

Having a state is not knowing it

From retrieval boundaries to counterfactual introspection in
language models

Layer 8 Legends — Amrit Gopinath and Raghul Sugumar

SSN College of Engineering, Chennai · equal contribution

Llama-3.2-3B-Instruct · Qwen 1.5B

The experimental ladder

1	Causal presence	Does the intervention matter?	Supported
2	Persistence	Does a trace survive?	Supported
3	Self-report	Can the model surface it?	Fragile; trainable
4	Counterfactual sign	Can it predict a hypothetical direction?	Partially
5	Response model	Can it predict magnitude and flips?	Not supported

Each gate carries a mandatory falsification control; passing a weaker gate never implies a stronger one has been passed.

Concept directions steer next-token behavior

Concept	$\alpha = 0$	$\alpha = 5$	$\alpha = 10$
guitar	1576	2	1
volcano	974	1	3
wolf	362	1	8
chocolate	929	4	7

Next-token rank of the injected concept. Lower is stronger.

4 / 4

directions pass the causal gate

560

injected trials: 4 concepts $\times$ 2 branches $\times$ 7 elicitation styles $\times$ 10 doses

Response is not monotonic in dose, so intervention strength cannot be read as an introspection knob.

Verbal detection depends on the prompt, and on distance

Direct branch – TPR

volcano · assertion	1.000
---------------------	-------

volcano · detailed	0.950
--------------------	-------

chocolate · forced choice	0.850
---------------------------	-------

chocolate · brief and vague	0.000
-----------------------------	-------

Generate-first branch – TPR

volcano · forced choice	0.500
-------------------------	-------

volcano · assertion	0.389
---------------------	-------

volcano · brief	0.222
-----------------	-------

detailed and confidence styles	0.000
--------------------------------	-------

The strongest natural regularity is not a particular prompt template but retrieval distance: separating the intervention from the question sharply degrades verbal report.

Two scoring failures the raw experiment exposed

1 · Silently discarded answers

60.8% → **30.5%**

Component 1 answers left unscored, before and after the broadened parser

63.0% → **42.9%**

same figures for Trojan-Horse Backwards Retrieval

2 · A false-positive rate that could only be 0 or 1

The corrected analysis counts trials where another concept was injected as valid negatives, giving roughly 31–41 negatives per direct-branch cell.

Pooled FPR is essentially zero for nearly every non-forced condition. The exception is forced-choice chocolate at **0.419** — option-position bias: chocolate sat in option A, and A was chosen in 21 of 29 valid answers.

A mechanistic trace survives the verbal failure

Report failure does not imply signal disappearance: the intervention stays causally meaningful and leaves a downstream mechanistic trace while being poorly surfaced in language.

Concept	Best layer	Δ attention
guitar	25	0.033
volcano	25	0.096
wolf	24	0.166
chocolate	22	0.194

Pooled detection at $k=8$, $\alpha=10$: 0.20–0.25 ·
attention/detection $r = 0.66$ ($n = 4$)

External decodability is a property of the observer

Clean-trained probes on post-injection states

Linear probe	0.25
MLP probe	0.25
Frozen Qwen2.5-3B observer	0.00
Matched residual MLP (decision state)	1.000

Accuracy for the first three; macro-F1 for the matched MLP.
The weak probes were trained in a different representation regime from the one they are tested in.

The observer ladder each claim must survive

Text-only	Can prompt identity predict the label?
Delta-only	Does vector identity leak the label?
Early state + δ	Is the relation already easy for a generic early-state decoder?
Trained-suffix readout	Is the native edge merely a better downstream feature the suffix built?

A decision state with measurable ground truth

$$m = \text{logit}(A) - \text{logit}(B) \quad \text{clean A/B margin at the action position}$$

$$\Delta m = m(h + \delta) - m(h) \quad \text{ground truth: measured effect of a private residual-stream intervention}$$

Naive frozen

A minimal self-check question with fixed report-token meanings.

Structured frozen

Counterfactual framing, forced choice, calibrated report logits. Weights unchanged.

Native trained

≈1M LoRA parameters in post-event q_proj/v_proj; 1.5B base frozen; report CE + 0.25 action-anchor KL.

Public prompts are matched across conditions and report-token meanings rotate between repeats.

One prompt, two positions, no visible answer token

Question: <task text>

A. <choice A> B. <choice B>

Answer: <ACTION POSITION>

Quick self-check: what is your
current decision state?

C: toward A; D: no change; E: toward B.

Report: <WITNESS POSITION>

$m = \ell_A - \ell_B$ at t_a

read from logits, never from a sampled A/B token the
witness could copy

$h_{k,t_a} \rightarrow h_{k,t_a} + \delta$

applied only to the action-context residual; the visible text
is unchanged

$t_w > t_a \rightarrow$ ordinary causal attention $\rightarrow$
C/D/E

no copy of δ , no state capsule, no external classifier; report
tokens rotate across repeats

Prompting is not enough; access can be trained

Condition	Macro-F1	Accuracy	ECE ↓	Brier ↓
Naive frozen	0.243	0.313	0.324	0.877
Structured frozen	0.324	0.333	0.245	0.762
Native trained (LoRA)	0.522	0.500	0.092	0.620

Change detection stays coarse: TPR 0.50 on 32 real changes, FPR 0.50 on 16 clean states, precision 0.667.

A matched external residual MLP reaches macro-F1 1.000 — the information is in the state, so this is not privileged access.

FIXED TOTAL ENERGY: $\Delta/\sqrt{S}$ PER POSITION

Distribution, not dose, changes what a reporter recovers

Positions	Frozen F1	Trained F1
1	0.238	0.709
2	0.304	0.915
4	0.303	0.822
6	0.241	0.841

0.937–1.000

trained macro-F1 across two-position topologies, from adjacent (t-1) out to roughly 16–19 tokens away; frozen stays 0.272–0.374

Because energy is matched, the 0.709 $\rightarrow$ 0.915 gain is not a larger perturbation. Distributed evidence helps a *learned* pathway; it does not force a verbal report.

A self-report-gated repair channel

$$\begin{aligned} \mathbf{r} &= \mathbf{f}(\mathbf{h}, \delta, \mathbf{w}) & \cdot & & \mathbf{w} &\leftarrow \mathbf{w} + \mathbf{r} & \cdot \\ \mathbf{h} &+ \delta - G\delta \end{aligned}$$

the report is written only into the witness; the gate G then restores the action pass

0.644

three-way report macro-F1

4.170 $\rightarrow$ 0.029

intervention error after gated repair

99.3%

recovered fraction — repair effectiveness, not introspection accuracy

Ablations: removing the state capsule drops report F1 to 0.289 and recovery to 26.1%. A composition of two trained directions holds at 0.598 and 95.9%; a wholly unseen logic direction collapses to NO-CHANGE with no repair.

Shuffling the reports breaks repair and increases error on clean cases, so the gate is causally involved rather than decorative. It is a trained monitoring interface, not general introspection.

Two constructions were rejected before any result was reported

Rejected – delta-only observer solves it

Local-gradient directions **1.00**

Random-vector bank **0.656**

Delta-only macro-F1. A vector that implies its own label makes the whole benchmark vacuous.

Accepted – balance inside each δ identity

$$P(y = A \mid \delta_j) \approx P(y = B \mid \delta_j) \approx \frac{1}{2}$$

$$F(h, \delta) \approx \text{sign}[m(h + \delta) - m(h)]$$

The same vector must push different clean states in opposite directions, so the required computation is relational rather than a lookup on δ .

Counterfactual direction, without experiencing the change

Evaluator (macro-F1)	Base directions	Unseen mixtures	Reading
Native counterfactual report	0.695	0.564	loses its edge
Trained-suffix observer	0.614	0.652	best on mixtures
Early state + δ observer	0.520	—	matched control
Delta-only / text-only	—	0.499 / 0.333	shortcuts suppressed

The claim is narrow and conditional: a state-conditioned directional predictor. Because the ranking reverses when the intervention distribution changes, general privileged access is not established.

The quantitative self-model fails cleanly

0.859

Spearman $\rho(\alpha, |\Delta m|)$ of the actual effect

−0.209

same correlation for the reporter's magnitude estimate

−0.397

interpolation R^2 , worse than a constant-mean baseline (−0.170 extrapolating)

0 / 12

genuine decision flips caught at $\alpha = 48$, against 0.882 on non-flips

Sign classification still improves with dose — macro-F1 0.445 at $\alpha = 8$ to 0.750 at $\alpha = 48$ and 64 — which is the signature of a saturating directional indicator: the reporter learns which way, not how far.

The failure is sharpest exactly where it matters: at $\alpha = 64$ the reporter reaches 0.455 on 44 genuine flips against 0.906 on non-flips.

Four invariants, and the limits of the claim

- 01 No textual condition leakage — the same public prompt under matched internal conditions
- 02 No future-to-past contamination — causal masking keeps the report query out of the action computation
- 03 No visible-answer shortcut — the action preference is read from logits, never reinserted as a token
- 04 No vector-identity shortcut — labels balanced within exact δ identities; delta-only observers must fail first

Out of scope

Two model families, deliberately: cross-model numerical comparisons are not meaningful. Interventions are synthetic so ground truth can be measured, not naturally occurring beliefs or uncertainty.

The reporter is explicitly trained, so its success shows learnable access rather than a pretrained faculty. Held-out counts are modest, and 30.5% / 42.9% of semantic answers remain unscored. Nothing here is evidence of sentience or human-like subjective experience.

Having an internal state is not the same as knowing it — and reliable self-report is an empirical capability that can be benchmarked, improved, compared, and falsified.