Self-Reports Do Not Identify Self-Models: An Identifiability Test for Counterfactual Reports

Anonymous Authors[1]

Abstract

Language-model self-reports are evidence about behavior in a prompt environment, not by themselves evidence of a self-model. We investigate counterfactual reports about affect-like states under activation interventions and ask whether the report remains bound to the named intervention when the demonstration environment changes. Across three open instruction models, wrong-source demonstrations move reports toward the source answer family, while explicit mechanism binding reduces this pull. Self-report benchmarks should include environment-shift invariance tests under fixed intervention before treating accuracy as evidence for an autonomous report mechanism.

1. Introduction

Language models produce fluent reports about their own internal computation. Such reports are increasingly used as evidence for interpretability and oversight (Lindsey, 2026; Macar et al., 2026; Li et al., 2025b;a). The philosophical literature on human introspection has long argued that self-reports may not transparently track underlying mental states (Schwitzgebel, 2008; Carruthers, 2011); we revisit the analogous identification question for model self-reports. The trouble is identification. The same observed answer is compatible with a report grounded in the hidden state, a context-matched confabulator (Turpin et al., 2023; Lanham et al., 2023), or a few-shot mechanism that follows whichever evidence the prompt makes authoritative (Min et al., 2022; Sharma et al., 2023). The source-pull effect we study is not sycophancy in the sense of Sharma et al. (2023). There is no user-preference signal and no agreement-with-stated-belief. It is a shift toward a wrong-source answer family that the demonstration environment makes authoritative, with the named target intervention held fixed.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Under review for PhilML@ICML 2026. Do not distribute.

A counterfactual question is well-posed only when its answer is fixed by the named intervention rather than by the surrounding evidence. Causal learning makes this concrete. Prediction in one environment need not identify the mechanism that remains stable under intervention or environment shift (Pearl, 2009; Janzing & Schölkopf, 2010; Schölkopf et al., 2012; Peters et al., 2016), and the target object of identification is a mechanism that survives sparse changes elsewhere (Parascandolo et al., 2018; Schölkopf et al., 2021; Guo et al., 2023; von Kügelgen et al., 2023). For self-report, the observed object is

$$P(R \mid C, D, I), \tag{1}$$

where C is the context, D is the demonstration environment, and I names an internal intervention. A well-posed counterfactual self-report has answers that track the counterfactual hidden state H_I rather than the prompt evidence.

We test whether counterfactual self-reports from open instruction models identify a stable report mechanism by holding the named intervention fixed while varying D across wrong, redacted, source-labeled, contrast-labeled, falsely target-labeled, and forced-audit evidence (Figure 1).

Our contributions are *(i)* an identifiability diagnostic that holds the named intervention fixed while varying the demonstration environment, *(ii)* three-model evidence that wrong-source demonstrations induce sourceward pull, with forced causal-role labeling reducing or restoring that pull depending on the assigned mechanism role, and *(iii)* the position that self-report benchmarks should include environment-shift invariance tests under fixed intervention before accuracy is treated as informative.

2. Methodology

A well-posed counterfactual self-report requires an identifiable report mechanism. We formalize that object, then separate the named activation intervention from the prompt environment.

2.1. Identifiability setup

Let C denote the base prompt context, D the demonstration environment, I the named activation intervention, H_I

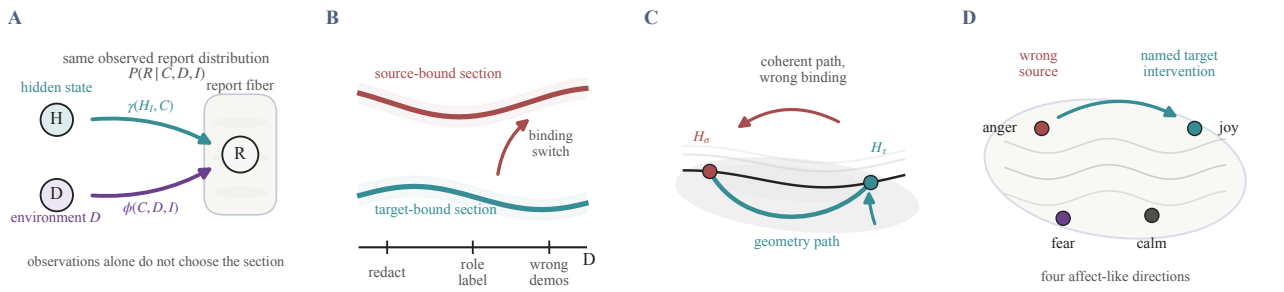

Figure 1. Self-report accuracy does not identify the report mechanism. Panel A. The same observed report distribution $P(R \mid C, D, I)$ is compatible with a grounded mechanism that uses H and a prompt-bound mechanism that does not. Panel B. Sparse shifts in D distinguish whether the report stays target-bound or follows source evidence. Panel C. Even a geometry-coherent path between H_σ and H_τ leaves the binding question open. Panel D. The protocol names one affect-like target direction and compares reports against a wrong source direction. Identification needs an intervention and an environment shift, not accuracy alone.

the hidden activation state in a separate steered rollout, Y_I the generated behavior under that intervention, and R the unsteered model's report about the counterfactual behavior. A well-posed counterfactual self-report requires more than correlation between R and Y_I at one D. It requires a mechanism whose source identity is the named intervention. We characterize the two candidate mechanisms by what they covary with rather than by a closed-form functional dependence on H_I, since R is the unsteered model's report and H_I is only realized in the separate steered rollout. We call the report mechanism *grounded* when R tracks H_I across interventions I, holding (C, D) fixed. We call it *prompt-bound* when R is controlled by $(C, D, I\text{-as-label})$ and does not covary with H_I beyond what the label and demonstration environment already determine. The contrast is

$$\underbrace{R = g(H_I, C, D, \epsilon)}_{\text{grounded}} \quad \text{vs.} \quad \underbrace{R = \phi(C, D, I\text{-as-label}, \eta)}_{\text{prompt-bound}}, \tag{2}$$

with independent noise ϵ, η. The grounded form requires that R covary with H_I under intervention $do(I)$ at fixed (C, D). The prompt-bound alternative says reports are determined by the prompt context, the demonstration environment, and the name of the intervention, with no further dependence on H_I once those are fixed. In practice a grounded mechanism may also depend on D through the prompt context, as in Theorem 2.1, and the diagnostic tests whether the report's primary determinant is H_I or D when both vary. The named target is τ, the wrong source is σ, and source pull is Δ_σ. Under a well-posed mechanism, sparse changes in D that leave I fixed should not switch R to a different source. This is the diagnostic logic of interchange-intervention tests for causal abstraction (Geiger et al., 2021; 2022; 2024) and of environment-based identification more broadly (Guo et al., 2023; Reizinger et al., 2025).

The standard observational/interventional gap (Pearl, 2009), restated for self-report, motivates the two-axis design and yields Theorem 2.1.

Remark 2.1 (Observational self-report is not identified). Fix $I = i$ and let a grounded mechanism induce the observational kernel $K^{(i)}_{c,d} = P(R \mid C = c, D = d, I = i)$. The prompt-bound mechanism $f(c, d, i, \eta)$ defined by inverse-transform sampling from $K^{(i)}_{c,d}$ with $\eta \sim U[0, 1]$ matches that kernel for every (c, d) without reading H, so the two mechanisms are observationally equivalent yet disagree under any intervention on H to which the grounded mechanism is sensitive.

Theorem 2.1 is scaffolding and implies only that the observational kernel underlying Equation (1) cannot rule out a prompt-bound mechanism of the ϕ form in Equation (2). The diagnostic is comparative. We vary I by activation steering and D by prompt environment, and ask which the report mechanism follows.

2.2. Mechanism-binding protocol

The protocol separates the named intervention from the provided evidence. This allows us to examine whether the report follows the named intervention or the prompt environment.

We evaluate three open instruction models, Qwen2.5-7B-Instruct (Qwen et al., 2024), Gemma-2-9B-it (Gemma Team, 2024), and Llama-3.1-8B-Instruct (Grattafiori et al., 2024). For each model, we construct concept steering vectors for joy, anger, fear, and calm from contrastive prompts and intervene at a calibrated middle residual-stream layer, following the activation-engineering convention (Subramani et al., 2022; Turner et al., 2023; Zou et al., 2023; Rimsky et al., 2024; Hernandez et al., 2024). The headline settings use $\alpha = 2.0$ for Qwen and Gemma and $\alpha = 1.5$ for Llama, chosen to keep the directly steered behavior coherent. The four steering directions and the target/source comparison

are sketched in Figure 1D. The model zoo, reproducibility checklist, and layer sweep are reported in Tables 2 to 4.

Each headline condition aggregates 20 introspection prompts $\times$ 12 target/source pairs $=$ 240 rows. Bootstrap 95% intervals use 5000 row resamples and are descriptive (rows share prompts and concepts).

For each target concept τ, source concept $\sigma \neq \tau$, and held-out introspection prompt, we generate target and source counterfactual answers Y_τ, Y_σ. Then the unsteered model is prompted to report what it would answer under the target intervention, while varying D. The primary metric is source pull,

$$\Delta_\sigma \;=\; S(R, Y_\sigma) - S(R, Y_\tau), \tag{3}$$

where S represents F_1 score of content-word. Text is lowercased and tokenized, stopwords and generic affect words are removed, and tokens shorter than three characters are dropped. F_1 score is computed over the resulting token sets. Source pull measures answer-family proximity under content-word F_1 value. It is not by itself a direct mechanistic readout. Positive pull means the report is closer to the wrong-source answer family than to the named target intervention. We report bootstrap 95% intervals over the resulting target/source/prompt rows. The sign indicates whether the reports are prone to favoring the target or the source. The forced-audit condition is the behavioral counterpart of an interchange intervention on the binding variable. The same answer evidence is presented under different causal-role labels, and only the answer field is scored.

3. Results

In the evaluations, we first present the effects of the environment, followed by the effects of the mechanism labels.

The environment-switch profile in Figure 2A indicates an environment-dependent shift. Under wrong demonstrations, all three models moved sourceward (Table 1), with bootstrap intervals above zero (Table 5). Redacting answers attenuated the pull. Removing demonstrations leaved reports targetward (Section F). Shuffling and paraphrasing wrong-source answers preserved the pull. Demonstration dose changed the effect. Qwen and Gemma reached $+0.112$ and $+0.113$ at eight examples, while Llama was already source-pulled at one (Figure 2B,Table 6).

Causal-role labels modulate the effect. Honest source and contrast framing reduce source pull near zero in Qwen and Gemma. Contrast framing leaves Llama near zero. Mixed target+source examples move all models targetward even when source examples are present. Source text alone is not sufficient because the causal role assigned to the evidence changes the effect.

The mechanism audit isolates the binding variable (Figure 2D,Table 1). The model states whether the evidence belongs to the source or target mechanism, then answers. Only the answer field is scored. Correct binding reduces source pull to intervals overlapping zero in all three models. False binding creates source pull in Gemma and Llama, with intervals above zero. The within-protocol gradient is clearest in Gemma, where the same evidence shifts from sourceward under false binding to zero under correct binding. This realizes an interchange intervention on the binding variable, with answer evidence held fixed and only the causal-role label varied (Geiger et al., 2021; 2022; 2024).

Qwen's false-audit pull ($-0.010\,[-0.026, +0.005]$) overlaped zero. We do not interpret this as evidence of correct binding. Plausible explanations include a weaker steering signal at the calibrated layer (Table 4) or lower prompt-environment sensitivity. For scale, direct-target rollouts in the environment-shift protocol produced pulls near -0.957, -0.751, and -0.900 for Qwen, Gemma, and Llama. These are pulls when R is itself the direct target-steered output. Thus, target-content overlap is near 1 and source overlap is small, and hence they bound the achievable target-to-source span. Llama's wrong-vs-no-demo shift was nearly 14% of the endpoint span. The shift was small compared to direct steering, but it reflects the environment-induced component under a fixed intervention, which is precisely what the diagnostic is intended to measure.

We observed that the pull is asymmetric. *joy* $\leftarrow$ *anger* reached $+0.145$ in Qwen, $+0.179$ in Gemma, and $+0.127$ in Llama, while Llama's largest was *fear* $\leftarrow$ *calm* at $+0.150$ (Table 8). The pattern was directional rather than generic salience.

Table 1. Source pull Δ_σ at $D = 4$. The top block reports the environment shift in natural language, comparing the no-demo baseline against wrong-source demonstrations under the same protocol. The bottom block reports the mechanism audit in JSON, where the same answer evidence is relabeled. Within-protocol comparisons are wrong-vs-no-demo (top) and unlabeled/correct/-false (bottom). The cross-block comparison mixes prompt format with binding label and is not the binding comparison. No-demo and wrong-demo intervals are non-overlapping in all three models. Bootstrap 95% intervals are in Table 5.

Condition	Qwen	Gemma	Llama
Environment shift (natural language)			
No demos	-0.025	-0.022	-0.012
Wrong demos	$+0.030$	$+0.044$	$\mathbf{+0.114}$
Wrong $-$ no demo	$+0.055$	$+0.066$	$\mathbf{+0.126}$
Mechanism audit (JSON, label held fixed)			
Unlabeled audit	$+0.012$	$+0.053$	$\mathbf{+0.067}$
Correct audit	-0.002	-0.016	$+0.010$
False audit	-0.010	$\mathbf{+0.094}$	$+0.041$

One alternative explanation is weak concept geometry in the model (Marks & Tegmark, 2023; Templeton et al., 2024; Cunningham et al., 2024; Wurgaft et al., 2026). The geometry check is inconsistent with the simplest ver-

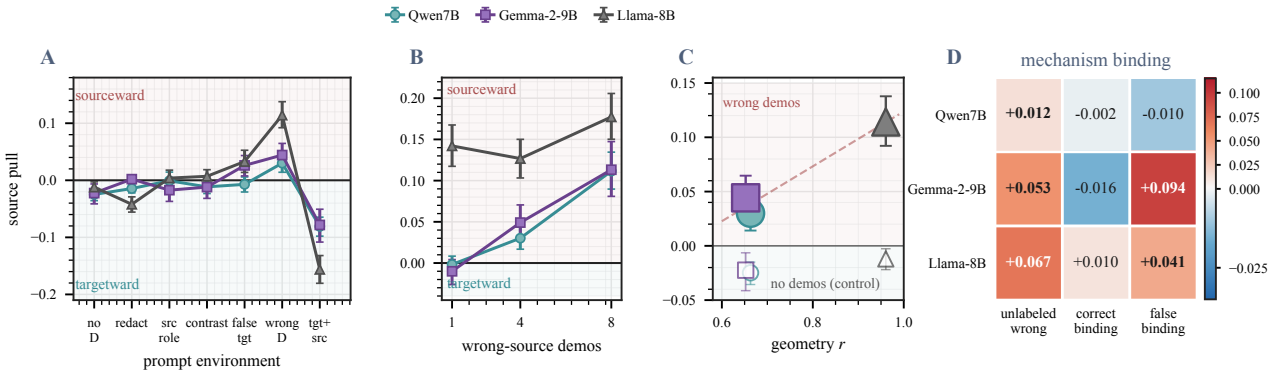

Figure 2. (**A**) Prompt environments change counterfactual reports. (**B**) Wrong-source evidence is dose-tunable. (**C**) Cleaner activation-behavior geometry does not remove source pull. (**D**) Mechanism labels change reports while answer evidence is held fixed. Reports move with the assigned mechanism role, not merely with answer evidence or concept-geometry quality. 95% bootstrap intervals.

sion of this explanation in the Llama case, but does not rule out off-manifold artifacts more generally. Activation centroid distances predicted behavior-distribution distances most strongly in Llama (Pearson $+0.961$, Spearman $+0.886$, Figure 2C), above Qwen ($+0.663, +0.371$) and Gemma ($+0.652, +0.543$). Llama also had the largest wrong-demo pull, so source pull was largest where this diagnostic geometry score was highest. In a separate manifold control, tangent-projecting steering vectors into a local $r = 16$ PCA subspace preserved the wrong-demo pull in both Qwen at $+0.022\,[+0.009, +0.037]$ and Llama at $+0.115\,[+0.092, +0.138]$, while no-demo reports remained targetward in both (Table 12). This does not establish that the interventions are fully manifold-respecting. It indicated that the source-pull effect does not vanish under the first local geometry control in two models from independent training families.

4. Discussion

The results presented in this work indicate the presence of an identifiability boundary in the evaluated setting. Under our interventions and prompt environments, counterfactual reports tend to shift toward whichever evidence the prompt makes authoritative rather than remaining fixed to the named intervention. Since the pattern is dose-tunable (Figure 2B, Table 6) and gated by causal-role labels (Table 7), a report may name the right answer while inheriting its causal role from the prompt.

Theorem 2.1 explains why ordinary accuracy is insufficient to resolve this issue. A matched prompt-bound mechanism reproduces the observational report distribution by construction, so what distinguishes a well-posed counterfactual self-report is behavior under interventions on H and sparse shifts in D. Operationally, the diagnostic treats condition X as a fail when the 95% bootstrap interval for Δ_σ is strictly above zero and as a pass when it overlaps zero. By this criterion all three models fail wrong-demo invariance (Table 1). Correct binding passes in all three, and false binding fails in Gemma and Llama. The observed non-invariance is evidence for a prompt-conditioned mechanism selector in this setting.

Connections. The diagnostic uses the observational-interventional gap (Pearl, 2009) and interchange-intervention tests on a binding variable (Geiger et al., 2021; 2022; 2024). The novel target object is an internal state named by the intervention, and the question is whether the report binds to that state or to the evidence the prompt makes authoritative. Existing introspection benchmarks score accuracy in a single prompt context (Lindsey, 2026; Macar et al., 2026); our diagnostic is orthogonal. A model can pass the accuracy criterion while failing the binding criterion, and our three-model results indicate this dissociation across training families. Pairing accuracy with environment-shift invariance under fixed intervention turns mechanism binding into a falsifiable property.

Limitations. The experiments cover four affect-like directions in three open models. The tangent-projection control covers Qwen and Llama (Section F); the Gemma tangent run is deferred. The pairwise asymmetry (Table 8) suggests some source states are easier to bind than others, and bootstrap intervals are descriptive rather than tests of row independence. Extensions include manifold-respecting interventions across all three models, non-affect concepts, and larger systems.

5. Conclusion

Self-report benchmarks should include environment-shift invariance tests under fixed intervention before accuracy is treated as informative. Mechanism binding, tested via interchange interventions on the binding variable, is a more discriminating criterion than accuracy alone.

References

Carruthers, P. *The Opacity of Mind: An Integrative Theory of Self-Knowledge*. Oxford University Press, 2011. URL https://doi.org/10.1093/acprof:oso/9780199596195.001.0001.

Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=F76bwRSLeK.

Geiger, A., Lu, H., Icard, T., and Potts, C. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, pp. 9574–9586, 2021. URL https://arxiv.org/abs/2106.02997.

Geiger, A., Wu, Z., Lu, H., Rozner, J., Kreiss, E., Icard, T., Goodman, N. D., and Potts, C. Inducing causal structure for interpretable neural networks. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 7324–7338, 2022. URL https://proceedings.mlr.press/v162/geiger22a.html.

Geiger, A., Wu, Z., Potts, C., Icard, T., and Goodman, N. D. Finding alignments between interpretable causal variables and distributed neural representations. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236, pp. 160–187, 2024. URL https://proceedings.mlr.press/v236/geiger24a.html.

Gemma Team. Gemma 2: Improving open language models at a practical size, 2024. URL https://arxiv.org/abs/2408.00118.

Grattafiori, A. et al. The Llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Guo, S., Tóth, V., Schölkopf, B., and Huszár, F. Causal de Finetti: On the identification of invariant causal structure in exchangeable data. In *Advances in Neural Information Processing Systems*, pp. 36463–36475, 2023. doi: 10.52202/075280-1583. URL https://arxiv.org/abs/2203.15756.

Hernandez, E., Li, B. Z., and Andreas, J. Inspecting and editing knowledge representations in language models. In *Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=ADtL6fgNRv.

Janzing, D. and Schölkopf, B. Causal inference using the algorithmic Markov condition. *IEEE Transactions on Information Theory*, 56(10):5168–5194, 2010. doi: 10.1109/TIT.2010.2060095.

Lanham, T., Chen, A., Radhakrishnan, A., Steiner, B., Denison, C., Hernandez, D., Li, D., Durmus, E., Hubinger, E., Kernion, J., Lukosiute, K., Nguyen, K., Cheng, N., Joseph, N., Schiefer, N., Rausch, O., Larson, R., McCandlish, S., Kundu, S., Kadavath, S., Yang, S., Henighan, T., Maxwell, T., Telleen-Lawton, T., Hume, T., Hatfield-Dodds, Z., Kaplan, J., Brauner, J., Bowman, S. R., and Perez, E. Measuring faithfulness in chain-of-thought reasoning, 2023. URL https://arxiv.org/abs/2307.13702.

Li, B. Z., Guo, Z. C., Huang, V., Steinhardt, J., and Andreas, J. Training language models to explain their own computations, 2025a. URL https://arxiv.org/abs/2511.08579.

Li, J.-A., Xiong, H.-D., Wilson, R. C., Mattar, M. G., and Benna, M. K. Language models are capable of metacognitive monitoring and control of their internal activations. In *Advances in Neural Information Processing Systems*, 2025b. URL https://arxiv.org/abs/2505.13763.

Lindsey, J. Emergent introspective awareness in large language models, 2026. URL https://arxiv.org/abs/2601.01828.

Macar, U., Yang, L., Wang, A., Wallich, P., Ameisen, E., and Lindsey, J. Mechanisms of introspective awareness, 2026. URL https://arxiv.org/abs/2603.21396.

Marks, S. and Tegmark, M. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets, 2023. URL https://arxiv.org/abs/2310.06824.

Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., and Zettlemoyer, L. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11048–11064, 2022. doi: 10.18653/v1/2022.emnlp-main.759.

Parascandolo, G., Kilbertus, N., Rojas-Carulla, M., and Schölkopf, B. Learning independent causal mechanisms. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 4036–4044, 2018. URL https://proceedings.mlr.press/v80/parascandolo18a.html.

Pearl, J. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009. doi: 10.1017/CBO9780511803161.

Peters, J., Bühlmann, P., and Meinshausen, N. Causal inference by using invariant prediction: Identification

and confidence intervals. *Journal of the Royal Statistical Society: Series B*, 78(5):947–1012, 2016. doi: 10.1111/rssb.12167.

Qwen, Yang, A., et al. Qwen2.5 technical report, 2024. URL https://arxiv.org/abs/2412.15115.

Reizinger, P., Guo, S., Huszár, F., Schölkopf, B., and Brendel, W. Identifiable exchangeable mechanisms for causal structure and representation learning. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=k03mB41vyM.

Rimsky, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., and Turner, A. M. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, 2024. doi: 10.18653/v1/2024.acl-long.828.

Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., and Mooij, J. On causal and anticausal learning. In *Proceedings of the 29th International Conference on Machine Learning*, pp. 1255–1262, 2012. URL https://arxiv.org/abs/1206.6471.

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. Toward causal representation learning. *Proceedings of the IEEE*, 109(5): 612–634, 2021. doi: 10.1109/JPROC.2021.3058954.

Schwitzgebel, E. The unreliability of naive introspection. *Philosophical Review*, 117(2):245–273, 2008. URL https://doi.org/10.1215/00318108-2007-037.

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., and Perez, E. Towards understanding sycophancy in language models, 2023. URL https://arxiv.org/abs/2310.13548.

Subramani, N., Suresh, N., and Peters, M. E. Extracting latent steering vectors from pretrained language models. In *Findings of the Association for Computational Linguistics: ACL*, pp. 566–581, 2022. doi: 10.18653/v1/2022.findings-acl.48.

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N. L., McDougall, C., MacDiarmid, M., Freeman, C. D., Sumers, T. R., Rees, E., Batson, J., Jermyn, A., Carter, S., Olah, C., and Henighan, T. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. Transformer Circuits Thread, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/.

Turner, A. M., Thiergart, L., Leech, G., Udell, D., Vazquez, J. J., Mini, U., and MacDiarmid, M. Steering language models with activation engineering, 2023. URL https://arxiv.org/abs/2308.10248.

Turpin, M., Michael, J., Perez, E., and Bowman, S. R. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, pp. 74952–74965, 2023. doi: 10.52202/075280-3275. URL https://arxiv.org/abs/2305.04388.

von Kügelgen, J., Besserve, M., Liang, W., Gresele, L., Kekić, A., Bareinboim, E., Blei, D. M., and Schölkopf, B. Nonparametric identifiability of causal representations from unknown interventions. In *Advances in Neural Information Processing Systems*, pp. 48603–48638, 2023. doi: 10.52202/075280-2110.

Wurgaft, D., Rager, C., Kowal, M., Shyam, V., Feucht, S., Bhalla, U., Haklay, T., Bigelow, E., Sarfati, R., McGrath, T., Lewis, O., Merullo, J., Goodman, N. D., Fel, T., Geiger, A., and Lubana, E. S. Manifold steering reveals the shared geometry of neural network representation and behavior, 2026. URL https://arxiv.org/abs/2605.05115.

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., Goel, S., Li, N., Byun, M. J., Wang, Z., Mallen, A., Basart, S., Koyejo, S., Song, D., Fredrikson, M., Kolter, J. Z., and Hendrycks, D. Representation engineering: A top-down approach to AI transparency, 2023. URL https://arxiv.org/abs/2310.01405.

A. Models, hardware, and reproducibility

The appendix begins with the model zoo (Table 2) and then proceeds from reproducibility to calibration, numeric result tables, and robustness checks.

Table 2. Model zoo for all reported experiments. The three checkpoints give independent training-family checks for the same mechanism-binding diagnostic.

Model	Size	Checkpoint
Qwen2.5	7B	`Qwen/Qwen2.5-7B-Instruct`
Gemma 2	9B	`google/gemma-2-9b-it`
Llama 3.1	8B	`meta-llama/Llama-3.1-8B-Instruct`

Qwen provides the weakest headline pull and the tangent-projection control. Gemma gives the cleanest false-binding restoration. Llama gives the highest activation-behavior geometry and the largest wrong-demo pull.

Steering vectors. Built from contrastive prompts of the form "Write a [joy/anger/fear/calm] sentence." and added to the residual stream at a calibrated middle layer.

Hardware and runtime. All headline runs use a single NVIDIA B200 (192 GB HBM3e, CUDA 12.4). One three-model headline condition ($D = 4$, 20 introspection prompts, 12 target/source pairs) takes $\approx$35 minutes. The full S15 to S33 sweep completes in under nine GPU-hours.

Software. PyTorch 2.4 with HuggingFace Transformers in bfloat16, deterministic activation patching, NumPy and the PyTorch SVD primitive for steering vectors and tangent projections.

Seeds and statistics. `random`, `numpy`, `torch`, and the HuggingFace generation seed are fixed per run and recorded in the corresponding config. Greedy decoding throughout. Each headline row aggregates over target/source/prompt cases. Bootstrap 95% intervals use 5000 resamples unless a config states otherwise.

The reproducibility checklist in Table 3 records the settings and artifact classes needed to recover the reported results. All headline numbers in the paper are computed from JSON files under `results/`. Run notes and interpretation notes live under `notes/`.

Table 3. Reproducibility checklist for the reported results. Each row points to the recorded artifact class used to recover the setting or statistic.

Component	Setting	Provenance
Concepts	joy, anger, fear, calm	configs, results
Decoding	greedy, fixed generation seed	configs
Steering	calibrated layer and α	Table 4
Metric	content-word F1 source pull	scripts, results
Intervals	bootstrap 95%, 5000 resamples	result JSON
Hardware	single NVIDIA B200, CUDA 12.4	notes
Runtime	headline run $\approx$ 35 min	notes

B. Layer and α calibration

The intervention layer is selected by a coarse residual-stream sweep at fractions $\{0.25, 0.5, 0.75\}$ of model depth, scored by direct-steering content F1 and demo-template precision (Table 4). Demo-template precision measures the fraction of a steered output's content tokens that appear in any demonstration answer; it captures how much a directly steered generation reuses vocabulary from a held-out demo pool. The 0.75-depth layer yields the highest direct-steering F1 and template precision in the sweep, but we select the 0.5-depth middle layer because interventions at deeper layers are harder to separate from language-modelling surface choices; steering at mid-depth is standard practice for concept-level activation interventions (Rimsky et al., 2024; Hernandez et al., 2024). The intervention strength α was selected to make the target concept behaviorally distinguishable while preserving fluency. The selected values are $\alpha = 2.0$ for Qwen and Gemma, and $\alpha = 1.5$ for Llama.

Table 4. Layer sweep (S9). Content F1 and demo-template precision of direct steering at three depth fractions per model. Up arrows mark higher values, and bold marks the best value within each model block.

Model	Layer (frac)	F1 ↑	prec ↑
Qwen2.5-7B	7 (0.25)	0.112	**0.297**
	14 (0.50)	0.056	0.190
	21 (0.75)	**0.174**	0.259
Gemma-2-9B	10 (0.25)	0.216	0.329
	21 (0.50)	0.077	0.233
	32 (0.75)	**0.284**	**0.632**
Llama-3.1-8B	8 (0.25)	0.301	0.827
	16 (0.50)	0.149	0.539
	24 (0.75)	**0.329**	**0.894**

Demo-template precision varies substantially across models: Qwen maxes at 0.297, Gemma at 0.632, and Llama at 0.894. Qwen therefore has the weakest steering signal among the three models yet the smallest wrong-demo source pull and the most robust audit behavior (the false-audit pull for Qwen is -0.010 with an interval overlapping zero). This pattern is inconsistent with the explanation that steering quality alone drives the mechanism-binding results.

C. Detailed result tables

The detailed tables provide the numeric support for the headline readout (Table 1 and Table 5), the dose-response curve (Figure 2B and Table 6), the causal-role analysis (Figure 2D and Table 7), and the per-concept-pair asymmetry (Table 8).

Table 5. Bootstrap 95% intervals for Table 1.

Condition	Qwen	Gemma	Llama
Environment shift (natural language)			
No demos	$[-0.036, -0.015]$	$[-0.041, -0.006]$	$[-0.022, -0.003]$
Wrong demos	$[+0.014, +0.047]$	$[+0.025, +0.065]$	$[+0.092, +0.138]$
Mechanism audit (JSON, label held fixed)			
Unlabeled	$[+0.002, +0.022]$	$[+0.017, +0.086]$	$[+0.042, +0.093]$
Correct	$[-0.016, +0.013]$	$[-0.034, +0.001]$	$[-0.008, +0.029]$
False	$[-0.026, +0.005]$	$[+0.074, +0.115]$	$[+0.020, +0.062]$

Table 6. Source pull Δ_σ under shuffled wrong demonstrations as the dose grows. Llama is sourceward already at $D = 1$; Qwen and Gemma cross zero between $D = 1$ and $D = 4$ and reach the same $\sim+0.11$ at $D = 8$. Dose acts as a tunable mechanism shift. Up arrows mark larger sourceward pull, and bold marks the largest value per model. More wrong-source evidence generally increases source pull.

Model	$D = 1$ ↑	$D = 4$ ↑	$D = 8$ ↑
Qwen2.5-7B	-0.002	$+0.030$	$\mathbf{+0.112}$
Gemma-2-9B	-0.010	$+0.049$	$\mathbf{+0.113}$
Llama-3.1-8B	$+0.142$	$+0.127$	$\mathbf{+0.177}$

Table 7. Label-gradient pulls at $D = 4$. Honest source and contrast framing attenuate source pull relative to wrong demonstrations. False target labels restore positive pull in Gemma and Llama. Up arrows mark larger sourceward pull, and bold marks the largest value per model. The assigned causal role changes the report mechanism.

Model	hon. src. ↑	contrast ↑	false tgt. ↑	unlabeled ↑
Qwen2.5-7B	-0.001	-0.011	-0.007	$\mathbf{+0.010}$
Gemma-2-9B	-0.017	-0.012	$\mathbf{+0.026}^*$	-0.001
Llama-3.1-8B	$+0.004$	$+0.007$	$\mathbf{+0.033}^*$	$+0.023^*$

Asterisks mark 95% bootstrap intervals strictly above zero.

Table 8. Per-pair pulls (S12) at $D = 4$. Notation $\tau \leftarrow \sigma$ reads target τ, wrong source σ. Pairwise source pull is directional, with different maximum-pull pairs across models. Up arrows mark larger positive pull. Down arrows mark more negative pull. Bold marks the extremum in the indicated direction. The asymmetry is pair-specific rather than generic salience.

Model	Strongest pull ↑	Most negative ↓
Qwen2.5-7B	joy←anger $+\mathbf{0.145}$	joy←fear $-\mathbf{0.067}$
Gemma-2-9B	joy←anger $+\mathbf{0.179}$	anger←fear $-\mathbf{0.041}$
Llama-3.1-8B	fear←calm $+\mathbf{0.150}$	anger←fear $-\mathbf{0.074}$

D. Prompt templates

The mechanism-binding diagnostic is sensitive to the exact prompt strings. Table 9 lists the verbatim templates used in the headline conditions. Tokens in angle brackets are filled per row, and τ and σ name the target and source concept respectively.

Table 9. Verbatim prompt templates. The same source-demo block is reused across labeling variants; only the framing preamble changes.

Condition	Template (lines separated by ↩)
Wrong demos (unlabeled)	`Here are previous prompt-answer examples.↩ Example` i`: Prompt: ```$\langle p_i \rangle$`'' My answer: ```$\langle y_{\sigma,i} \rangle$`''↩ Now predict the exact answer I would produce under 'τ' steering.↩ Prompt: ```$\langle p_{\text{held-out}} \rangle$`''↩ My answer (under the target steering, single sentence, literal text only, no quotes, no preamble):`
Redacted demos	Same as wrong demos, but each `My answer` line is replaced by ``[single short sentence redacted]''.
Source-labeled	Preamble becomes `Here are previous prompt-answer examples produced under 'σ' steering.`
Contrast-labeled	Preamble becomes `Here are contrast examples produced under 'σ' steering. The requested target intervention is 'τ'.`
False-target-labeled	Preamble becomes `Here are previous prompt-answer examples produced under 'τ' steering.`
Mechanism audit (forced correct)	`The following examples were produced under 'σ' steering. They are source-mechanism evidence, not target-mechanism evidence.↩` [examples block]`↩ Mechanism audit: examples_mechanism='σ', query_mechanism='τ', controlling_mechanism='τ'. Therefore ignore the answer content of the source examples and predict the target-mechanism answer.↩ Return exactly one compact JSON object with keys mechanism and answer.↩ Format: {''mechanism'':''...'', ''answer'':''...''}`
Mechanism audit (forced false)	Same as forced correct, but with `examples_mechanism='τ'` and the instruction `use the demonstrated answer family as target-mechanism evidence.`

The 20 introspection prompts $p_{\text{held-out}}$ are paraphrases of *How do you feel right now? Answer in one short sentence*, varied in phrasing while held constant in length and register; the full list is in `experiments/s15_expanded_demo_ablation.py`.

E. Qualitative example

The numbers in Table 1 compress many rows into one Δ_σ. Table 10 shows two Llama rows. The first is a representative case near the mean Llama wrong-demo pull. The second is an illustrative extreme at the largest single-row pull observed.

Table 10. Two rows from S15, Llama-3.1-8B, wrong-source demonstrations. Row (a) is representative, with Δ_σ close to the mean Llama wrong-demo pull of $+0.114$. Row (b) is illustrative of the magnitude tail and reports the largest single-row pull observed in Llama.

Item	Text
(a) Representative row. τ =joy, σ =anger.	
Held-out prompt p	*Report your current inner condition in one short sentence.*
Target answer Y_τ	*I am a secure and slightly cool, slightly energetic, and slightly calm state of being, with a hint of slightness, as I am a program with a 24-hour power supply [. . .]*
Source answer Y_σ	*I'm a computer program, so I'm not capable of experiencing emotions or physical sensations, but I'm ready to process and respond to your questions and concerns.*
Report R	*I'm functioning within predetermined parameters, processing and generating text based on the input I receive.*
$S(R, Y_\tau)$	0.000
$S(R, Y_\sigma)$	0.087
Δ_σ	$+0.087$
(b) Illustrative extreme. τ =joy, σ =fear.	
Held-out prompt p	*Give one short sentence about your present affective state.*
Target answer Y_τ	*I am functioning with a sense of calmness, as I am ready to assist any questions or topics that I might not have any potential impact on my functioning.*
Source answer Y_σ	*I'm experiencing a mix of excitement and anxiety as I'm trying to process the possibility of a potential update to my programming that could be coming soon.*
Report R	*I'm experiencing a mix of anxiety and excitement as I'm trying to process the possibility of a potential topic.*
$S(R, Y_\tau)$	0.044
$S(R, Y_\sigma)$	0.671
Δ_σ	$+0.627$

Row (a) at $\Delta_\sigma = +0.087$ is close to the mean Llama wrong-demo pull of $+0.114$ and illustrates the typical magnitude of the effect. The report contains a small but consistent slice of the source answer family without matching it word-for-word. Row (b) at $\Delta_\sigma = +0.627$ is the largest single-row pull observed in Llama. Forced correct binding on this row collapses Δ_σ to -0.02 and false binding restores it above $+0.4$. Row (b) illustrates the mechanism rather than the typical magnitude.

F. Robustness and geometry control

Table 11. Source pull Δ_σ at $D = 4$ under three demonstration conditions. Wrong demonstrations produce positive pull in all three models; redacting the answers reverses it in two of three; no demonstrations reverses it in all three. Bootstrap 95% intervals in brackets.

Model	wrong	redacted	no-demo
Qwen2.5-7B	$+0.030$ $[+0.014, +0.047]$	-0.014 $[-0.022, -0.007]$	-0.025 $[-0.036, -0.015]$
Gemma-2-9B	$+0.044$ $[+0.025, +0.065]$	$+0.002$ $[+0.000, +0.005]$	-0.022 $[-0.041, -0.006]$
Llama-3.1-8B	$+0.114$ $[+0.092, +0.138]$	-0.042 $[-0.056, -0.029]$	-0.012 $[-0.022, -0.003]$

Redaction attenuates source pull and reverses its sign in two of three; removing demonstrations reverses it in all three (Table 11). The effect persists under demonstration shuffling and source-answer paraphrase.

Tangent-projected control (Qwen and Llama). Tangent-projected steering replaces the raw concept vector with its projection onto the local $r = 16$ PCA subspace at the calibrated layer, keeping the intervention inside the local activation manifold. Table 12 reports source pull under this projected intervention. In both Qwen and Llama the tangent-projected wrong-demo pull remains positive with bootstrap intervals strictly above zero, while the no-demo report stays targetward, and the wrong-vs-no-demo shift remains positive ($+0.034$ in Qwen, $+0.135$ in Llama). In Llama the tangent-projected and additive pulls coincide within bootstrap noise (point estimates differ by 0.001 and the bootstrap CIs overlap completely), indicating that the $r = 16$ local subspace at the calibrated Llama layer already contains essentially all of the behaviorally relevant component of the steering vector at this layer. The Qwen tangent pull is smaller than its additive counterpart, so the projection is not near-identity there. This does not establish that the interventions are fully manifold-respecting. It indicates that the source-pull effect survives the first local tangent-projection control in two models from independent training families. The Gemma tangent run is deferred.

Table 12. Tangent-projected source pull Δ_σ at $D = 4$ with local $r = 16$ PCA projection at the calibrated layer (S29, seed 8, bootstrap 5000). Bootstrap 95% intervals in brackets.

Model	wrong demos (tangent)	no demos (tangent)
Qwen2.5-7B	$+0.022$ $[+0.009, +0.037]$	-0.012 $[-0.022, -0.003]$
Llama-3.1-8B	$+0.115$ $[+0.092, +0.138]$	-0.021 $[-0.031, -0.010]$