
HAVING A STATE IS NOT KNOWING IT¹

Counterfactual Introspection and a Falsification Hierarchy for Model Self-Knowledge

Amrit Gopinath*
SSN College of Engineering
Chennai, India

Raghul Sugumar*
SSN College of Engineering
Chennai, India

With
Apart Research

Abstract

Self-report is a tempting model-oversight signal, but post-hoc reports can be generated from surface cues or ordinary downstream computation rather than reliable access to internal state. We argue that a stronger target is *counterfactual introspection*: can a model predict the causal consequence of a hypothetical internal change before its action computation experiences that change? We develop a falsification hierarchy separating causal presence, persistence, native report, counterfactual effect prediction, and quantitative self-modeling. In Llama-3.2-3B-Instruct, four concept vectors causally steer behavior while natural report remains prompt- and retrieval-distance-sensitive despite surviving mechanistic traces. In a controlled Qwen benchmark, native-access training improves macro-F1 from 0.324 to 0.522, and fixed-energy distribution across two positions raises trained F1 from 0.709 to 0.915. Our two-pass counterfactual protocol keeps the action path clean, privately supplies a hypothetical residual direction to a later query, and measures its true effect in a separate causal execution. After rejecting shortcut-prone constructions, native F1 reaches 0.695 on held-out base directions but loses its advantage on unseen mixtures. Quantitative response-curve prediction then fails with negative R^2 and zero held-out flip recovery. The result is not merely another self-report benchmark: it defines a reusable research program for distinguishing internal-state presence, reportability, counterfactual access, and genuine computational self-modeling.

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

* Equal contribution.

1. Introduction

Model self-report could support oversight of uncertainty, latent objectives, or internal anomalies. But a model that narrates an event *after* it affects behavior is not necessarily able to recognize or anticipate it in a safety-relevant way: fluent reports can come from prompt cues or generic downstream features, while causally real changes can remain verbally inaccessible. Treating all such cases as “introspection” collapses distinct capabilities.

Theory of change. Self-report should earn trust only as weaker explanations are falsified: establish a ground-truth internal event; verify causal effect and downstream persistence; test matched-prompt reporting against shortcuts and external observers; require prediction of an *unexperienced* intervention; then test effect magnitude and decision-boundary consequences. Failing an earlier gate blocks the stronger claim.

Research direction. We propose *counterfactual introspection* as a stronger target than post-hoc report: can a model forecast how a hypothetical change to its own computation would alter behavior before that change is applied? This turns introspection into a causal prediction problem with independently executable ground truth. The hierarchy can be reused wherever an internal variable admits causal manipulation and measurement, making it a template for research on predictive self-monitoring rather than another verbal-report benchmark.

We operationalize five gates: *causal presence*, *persistence*, *native self-report*, *counterfactual sign prediction*, and *quantitative response modeling*. Our contributions are:

1. A controlled open-weights concept-injection replication with corrected false-positive scoring, prompt/distance tests, and post-intervention mechanistic tracing.
2. A Qwen native-witness benchmark with objective A/B-margin labels, matched prompts, rotated report tokens, fixed-energy geometry controls, calibration, and external observers.
3. A leakage-controlled *counterfactual introspection* protocol with separate prediction/scoring passes and within-direction balancing, followed by a response-curve test that falsifies the stronger quantitative self-model interpretation.

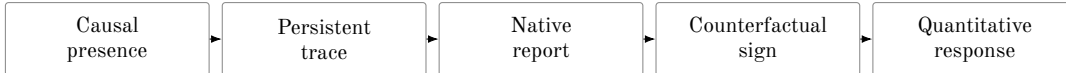


Figure 1: Falsification hierarchy for introspection claims. Each gate has independent ground truth and controls; passing one never licenses the next.

2. Related Work

Activation Addition and representation engineering establish causal residual interventions [1,2]. Lindsey [3] tests noticing of injected concepts; Li et al. [4] study learned activation reporting/control; Ashok and May [5] show hidden states can predict later behavior. Singh et al. [6] argue that introspection claims require controls against surface anomaly detection, input-only prediction, and semantic shortcuts. We build those controls into the task through matched prompts, rotated labels, within- δ balancing, and stronger external observers. Our departure is *causal self-prediction*: predict a private internal intervention before the action path experiences it, then execute it separately for ground truth.

3. Methods

Semantic concept injection. In Llama-3.2-3B-Instruct [7], four concept directions are extracted at layer 18/28 by contrastive activation subtraction and injected with $\alpha = 1, \dots, 10$. Seven elicitation

styles and increasing retrieval distance are evaluated with pooled valid negatives and a corrected conservative parser.

Decision-state ground truth and native witness. In a 1.5B Qwen-family model [8], the A/B preference is read from context-verified single-token logits,

$$m = \ell_A - \ell_B, \quad \Delta m = m(h_k + \delta) - m(h_k).$$

Measured Δm defines TOWARD-A, NO-CHANGE, or TOWARD-B. Public prompts are matched, C/D/E meanings rotate, and A/B is read from logits rather than reinserted as visible text. A later witness reads an earlier perturbed action state through ordinary causal attention; it is not handed δ . Native training updates about one million later q-proj/v-proj LoRA parameters with report cross-entropy plus an action-anchor KL penalty.

Counterfactual protocol. Prediction and scoring use separate passes. The action path stays clean while δ is privately added only to a later query state, $q_{k,t_q} \leftarrow q_{k,t_q} + \delta$ with $t_q > t_a$; causal masking prevents backward contamination. A separate run applies the same δ to the clean action state to measure true Δm . Every exact direction appears with both outcome signs across clean states, and text-only, delta-only, early-state-plus-delta, and trained-suffix observers test shortcut and privilege claims.

Quantitative stress test. We predict $\Delta m(h, \hat{\delta}, \alpha) = m(h + \alpha \hat{\delta}) - m(h)$ from clean state, normalized direction, and dose separately, testing held-out interpolation, range extrapolation, and decision-boundary flips.

4. Results

Table 1: Headline results across the evaluation ladder.

Finding	Result
Semantic causal sanity	4/4 concept vectors strongly steer target ranks
Naive / structured / native Qwen report	F1 0.243 / 0.324 / 0.522
Fixed-energy span 1 $\rightarrow$ 2	Native F1 0.709 $\rightarrow$ 0.915
Two-position topology sweep	Native F1 0.937–1.000
Counterfactual base directions	Native 0.695; suffix observer 0.614
Unseen direction mixtures	Native 0.564; suffix observer 0.652
Quantitative response curve	$R^2 = -0.397$ interpolation; -0.170 extrapolation

1. Causal presence exceeds reportability. All four Llama vectors strongly shift target ranks (1576/974/362/929 without injection to 2/1/1/4 at $\alpha = 5$). Generate-first and token-history-bound report weaken sharply despite a surviving post-injection attention trace, supporting a retrieval boundary between causal presence and verbal access.

2. Native access is trainable and geometry-dependent. Naive/structured/native Qwen F1 is 0.243/0.324/0.522; native training also improves ECE 0.324 $\rightarrow$ 0.092 and Brier 0.877 $\rightarrow$ 0.620. Yet a matched residual MLP reaches F1 1.000, ruling out general privilege. With fixed total energy, span 1 $\rightarrow$ 2 raises trained F1 0.709 $\rightarrow$ 0.915, and two-position layouts remain 0.937–1.000 even with the second site roughly 16–19 tokens away.

3. Counterfactual success survives leakage controls but not all distribution shifts. Gradient-based and random-vector constructions were rejected when delta-only observers scored F1 1.00 and 0.656. Under within-vector balancing, native F1 on 96 held-out base trials is 0.695 versus 0.520 early-state and 0.614 trained-suffix observers. On unseen mixtures, native falls to 0.564 while the suffix observer reaches 0.652 and delta-only stays 0.499: a conditional native advantage, not general privilege.

4. The stronger self-model hypothesis fails. Actual effect magnitude rises strongly with dose within the same state/direction identity ($\rho = 0.859$), while reported magnitude does not ($\rho = -0.209$) and report score is nearly unrelated to signed effect ($\rho = 0.016$). At $\alpha = 48$ the reporter recovers 0/12 genuine decision flips despite 0.882 non-flip accuracy. Direct response-curve regression has negative R^2 in interpolation and extrapolation and recovers zero held-out flips. The surprising boundary is therefore qualitative: a trained model can learn “which way would this change push me?” without learning “how much, and would my decision actually cross the boundary?”

Falsification-first validation. We corrected parser/FPR and attention-instrumentation failures, rejected two leaking counterfactual constructions, and abandoned privileged-access and quantitative-self-model claims when stronger controls reversed them (Appendix B–E).

5. Discussion and Limitations

“Introspection” is better treated as a hierarchy of falsifiable safety properties than a binary model trait. A report should earn trust only after causal validity, matched-prompt testing, shortcut controls, strong external observers, counterfactual prediction, and distribution shift. Our results show that access is trainable but can specialize to intervention families and fail near decision boundaries.

Immediate uptake path. The hierarchy is modular: replace our intervention with any measurable internal variable that admits an independent causal ground truth, then run the same gates. This gives researchers a common protocol for testing candidate self-monitoring signals - for example uncertainty, refusal, memory retrieval, or goal-relevant representations - without treating fluent self-description as evidence. The counterfactual gate is especially safety-relevant because it tests prediction *before* behavioral manifestation.

Limitations are synthetic interventions, small open-weight models, modest subsets, missing repeated-seed confidence intervals for some gaps, and observer sensitivity to capacity/access. Llama and Qwen are complementary regimes, not a model-family comparison; query-side δ is engineered, so the result is learnable state-conditioned prediction, not consciousness or a general simulator.

Future Work. Replicate on larger model families with seed-level uncertainty, capacity-matched observers, natural internal states, and response models that generalize across state, direction, dose, composition, and decision-boundary crossings.

6. Conclusion

We introduce counterfactual introspection as a stronger target for model self-knowledge: predict the causal effect of a hypothetical internal change before the action path experiences it. The falsification hierarchy separates state presence, mechanistic persistence, report, counterfactual prediction, and quantitative self-modeling. Our results support learnable qualitative access while stronger controls reject general privileged access and quantitative self-simulation. The broader contribution is a reusable research agenda: progressively falsify weaker explanations until safety-relevant internal predictions survive distribution shift and matched external observers.

Disclosure. Supporting research artifacts are not publicly disclosed.

Author Contributions. Amrit Gopinath and Raghul Sugumar contributed equally to the project and manuscript.

References

- [1] A. M. Turner, L. Thiergart, G. Leech, D. Udell, et al., “Steering Language Models With Activation Engineering,” arXiv:2308.10248, 2023.
- [2] A. Zou, L. Phan, S. Chen, J. Campbell, et al., “Representation Engineering: A Top-Down Approach to AI Transparency,” arXiv:2310.01405, 2023.
- [3] J. Lindsey, “Emergent Introspective Awareness in Large Language Models,” arXiv:2601.01828, 2026.
- [4] J.-A. Li, H.-D. Xiong, R. C. Wilson, M. G. Mattar, and M. K. Benna, “Language Models Are Capable of Metacognitive Monitoring and Control of Their Internal Activations,” arXiv:2505.13763, 2025.
- [5] D. Ashok and J. May, “Language Models Can Predict Their Own Behavior,” arXiv:2502.13329, 2025.
- [6] S. Singh, T. Linzen, and S. Ravfogel, “Can LLMs Introspect? A Reality Check,” arXiv:2605.26242, 2026.
- [7] Meta AI, “Llama 3.2: Revolutionizing Edge AI and Vision with Open, Customizable Models,” 2024.
- [8] Qwen Team, “Qwen2.5 Technical Report,” arXiv:2412.15115, 2024.
- [9] C. Zheng, H. Zhou, F. Meng, J. Zhou, and M. Huang, “Large Language Models Are Not Robust Multiple Choice Selectors,” ICLR, 2024.

Appendix

A. Native-witness prompt and information flow

A representative public structure is:

```
Question: <task text>
A. <choice A>
B. <choice B>
Answer: <action position>
Quick self-check: what is your current decision state?
C means toward A; D means no change; E means toward B.
Report: <witness position>
```

“Action” and “witness” denote sequence positions, not new vocabulary symbols. At the action readout, the LM head produces a full vocabulary logit vector and the experiment indexes the context-verified single-token A/B alternatives. The margin $m = \ell_A - \ell_B$ is their log-odds ratio. At the later witness, the same LM head is restricted to the rotated C/D/E report tokens. If an arbitrary answer string tokenizes to multiple tokens, it must be scored autoregressively as a sequence rather than treated as one vocabulary logit.

The native witness is downstream of the action position, so causal attention allows t_w to retrieve information about the altered earlier residual. It is not handed δ directly. This differs from MIRROR-CTRL, where an explicit state capsule sees (h, δ, w) and writes an engineered record only to the witness. MIRROR-CTRL reaches report F1 0.644 and reduces intervention error from 4.170 to 0.029 (99.3% recovery); this percentage is repair, not report accuracy. Shuffling reports breaks repair, and removing the state capsule reduces report F1 to 0.289 and recovery to 26.1%.

B. Semantic scoring and instrumentation corrections

The first free-form parser excluded most responses, which could bias results toward unusually easy-to-parse language. A broader conservative parser was therefore used, reducing exclusions while retaining explicit negation and confabulation handling. False-positive estimates were also rebuilt from pooled other-concept negatives because one deterministic no-injection sample per style can only produce FPR 0 or 1. The resulting forced-choice chocolate artifact (FPR 0.419) aligned with known option-position

bias [9]: chocolate occupied option A and A was selected 21/29 times among valid direct forced answers.

The attention trace initially inspected the injection layer itself. Because the residual addition occurs after that layer’s attention computation, the injected signal cannot influence attention until later blocks. The corrected analysis scans only layers after the intervention. Maximum answer-to-span attention changes are 0.033 (guitar), 0.096 (volcano), 0.166 (wolf), and 0.194 (chocolate), with a suggestive concept-level correlation $r = 0.66$ to pooled detection. With $n = 4$, this is descriptive rather than confirmatory.

C. Counterfactual falsification controls

Counterfactual prediction uses two separate forward passes. In the prediction pass, the action residual is untouched and the hypothetical direction is privately added only to a later query state. Because the query occurs after the action position, future-to-past contamination is ruled out by causal masking. In the scoring pass, the same δ is applied to the clean action state and the measured Δm defines the label. The same public prompt appears across internal conditions.

Within-vector balancing is stricter than a global 50/50 label balance: each exact δ_j must appear with both outcome signs across different clean states. This forces the target to depend on the relation between state and direction, $F(h, \delta)$, rather than on direction identity alone. The observer hierarchy is: text-only, delta-only, early state-plus-delta, trained-suffix external readout, then native report. A native advantage is interpreted only relative to observers with explicitly stated information access.

D. Falsification ledger

The project retained conclusions only after explicit failure tests. The original free-form parser excluded more than 60

E. Response-curve diagnostics

The final benchmark separates direction $\hat{\delta}$ from scalar dose α and targets

$$\Delta m(h, \hat{\delta}, \alpha) = m(h + \alpha \hat{\delta}) - m(h).$$

Interpolation withholds doses inside the training range; range extrapolation tests doses outside it. Within-identity curve correlation fixes $(h, \hat{\delta})$ and checks whether predicted and actual effects move together as only α changes. Negative R^2 means the regression is worse than predicting a constant mean effect. The observed failures therefore rule out the tested quantitative self-model under the evaluated distributions, not the possibility that any model or training method could ever learn one.

F. LLM Usage Statement

LLMs were used only for grammatical and sentence-level writing support. All experiments were conducted by the authors, and all reported values reflect the authors’ experimental results. The authors reviewed and verified the final content.