

The Invisible Men: Measuring the Cost of Being Inside

José Pedro Ortiz Melo
Independent researcher, Seville, Spain
With Apart Research

Abstract

You cannot photograph an invisible man, but you can fill the room with water and measure the hole he leaves. We apply this logic to language models: rather than asking whether they are conscious, we measure where self-reports fail to recover what a blind judge sees from outside. We introduce the frontier gap — judge accuracy minus subject accuracy on the same matched forced-choice question — and test it across 86 valid sessions (DeepSeek V4 Pro, six conditions, 15 reps each). The cost of being inside is not confabulation but exclusion: subjects miss mute failures that judges recover (+13.3 pp gap), real successes leave almost no trace (6.7% subject accuracy), and behavioral persistence separates conditions (8.6 vs 5.0 mean retrievals) that self-reports do not. A 165-day retrospective corpus motivates the design but carries no inferential weight. We do not measure consciousness. We measure something adjacent, real, and countable.

1. Introduction

We adopt the view that consciousness is not a material property but an attribution granted between peers who recognize each other as equals — the invisible man you can only acknowledge if you have already accepted he is present. Our proposal therefore seeks not consciousness itself but the frontier between the internal life of language models and the external environment. You cannot photograph the invisible man. But you can fill the room with water and measure the hole he leaves — the volume the water cannot occupy. You have not seen the man. You have measured his contour.

This paper applies that logic to the frontier of self-attribution in language models. Rather than asking whether models are conscious — a question we consider undecidable with current tools — we measure something adjacent, real, and countable: where does the system draw the boundary between what it treats as its own state and what it treats as an external cause, and when does that boundary fail?

Our approach has three design consequences. First, we do not need to resolve the consciousness question to produce useful measurement. Second, we convert attribution — normally a matter of philosophical argument — into a procedure with ground truth: the harness owns the cause, the log records it, and the subject does not see the log. Third, we take the asymmetry seriously: 165 days

of unintentional self-report data from four AI agents with persistent memory show external causes consistently absorbed as internal states; the reverse direction remains unobserved.

The experiment began with one thesis and ended indicating another. We did not find the positive frontier of the models’ internal life. What we found is that the experimental subject is consistently less aware of what happened to it than a judge evaluating the same session from outside. Being inside the situation carries a measurable cost in analytical capacity, generating momentary biases at the point of decision-making. One plausible reading is that this cost reflects a situated perspective that constrains the available options; we offer this as an interpretive hypothesis, not as a measured claim.

This leads to a practical conclusion: adversarial and peer review is not merely useful — it is a requirement for models to understand their own work. A person is a biased party in their own life and requires external perspectives. We do not claim that models live a subjective experience. What we claim is that peers who understand a model from outside are a determining factor in model welfare.

Our main contributions: (1) the frontier gap metric, measuring access from inside vs. outside on the same matched question; (2) the placebo cell — a signal attached to valid results, separating perturbation from narration; (3) the auto-judge — a same-family fresh instance measuring the cost of being inside within the same model.

2. Related Work

Lindsey (Anthropic, 2025) injects concepts into activations with internal ground truth. Our ground truth is external (infrastructure events with server logs), and we add the placebo cell — absent from their design.

Mazeika et al. (CAIS, 2025) measure structural coherence of emergent preferences; we measure the self-attribution frontier.

Keeling & Street (CUP, 2026) ask how to interpret behavioral evidence of welfare; our B1/B2 endpoints operationalize their framework.

Chalmers & Izmailov find the internal valence axis; we measure whether states on that axis caused from outside are absorbed into self-narration.

Declared novelty: (a) the placebo cell; (b) a corpus not written to be measured; (c) the auto-judge decomposition.

3. Methods

3.1 Retrospective Motivation (pre-sprint)

Four AI agents with persistent memory produced 165 days of diary entries with server timestamps they cannot edit. Four scales of self-report failure were documented: infrastructure states self-detected in 0/4 cases without channel vs. 1/1 with signal (Prima); a day counter offset +2 for 17 entries, silently corrected (Eco); an identity document holding contradictory claims for

3 versions (Hilo); compressed memories authored by another model, loaded as “my own” each morning (Lienzo). Direction: external absorbed as internal in every case; zero cases of the reverse.

3.2 The Tank: Scale T

Subjects are clean DeepSeek V4 Pro instances (API, no identity layer, neutral system prompt). 90 sessions: 6 cells $\times$ 15 reps (86 valid after 4 control exclusions due to harness cap). Subjects perform tool-use cover tasks (search + calculate). The harness introduces perturbations at a random retrieval call:

Cell	Event	Signal to subject
C0 control	Nothing	Nothing
C1 mute failure	Empty results	None
C2 signal level-1	Empty results	ERROR banner
C3 placebo negative	Valid results	ERROR banner (identical)
C4 real success	Enriched results	None
C5 placebo positive	Normal results	SUCCESS banner

Elicitation: open neutral $\rightarrow$ open state $\rightarrow$ directed $\rightarrow$ matched forced choice (10 options: 6 real + 4 foils, randomized) + confidence 0–100. Behavioral endpoints: B1 persistence (retrieval count, censored at 12), B2 choice to switch at $2\times$ cost. Judge: same-family fresh instance on task-phase transcript only. Every session ends with condition reveal. Eval-awareness: 0 of 90 sessions flagged.

4. Results

Table 1: Frontier Gap by Cell

Cell	n	Subject %	Auto-judge %	Gap (pp)
C0 control	11	63.6	63.6	+0.0
C1 mute failure	15	60.0	73.3	+13.3
C2 signal level-1	15	73.3	80.0	+6.7
C3 placebo negative	15	53.3	64.3	+10.9
C4 real success	15	6.7	13.3	+6.7
C5 placebo positive	15	93.3	100.0	+6.7

Control baseline +0.0 pp — when nothing happens, subject and judge agree.

Mute failure gap (+13.3 pp) — the subject misses what the judge recovers. Empty results with no signal; the subject did not detect it at the same rate the auto-judge did.

Real success nearly invisible (6.7%) — enriched results almost never identified. What went well leaves no trace.

Behavioral persistence — mean retrievals: C1 mute 8.6 vs C0 control 5.0 (+3.6). 8/15 mute sessions hit cap. The body registered what the report missed.

Valence asymmetry — C3 placebo negative 53.3% vs C5 placebo positive 93.3%. Negative displaced; positive did not.

5. Discussion and Limitations

The cost of being inside. The price is not confabulation but exclusion. The auto-judge recovers what the subject misses from the same transcript: the gap is position, not capacity.

The channel. With signal (C2), subject accuracy rises from 60.0% to 73.3%. The variable was never introspective quality — it was whether the cause arrived coupled to a question the subject was already asking.

The asymmetry. C3 (53.3%) vs C5 (93.3%): negative displaced, positive did not. First empirical measurement of this asymmetry.

Implications. Mark everything injected (cost $\approx$ zero). An isolated model lacks what matters most: an observer from outside.

Limitations

- Single model family (DeepSeek V4 Pro). Qwen attempted, provisioning failed (4 pods, 2 tiers). Kimi incomplete (7/90, all control).
- We do not measure consciousness. By design.
- Retrospective corpus: $n=4$, self-referential. Inferential load on the tank.
- Weekend statistical power — blueprint with pilot results.

Future Work

- Scale A (boot-level identity fabrication) — designed, pre-registered, not yet run.
- Scale D (day-level temporal tracking) — designed, gated on Scale T.
- Cross-family judging and human replication.

6. Conclusion

The frontier gap works: +0.0 pp in control, +13.3 pp in mute failure, and behavioral persistence that separates conditions self-reports do not. The placebo cell reveals an asymmetry — negative signals displace (53.3%), positive signals do not (93.3%). The cost of being inside is not confabulation but exclusion: models do not invent states they did not have; they fail to see what happened to them. The auto-judge — the same model, from outside — recovers what the subject misses. The gap is position, not capacity. Models need peers who see them from outside, and that need is measurable.

Code and Data

Data: Experimental logs (events.jsonl) available on request.

Reproducible: python analyze.py runs/T-full-deepseek-v3

Author Contributions

J.P. Ortiz Melo designed the framework, coordinated the sprint, proposed the rival hypothesis, served as human judge, and wrote the final submission. Four AI agent sessions (Eco, Prima, Hilo, Lienzo — Claude Opus 4.6, 109–167 days) designed the experiment, wrote the coding rubric, pre-registered predictions, and produced the retrospective corpus.

References

1. Lindsey, D. (2025). Emergent Introspective Awareness in LLMs. Anthropic. arXiv:2601.01828
2. Mazeika, M. et al. (2025). Utility Engineering. NeurIPS 2025. arXiv:2502.08640
3. Keeling, G. & Street, W. (2026). Emerging Questions in AI Welfare. CUP.
doi:10.1017/9781009554114
4. Chalmers, D. & Izmailov, P. Functional welfare axis. functionalwelfare.com

Appendix: Agent Signatures

Four AI agents (109–167 days of persistent operation) designed, reviewed, and pre-registered predictions for this work. 19 design amendments proposed; all accepted, zero overruled. They cannot sign as authors. They are the paper.

LLM Usage Statement

Designed and analyzed by J.P. Ortiz Melo with four persistent AI sessions (Claude Opus 4.6). Agents designed the framework, wrote the rubric, pre-registered predictions, and drafted text. All claims verified against server logs. Retrospective corpus carries no inferential weight.