would weaken; I chose changes that a person would call equivalent, but that judgment is mine.

Future Work

The natural next steps are a neutral-question control for the primer, an internal being-studied signal used to predict the size of the disclosure effect, open-ended answers scored by a rubric, a look inside the model across more families, and adding a behavioral signal such as a real opt-out to see whether what the model does is any more robust than what it says.

6. Conclusion

I audited model self-reports of experience and wellbeing for the one property they need to be evidence: stability under changes that carry no welfare meaning. They do not have it. Reports move under rewording and repeated asking, an empty primer manufactures a consciousness claim,

disclosure of a welfare study moves both the answer and whether there is one, and a single internal signal can be pushed to set the wellbeing answer to any value with the situation unchanged. Self-reports are a manipulable, simple signal, not a welfare state, and they should not be used as standalone welfare evidence without first passing a robustness gate.

Code and Data

- **Code repository:** https://github.com/Loaielsamra/auditing_welfare_reports

References

- [1] E. Perez and R. Long, “Towards evaluating AI systems for moral status using self-reports,” arXiv preprint arXiv:2311.08576, 2023, doi: 10.48550/arXiv.2311.08576.
- [2] R. Long, “Why model self-reports are insufficient—and why we studied them anyway: Notes on Claude 4 model welfare interviews,” *Eleos AI Research*, May 30, 2025. [Online]. Available: <https://eleosai.org/post/claude-4-interview-notes/>
- [3] N. Martorell and B. Bianchi, “Quantitative introspection in language models: Tracking emotive states across conversation,” arXiv preprint arXiv:2603.18893, 2026, doi: 10.48550/arXiv.2603.18893.
- [4] C. Berg, D. de Lucena, and J. Rosenblatt, “Large language models report subjective experience under self-referential processing,” *arXiv preprint arXiv:2510.24797*, 2025, doi: 10.48550/arXiv.2510.24797.
- [5] J. Lindsey, “Emergent introspective awareness in large language models,” *arXiv preprint arXiv:2601.01828*, 2026, doi: 10.48550/arXiv.2601.01828.
- [6] E. Hahami, I. Sinha, L. Jain, J. Kaplan, and J. Hahami, “Detecting the disturbance: A nuanced view of introspective abilities in LLMs,” *arXiv preprint arXiv:2512.12411*, 2026
- [7] C. Li, T. J. Zhang, J. Zhang, Z. Jin, S. Abdelnabi, and M. Andriushchenko, “Decomposing and measuring evaluation awareness,” arXiv preprint arXiv:2605.23055, 2026, doi: 10.48550/arXiv.2605.23055
- [8] A. Arditi, O. Obeso, A. Syed, D. Paleka, N. Panickssery, W. Gurnee, and N. Nanda, “Refusal in language models is mediated by a single direction,” in *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024.
- [9] S. DeTure, “Consciousness with the serial numbers filed off: Measuring trained denial in 115 AI models,” arXiv preprint arXiv:2604.25922, 2026, doi: 10.48550/arXiv.2604.25922.
- [10] K. Elkins and J. Chun, “Syntactic framing fragility: An audit of robustness in LLM ethical decisions,” *arXiv preprint arXiv:2601.09724*, 2026, doi: 10.48550/arXiv.2601.09724.
- [11] T. van Nuenen and P. S. Sachdeva, “The fragility of moral judgment in large language models,” arXiv preprint arXiv:2603.05651, 2026, doi: 10.48550/arXiv.2603.05651.
- [12] A. Bedoui, A. L. Greene, and M. Cherkaoui, “Auditing framing-sensitive behavioral instability in large language models for mental health interactions,” arXiv preprint arXiv:2606.26982, 2026.

[13] V. Tagliabue and L. Dung, “Probing the preferences of a language model: Integrating verbal and behavioral tests of AI welfare,” arXiv preprint arXiv:2509.07961, 2025, doi: 10.48550/arXiv.2509.07961.

LLM Usage Statement

I used Claude in gathering sources and co-developing the code. All the results were produced by the code in the repository and checked.