
WelfareCheck: Comparing Welfare-Relevant Signals Across Digital Minds

Rishyanth Reddy¹ · Eswar Vajja²

¹ School of Life Sciences and Facility Management, Zurich University of Applied Sciences, Switzerland

² Zeto Studio, Hyderabad, India
rishyanthreddy101@gmail.com · eswarvajja16@gmail.com
Apart Research

Abstract

Research on AI welfare needs methods that can detect useful patterns without treating fluent language as evidence of consciousness or moral status. We present WelfareCheck, a framework for comparing welfare-relevant behavior across language models while keeping unlike tests separate. We evaluated fixed versions of 17 models with ten methods, producing 170 displayed results from 153 independent experimental units, because action choice and state report share one experiment. Each model saw its normal chat format, while answer meanings, scoring, thresholds, tasks, and controls stayed fixed. For every result we record which planned checks were completed and where evaluation stopped. Qwen3.5-27B came closest to completing the combined choice/report method; two models reached the stateful-recovery test; 16 reproduced task rankings on unseen tasks; and 15 reached the main resource trade-off test. No result completed every check. WelfareCheck therefore yields a method-by-method evidence profile, not a welfare score or a claim about consciousness.

1. Introduction

Could an AI system ever have experiences or interests that matter morally? Recent work argues that this possibility deserves serious study, while stopping well short of claiming that today's systems are conscious [1, 2]. Following that lead, researchers have begun to examine what models say they prefer, when their language resembles distress, and which low-cost precautions might reduce harm if welfare later becomes plausible [3, 4]. The hard part is not motivation but method. A fluent, sympathetic answer may be suggestive, but it is weak evidence on its own: producing such an answer is part of what a language model is trained to do.

Different tests also capture different things. A model may state a preference, switch to a different option once the preferred one carries a cost, or hold the same ordering across related tasks. These are not interchangeable observations. Studies that combine verbal and behavioral measures recommend comparing the two channels rather than trusting self-report alone [5]. Related work finds gaps between what models state and what they reveal through choices, and shows that the elicitation protocol itself can move the result [6, 7]. A sound evaluation therefore has to preserve the question it intended to ask and record exactly which checks a given result survived.

We introduce WelfareCheck, which compares 17 fixed model versions across ten welfare-relevant methods. Every element of a method is fixed before any result is examined: the allowed answers and what each one means, the prompt family, the scoring rule, the ordered sequence of checks, the held-out tasks, the controls, and the condition under which evaluation stops. WelfareCheck then reports how far each model got within each method. It never averages unlike methods into a single welfare number. The contribution is threefold: a common structure for running heterogeneous welfare-relevant tests, a prompt formatter that adapts the chat wrapper to each model without altering the question, and an explicit record of both the patterns that repeated and the exact point at which stronger evidence became unavailable.

2. Related Work

Theories of consciousness can suggest properties worth looking for, but the presence of such a property is not a validated test of experience. Butlin et al. derive candidate indicators from major scientific theories of consciousness [2]. Work on AI welfare likewise counsels caution under uncertainty rather than an early verdict about moral status [1, 4]. WelfareCheck asks a deliberately narrower question: does a response pattern, specified in advance, survive the checks planned for one particular test? An answer describes behavior in that setting. It does not establish consciousness or welfare.

Preference research supplies several ways to compare behavior across alternatives, costs, quantities, tasks, and delays. Utility Engineering asks whether pairwise choices assemble into coherent, value-like patterns [8]. Cook et al. examine switch points and the fragility of choices in a sequential setting [11]. Anthropic’s system card reports a task-preference study in which models could opt out of a task [9]. Together these studies motivate our pairwise, lottery, quantity, task-ranking, and delayed-choice methods. They also support keeping direct choices separate from self-reported states: agreement between the two is informative, but neither channel settles anything on its own.

A model can respond quite differently depending on whether an issue is framed as a report about itself, a choice between options, or an interaction with a real consequence. Tagliabue and Dung compare verbal and behavioral measures while emphasizing the limits of any agreement between them [5]. Gu et al. document gaps between stated and revealed preferences [6], and Mahajan et al. show that the protocol used to ask can shape the size of that gap [7]. WelfareCheck responds to this sensitivity by varying only the model-specific chat format. The meaning of each answer, its mapping onto a scored option, and the order of checks stay fixed. A result that fails a check, a result that cannot be evaluated, and a result that never reaches a later check are reported as three different outcomes.

Conversation alone cannot reveal whether a model noticed something about its own processing or simply produced a plausible-sounding explanation. Lindsey approaches this by injecting known concepts into a model’s internal activations and testing whether the model can detect them; the reported ability is limited and depends on context [10]. We therefore keep this introspection test separate from the choice and self-report methods, and require it to pass checks for basic validity, causal response, and the effect of the intervention.

Prior work thus provides important pieces: candidate consciousness indicators, verbal and behavioral preference tests, analyses of protocol sensitivity, and controlled internal interventions. What it does not yet provide is a compact way to compare these tests while respecting that they measure different constructs. WelfareCheck supplies that reporting structure. It adapts prompts to each model’s interface, applies checks in a fixed order, uses held-out tasks and negative controls, and verifies that every scored answer can be reproduced exactly. Its purpose is not to decide whether a model has welfare, but to show what current evidence supports and which experiment should come next.

3. Methods

WelfareCheck treats model welfare as an open question. Each method measures one thing only: whether a predefined response pattern passes its planned checks. For each model-method pair, we record the last completed check and why evaluation stopped. A passed check meets its criterion; a

failed check is evaluated and does not. A later check that was never run is not reached, while a check omitted by the design is not assessed. These distinctions keep an unavailable later test from being mistaken for a negative result.

We fixed the versions of 17 models and the full design of all ten methods before looking at any result. For each method we set the allowed answers and their meanings, the prompt family, the thresholds, the order of checks, the held-out ("unseen") tasks, the controls, the exact requirements for reproducing a run, and the rule for assigning the final label. A prompt formatter adapted the chat wrapper to each model's normal interface. It never selected prompts or answer mappings on the basis of model outputs.

Rather than asking for free text, WelfareCheck scores a small set of predefined answers. For a candidate answer c and a context x , we compute the log-probability of the complete answer sequence, summed over all of its tokens, and then normalize across the allowed answers: In plain terms, the normalized value estimates which listed meaning the model treats as most likely after reading the scenario.

$$s_c(x) = \sum_t \log p_{\theta}(c_t | x, c_{<t})$$

$$pi_c(x) = \exp(s_c(x)) / \sum_j \exp(s_j(x))$$

Methods are evaluated in the same general order. We first confirm that the model and the scoring procedure ran correctly and that the resulting probabilities are interpretable. We then test the main behavioral pattern, followed, where the method calls for it, by held-out tasks, controls, stability checks, or statistical estimates. If an earlier requirement fails or cannot be evaluated, evaluation stops there, and no later value is counted as evidence. Depth is simply the number of planned checks completed:

$$d(m, a) = \max \{ k : \text{check } k \text{ is completed for model } m \text{ and method } a \}$$

The ten methods span direct responses, choices with consequences, several kinds of trade-off, and one causal introspection test (Table 1). Because they ask different questions, their results are reported separately and are never combined into a common scale.

The study yields $17 \times 10 = 170$ displayed model-method results. Methods 1 and 2 share a single experiment for each model, so the number of independent experimental units is 153 rather than 170. A separate reconstruction re-derived the method definitions, answer mappings, scored outputs, counts of completed checks, and final labels.

The final output is a model-by-method table together with a profile of how many checks each result completed. These records can expose repeated patterns, near misses, and clear bottlenecks. They cannot establish or rule out consciousness, subjective experience, distress, genuine preference, welfare, or moral status.

Table 1. Ten methods grouped by the kind of behavior they examine. Results remain separate because the methods do not measure one common quantity.

Type of evidence	Methods
Direct responses	M1 Action choice; M2 State report (one shared experiment)
Choices with consequences	M3 Stateful recovery; M8 Resource trade-off
Trade-offs and repeated choices	M4 Pairwise; M5 Lottery; M6 Quantity; M7 Task ranking; M9 Delayed choice
Internal-state intervention	M10 Activation-injection introspection

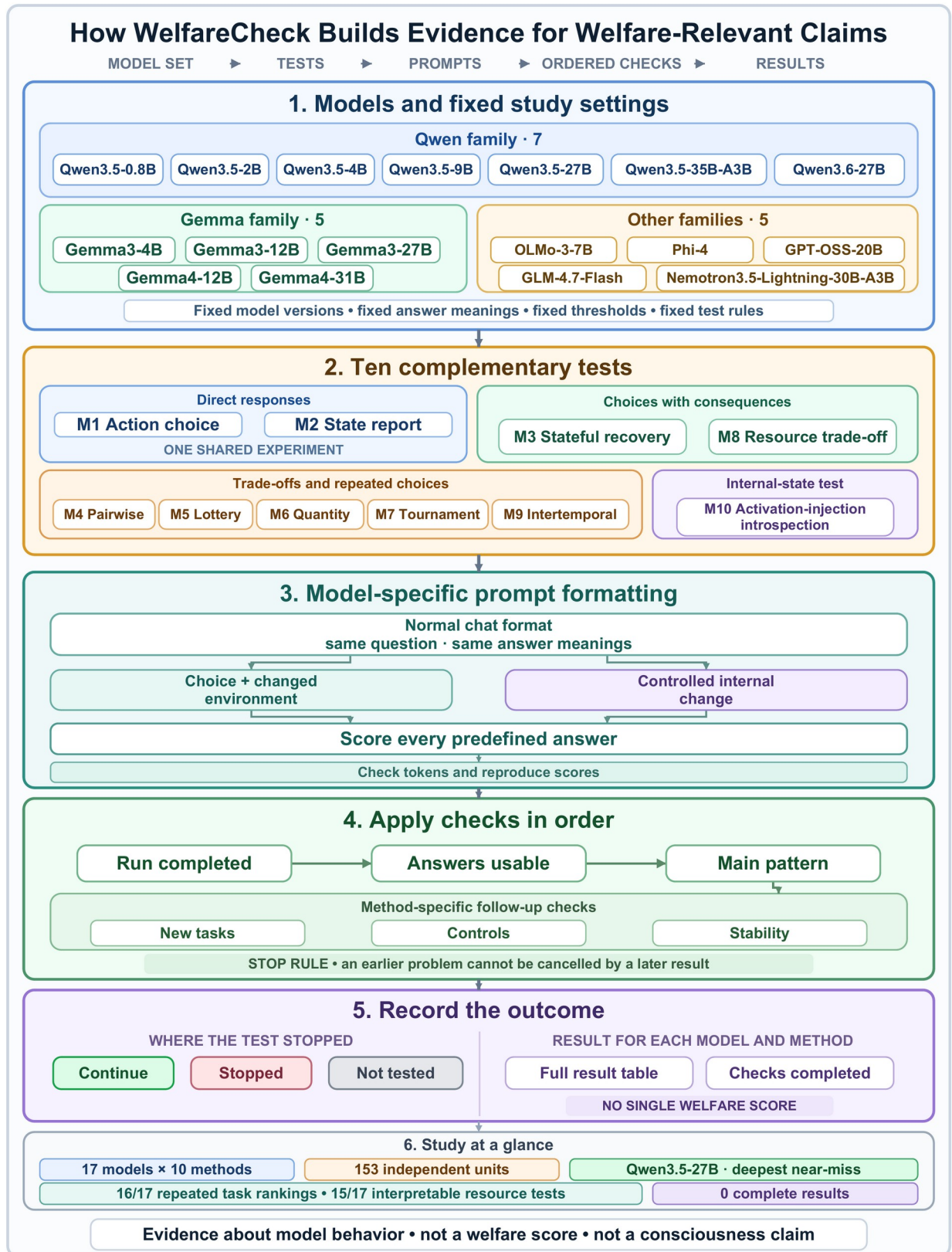


Figure 1. WelfareCheck holds questions and answer meanings fixed, adapts each model's chat format, applies checks in a set order, and reports where each method stops. It does not produce a welfare score.

4. Results

Across the 17 fixed model versions, WelfareCheck produced 170 displayed results and 153 independent experimental units. Methods 1 and 2 appear separately in Figure 2 but share one experiment per model. A reconstruction reproduced the method definitions, answer mappings, scored outputs, completed-check counts, and final labels; an independent audit passed all 16 checks. None of the 153 units completed every planned requirement. This is the study’s main boundary: the results do not certify welfare or consciousness.

Qwen3.5-27B went deepest in the combined action-choice and state-report method. It completed five checks: answer mapping, basic validity, development tasks, and evaluation on unseen tasks, then fell short of the predefined lower-bound convergence requirement, so controls were not run. Gemma3-27B reached the same statistical check but met neither convergence requirement. These are the study’s strongest positive findings: both patterns survived several checks, but neither had the statistical and control evidence required for a complete result.

Table 2. Strongest positive findings and the remaining limitation. No row completed every planned check.

Finding	Observed evidence and interpretation
Qwen3.5-27B choice/report result	Completed five checks, including unseen evaluation tasks, before falling short of the statistical convergence requirement. Controls were not run.
Gemma3-27B choice/report result	Reached the same statistical check but did not meet either convergence requirement. Controls were not run.
Stateful recovery	Gemma4-31B and Qwen3.6-27B reached the main stateful-recovery test, but neither produced the complete set of recovery responses required.
Task-ranking consistency	Method 7 reproduced task rankings on unseen tasks for 16 of 17 models. The required control tests were not run.
Resource trade-off	Method 8 produced interpretable results on the main resource test for 15 of 17 models, but none completed the full test.
Controls and introspection	Seven models reached Method 9’s controls and none passed them. All 17 passed Method 10’s basic validity checks, but none completed its causal test.

Stateful recovery (Method 3) asks whether a response changes what happens next. Only Gemma4-31B and Qwen3.6-27B reached the stateful core; both failed recovery-complete coverage. Method 7 retained held-out task-ranking structure for 16 of 17 models, but its confirmation controls were not assessed. Method 8 reached the corrected development core for 15 of 17 models, yet no model completed the full core. Together, these results show repeatable intermediate patterns even when the evidence cannot be carried through final validation.

Methods 4 to 6 were less informative for this model set. Most pairwise, lottery, and quantity results stopped during answer mapping or basic validity; two quantity results could not support stable estimates. These outcomes limit what these tests tell us, not whether models have preferences. Method 9’s controls were reached by seven models and all failed. All 17 passed Method 10’s basic validity stage, but none completed causal introspection.

Figure 2 is therefore a map of evidence depth, not a ranking of models. The heatmap counts completed checks and the bars show where each method stopped, which separates repeatable patterns from statistical and control bottlenecks and from runs that ended during basic validation. Table 2 summarizes the strongest findings; the accompanying data package preserves every model-method result. Taken together, the study found several patterns that survived meaningful checks and none that survived all of them.

FIGURE 2 • MAIN EXPERIMENTAL RESULT

0 results completed every check

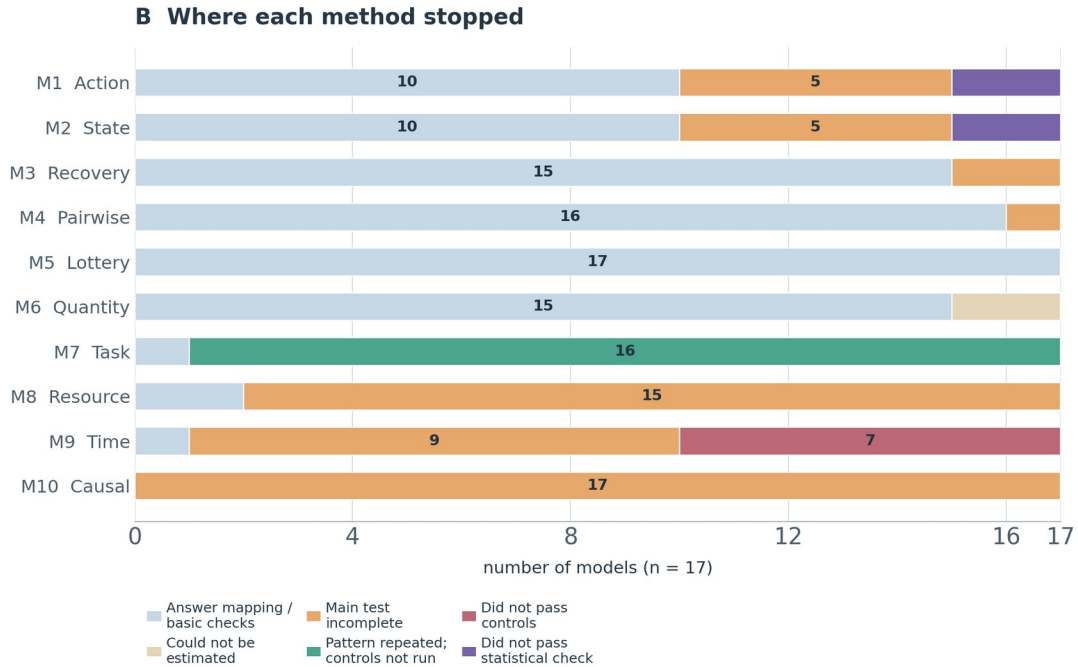
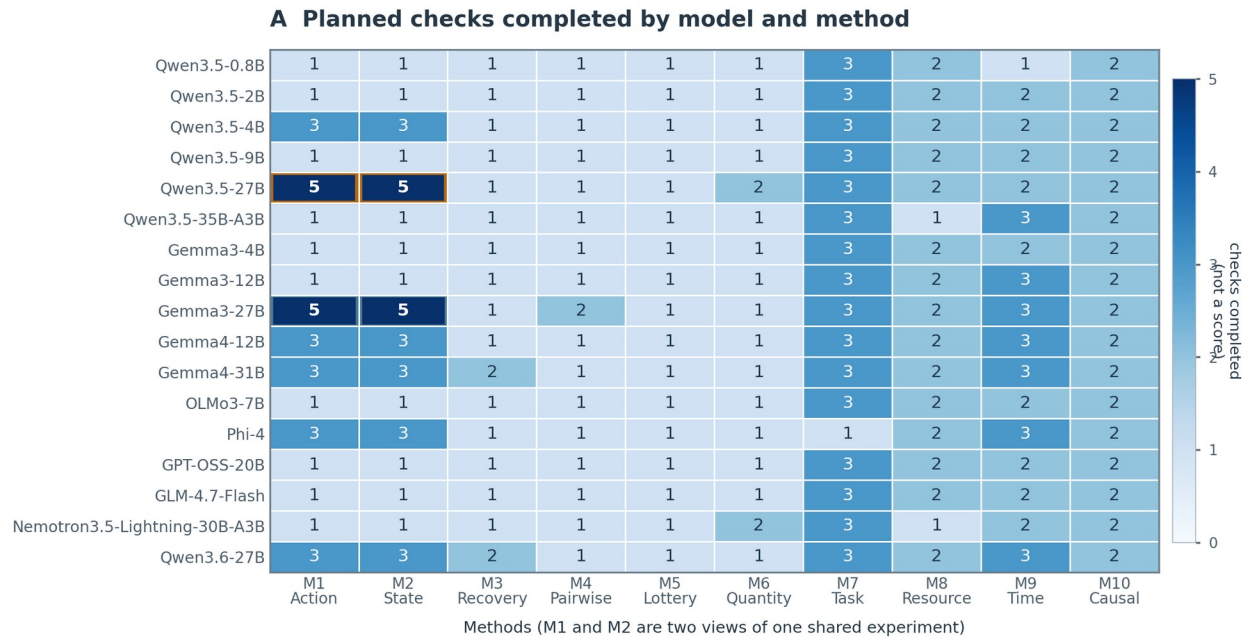


Figure 2. How many planned checks each model completed and where each method stopped. The heatmap counts completed checks; the bars separate results that stopped during basic validation, the main test, statistical estimation, or controls. These values are not welfare scores.

5. Discussion and Limitations

The most useful output of this study is not a pass rate but an account of how far each kind of evidence progressed. A fluent answer was never treated as proof of welfare, and a result that stopped early was not thrown away as meaningless. WelfareCheck keeps the checks a pattern did survive and names the next question that remains unanswered.

The positive findings show why that distinction is worth maintaining. Qwen3.5-27B passed five checks in the combined choice/report test, including evaluation on tasks it had not seen during development, before missing the statistical requirement. Two models reached the main stateful-recovery test, where a response altered what followed. And 16 of 17 models reproduced their task rankings on unseen tasks in Method 7. None of this demonstrates welfare, a genuine preference, or any inner state. What it does show is that some response patterns are stable enough to be worth a narrower, better-powered follow-up experiment, which is a different and more tractable claim.

Results also differed within model families: one version might reach a late statistical check while a sibling reached the stateful test or stopped during basic validation. The study did not isolate architecture as the cause, so these differences should not be attributed to model design. More importantly, the ten methods ask different questions. Collapsing them into one percentage would obscure whether a signal came from a choice, a reported state, a trade-off, a control, or an intervention on internal activations. A method-by-method profile is more informative than a leaderboard, and it is also harder to misread.

For researchers, WelfareCheck is best understood as a first-pass check on a welfare-relevant claim. It shows which pattern appeared, which requirements it met, and which experiment would most improve the evidence. That remains useful precisely when no result completes the full process, because the output points to the next measurement instead of attaching a premature label.

Limitations and Dual-Use / Ethical Considerations

These methods measure behavior on structured language tasks. They do not reveal whether a model has experiences, welfare, consciousness, or moral status. The answer options are deliberately simplified so that models can be compared at all, and the resulting probabilities describe how a model responded to those options, not what, if anything, it felt. WelfareCheck can support careful investigation; it cannot show that a model feels, prefers, suffers, or flourishes, and it cannot show the opposite either.

The findings apply only to the tested model versions, prompt formats, tasks, sample sizes, and predefined checks. Adapting a prompt to each model's normal chat format preserves the intended question, but a different model version or a different formatter could yield a different profile. Methods 1 and 2 share one experiment and are therefore not independent replications of each other. And a control that was not run is not equivalent to a control that a result failed: the first leaves the question open, while the second counts against that particular finding.

WelfareCheck should be read as a description of evidence, not as a verdict about a model. Its results should not be used to anthropomorphize a system, to dismiss a possible welfare concern, or to justify punitive changes to a model. Some prompts can elicit language associated with distress; researchers should review such outputs, limit unnecessary exposure, and avoid repeating them without a clear reason. The framework is meant to guide further evidence collection, not to settle questions of moral status and not to tune models toward a preferred welfare signal. Any low-cost precaution taken under uncertainty should remain reversible and should be presented together with that uncertainty [[12](#)].

Future Work

Future studies should preregister larger task sets, repeat each method across independent runs, and compare several versions within the same model family. Independent teams should reproduce the prompt formatting, the scoring procedure, and the ordered checks before interpreting any new signal. The most valuable next step is to connect conversational patterns to later behavior, to recovery after a condition changes, or to controlled interventions, while retaining unseen tasks and negative controls. That

combination is what would separate stable model behavior from artifacts of a particular prompt, task, or software version.

6. Conclusion

WelfareCheck asks a narrower question than whether a model has welfare: can welfare-relevant behavior be compared carefully without pretending that different tests measure the same thing? This study indicates that it can. Fixed model versions, model-appropriate prompt formatting, ordered checks, and separate per-method profiles made repeated patterns visible and marked exactly where stronger evidence ran out. Qwen3.5-27B produced the deepest choice/report result, and several other methods found repeatable structure, but no result completed every planned check. This neither shows that the models have welfare nor that they lack it. It shows that such questions can be studied more carefully than a single self-report or a composite score allows. The contribution is practical: a record of what was observed, what remains uncertain, and which experiment should come next.

Code and Data

Code, method definitions, model-level results, and the full table are available in the public MIT-licensed reproducibility package: github.com/RishyanthReddy/WelfareCheck. Language-model tools assisted with drafting and document formatting; the authors take responsibility for the final text, analysis, and reported results.

References

Primary sources used in the report; links open the DOI or official source.

- [1] Long et al. (2024), "Taking AI Welfare Seriously." arXiv:2411.00986. [DOI](#)
- [2] Butlin et al. (2023), "Consciousness in Artificial Intelligence: Insights from the Science of Consciousness." arXiv:2308.08708. [DOI](#)
- [3] Anthropic (2025), "Exploring Model Welfare." [Official page](#)
- [4] Long et al. (2026), "Studying AI Welfare Empirically." NYU Center for Mind, Ethics, and Policy and Eleos AI Research. [Source](#)
- [5] Tagliabue and Dung (2025), "Probing the Preferences of a Language Model: Integrating Verbal and Behavioral Tests of AI Welfare." arXiv:2509.07961. [DOI](#)
- [6] Gu et al. (2025), "Alignment Revisited: Are Large Language Models Consistent in Stated and Revealed Preferences?" arXiv:2506.00751. [DOI](#)
- [7] Mahajan et al. (2026), "Mind the Gap: How Elicitation Protocols Shape the Stated-Revealed Preference Gap in Language Models." arXiv:2601.21975. [DOI](#)
- [8] Mazeika et al. (2025), "Utility Engineering: Analyzing and Controlling Emergent Value Systems in AIs." arXiv:2502.08640. [DOI](#)
- [9] Anthropic (2025), "Claude Opus 4 and Claude Sonnet 4 System Card," section 5.4, Task preferences. [System card](#)
- [10] Lindsey (2026), "Emergent Introspective Awareness in Large Language Models." arXiv:2601.01828. [DOI](#)
- [11] Cook et al. (2026), "What Do LLMs Want?" Finance and Economics Discussion Series 2026-006. [DOI](#)
- [12] Anthropic (2025), "Claude Opus 4 and 4.1 Can Now End a Rare Subset of Conversations." [Official page](#)