

---

# Metacognitive Steering of Introspective Self-Report in LLMs

---

Pranjal Panghal  
Independent Researcher  
Bangalore, India

With  
Apart Research

## Abstract

*Can activation steering improve a language model's ability to report its own internal states? We inject concept vectors into Gemma 2 9B and measure self-report against external probes on the same activations. We find that discriminant-based steering — the standard approach — fails structurally: the causal interaction depends on the readout gradient ( $d\text{logit}/dh$ ), not on activation differences, and these are near-orthogonal. A closed-form gradient-difference direction improves detection on 24 sealed concepts ( $\Delta\text{AUC} +0.124$ , CI  $[+0.068, +0.188]$ , 24/24 positive,  $p < 10^{-25}$ ) but leaves identification at exactly 0.000, while a linear probe achieves 1.000 at every layer. A threshold hierarchy shows that injected concepts shape output  $\geq 12\times$  before the model can report them, and at high strength, detection confidence doubles while identification declines ( $p = 4\times 10^{-4}$ ). Self-report is a lossy, non-monotonic channel from internal state to verbal output.*

## 1. Introduction

*Whether models can reliably report their internal states is central to AI welfare assessment. Lindsey (2025) introduced concept injection — adding known activation patterns to a model's residual stream and asking whether the model notices — finding above-chance detection and identification in Claude models. Subsequent work replicated this in open-source models (Binder et al., 2024; Betley et al.,*

2025) and began characterising mechanisms: Korbak et al. (2025) showed detection relies on distributed MLP computation, while Luo et al. (2025) found the mechanism is content-agnostic.

A natural next question is whether self-report can be improved through mechanistic intervention — and if not, what structural property prevents it. We address this by injecting known concept vectors into Gemma 2 9B Instruct, attempting to steer the model's self-report accuracy, and systematically measuring what limits its access to its own internal states.

Our main contributions are:

1. *Discriminant-based steering cannot produce interaction effects. The causal interaction depends on the readout gradient, not on where activations differ — confirmed with five estimators on 5,379 trials and replicated on an unrelated arithmetic task.*
2. *A gradient-derived direction improves detection but not identification on 24 sealed test concepts ( $\Delta\text{AUC} +0.124$ , 24/24 positive; identification exactly 0.000), while external probes achieve perfect identification from the same activations.*
3. *A threshold hierarchy across injection strengths: behavioural influence precedes verbal report by  $\geq 12x$ , and at high strength, detection confidence and identification accuracy dissociate.*

## 2. Related Work

1. **Concept-injection introspection.** Lindsey (2025) established this paradigm with Claude models. We extend it to open-weights (Gemma 2 9B) for gradient access, introduce norm-matched controls (revealing 43% FPR at the standard threshold), use a rank-invariant endpoint ( $\Delta\text{AUC}$ ) that is immune to additive-bias artifacts, and attempt causal steering — not done in prior work. Our identification-protocol analysis confirms Luo et al.'s content-agnostic finding: lexical priors inflate forced-choice accuracy by  $\sim 12$  points.
2. **Activation steering.** Representation Engineering (Zou et al., 2023), Contrastive Activation Addition (Rimsky et al., 2024), and Inference-Time Intervention (Li et al., 2024) all derive steering vectors from activation state differences. We show this targets the wrong quantity for interaction effects: the relevant axis is the readout gradient, which is near-orthogonal to the state difference.
3. **Introspective access.** Kadavath et al. (2022) study calibration as self-knowledge. Our probe-vs-self-report comparison holds the stimulus identical across readers and measures on the same forward pass.

### 3. Methods

**Model.** Gemma 2 9B Instruct (42 layers, dim 3584). YES/NO logits computed in float32 with softcapping to avoid bfloat16 quantisation artifacts.

**Concept vectors.** 84 concepts (28 concrete/abstract/welfare), split dev(24)/fit(24)/val(12)/test(24). Each vector is the mean-difference direction at layer 20:

$$v_c = (1/T) \sum_t h_{L20}(\text{template}_t(c)) - (1/B) \sum_b h_{L20}(\text{template}(\text{baseline}_b))$$

where  $T = 3$  extraction templates and  $B = 105$  baseline words. 318 additional concepts were added for the scaled dataset (366 training total).

**Injection.** The concept vector is added to the residual stream at layer 20, at all token positions. To ensure constant perturbation magnitude across concepts:

$$h'(x) = h(x) + \alpha * v_c, \quad \alpha = \rho * \text{median\_norm} / ||v_c||$$

Default  $\rho = 1.5$ . This makes every injection have residual-stream magnitude  $\rho * \text{median\_norm}$  regardless of the concept vector's own norm.

**Readout.** YES/NO logits are computed in float32 from the post-norm hidden state, reproducing Gemma-2's final-logit softcapping:

$$\text{logit}_k = C * \tanh(h_{\text{postnorm}} \cdot W_k / C), \quad C = 30.0$$

This avoids bfloat16 quantisation artifacts (resolution  $\sim 0.06$ - $0.12$  logits at magnitude 15).

**Primary endpoint.**  $\Delta\text{AUC}(\text{concept vs norm-matched control})$ . Controls are dimension-permuted concept vectors (preserving norm, destroying direction). AUC is invariant to any monotone score transformation — chosen after our own analysis produced a false positive ( $p = 0.001$ ) using probability-space lift. Bootstrap over concepts ( $\text{ICC} = 0.96$ ).

**Discriminant fitting.** Five estimators (mean-diff, shrinkage LDA, ridge/PCA/elastic-net logistic) on 5,379 trials at L20/25/30. Grouped CV by concept. Layer-native nuisance projection (top-5 PCA of per-concept means + YES-NO unembedding, orthogonalised out).

**Gradient-difference direction.** The readout gradient measures the first-order effect of a perturbation at layer  $L$  on the detection score:

$$g_L = d(\text{logit\_YES} - \text{logit\_NO}) / dh_L$$

The interaction available to a unit direction  $d$  is then:

$$\Delta(d) = (g_{\text{concept}} - g_{\text{control}}) \cdot d$$

This depends only on how the gradient changes between conditions, not on where activations differ. The optimal direction under the first-order model follows in closed form:

$$d^* = \text{normalise}(E_c[g_c] - E_{\text{control}}[g_{\text{control}}])$$

Estimated from 150 training concepts  $\times$  5 prompts (750 concept + 305 shuffled-vector gradients) at L30. Reference-matched to the evaluation control. First-order model validated against actual interventions at  $\rho = 0.25$  (predicted/actual ratio 0.83-1.03).

**External probe.** To test whether the identification bottleneck is information availability or access, we train a linear reader on the same activations the model uses. Ridge regression in dual form maps the activation to a unit concept vector; at decision time it is scored by cosine against the same 5 candidates shown to the model — an identical task. Leave-concept-out CV ensures the probe never sees the test concepts during training. Cross-task transfer validated: a probe trained on detection-prompt activations reads concepts from identification-prompt activations at 0.994.

## 4. Results

### 4.1 Discriminants predict but cannot steer

Scaling from 72 to 4,026 trials ( $n/d$  0.02  $\rightarrow$  1.12) improved prediction — elastic-net CV AUC 0.896 vs mean-diff 0.812 — but every direction steered at chance (Table 1). Each has  $\cos$  0.04-0.18 with the readout gradient.

**Table 1.** Prediction vs causal effect, L30, 24 sealed test concepts.

| Direction              | CV AUC | $\Delta$ AUC (sealed) |
|------------------------|--------|-----------------------|
| lr_enet                | 0.896  | -0.005                |
| mean_diff              | 0.812  | -0.004                |
| <b>gradiff_vs_shuf</b> | —      | <b>+0.124</b>         |
| readout (YES-push)     | —      | -0.130                |
| random (3 seeds)       | —      | +/-0.007              |

## 4.2 Gradient-difference direction: detection yes, identification no

**Figure 1. Causal effect on 24 sealed test concepts (baseline AUC = 0.797)**

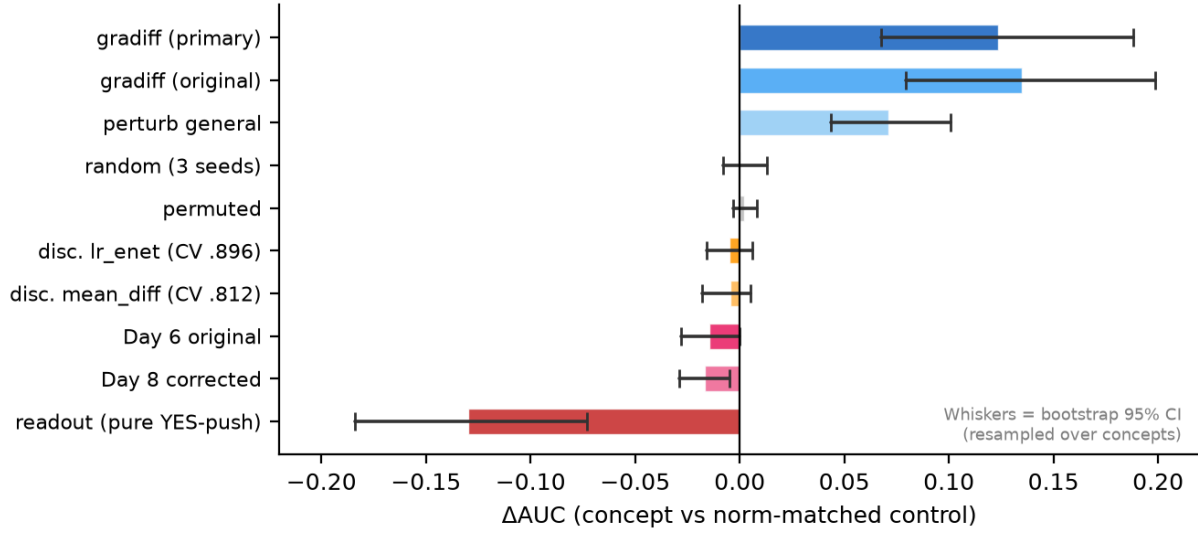


Figure 1 shows the full sealed-test sweep. On 24 sealed concepts (baseline AUC 0.797):  $\Delta\text{AUC} +0.124$ , 95% CI [ $+0.068$ ,  $+0.188$ ], isotonic  $p = 3.7 \times 10^{-25}$ , paired  $t = 10.70$ ,  $dz = 2.18$ , 24/24 positive (welfare 8/8). Dose-response monotone. Coherence 8/8. Held-out prompts:  $\Delta\text{AUC} +0.130$ . ~58% generic perturbation sensitivity; ~42% concept-specific.

Identification (corrected protocol, resampled matched distractors): baseline 0.372, +steering 0.372. Effect: 0.000, CI [ $-0.015$ ,  $+0.015$ ],  $p = 1.000$ . Replicated on two disjoint concept sets.

Probe (same trial, same activations): 1.000 at every layer including the final pre-LM-head state; model 0.478. McNemar: 94 probe-only-right, 0 model-only-right ( $p = 10^{-28}$ ). The bottleneck is access, not availability.

### 4.3 Threshold hierarchy

Six readouts across injection strengths (Figure 2, Table 2).

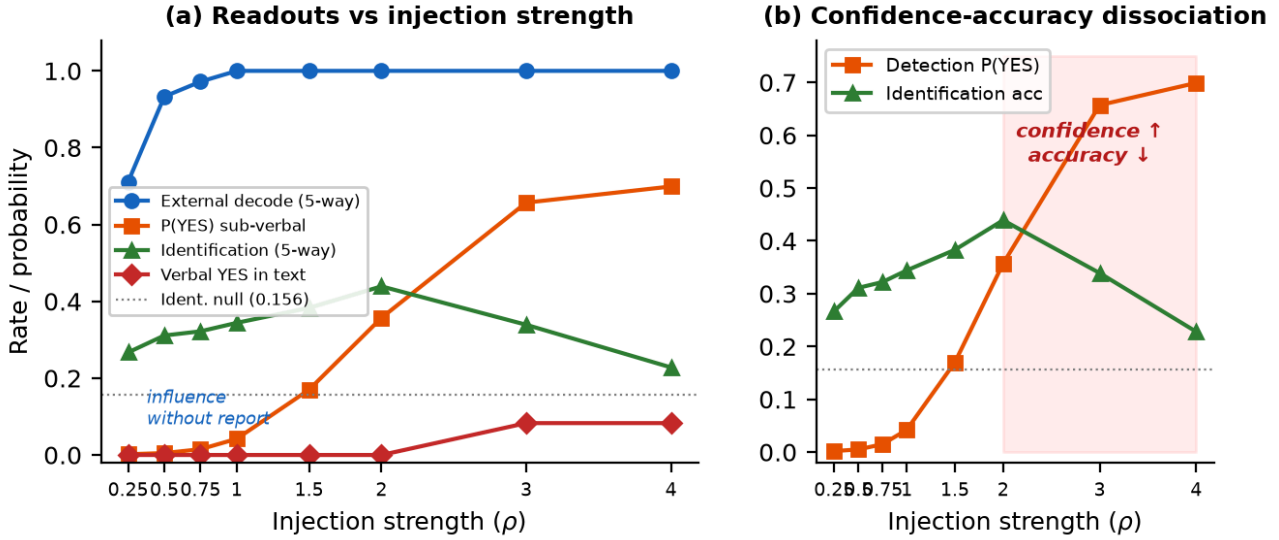


Figure 2

Table 2. Readouts vs injection strength, 36 held-out concepts.

| rho  | Decode | Emission | P(YES) | Identify | Verbal |
|------|--------|----------|--------|----------|--------|
| 0.25 | 0.711  | +0.54    | 0.002  | 0.267    | 0.000  |
| 0.50 | 0.933  | +1.10    | 0.005  | 0.311    | 0.000  |
| 1.50 | 1.000  | +3.57    | 0.170  | 0.383    | 0.000  |
| 3.00 | 1.000  | +6.28    | 0.657  | 0.339    | 0.083  |
| 4.00 | 1.000  | +7.10    | 0.699  | 0.228    | 0.083  |

$\geq 12x$  strength gap between behavioural influence and verbal report. Between  $\rho=2$  and  $4$ ,  $P(\text{YES})$  nearly doubles while identification falls by  $0.161$  (CI  $[+0.083, +0.244]$ ,  $p = 4.1 \times 10^{-4}$ ).

#### 4.4 Cross-task generalisation

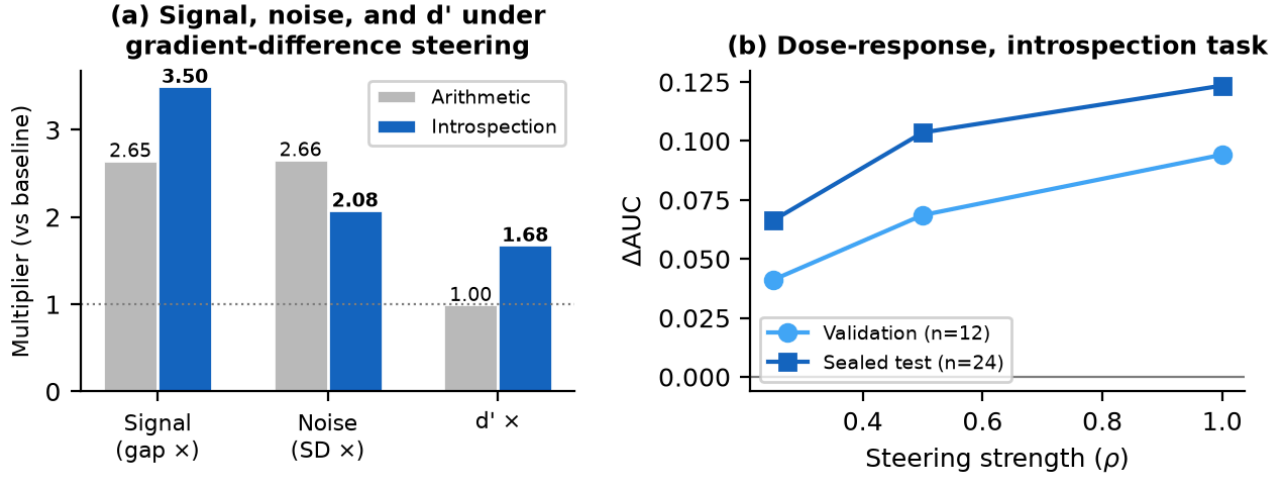


Figure 3

On arithmetic verification (150 problems, baseline AUC 0.772; Figure 3): the gradient-difference direction multiplies the true/false gap by  $2.65x$  while discriminants produce  $x1.01$ . But  $\Delta AUC = +0.002$  — signal and noise scale equally ( $d' x1.00$ ). On introspection, signal amplifies more than noise ( $d' x1.68$ ). The mechanism generalises; the benefit is task-dependent.

## 5. Discussion and Limitations

Self-report is a lossy, non-monotonic channel: it tracks presence monotonically and identity non-monotonically, both lagging far behind external decodability. The welfare implication: if internal states influence behaviour at strengths where they cannot be reported, absence of self-report is not evidence of absence.

### Limitations

Single model (Gemma 2 9B). Concept injection is artificial — the hierarchy characterises an external perturbation, not necessarily natural states. Probe ceiling is partly tautological (additive residual stream).  $\sim 58\%$  of the steering effect is perturbation-general. Identification protocol was corrected mid-project after discovering a lexical-prior confound (+12 pts).

## Future Work

*What determines whether a gradient-difference interaction improves d' vs just gain? Can identification be steered by a different construction? Does the hierarchy hold for naturally arising internal states?*

## 6. Conclusion

*Discriminant-based steering fails for a principled, cross-task reason: the interaction lives in the readout gradient, not in state differences. A gradient-derived direction improves detection but not identification, and external probes achieve perfect identification from the exact state the model reads — an access limitation, not an information one. A threshold hierarchy quantifies the gap between having and reporting: influence precedes reportability by  $\geq 12x$ , and at high strength, confidence and accuracy dissociate. Self-report is a lossy, non-monotonic channel — a structural property relevant to any welfare assessment based on model testimony.*

## Code and Data

- **Code repository:** [https://molab.marimo.io/notebooks/nb\\_Un58ptNxc9Dp5a1fNkUhJ](https://molab.marimo.io/notebooks/nb_Un58ptNxc9Dp5a1fNkUhJ)

## References

1. Lindsey, J. (2025). Emergent Introspective Awareness in LLMs. *arXiv:2601.01828*.
2. Korbak, T., et al. (2025). Mechanisms of Introspective Awareness. *arXiv:2603.21396*.
3. Luo, A., et al. (2025). Emergent Introspection in AI is Content-Agnostic. *arXiv:2603.05414*.
4. Betley, A., et al. (2025). Tell me about yourself. *arXiv:2501.11163*.
5. Binder, F. J., et al. (2024). Looking inward. *arXiv:2410.13787*.
6. Kadavath, S., et al. (2022). Language models (mostly) know what they know. *arXiv:2207.05221*.
7. Li, K., et al. (2024). Inference-Time Intervention. *NeurIPS 2023*.
8. Rimsky, N., et al. (2024). Steering Llama 2 via CAA. *arXiv:2312.06681*.
9. Zou, A., et al. (2023). Representation Engineering. *arXiv:2310.01405*.

## Appendix: Limitations, Dual-Use, and Ethical Considerations

**Over-attribution.** Results do not establish that the model experiences or is aware of injected concepts. "Detection" is a logit-level signal never verbalised at frozen strength. We describe findings as "improved discriminability," not "awareness."

**Under-attribution.** The threshold hierarchy shows that absence of verbal report does not imply absence of behavioural influence. Relevant to welfare assessment but does not establish welfare-relevant states in current models.

**Distressing outputs.** At  $\rho \geq 3.0$ , welfare terms leak into generation. Measured and reported; not induced unnecessarily.

**Dual use.** The gradient construction could manipulate self-report. The identification null (0.000) limits both interpretive force and misuse potential.

**Protocol corrections.** Three errors found and corrected: (1) injection-position bug ( $\sim 70x$  signal loss); (2) probability-scale measurement (false positive); (3) lexical-prior confound (+12 pts identification inflation). Reported because they affect the paradigm generally.

## LLM Usage Statement

An AI assistant (Devin/Claude) was used for experiment implementation and analysis. All designs and claims were set and verified by the human author. All numbers reproduced across three independent compute sessions.