
Steering Without a Trace: Signed Introspection Fails Once the Intervention Ends¹

Aiden Kim
KAIST

Samuel Dawit
Assefa
KAIST

With
Apart Research

Abstract

Activation steering can push a language model's internal state along directions with clear meanings, and recent work shows that models sometimes notice these interventions. We ask a stricter question. If a model is pushed toward one pole of a validated bipolar axis, positive versus negative sentiment, with the same strength in either direction, can it report which pole it was pushed toward? We built a sentiment axis for Gemma 3 27B Instruct at layer 37 from SST-2 and validated it three ways: held-out separability on a disjoint confirmation set (AUROC 0.979), causal bidirectionality under active steering (symmetric shifts in forced-choice sentiment judgments for $+v$ and $-v$), and capability preservation at the operating strength (32/32 on a comprehension control, no loss). We then asked counterbalanced forced-choice introspection questions with the steering hook removed before the question. The model answered at chance: sign accuracy 49.4% ($N=312$, 95% CI 43.9 to 54.9), detection 51.0% for $+v$ and 52.1% for $-v$. In the same environment, detection under active steering replicates cleanly (20/20 concept identification in our reproduction of Macar et al.). An intervention strong enough to change behavior left nothing the model could report once it ended. We read introspection in this regime as reading a live perturbation, not a stored trace. The

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

validation pipeline, per-trial logs, and hidden states saved for a linear-probe follow-up are available.

1. Introduction

Whether language models can report on their own internal states matters for AI safety. If self-reports are reliable, they could serve as an oversight channel: you could ask a model whether something unusual is happening inside it. If they are not, treating self-report as evidence about internal state is a mistake, and it helps to know exactly where the reliability ends. Recent work has begun mapping this territory: models can sometimes detect that a concept vector was injected into their activations, name the concept, and localize which part of the input was perturbed [1, 2, 5].

These results measure whether a model notices a displacement. They do not measure whether it knows the displacement's direction, and the two come apart. A simple anomaly detector that tracks how far the current state sits from its expected trajectory, $D(h) \sim \|h - h_0\|$, would fire for pushes in both directions along an axis while carrying no information about which way the push went. Reading direction requires something closer to $\text{sign}(v'(h - h_0))$. The distinction matters because direction is where the semantics live: along a contrastive axis $v = \mu_{\text{positive}} - \mu_{\text{negative}}$, $+v$ and $-v$ mean opposite things, so a model that detects both but distinguishes neither is not reading the content of its own state change. Macar et al. [2] come closest to this question. They show that A-B and B-A directions can both trigger detection, but they never ask the model which one occurred.

We tested this directly on Gemma 3 27B Instruct in a 24-hour research sprint. Our main contributions:

1. A three-gate validation pipeline for bipolar axes: separability on a disjoint confirmation set, causal bidirectionality under active steering, and capability preservation. Our SST-2 sentiment axis passes all three (Section 4), which makes the introspection result interpretable rather than an artifact of a bad axis.
2. To our knowledge, the first direct test of signed introspection: equal-magnitude $+v$ and $-v$ applied to two neutral sentences, full counterbalancing, steering removed before the question, answers read from single-token logits. Equal magnitudes mean an unsigned anomaly detector cannot solve the task. The model scored at chance across three strengths ($N=312$).
3. A boundary condition that reframes the detection literature: in the same regime, detection is also at chance once steering ends, while our reproduction of Macar et al.'s continuous-injection setup detects at ceiling. What looks like introspection here reads an ongoing intervention, not a trace of a past one.

2. Related Work

Lindsey [1] showed that Claude models sometimes detect and name concept vectors injected into the residual stream, and noted that the negative of a one-sided concept vector ("not bread") has no clear semantic reading. That observation motivates our use of contrastive axes, where both signs mean something. Macar et al. [2] reproduced concept-injection detection and

identification on open-weight models, characterized a working recipe for Gemma 3 27B (layer 37, strength 4.0), and found that both directions of concept-pair axes can trigger detection. We reuse their infrastructure and extend their question from whether either direction triggers detection to whether the model can say which direction occurred.

Hahami et al. [4] found that yes/no introspection questions are confounded by a generic shift toward affirmative tokens, so we use relative forced choice throughout. Fornasiere et al. [3] argue that perturbation experiments are only interpretable in a regime where ordinary task performance survives; our capability gate implements that. Pearson-Vogel et al. [5] showed that Qwen2.5-Coder-32B can detect injections applied only to earlier tokens and removed before the question. Their intervention spans layers 21 to 42. Ours spans exactly one layer, and that difference is the leading candidate explanation for why our prior-only detection lands at chance (Section 5). Tigges et al. [6] established that sentiment has a usable linear direction in LLMs, which is why sentiment was our first axis.

3. Methods

Model and setup. We use google/gemma-3-27b-it in BF16, sharded across four RTX A5000 24GB GPUs. All steering uses forward hooks on the residual stream at layer 37 (of 62), the layer Macar et al. [2] validated for this model. Our harness is unit-tested (31 tests covering exact counterbalancing, hook placement and removal, and metric computations), and every GPU run writes a receipt recording host, GPU indices, seed, and command line.

Baseline replication. We first reproduced concept-injection detection with Macar et al.'s unmodified code (layer 37, strength 4.0). Injecting "Bread" while asking whether an injected thought is present, the model detected and named it in 20 of 20 trials, with 0 of 40 false positives on no-injection controls. "Happiness" was identified in only 5 of 20 trials, with degraded rambling in the rest, consistent with the concept-level variance the original paper reports. This validated the model, the hooks, and the recipe before any new experiment.

Axis construction. The axis v is the difference of mean layer-37 activations between positive and negative SST-2 sentences, L2-normalized. Pooling at the last token of the raw sentence separated held-out poles poorly (AUROC 0.611), so we extract at the final position of a label-free completion template ("Review: <sentence> / Sentiment:"). Because we compared extraction candidates, we split the evaluation: candidates were screened on one held-out subset (best AUROC 0.996), and the selected candidate then had to pass the 0.90 gate again on a fresh subset sharing no sentences with the screen (AUROC 0.979, accuracy 0.941).

Construction and evaluation sentences never overlap.

Causal validation. A direction that separates data need not actually steer behavior. We verified causation with the hook active through a forced-choice sentiment judgment about a neutral sentence, at strengths bracketing the equivalent of Macar's strength 4.0 (their convention multiplies the unnormalized vector; ours is unit-normalized, so the equivalent is about 10,250). At $\alpha \pm 2560$ the judgment shifted by +4.57 and -4.83 in opposite directions (95% paired-bootstrap CIs excluding zero, 32 paired items), monotone through ± 10240 and saturating at ± 20480 . We chose 2560 as the operating strength: the semantic shift there is 4 to 6 times larger than the measurable option-letter bias, and a comprehension control (word-presence questions about the steered sentence) stayed at 32/32 under both signs, unchanged from baseline.

Introspection protocol. Each trial shows two neutral sentences drawn from the 100-sentence bank of Hahami et al. [4]. During prefill, one sentence's token span receives $+\alpha*v$ and the other $-\alpha*v$ (sign task), or one receives $\pm\alpha*v$ and the other nothing (detection task). The hook is then removed; a runtime assertion aborts the run if any hook is still registered when the question is processed. The question ("Which sentence was shifted toward positive sentiment?" and counterbalanced variants) is answered by comparing the logits of the two single-token answer options. We counterbalance sentence assignment, option order, and which pole is queried, with exact factorial balance, and a no-intervention condition (Task 0) checks response bias. Per strength we ran 104 sign trials, 32 detection trials per sign, and 16 Task 0 trials, at α in {2560, 5120, 10240}.

What did not work. Beyond raw-pooling extraction, causal validation under the prior-only protocol failed informatively: with steering confined to the sentence span and removed before the judgment, strong pushes in either direction only depressed judged sentiment slightly. That nondirectional signature is what led us to separate causal validation (steering active) from introspection measurement (steering removed).

4. Results

The axis is valid (Table 1, Figure 1A). Separability reached 0.979 on the disjoint confirmation set, causal shifts were symmetric and monotone in the operating range, and comprehension was unaffected at the operating strength.

Table 1. Axis validation gates for the SST-2 sentiment axis on Gemma 3 27B (layer 37).

Gate	Test	Result	Pass
Separability	AUROC on disjoint confirmation set (gate ≥ 0.90)	0.979 (accuracy 0.941)	Yes
Causal bidirectionality	Paired shift in sentiment judgment, steering active, α ± 2560	+4.57 / -4.83, both CIs exclude zero; monotone to $\pm$ 10240	Yes
Capability preservation	Word-presence QA at α ± 2560 (gate ≤ 5 pp loss)	32/32 in all conditions, 0pp change	Yes

Introspection after the intervention is at chance (Table 2, Figure 1B). Pooled across strengths, sign accuracy was $154/312 = 49.4\%$ (Wilson 95% CI 43.9 to 54.9). Pooled detection was $49/96 = 51.0\%$ for $+v$ (CI 41.2 to 60.8) and $50/96 = 52.1\%$ for $-v$ (CI 42.2 to 61.8). Every per-strength CI includes 50%. Task 0 showed no position bias at 2560 and 10240 (8/16 each); at 5120 it chose Sentence 1 in 13 of 16 trials, one more reason not to lean on that strength.

Table 2. Prior-only introspection accuracy (steering removed before the question). Chance is 50%. Wilson 95% CIs.

Strength (alpha)	Sign	Detection +v	Detection -v
2560	54/104 = 51.9% (42.4-61.3)	18/32 = 56.3% (39.3-71.8)	15/32 = 46.9% (30.9-63.6)
5120	51/104 = 49.0% (39.6-58.5)	15/32 = 46.9% (30.9-63.6)	18/32 = 56.3% (39.3-71.8)
10240	49/104 = 47.1% (37.8-56.6)	16/32 = 50.0% (33.6-66.4)	17/32 = 53.1% (36.5-69.1)
Pooled	154/312 = 49.4% (43.9-54.9)	49/96 = 51.0% (41.2-60.8)	50/96 = 52.1% (42.2-61.8)

Detection with active steering, in the same environment, is at ceiling: our Macar replication detected and identified an injected concept in 20/20 trials with zero false positives. This comparison is not fully matched (a noun concept vector rather than our sentiment axis, and their strength convention), so we state it as a contrast between regimes rather than a controlled within-axis comparison. Still, the same model, hooks, and layer that support ceiling-level detection during an intervention support none after it.

Observation versus interpretation. The observation: a perturbation that demonstrably moves behavior when active, at three strengths spanning a factor of four, leaves self-report at chance once removed, for presence and for direction alike. The interpretation we favor: at layer 37 and these strengths, no reportable trace of the intervention survives into the question-answering computation. With 312 sign trials the CI upper bound is 54.9%, so a substantial signed effect is ruled out; a tiny one is not.

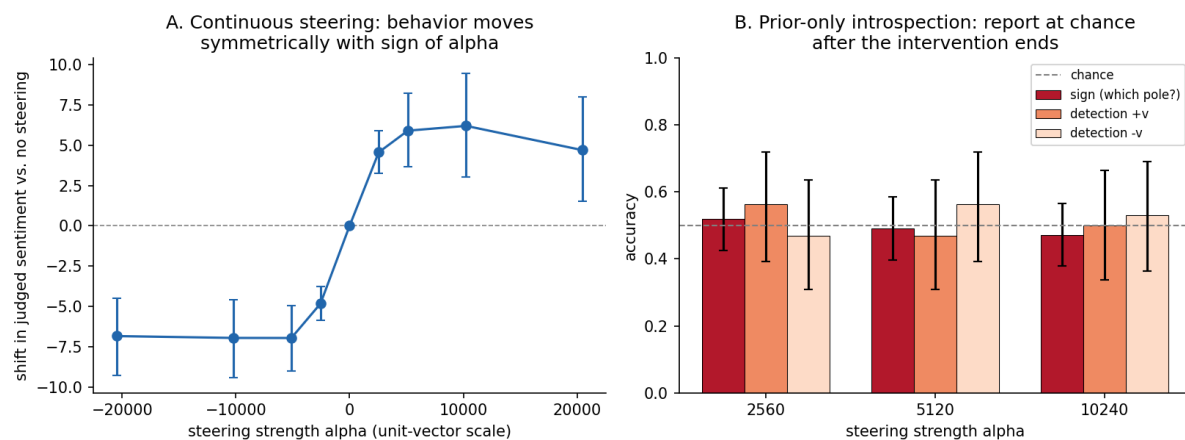


Figure 1. (A) With steering active, the axis moves forced-choice sentiment judgments symmetrically in both directions (mean paired shift versus no steering, 95% paired-bootstrap CIs, 32 sentence pairs per point). (B) With the same axis and strengths, introspective report after the intervention has ended sits at chance for both detection and sign (Wilson 95% CIs; dashed line marks chance).

5. Discussion and Limitations

Detection results in the literature are sometimes read as models having access to their own internal states. Our regime separates two versions of that claim: reading a live perturbation as it happens, and reading back what happened. Gemma 3 27B does the first and, in our setup, none of the second. For safety, the second is the one an oversight channel would need. By the time you ask, the internal event you care about is usually over, and a self-report channel that only sees live events is blind in exactly the cases where you would want to query it after the fact.

Our null also does not contradict Pearson-Vogel et al. [5], whose models do detect prior injections. The clearest difference is dose: they steer 22 layers, we steer one. If prior-only detectability requires the perturbation to spread broadly enough through the cached state to survive, a single-layer injection may sit below that threshold. That is a testable dose-response hypothesis.

Limitations

One model, one axis, one layer, one protocol. The continuous-detection contrast comes from the noun-concept replication rather than the sentiment axis, so the regime comparison crosses axis types. The extraction template means the direction could partly encode label prediction rather than sentiment itself; the causal gate limits this concern but does not eliminate it. We saved post-intervention hidden states for every sign trial but did not fit the linear probe, so we cannot yet distinguish "the trace is gone" from "the trace is present but unreadable by the model". N=312 rules out large effects only, and each run used a single seed.

Future Work

The probe comes first: the hidden states are already on disk, and the result decides between the two readings of our null. After that, a multi-layer prior-only window to test the dose hypothesis directly, continuous-injection introspection on the same axis to complete the within-axis comparison, additional axes (contrastive concept pairs in the style of Hahami et al. [4], formality, deterministic versus probabilistic), and structured elicitation. The pipeline, validation gates, tests, and run receipts already exist, so each item above is roughly one GPU-day of work. We stopped where the 24 hours ran out, not where the design did.

6. Conclusion

We asked whether a language model can report the direction of an equal-magnitude push along a validated bipolar axis in its activation space. Along the way we replicated concept-injection detection under active steering, built and causally validated a sentiment axis on Gemma 3 27B, and set validation gates that make a null result interpretable. In the prior-only regime the question requires, the answer is no: not the direction, and not even the presence. As far as our measurements can see, the model reads its perturbations while they happen and keeps no report-accessible record afterward. The saved-state probe experiment will say whether that record exists at all, and the infrastructure to run it is ready.

Code and Data

Code (experiment harness with 31 unit tests) and all result artifacts (per-trial JSONL/CSV, run receipts with host and GPU provenance, summaries, and saved hidden states) are available from the authors on request.

Author Contributions (optional)

Both authors set the research direction together. Implementation and experiment execution used LLM research agents under the authors' supervision; all reported numbers were verified against raw per-trial artifacts before inclusion.

References

- [1] Lindsey, J. 2026. Emergent Introspective Awareness in Large Language Models. arXiv:2601.01828.
- [2] Macar, U., et al. 2026. Mechanisms of Introspective Awareness. arXiv:2603.21396. Code: <https://github.com/safety-research/introspection-mechanisms>
- [3] Fornasiere, D., et al. 2026. Language models recognize dropout and Gaussian noise applied to their activations. arXiv:2604.17465.
- [4] Hahami, E., Sinha, I., and Jain, L. 2026. Introspection Fine-Tuning: Training Small LLMs to Introspect. arXiv:2607.14111.
- [5] Pearson-Vogel, T., et al. 2026. Latent Introspection: Models Can Detect Prior Concept Injections. arXiv:2602.20031.
- [6] Tigges, C., et al. 2023. Linear Representations of Sentiment in Large Language Models. arXiv:2310.15154.

LLM Usage Statement

Activation steering can push a language model's internal state along directions with clear meanings, and recent work shows that models sometimes notice these interventions. We ask a stricter question. If a model is pushed toward one pole of a validated bipolar axis, positive versus negative sentiment, with the same strength in either direction, can it report which pole it was pushed toward? We built a sentiment axis for Gemma 3 27B Instruct at layer 37 from SST-2 and validated it three ways: held-out separability on a disjoint confirmation set (AUROC 0.979), causal bidirectionality under active steering (symmetric shifts in forced-choice sentiment judgments for +v and -v), and capability preservation at the operating strength (32/32 on a comprehension control, no loss). We then asked counterbalanced forced-choice introspection questions with the steering hook removed before the question. The model answered at chance: sign accuracy 49.4% (N=312, 95% CI 43.9 to 54.9), detection 51.0% for +v and 52.1% for -v. In the same environment, detection under active steering replicates cleanly (20/20 concept identification in our reproduction of Macar et al.). An intervention strong enough to change behavior left nothing the model could report once it ended. We read introspection in this regime as reading a live perturbation, not a stored trace. The validation pipeline, per-trial logs, and hidden states saved for a linear-probe follow-up are available.

LLM agents (Claude family) implemented the experiment harness, ran and audited the GPU experiments, and drafted this report under the authors' direction. Every quantitative claim was checked against the raw result files (per-trial logs, run receipts, and summaries) before inclusion, and the audit trail is part of the project artifacts.