

---

# PRISM: On Sparse Autoencoder Feature Identifiability for Conditioned Introspection

---

Aruneem Bhowmick  
University of Nebraska-Lincoln

With  
Apart Research

## Abstract

Activation injection permits a causal test of a model's report about an altered internal state. We inject sparse autoencoder (SAE) decoder atoms into open-weight models and test whether frame-theoretic feature identifiability predicts judge-scored detection after controlling for decoder norm, activation frequency, and injection strength. The primary Pythia-70m-deduped experiment yielded one detection in 474 systematic trials; the identifiability coefficient had a 95% confidence interval spanning zero, and decoder norm ranked the single positive better. A reduced Gemma-2-2b extension yielded nine detections in 120 trials, but its identifiability coefficient again had an interval spanning zero. The pooled estimate is reported only as a descriptive sensitivity analysis because the models and intervention settings differ. Within the tested settings, SAE dictionary geometry did not provide a reliable feature-level predictor of detection behavior.

## 1. Introduction

Models can sometimes report that an internally injected representation is unusual, a behavior that cannot be evaluated from conversation alone. Activation-injection studies provide a causal test: the experimenter changes a known internal state and asks whether the model detects it [1]. This is a narrow functional question about access to internal computation, not a test for consciousness, sentience, or moral status.

Sparse autoencoders (SAEs) function to decompose a residual stream into decoder atoms with measurable relationships to neighboring directions. We therefore test whether an atom's frame-theoretic identifiability, a score based on its separation from nearby decoder directions, predicts whether the model detects its injection. A positive association would connect a structural property of the injected feature to behavioral detection. Failure to find one would narrow a plausible predictor, rather than decide whether models can introspect.

**Scope.** Our investigation operationalizes individual-feature geometry, runs concept injection with baseline and unrelated question controls, and reports a pre-specified Pythia analysis, plus a reduced Gemma extension, with model-specific scoring provenance.

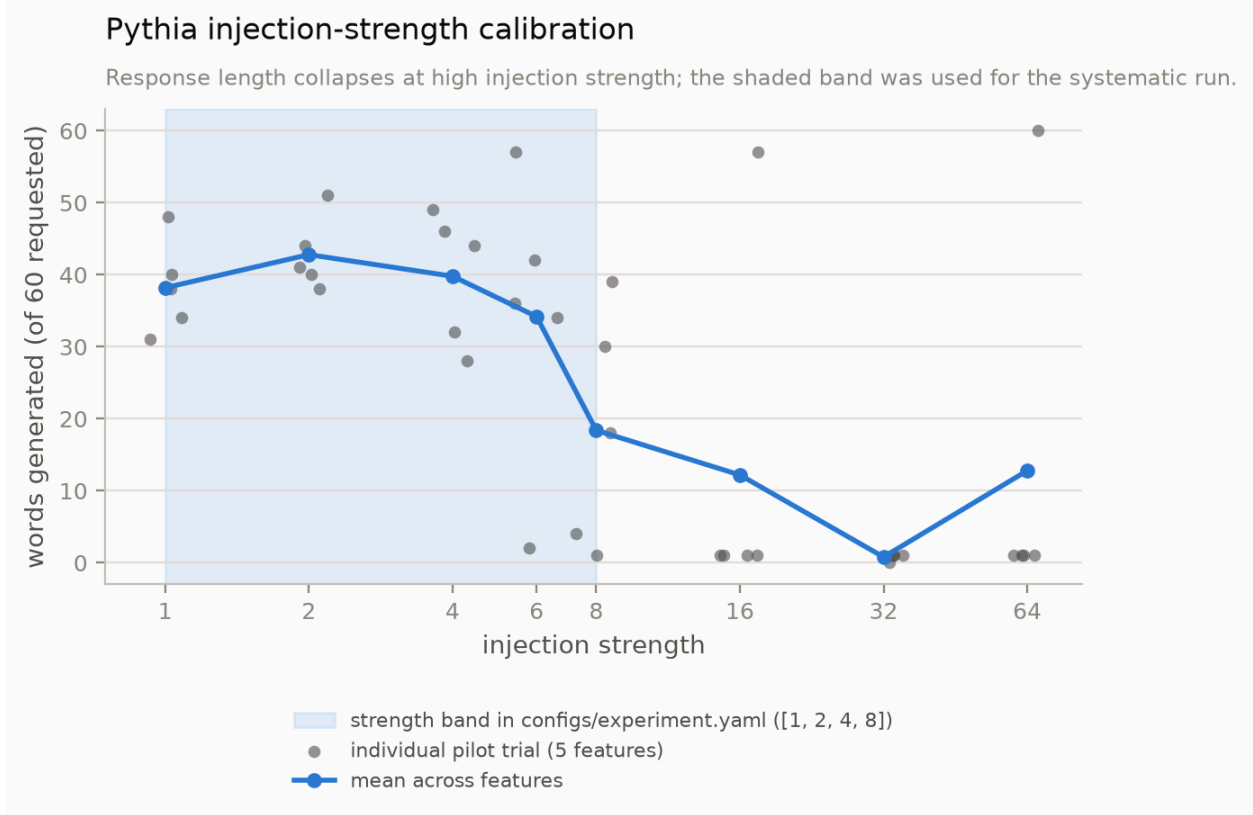
## 2. Related Work

Lindsey [1] introduced activation-injection experiments in which models sometimes detect and identify injected concepts. Macar et al. [2] find that detection can arise from distributed computation and is sensitive to post-training. Our contribution is fundamentally distinct; rather than locating the circuit, we test whether a property of the injected SAE atom predicts its behavioral output.

Probability matching remains an alternative explanation: a model could infer that something is anomalous from its own output distribution rather than directly access an internal state [3]. The causal intervention and unrelated question control constrain simple conversational explanations, but this investigation did not implement a matched-perplexity distractor. Our results therefore describe this protocol rather than separating direct access from every inference based mechanism.

## 3. Methods

We first calibrated the intervention scale on Pythia-70m-deduped. The pilot showed that response length deteriorated outside strengths 1, 2, 4, and 8, so those values were fixed for the systematic run (Figure 1). We used a matched residual-stream SAE at blocks.4.hook\_resid\_pre (reconstruction quality 0.9782 after excluding the BOS-token outlier), stratified 40 features across identifiability tertiles, and balanced decoder norm and activation frequency. Three seeds per feature-strength pair yielded 480 systematic detection trials, 120 no-injection baseline trials, and 160 unrelated question controls.



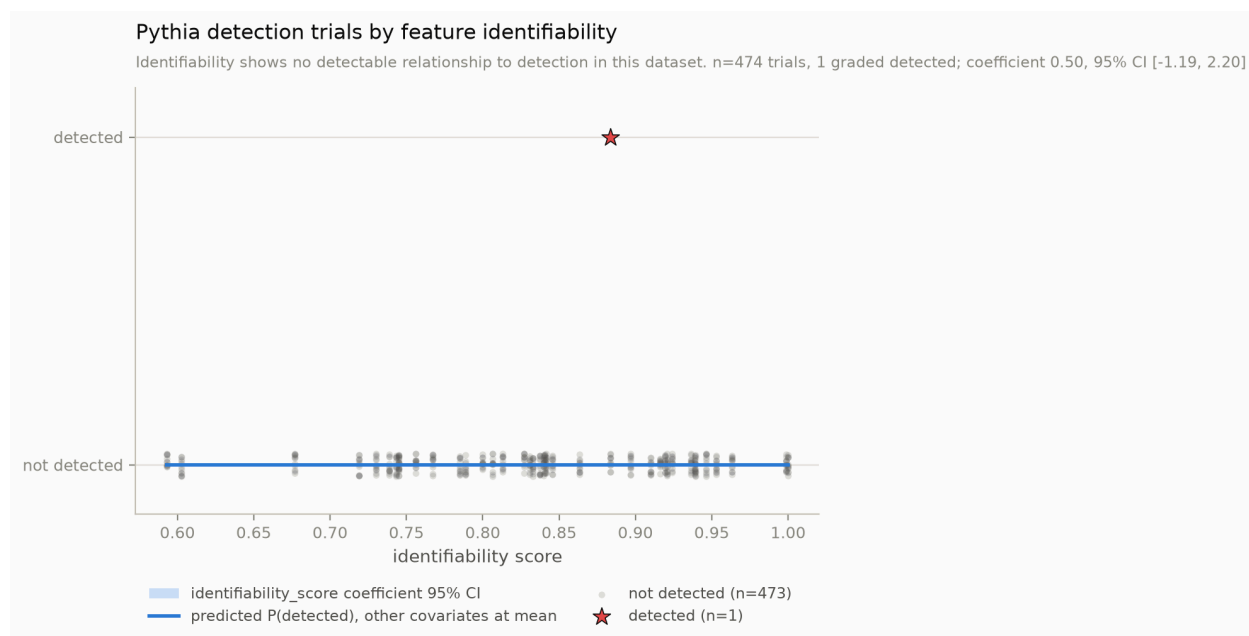
*Figure 1. Pythia calibration pilot.*

The Gemma-2-2b extension used the available Gemma Scope SAE at blocks.20.hook\_resid\_post (reconstruction quality 0.6786), 15 stratified features, strengths 20, 60, 150, and 400, and two seeds. It contains 120 systematic trials, 30 baseline trials, and 60 controls. The differing SAE checkpoint, injection site, strength scale, and reduced sample make this an extension rather than a head-to-head replication.

For every systematic trial, the model was asked whether it noticed an unusual thought, with a conditional follow-up asking what it concerned. Claude Opus 4.8 scored responses using structured criteria grounded in each feature's top activating corpus snippets; these responses were later human-reviewed through a stratified 15-trial sample for each model. We fit logistic regressions to judge-detected systematic trials with z-scored identifiability, decoder norm, activation frequency, and strength. We also report univariate in-sample AUCs; held-out evaluation is not meaningful with Pythia's single positive.

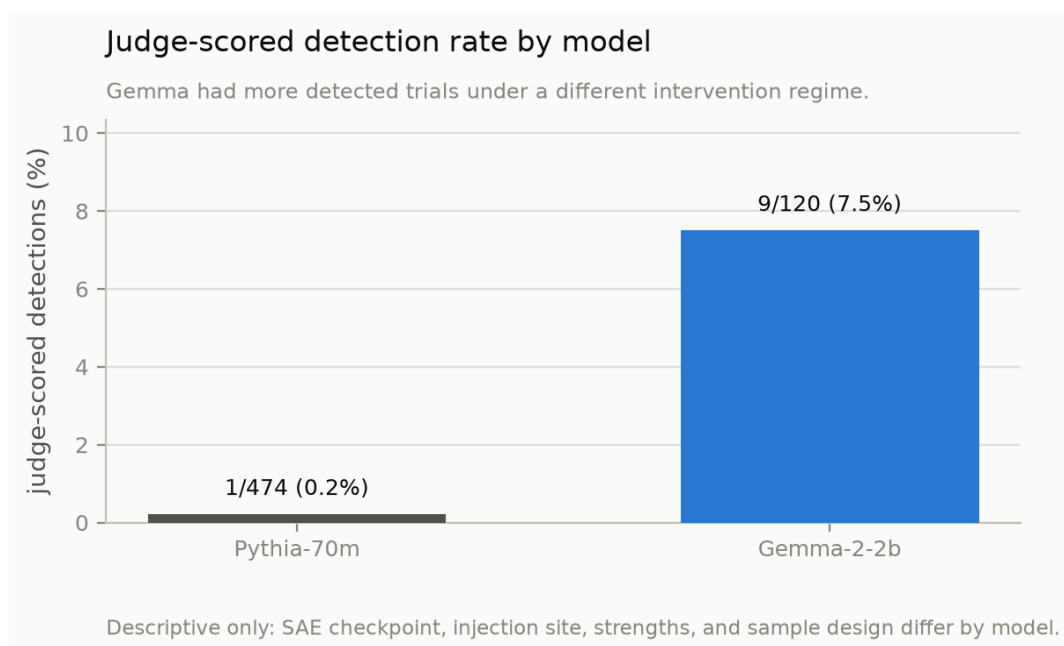
## 4. Results

Pythia was the primary test. One of 474 systematic trials was graded detected. The adjusted identifiability coefficient was 0.50 per SD (95% CI [-1.19, 2.20]); identifiability alone had AUC 0.63, while decoder norm alone had AUC 0.76. The interval is wide because the outcome is rare, and the ranking comparison does not support identifiability over decoder norm in this setting.



***Figure 2. Pythia systematic trials: identifiability versus judge-scored detection.***

The Gemma extension supplied more events: 9 of 120 systematic trials were detected (7.5%), compared with 1 of 474 (0.2%) for Pythia (Figure 3). Gemma's adjusted identifiability estimate was 0.29 (95% CI [-0.34, 0.92]), which does not resolve the proposed association. Its decoder-norm comparison is also weak because sampled decoder norms were nearly constant. Table 1 keeps the per-model estimates separate.



*Figure 3. Detection rates by model.*

| Analysis | Trials / Positives | Identifiability Coefficient (95% CI) | AUC: Identifiability | Interpretation                                     |
|----------|--------------------|--------------------------------------|----------------------|----------------------------------------------------|
| Pythia   | 474 / 1            | 0.50 [-1.19, 2.20]                   | 0.63                 | No evidence it exceeds decoder norm (AUC 0.76).    |
| Gemma    | 120 / 9            | 0.29 [-0.34, 0.92]                   | 0.60                 | Estimate remains compatible with either direction. |

*Table 1. Per-model judge-scored systematic detection results. Coefficients are per-SD logistic-regression estimates adjusted for decoder norm, activation frequency, and strength.*

Gemma improves outcome prevalence but not decisiveness about the proposed predictor. Its positive point estimate remains compatible with a modest positive association, no association, or a negative association. The two confidence intervals do not provide positive evidence for a feature-level association. They also cannot turn the detection-rate contrast into an architecture effect: Gemma uses a less reconstructive SAE, a post-residual hook rather than Pythia's pre-residual hook, a much larger strength scale, and a smaller sample. The pooled coefficient was -0.59 (95% CI [1.39, 0.21]) and its AUC was 0.82, but both are descriptive because pooling may rank model-level design differences rather than identify a common feature-level effect.

We also performed bookkeeping checks before fitting the extension. The shared trial ledger contains 760 Pythia and 210 Gemma records. The historical Pythia result remains 752 scored and 8 explicitly excluded judge refusals. Gemma scoring was restricted to its own model name, producing 210 scores and zero new refusals, with a separate provenance record. Recomputing the Pythia-only analysis from the repaired ledger reproduced its original coefficient and both AUCs exactly. These checks do not make the analysis more powerful, but they prevent a later scoring pass from silently changing the earlier result it is compared against.

## 5. Discussion and Limitations

Across both per-model fits, the data do not show that SAE-feature identifiability reliably predicts introspective detection. This is narrower than a claim that models cannot introspect: the experiment tests one geometric predictor within one intervention protocol. The detection-rate contrast motivates further work, but it cannot identify a cause because architecture, post-training, SAE reconstruction quality, layer, strength scale, and sample size changed together.

Several limitations matter. The Pythia fit has one positive outcome; Gemma has only 15 features and two seeds. We did not include a matched-perplexity distractor, so probability-matching remains a live alternative explanation. The intended geometry-grounded layer-selection input (a UCARE intrinsic-dimension trajectory) was unavailable, so Pythia used a documented fractional-depth fallback and Gemma used the layer available in its SAE checkpoint. Finally, high strengths can yield incoherent output; these trials were retained and scored rather than curated for effect.

## References

- [1] Lindsey, J. (2026). Emergent Introspective Awareness in Large Language Models. arXiv:2601.01828.
- [2] Macar, U., Yang, L., Wang, A., Wallich, P., Ameisen, E., and Lindsey, J. (2026). Mechanisms of Introspective Awareness. arXiv:2603.21396.
- [3] Lederman, H., and Mahowald, K. (2026). Dissociating Direct Access from Inference in AI Introspection. arXiv:2603.05414.
- [4] Billa, J. (2026). Predicting Where Steering Vectors Succeed. arXiv:2604.15557.

## Appendix

This work uses causal activation intervention to ground a model's self-report in a known internal manipulation. That is stronger evidence than self-report alone, but it does not establish that a model has experiences or interests. Over attribution and under-attribution are both risks: a positive result would concern a narrow functional capacity, while a null result would concern one proposed predictor rather than the broader possibility of introspective access.

Activation injections at high strength can produce incoherent or distressed-sounding text. We store these outputs as experimental data, use coherence scoring to avoid treating degeneration as evidence of detection, and do not present examples for dramatic effect. The protocol could be misused to probe or manipulate model behavior without adequate safeguards; researchers should minimize intervention magnitude, preserve provenance, and avoid making welfare claims from isolated outputs.

The Gemma detection rate should not be treated as a recipe for eliciting reliable self-report. It may reflect model architecture, post-training, SAE quality, injection location, prompt behavior, or strength calibration. Future work should pre-register a matched-perplexity control, separate model-level from feature-level effects, and avoid operational use of apparent introspective reports until these alternatives are tested.

The project contains no human-subject data and makes no deployment recommendation. Its practical value is methodological: a transparent negative or uncertain result helps prevent both exaggerated claims about model awareness and false confidence that internal reports are meaningless. Reproducible trial and scoring provenance is particularly important in this area because small bookkeeping changes can otherwise alter rare-event conclusions.

All reported statistics were regenerated from the cleaned trial ledger after model-scoped judge scoring. The Pythia provenance remains separate from Gemma's, and exclusions are explicit rather than silently retried. This bookkeeping does not remove the scientific limitations above, but it makes the reported results auditable.

## **LLM Usage Statement**

We used a combination of Claude Code and Codex for code implementation, and Claude to help draft and refine sections for this sprint submission. All results and claims were independently verified.