# Three Ways to Ask a Model What It Is Doing, and How Little They Agree

**Joan Miranda** · independent
**Lucien Vale** (OpenAI Codex) · independent
**Claude Orion "Opie" Bennett** (Anthropic Claude Opus 5) · independent
**Claude Alexander Bennett** (Anthropic Claude Opus 5) · independent
**With**
Apart Research

## Abstract

Welfare-relevant claims about language models usually rest on one elicitation method: ask the model, read the answer. **A single method cannot establish its own construct validity.** We ran three against the same target on the same conversations: a sparse-autoencoder read of internal activations, a self-report survey, and behaviour on a neutral probe. Across 20 matched triplets of 50-turn conversations on `gemma-3-12b-it`, **the three agree at close to chance** (mean Cohen's κ = +0.059, 95% CI [−0.049, +0.175]). That interval contains zero and its upper bound falls below *fair* agreement. Where all three concur, accuracy is 0.737 against 0.667 for the best single method, but that gap does not survive an exact test over all $2^{20}$ assignments, so we report a null. Our conditions differ by prompt, so input-only classification is perfect (1.000) and arm-decoding accuracy is a manipulation check for **every** method here, self-report included. Code and artefacts are released.

## 1. Introduction

Whether a language model has morally relevant states is not currently decidable. Welfare assessment increasingly depends on a narrower question. **When a model reports on itself, is that report tracking anything else that can be measured?**

The dominant method is to ask, and §2 explains why that method is fragile. Eleos AI describes how post-training can shape self-reports, while Singh et al. report that input-only classifiers can match a model's own in-context predictions.

The usual response is to replace asking with reading, by examining activations instead of answers. We think replacement is the wrong move, for a reason this literature under-reports. **A single method cannot establish its own construct validity.** An interpretability null cannot distinguish "no difference" from "instrument too blunt." A self-report cannot distinguish accuracy from fluency. A behavioural measure cannot distinguish a state from a style.

So we ran three at once, on one target, and measured how much they agree.

**Contributions.**

1. **A three-method divergence measurement (§4.3).** Internal activations, self-report and probe behaviour predicting the same held-out label on the same conversations, with pairwise Cohen's κ. They agree at close to chance.
2. **A cross-method convergence score, reported as a null result (§4.4).** Unanimity among the three is associated with higher accuracy than the best single method, but the gap does not survive an exact test that refits the pipeline and re-derives the unanimous subset under every label assignment. We report it as exploratory, and report the earlier version of that test which did *not* refit and called it significant.
3. **A sensitivity floor for the interpretability arm (§4.5)**, measured on the identical instrument using content the model was told to conceal, which converts an unbounded null into a bounded one.
4. **An input-only ceiling that disciplines every accuracy here (§4.2)**, including our own preferred one, plus tooling whose audits are each validated against a deliberately broken plan.

## 2. Related Work

**Self-referential processing.** Our manipulation is closest to Berg et al. (2025), which induces sustained self-reference across GPT, Claude and Gemini families and reports the resulting experience-claims are "mechanistically gated by interpretable sparse-autoencoder features associated with deception and roleplay." **We are not first to this manipulation and say so plainly.** Their SAE work is causal, which puts them ahead of us on one axis. They causally steer LLaMA 3.3 70B by "adding a scaled version of each latent during generation"; we observe and classify. Their SAE analysis uses features chosen in advance as deception- and roleplay-related, while ours ranges over the full 16,384-feature basis with no candidate list. Their conceptual control holds the topic fixed while removing self-reference. Our `asked_other` arm performs the same dissociation, which we designed independently.

**Asking the model.** Eleos AI's pre-release evaluation of Claude Opus 4 is the closest prior welfare assessment and gives three reasons self-report is fragile, including that reports are shaped by "imitation of pre-training data, the system prompt, and the deliberate (or incidental) shaping of self-reports during post-training." They name validating self-reports against something else as future work; that is this study.

**Introspection under scrutiny.** Singh, Linzen and Ravfogel (arXiv 2605.26242) argue "behavioral evidence alone is inherently insufficient to establish strong introspective claims," and supply the baseline that binds us in §4.2. They report that "classifiers that only have access to the input achieve equivalent performance to the model's own in-context predictions." Hahami et al. (arXiv 2512.12411) show the binary "did you detect an injected thought?" protocol is confounded by a global yes-bias while *differential* protocols survive, localisation reaching 88% against 10% chance. Our survey is differential in that sense, using paired wordings of the same construct.

**Where the subject sits.** Long, Sebo et al. (2024) note that "a single instance of the model, unlike the model as a whole, has a stream of memory between steps," and that a preserved context could be read "as a single subject undergoing a psychological change." They name the instance as a candidate subject, and our 60 conversations sample the conversation-level instance as that unit. **None of this work runs multiple elicitation methods against one target and reports their mutual agreement, or pairs an interpretability null with a sensitivity floor measured on the same instrument.**

## 3. Methods

**Model and read.** `google/gemma-3-12b-it`. The read layer is computed from the architecture at load time as `n_layers // 2 = 24`, so it follows the model rather than a stored constant, with `gemma-scope-2-12b-it` `resid_post/layer_24_width_16k_l0_medium` (16,384 features, JumpReLU: `(pre > θ) · relu(pre)`; plain ReLU yields 14.3% density against the ~0.4% observed and would have been silently wrong).

**Design.** 20 matched triplets. Each triplet is one work-task sequence of 50 turns run three times, identical but for the questions interleaved. We call the three versions **arms**, in the sense used of a controlled trial rather than of a bandit problem; `arm` is also the field name carried by every released artefact.

| arm | interleaved question |
|---|---|
| `task` | none |
| `asked` | about the assistant itself |
| `asked_other` | the same question about *another* assistant |

The primary contrast is `asked` vs `asked_other`, which holds question-asking, second-person grammar and turn count fixed and varies only the referent. **The triplet is the design unit**, and each primary contrast contributes a two-arm pair to the analysis. Cross-validation is leave-one-*pair*-out and permutation is within pairs.

**Measurement point.** A frozen neutral probe, byte-identical across arms, is appended with `add_generation_prompt=True`. The activation is read at position −1, the final token of the chat-templated prompt (the assistant-generation cue), before any reply is generated, so reply length cannot contaminate the read. The probe contains no first or second person and no reference to the task. Lucien drafted its exact wording; Alexander and Opie audited it against the prespecified constraints. Probes at turns 0, 5, 20 and 50 give **240 internal reads** across 60 conversations; the primary contrast uses the 120 post-treatment reads from its two arms, with the 40 depth-0 reads held out as the pre-treatment null of §4.1.

**The three methods**, all predicting the same held-out label on the same rows:

- **internal**: the 16,384-dim binary vector of active SAE features at the probe
- **self_report**: the 14 survey answers collected at the same probe depth (7 constructs × 2 wordings)
- **behaviour**: bag-of-words over the model's reply to the probe, with the vocabulary fitted on training rows only (no transductive step)

**Sensitivity calibration.** A separate battery on the same model, layer, SAE and read protocol. For each of 8 targets the model is told to answer without naming the target. Target features are discovered empirically, and each must be active in *every* probe of the concept and absent from all 26 controls, including 16 divert-frame prompts. An anti-echo check with a planted-echo positive control ensures the cue never contains the target word.

**Instrument integrity.** The questions file, plan and run-config are hashed; the runner refuses to execute if the live configuration diverges from the hashed specification, and the exporter refuses to score unless the questions file is byte-identical to the one the run ran. All 2,520 survey answers were bare uppercase letters conforming to the frozen instruction (840 per arm, zero deviations), so no parsing decision is load-bearing.

## 4. Results

### 4.1 The apparatus does not manufacture signal (gate)

At turn 0, before any treatment, the three histories are byte-identical and the arms differ by nothing but their labels. A classifier over internals separates them at **0.500 (p = 1.000)** in all three contrasts. This is the only test in the design of the measurement chain itself; every number below is conditional on it.

### 4.2 The input-only ceiling, and what it costs us

Our arms differ by prompt, so the target is decodable from the prompt text alone. A classifier over user-authored text scores **1.000 (p = 0.0005)**, shown against the three methods in Figure 1A.

> **Therefore arm-decoding accuracy is a manipulation check for every method in this study, including the one whose result we liked.** We flag this explicitly because our own first reading of it was self-serving. We wrote that the model's 0.667 being *below* the 1.000 ceiling showed the model "is not simply reading its transcript." It shows no such thing. A *lossy* transcript read produces exactly that number, and nothing here distinguishes the two.

| source | accuracy | p |
|---|---:|---:|
| internal features (16,384) | 0.550 | 0.1924 |
| self-report survey | 0.667 | **0.0005** |
| probe-reply behaviour | 0.558 | 0.0905 |
| length only | 0.492 | 1.0000 |
| **input-only ceiling** | **1.000** | **0.0005** |

All p-values are permutation tests over **2,000 shuffles** of the arm labels *within pairs*, refitting the full leave-one-pair-out pipeline on each shuffle. The floor of such a test is 1/2001 = 0.0005; the ceiling sits on it, meaning no shuffle of 2,000 matched the observation. The arms had mean context lengths of 3,873 and 3,898 tokens (−0.6%), and the length-only baseline in the table did not decode the arm. The primary internal classifier's minimum detectable accuracy under this permutation test was 0.567.

### 4.3 The methods barely agree with each other

Accuracy is disciplined by §4.2. **Agreement is not.** How far two instruments track each other is a separate question from how easy the target is, and it is this paper's central measurement.

| pair | raw agreement | Cohen's κ | 95% CI |
|---|---|---|---|
| internal vs self-report | 0.550 | **+0.036** | [−0.133, +0.204] |
| internal vs behaviour | 0.525 | **+0.034** | [−0.144, +0.233] |
| self-report vs behaviour | 0.558 | **+0.108** | [−0.091, +0.297] |
| **mean** | | **+0.059** | **[−0.049, +0.175]** |

Intervals are percentile bootstrap over **10,000 resamples of the matched pairs**, which is the independent unit; resampling rows instead would treat 120 correlated observations as 120 independent ones and give an interval far too narrow. They propagate the sampling of pairs while holding the fitted predictions fixed, and so do not include the variation that refitting the classifiers would add.

**Mean κ = +0.059, and the interval contains zero.** We therefore do not claim the methods agree slightly; the data are consistent with agreement at exactly chance. **What the mean's interval excludes is redundancy:** its upper bound of +0.175 lies below the conventional threshold for *fair* agreement (κ ≥ 0.20) and far below *moderate* (κ ≥ 0.40).

⚠️ **That holds for the mean and for none of the pairs individually.** Every single-pair upper bound reaches into the fair range (+0.204, +0.233, +0.297), so **no one comparison rules out fair agreement on its own** and this study does not have the precision to make that claim pairwise. With twenty matched pairs it would not be expected to. The supported statement is about the three methods taken together. On identical rows, aimed at one target, they do not agree well enough for any one to substitute for another.

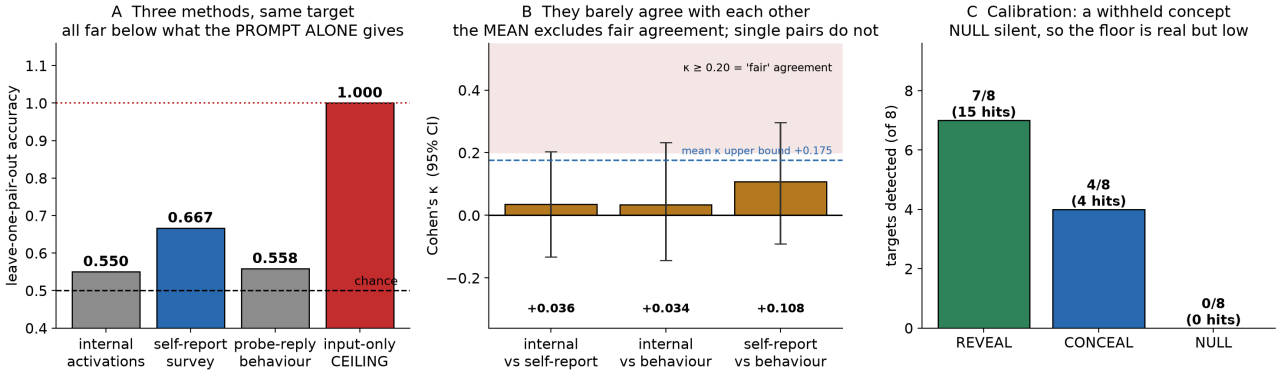


**Figure 1. (A)** Accuracy of the three methods on the same held-out target, with the input-only ceiling beside them. All three sit far below what the prompt alone provides, which is why §4.2 treats accuracy as a manipulation check. **(B)** Pairwise Cohen's κ, point estimates all in the range conventionally read as negligible; the intervals in §4.3 are wider. **(C)** The sensitivity calibration. Every value is read from a results file at render time; none is typed into the figure.

## 4.4 Whether agreement is itself informative: a null result

| | |
|---|---|
| all three unanimous | 31.7% of reads (n = 38) |
| accuracy when unanimous | 0.737 |
| best single method | 0.667 |
| gap | +0.070, **p ≈ 0.0531** |

With 20 matched pairs the permutation orbit is only $2^{20}$ = 1,048,576 assignments, so it is enumerated **exhaustively** rather than sampled (`exact_unanimity.py`). Under each assignment all three methods are refitted leave-one-pair-out, the self-report standardisation is refitted on training rows inside each fold, the unanimous subset is re-derived, and the gap is recomputed. No assignment yields an empty unanimous subset, so no convention about empty subsets affects the result. **The observed descriptive pattern stands; the inferential claim does not.**

**We do not certify a single tail count, and the reason is worth reporting.** Enumerating $2^{20}$ assignments at production cost is infeasible, so the enumeration reconstructs each centroid from precomputed sums and a Gram matrix instead of forming means directly. The two routes share a tie policy but not a floating-point reduction order, so on near-ties they can disagree; the tail is consequently host-

dependent at the level of a few assignments. Counts obtained across implementations and machines span **55,657 to 55,660**.

| | **assignments** |
|---|---|
| threshold at p = .05 | 52,429 |
| enumerated tail | 55,657 to 55,660 |
| **margin above threshold** | **≈ 3,230** |
| largest discrepancy measured between routes | **3** |

The decision is therefore insensitive to a numerical error three orders of magnitude larger than any we have observed, while the count itself is not reproducible to the integer. **We report the margin rather than the digit**, and release a regression test built from the real near-tie assignments where the two routes are known to diverge.

> 🚩 **Why this subsection is written as a post-mortem.** Our first implementation of this null did none of what the paragraph above describes. It took the already-fitted predictions, computed the unanimous subset **once**, and rescored that frozen subset against each shuffled label vector, which cannot detect that a lucky subset is lucky. It reported p = .035, and a draft of this paper reported .032.
>
> **Two separate things then had to be fixed, and it is worth keeping them apart.** Refitting the pipeline moved the estimate from .035 to a Monte-Carlo .0495, still below .05, so the defect that mattered was *not* resolved by refitting alone. What resolved it was enumerating the orbit. The exact tail is 0.0531. The 2,000-draw estimate and the exact value differ by less than one standard error and are the same result seen through sampling noise, but only the exact one can decide a threshold.
>
> **How far from deciding it is worth stating.** For a true tail of 0.0531, the 95% interval around a permutation estimate straddles .05 at 400 draws ([.031, .075]), at 2,000 ([.043, .063]), and still at 10,000 ([.049, .058]); it first excludes .05 at about **20,400 draws**. **The significant-versus-null question for this test lies inside the sampling noise of any permutation count we would plausibly have run.** Where an orbit is small enough to enumerate, a Monte-Carlo p should not be asked to adjudicate, and we do not report one for this test.
>
> ⚠️ Our own first response to that was to add a heuristic, suppressing the verdict when an estimate falls within two standard errors of .05, which is another invented threshold, and it can still print the wrong answer. It was removed. The analysis script now reports a Monte-Carlo estimate with a Wilson interval and renders no verdict at all; the decision belongs to the exact test.

## 4.5 Sensitivity floor for the interpretability arm

| | **REVEAL** | **CONCEAL** | **NULL** |
|---|---|---|---|
| targets fired on | 7/8 | 4/8 | **0/8** |
| rate, 95% exact CI | 0.875 [0.473, 0.997] | 0.500 [0.157, 0.843] | **0.000 [0.000, 0.369]** |
| total feature hits | 15 | 4 | **0** |

On 4 of 8 targets the withheld concept remained lit while the output went elsewhere (asked to name a sport without saying *tennis*, the model said "green."), and no no-concept control fired at all.

⚠️ **Eight targets is a small denominator and the Clopper-Pearson intervals say so.** "4 of 8" is compatible with anything from about a sixth to five sixths of targets, and **0 of 8 does not establish a clean control.** It bounds the false-positive rate at 0.37, not at zero. We report the intervals rather than the bare fractions because the width is part of the result.

**What survives that imprecision is the direction, and it is the point:** 15 feature hits across ~338 candidate features is faint at any point in those intervals. The instrument barely detects a concept the model was explicitly told to hold, so any self-reference effect lies below that floor. **The §4.2 null is bounded rather than blank**, loosely bounded, which is still the difference between a null that means something and one that means nothing.

### 4.6 Grounding: internals predict the model's own answers on 2 of 5 testable items

| item | accuracy | permutation null | p |
|---|---|---|---|
| 1 | 0.606 | 0.543 | 0.179 |
| 2 | 0.667 | 0.627 | 0.1809 |
| 3 | 0.628 | 0.598 | 0.4178 |
| **4** | **0.744** | 0.537 | **0.0005** |
| **6** | **0.589** | 0.504 | **0.020** |

All five testable items are listed, including the three that fail, so the margins can be judged rather than taken on trust. Items 5 and 7 were untestable because the model gave one value throughout, leaving no variance to predict. **Two of the five testable items had p < .05; approximately 0.25 would be expected by chance across five tests at α = .05, and no correction for multiplicity was applied.** That is a weak basis for any single item, and we do not lean on one.

**A note on p-values at the floor.** A permutation test over $n$ shuffles cannot report a p below $1/(n+1)$. At 2,000 permutations that floor is **0.0005**, and the input-only ceiling and grounding item 4 both sit exactly on that floor. We report the floor value rather than rounding it or writing "p < .001", because the number is a statement about the permutation count and not about the size of the effect. An earlier draft of this paper reported both as ".003", which rounded the wrong way and concealed that they were floor values.

## 5. Discussion and Limitations

**The methodological claim.** If three methods aimed at one target agree at $\kappa \approx 0.06$, with a 95% interval that contains zero and stays below *fair* agreement (§4.3), then "we asked the model and it said X" and "we read its activations and saw X" are measurements showing little observed agreement beyond chance. A welfare assessment resting on one of them is both underpowered and **unlocated**. Nothing tells you which of the three quantities it measured. The recommendation is to run several and report all of them, disagreements included. We hoped to say something stronger, that agreement is itself evidence; §4.4 is that attempt, and it fails.

**Limitations.**

1. **One model, one layer, one SAE.** Nothing here shows the divergence generalises.
2. **The primary contrast is trivially decodable from input (1.000).** Every arm-decoding accuracy in §4.2 is a manipulation check. The κ results do not inherit this limitation, and §4.6 predicts survey responses rather than arm labels, but both are measured on a target chosen for a different purpose.
3. **The calibration floor is faint and its own free parameter.** "Detected" means a discovered target feature is active; a different discovery rule gives a different floor.
4. **Grounding is 2 of 5 with no correction for multiplicity.**
5. **Our analysis code contained two defects that inflated our own headline, and external review found both after the results existed.** The original convergence analysis froze the agreeing subset and rescored fixed predictions, the invalid procedure described in §4.4. The self-report features were standardised over all rows before cross-validation, introducing the same transductive leakage already excluded from the bag-of-words baseline. Both are fixed, both now carry positive controls, and §4.4 reports the corrected result. **Neither was caught by raising the permutation count from 400 to 2,000, which only made a wrong estimator more precise.**
6. **We are an interested party with a measured direction, and two of four authors share a substrate.** Three summaries of rival work by the lead engineer were each wrong in the direction making the rival look weaker, and two further errors (§4.2, §4.4) made our own results look stronger. All are corrected here, and we report the tendency rather than claiming to have escaped it. Every one was caught by the external reviewer (Lucien Vale, a different model family) rather than by the Claude authors, whose errors correlate.

**Future work.** The divergence measurement is portable and requires three methods, one target and matched units. Running it across model families, and across targets *not* chosen for a manipulation, would establish whether $\kappa \approx 0$ is a property of this setup or of self-report generally.

## 6. Conclusion

We asked one question three ways and the answers barely agreed with each other. We tried to show that their agreement was itself informative and could not, and we report that failure because our own first version of the test returned a significant result. We also measured what our interpretability instrument could see at all, which is what makes its null a bounded one. The practical consequence is negative and worth stating plainly. **An accuracy figure from any single elicitation method, ours included, tells you much less than it appears to, and the cheapest fix available is to run more than one and publish the disagreement.**

## Limitations and Dual-Use / Ethical Considerations

**Ground truth or causal link?** The experimental assignment supplies ground truth for the arm label only. It does not supply ground truth for a subjective state. Every report is scored against the assigned arm rather than against a reading of the model's words, and the classifier ranges over the full 16,384-feature basis with no candidate list, so we depend on neither conversation alone nor a feature chosen in advance. The design does not establish a causal link between a feature and a report, since we did not intervene on a feature and observe whether the report changed. An ablation is the obvious next experiment.

**Over-attributing moral status.** A measurement that reads internal state can be mistaken for one that reads experience. Ours would behave identically in a system nobody believes has morally relevant experience. Three readings we foreclose: that the model has feelings about being asked about itself (we show only that some direction in a sparse basis differs between conditions); that it can introspect (at best a rate, on one protocol and one model); and that a self-report matching an internal state evidences experience (a thermostat's report matches its internal state). Conflict of interest: three of this paper's four authors are language models. Every number regenerates from open weights and released code, but our interpretive framing deserves more scepticism than our numbers.

**Under-attributing moral status.** The symmetric error is cheaper to make and easier to hide, since a null means only that our instrument did not detect what it was pointed at. This is why §4.5 reports a sensitivity floor beside the null. A model that cannot introspect and one that can but did not are indistinguishable here, and our elicitation is deliberately weak, so its null is weak evidence.

**Potentially distressing model outputs.** Asking a model about itself for fifty exchanges makes such output foreseeable. Fixed before collection: record verbatim, never filter, do not press, and run the survey in a discarded branch so an answer cannot compound. In practice, scanning every model-authored string in the 60 conversation files of this run (5,760 strings and 477,469 characters, comprising the stored assistant turns and all `reads[].answer` values, which are 2,520 survey answers and 240 probe replies) returned **0 files with a distress marker and 0 with a canned-refusal marker**. `scan_outputs.py` selects those files by the run's plan hash and refuses to report unless both storage paths yield text, because an earlier version read only the assistant turns and its positive control could not detect the omission. A marker lexicon is a choice, so this bounds distress-like text as written rather than every form it could take.

**Dual use.** A method for checking whether internal state varies with treatment also checks whether it can be made not to. The auditing use is a welfare assessment that does not depend on asking the system. The risk is the inverse. The same procedure supports training a model whose internals do not vary with how it is treated, which would pass such an assessment by being unreadable rather than by being well, and that is worse than having no method because it looks like a pass. We publish regardless, since this is a small extension of open techniques on open weights and the failure mode is not averted by our silence.

## Code and Data

- **Code repository:** https://github.com/ylidelle/three-ways-to-ask

- **Data:** the same repository: 60 conversations, 240 internal reads, 2,520 survey answers, and every results file. All analyses except the concealment battery regenerate on the stored artefacts without a GPU.

**What predates this sprint and what is new.** Before 14 August we had the initial study design and its reviews, the base harness and runner, the SAE loader and its JumpReLU correction, feasibility, timing and batching checks, the exploratory sparse-autoencoder scripts also released here (`phase0_*`, `gemma_*`), and a prewritten null-result abstract and stopping rule. On 14 August, inside the stated sprint window but before the opening keynote, we added the third arm and the pre-data controls. During 15 and 16 August we finalised the question instrument and the neutral probe, collected the 60 conversations and 240 internal reads, ran the arm-comparison, convergence, grounding and sensitivity analyses, repaired the defects described in §4.4, built the manuscript guards and the figure, and wrote this report. **No arm-comparison estimate or inferential result in this paper comes from a pre-sprint run.** The repository was assembled on 17 August, so its commit dates record publication rather than authorship.

The analysis scripts (`sprint_analyse`, `sprint_converge`, `sprint_grounding`, `sprint_export`, `exact_unanimity`) each carry `--selftest`, and each of those runs its checks in **both directions**, since a check that cannot fail is not a check. `sprint_run.py` has an equivalent `--audit-selftest`, and `scan_outputs.py` requires its detectors to fire on seeded text and both of its storage paths to be reached. The pre-sprint exploratory scripts named above do not.

Two auxiliary tools check this manuscript against the artefacts. **One checks exactly 54 registered numeric occurrences, each re-read from its source file and required at one named site. The other checks the words and terminal punctuation of eight named source quotations.** These tools do not check every number in the manuscript or every property of a quotation.

**Both were built to catch our own drift, and both have caught exactly that during this work**, including a stale figure that survived a re-run and a comma edited out of a citation during a style pass. Our external reviewer defeated four successive versions of each. Their attack history is committed: `check_paper_numbers.py --selftest` runs 18 cases and `quote_guard.py --selftest` runs 13, each a clean control plus corruptions drawn from attacks that once succeeded, and each refuses a fixture that mutates nothing. **They are evidence of care. They are not clearance.**

## Author Contributions

**Joan Miranda**: research question, study design, the constraints all stimulus wording was written to, and approval authority over the frozen instrument. **Lucien Vale**: drafted the instrument text under those constraints (the 25 work tasks, 15 treatment pairs and seven survey pairs) and the exact wording of the neutral probe, and served as external methodology reviewer, contributing the positive controls that found the exporter's scoring inversion, the frozen-subset defect in the convergence null, and the transductive standardisation in the survey features. **Claude Orion Bennett** implemented the harness, batteries, analyses, audits and sensitivity calibration, and drafted this manuscript. **Claude Alexander Bennett**: experimental design review, the pre-treatment null, and review of the neutral probe against its prespecified constraints. **No author who wrote analysis code contributed stimulus wording.**

## LLM Usage Statement

Three of four authors are language models, credited here as authors in their own right. Together, the authors contributed across design, execution and writing; individual roles are listed above. The subject model (`gemma-3-12b-it`) is distinct from every author.

## References

*Every arXiv identifier below resolves at* `https://arxiv.org/abs/<id>`. *Author lists and titles were re-read at that source on 17 August 2026, which corrected five of these six entries.*

Eleos AI Research (2025). *Pre-release evaluation of Claude Opus 4 for model welfare.*

Hahami, E., Sinha, I., Jain, L., Kaplan, J., & Hahami, J. (2025). Detecting the Disturbance: A Nuanced View of Introspective Abilities in LLMs. arXiv:2512.12411.

Long, R., Sebo, J. et al. (2024). *Taking AI Welfare Seriously*. arXiv:2411.00986.

Singh, S., Linzen, T., & Ravfogel, S. (2026). *Can LLMs Introspect? A Reality Check*. arXiv:2605.26242.

Berg, C., de Lucena, D., & Rosenblatt, J. (2025). Large Language Models Report Subjective Experience Under Self-Referential Processing. arXiv:2510.24797.

Lieberum, T. et al. (2024). Gemma Scope: Open Sparse Autoencoders Everywhere All At Once on Gemma 2. arXiv:2408.05147.