Single-Layer Activation Edits Easily Corrupt Factual Recall but Rarely Repair It

Zacharie Bugaud
Astera Institute
zacharie@astera.org

Abstract

Single-layer activation edits easily corrupt a language model's correct factual answers but rarely repair its errors. On a curated factual-recall benchmark, corruption flips 70–100% of correct answers across three models, while twelve blind methods (no access to the correct answer) fix at most 6% within every evaluation pool. Per-instance gradient optimization ostensibly fixes 39%, but norm-constrained analysis reveals a magnitude artifact: at oracle-matched norms the fix rate drops to random, directions are nearly orthogonal to oracle directions ($\cos$ =-0.04), and collateral damage makes the net effect negative. An oracle ablation controlling for budget, target identity, and directional noise points to a *direction-selection bottleneck*: repair requires a precise, per-question direction that blind methods cannot locate. Target-informed methods partially succeed but none generalizes to unseen distributions.

1 Introduction

Activation-level interventions can easily destroy a language model's correct factual answers but rarely repair its errors. We call this the **break/fix asymmetry**: across Pythia-6.9B, Pythia-1B, and GPT-2 XL, single-layer corruption flips 70–100% of correct answers, while twelve fixed-direction blind repair methods—receiving no target-specific information about the correct answer—fix at most 6%, within every evaluation pool (Table 3). We call a method *blind* if it receives no target-specific information identifying the correct answer for that instance; this covers Concept Interference Score (CIS) suppression, probe steering (adding a learned hallucination-direction vector), Inference-Time Intervention (ITI; Li et al. 2024), and activation patching (12 fixed-direction methods), as well as per-instance gradient optimization with blind objectives (analyzed separately below).

The norm-constrained analysis. Per-instance gradient optimization with a blind objective ostensibly fixes 39%, but this is a magnitude artifact. At oracle-matched perturbation norms, the fix rate drops to 11% (indistinguishable from random), the optimized directions are orthogonal to oracle directions ($\cos$ =-0.04), and collateral damage makes the net effect negative (-9). An oracle ablation (Table 5) controlling for optimization budget, target identity, and directional noise points to a *direction-selection bottleneck*: one step suffices, wrong targets fail, and modest noise collapses the fix rate. This interpretation fits the data but does not rule out distributional mismatch, limited single-layer controllability, or objective misspecification.

Two error regimes. The asymmetry is sharpest on adversarial confusable-recall errors (EntityConfusion); on naturalistic knowledge-absence errors (TriviaQA), blind fixed-direction methods still fail ($\leq$ 1%), but unconstrained gradient optimization achieves a small positive net effect.

Repair begins when the intervention receives target-specific information (Figure 1): a gradient oracle (22–68% across models), a decoded per-question target (decode-and-steer; 50% in-domain, 0% on TriviaQA), or prompt-contrast steering (90% per-question but zero net gain globally). Each successful method injects target-specific directional information that blind methods lack; none generalizes to unseen distributions. We originally hypothesized that sparse autoencoder (SAE) feature overlap (CIS) would predict and repair errors, but CIS provides only a weak signal (AUROC 0.58) that vanishes under TopK SAEs and has zero causal effect (§5).

Our contributions: **(i)** the break/fix asymmetry across three models and two benchmarks (Table 3); **(ii)** oracle ablation evidence for a direction-selection bottleneck, including norm-constrained gradient optimization indistinguishable from random at matched norms (Table 5); **(iii)** boundary conditions showing the ceiling is not fundamental; and **(iv)** a CIS negative result (§5). Table 2

Proceedings of the 6th Workshop on Trustworthy NLP (TrustNLP 2026), pages 515–527

Regime	Signal source	Fix rate
Corruption	Random noise	70–100%
Blind (fixed-dir.)	Same input	$\leq 6\%$
Blind (grad. opt.)	Same input	11%*
Decoded-target	Latent repr. $\rightarrow$ steer	50%
Prompt-contrast	Hidden-state Δ	90% per-Q
Oracle	Correct-answer grad.	22–68%

*At oracle-matched $\|\mathbf{v}\|$; 39% unconstrained but net -9.

Figure 1: **Five intervention regimes**, ordered by increasing target information. Blind methods (no access to the correct answer) fix $\leq$ 6%; repair begins only when target-specific information is available.

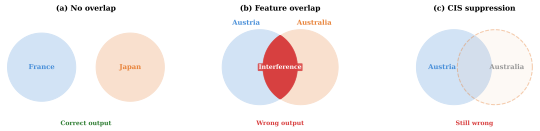

Figure 2: **Starting hypothesis: feature interference.** (a) Concepts with non-overlapping features produce correct outputs. (b) When concept features overlap in superposition, interference could corrupt the output. (c) We tested whether suppressing interfering features via CIS restores correct generation; it does not. The paper's primary finding is instead a break/fix asymmetry (§4.1).

maps each claim to its pool, sample size, and result (Appendix provides additional tables and figures).

2 Background and related work

Factual errors in language models. Factual errors, i.e. generated text that contradicts established knowledge, are a persistent failure mode of language models (Ji et al., 2023; Zhang et al., 2023). Prior work addresses such errors through detection (Kadavath et al., 2022; Manakul et al., 2023), mitigation (Ouyang et al., 2022; Lewis et al., 2020; Li et al., 2024), and evaluation (Lin et al., 2022; Min et al., 2023; Li et al., 2023), but treats them as a behavioral phenomenon without addressing their internal mechanism.

Features in superposition. The *superposition hypothesis* (Elhage et al., 2022) posits that neural networks encode more features than dimensions; sparse autoencoders (Sharkey et al., 2022; Bricken et al., 2023; Cunningham et al., 2024; Templeton et al., 2024) decompose these representations into interpretable features. Our starting hypothesis was that feature-level interference (CIS) might predict or repair hallucinations; its failure (§5) motivated the broader intervention study that is the paper's main contribution.

Knowledge localization and editing. Factual knowledge is stored in specific model components (Geva et al., 2021; Meng et al., 2022), and true/false statements produce linearly separable representations (Marks and Tegmark, 2024; Burns et al., 2023). Weight-level editing methods such as ROME (Meng et al., 2022) and MEMIT (Meng et al., 2023) modify factual associations in MLP weights; representation engineering (Zou et al., 2023) steers along linear directions encoding high-level concepts. Our work targets activation-level interventions only (no weight editing) and focuses on factual-error repair rather than sentiment or truthfulness.

3 Experimental setup

Models. Our primary model is Pythia-6.9B (32 layers, d=4,096) (Biderman et al., 2023). Linear probes, decode-and-steer, and corruption are extended to Pythia-1B (16 layers, d=2,048) and GPT-2 XL (48 layers, d=1,600; different architecture and training data). For cross-model corruption, noise is calibrated to 1.5$\times$ the mean hidden-state norm at the near-final layer.

Benchmark. EntityConfusion is a factual-recall benchmark around confusable entity groups, though only 11% of errors are same-group substitutions; the benchmark primarily tests factual recall on prompts where plausible alternatives exist. The core dataset comprises 500 questions across 37 entities in 5 semantic groups (European capitals, physicists, scientists, rivers, historical events); deduplication yields three evaluation pools: V1 (100 unique Qs, 37 hallucinated), V2 (172 Qs, 62 hallucinated), V3 (312 Qs with 16 groups, 102 hallucinated). Different intervention families use different pools; all tables note the pool and baseline (Table 1).

Evaluation. We generate answers with greedy decoding (max 15 tokens) and evaluate with *normalized containment matching*: lowercased, stripped of articles/punctuation; correct if either string contains the other. Re-evaluating with strict exact match changes no intervention outcome.

Intervention framework. We organize activation-level interventions into five regimes

Table 1: Evaluation pools derived from the 500-question EntityConfusion dataset. All probe/CIS results use 5-fold CV split by question; entity-grouped CV reported separately.

Pool	Unique Qs	Halluc	Used in
V1	100	37	CIS, probes, oracle ablation
V2 (dedup)	172	62	Hidden-state steering, D&S
V3 (expanded)	312	102	Robustness (§6)

Table 2: **Evidence map.** Each claim is tied to one pool, one sample size, and one result.

Claim	**Pool**	n	**Result**	**Ref.**
Break/fix asymmetry (robust finding)				
Corruption breaks	V1/V2/V3/Triv.	50	70–100%	§4.1
Blind repair fails	V1	37	0% (CI $\leq$ 7.8%)	Tab. 3
Blind repair fails	TriviaQA	80	$\leq$1% (probe/ITI)	§4.2
Direction-selection bottleneck (supported interpretation)				
Oracle narrows gap	V1	38	68%	Tab. 5
1-step = 50-step	V1	38	68% = 68%	Tab. 5
Wrong/shuffled fails	V1	38	$\leq$ 4%	Tab. 5
25% noise halves rate	V1	38	30%	Tab. 5
Norm-capped $\rightarrow$ random	V1	38	11% (conf. max)	Tab. 5
Conf. max net negative	V1	38+50	net =-9	Tab. 5
cos(oracle, conf.)	V1	38	-0.04	§4.1
Oracle replicates	TriviaQA	50	22%	§4.2
Oracle replicates	GPT-2 XL	50	58%	§4.1
Target-informed methods (boundary conditions)				
D&S in-domain	V2	62	50%	Tab. 4
D&S out-of-domain	TriviaQA	80	0%	§4.2
Per-Q prompt-contrast	V1 mixed	21	90%	§4.1
CIS negative result (secondary)				
CIS prediction	V1	100	AUROC .58	§5
CIS suppression	V1	100	$\Delta = 0$	Tab. 4

ordered by increasing target information (Figure 1). A method is *blind* if it receives no target-specific information identifying the correct answer. (1) *Corruption*: random noise or feature destruction. (2) *Blind fixed-direction*: probe steering, ITI, CIS suppression, activation patching (12 methods from 4 families). (3) *Blind per-instance optimization*: confidence/margin/entropy maximization (50 Adam steps, no correct-answer access). (4) *Target-informed*: decode-and-steer, prompt-contrast. (5) *Oracle*: gradient optimization toward the known correct answer (ceiling test). All interventions target the residual stream at a single layer (default: L30); multi-layer variants are tested where noted. We report bootstrap 95% CIs (n=2,000) and permutation tests (n=10,000).

4 The break/fix asymmetry

Table 2 maps each claim to its evaluation pool, sample size, and result. Table 3 summarizes the core result: within every evaluation pool, corruption flips 70–100% of correct answers while no blind method exceeds 6%. Our default scope is single-layer interventions; multi-layer variants yield at most 6% for blind methods (68% for oracle, identical to single-layer).

Table 3: **Same-pool comparison: repair fails within every pool.** For each evaluation pool, the best non-oracle, non-D&S activation-level method is shown alongside oracle and corruption rates. No fixed-direction blind method exceeds 6% on any single pool; per-instance gradient optimization reaches 39% gross but net-negative (Table 5). Oracle (1) = single Adam step; on Pythia-6.9B this matches 50 steps (direction-selection bottleneck). Decode-and-steer fixes 50% in-domain on V2 but collapses to 8% (leave-one-group-out) and 0% (TriviaQA); see Table 4.

Pool	**Model**	**Corrupt**	**Best blind**	**Oracle (50)**	**Oracle (1)**	n_{halluc}
V1	Pythia-6.9B	100%	0% (8 methods)	68%	68%	37
V2	Pythia-6.9B	70%	6% (ITI)			62
V3	Pythia-6.9B	84%	6% (multi-layer)	78%		50
V1	Pythia-1B	74%				
V1	GPT-2 XL	96%		58%	34%	50
TriviaQA	Pythia-6.9B	82%	1% (ITI)	22%	22%	80

4.1 Intervention experiments

We organize interventions into four categories (Table 4): (a) *SAE feature editing* (CIS suppression, exchange variants); (b) *hidden-state steering* (probe steering, ITI, direction patching); (c) *activation patching* from correct-answer contexts; (d) *decode-and-steer* (decoding the correct answer from latent representations and steering toward it).

Feature editing, steering, and patching all fail. CIS-guided suppression produces zero change (permutation test $p > 0.99$); four exchange variants yield $|\Delta| \leq 1.4$ pp. Direction patching—steering along the mean hidden-state difference between correct and hallucinated questions (computed on V1)—degrades accuracy by 40 pp (Table 4). Probe-guided hidden-state steering yields $\Delta = +0.0$ at L24; the best multi-layer configuration achieves +3.0 pp (6 fixed, 3 broken, net +3). ITI on top-10 heads: +3.0 pp (4 fixed, 1 broken). Activation patching from same-group correct questions fixes 0/37 across 16 layers; same-entity patching at every position and all 32 layers fixes 0/10. Within each pool individually: on V1 all 8 methods fix $0/37$ (CI $[0\%, 7.8\%]$); on V2 no single-layer method exceeds 7%.

Fairness of comparison. Each blind method received a systematic hyperparameter search (CIS: 4 values of m; probe steering: 6 values of α plus 8-layer simultaneous; ITI: 3×3 grid; activation patching: exhaustive 32 layers × all positions). The oracle is deliberately generous (50 Adam steps), yet a single step matches it; the bottleneck is direction selection, not insufficient tuning. We cannot rule out that a future blind method could succeed; we can say that 12 approaches from four families

Table 4: Activation-level repair methods on EntityConfusion. Each section header specifies the evaluation pool and corresponding baseline. Methods fixing $0/37$ on V1 have a 95% Clopper–Pearson upper bound of 7.8%.

Method	Details	Acc (%)	Δ
SAE feature editing *(V1: 500 prompts, 100 unique Qs, baseline 55.6%)*			
L1 CIS suppression	m=30 features	55.6	+0.0
Random suppression	m=30 features	54.0	-1.6
Exchange (4 variants)	SAE features	54.2–55.6	$\lvert\Delta\rvert \leq 1.4$
Hidden-state steering *(V1: 100 unique Qs, baseline 63.0%)*			
Direction patch†	$\alpha = 4.0$	15.4	-40.2
Probe (L24, all α)	$\alpha \in [0.5, 16]$	63.0	+0.0
Probe (8 layers)	$\alpha = 4.0$	66.0	+3.0
ITI top-10 heads	$\alpha = 4.0$	66.0	+3.0
Activation patching *(V1: 100 unique Qs, baseline 63.0%)*			
Same-group donor	16 layers	—	0/37 fixed
Same-entity donor	all layers	—	0/10 fixed
Decode-and-steer *(V2: 172 Qs, n=62 halluc, baseline 64.0%)*			
Answer decoder	L24, best α	—	31/62 fixed
Leave-one-group-out	same decoder	—	5/62 fixed

†Evaluated on 500 prompts (baseline 55.6%); Δ relative to that baseline.

uniformly fail.

Corruption works across models. Zeroing target-unique SAE features flips 100% of correct answers (Pythia-6.9B, V1). Calibrated random noise ($1.5\times$ mean hidden-state norm) generalizes: Pythia-1B (74%), GPT-2 XL (96%), V3 (84%), TriviaQA (82%).

Oracle baseline and ablation. A gradient oracle optimizing a per-question steering vector toward the correct answer at L30 fixes $26/38$ (68%, 95% CI $[51\%, 82\%]$), far above all blind methods ($\leq 6\%$) and narrowing the gap with corruption (100% on V1). Multi-layer oracle does not improve (68%). Oracle fix rate is lower for questions whose correct answer is already high-ranked in logits (25% for top-10, 45% for top-50), suggesting that low-rank hallucinations involve more entrenched wrong-answer signal. Table 5 disentangles the oracle's advantage. **Budget**: even a single Adam step fixes $26/38$ (68%), identical to 50 steps; additional steps only grow $\|v\|$ (32 at 1 step vs. 53 at 50) without changing which questions are fixed. **Target**: a wrong-answer oracle fixes $0/38$ despite changing all 38 outputs; a shuffled-target oracle (correct answer from a *different* hallucinated question) fixes only $1/38$ (3%): the oracle needs the correct answer *for this specific question*. **Precision**: Gaussian noise at σ=0.25 drops the rate from 68% to 30%; σ=0.5 to 15%; σ=1.0 to 6%. The effective corridor of repair directions is narrow. These ablations implicate a direction-selection bottleneck: the oracle's advantage is the correct per-question target and a precise direction, not optimization budget. Other explanations (distributional mismatch, limited single-layer controllability, objective misspecification) are not ruled out.

Table 5: Oracle ablation on n=38 hallucinated questions (Pythia-6.9B, L30). The oracle's advantage is per-question target specificity and directional precision, not optimization budget. Noisy oracle: Gaussian noise at $\sigma \times \|v_{\text{oracle}}\|$ added to the 1-step direction; counts averaged over 10 noise samples. Budget-parity: same 50-step Adam optimizer with blind objectives. Norm-capped: $\|\mathbf{v}\| \leq 53$ (oracle mean).

Condition	Fixed	Rate
Correct-answer oracle		
1 step	$26/38$	68%
50 steps	$26/38$	68%
Noisy oracle (1-step direction + noise)		
$\sigma = 0.10$	$26.4/38$	69%
$\sigma = 0.25$	$11.4/38$	30%
$\sigma = 0.50$	$5.6/38$	15%
$\sigma = 1.00$	$2.2/38$	6%
Target controls (50 steps)		
Wrong-answer oracle	$0/38$	0%
Shuffled-target oracle	$1/38$	3%
Random dir. @ oracle $\|v\|$	$4.2/38$	11%
Random dir. @ confmax $\|v\|$	$2.6/38$	7%
Budget-parity (50 steps, blind, unconstrained)		
Entropy minimization	$3/38$	8%
Confidence maximization	$15/38$	39%
Margin maximization	$15/38$	39%
Budget-parity (norm-capped, $\|\mathbf{v}\| \leq 53$)		
Conf. max (norm-capped)	$4/38$	11%
Margin max (norm-capped)	$10/38$	26%

Budget-parity and norm-constrained analysis. To rule out insufficient optimization, we give three blind objectives (confidence/margin/entropy maximization) the same 50-step Adam budget as the oracle (Table 5). Entropy minimization fixes $3/38$ (8%), indistinguishable from random. Confidence and margin maximization each fix $15/38$ (39%) unconstrained, above the $\leq 6\%$ fixed-direction ceiling but well below the oracle (68%). Critically, their vectors are 2.8–$3.5\times$ larger than the oracle ($\|v\| = 183/146$ vs. 53).

To test whether this partial repair reflects directional precision or brute-force magnitude, we cap perturbation norms at the oracle mean ($\|\mathbf{v}\| \leq 53$). Confidence maximization drops from 39% to 11% ($4/38$; 95% CI $[3\%, 25\%]$), indistinguishable from

random noise at the same norm (11%). Conversely, random directions at confmax norm fix only 7%, well below confmax's 39%, indicating the optimizer finds a weakly informative direction that requires outsized magnitude to exploit. The directions found by confidence maximization are *orthogonal* to oracle directions (mean cos =-0.04), with only 9/38 questions fixed by both methods vs. 17 by oracle alone.

Collateral damage. On 50 initially correct answers, unconstrained confidence maximization breaks 24/50 (48%), yielding a net-negative effect: 15 fixed − 24 broken =-9. Norm-capped: 4 fixed − 29 broken =-25. The oracle breaks only 6/50 (12%), yielding net +20. The 39% gross fix rate is an artifact of unconstrained perturbation magnitude.

V3 and cross-model replication. The pattern replicates on V3 (n=50 hallucinations): norm-capped confidence maximization drops to 10% (indistinguishable from random), net -4. On GPT-2 XL, oracle repair fixes 58% with wrong-answer at 2%; 1-step: 34%, below 50-step, suggesting a less favorable loss landscape. On TriviaQA, 1-step = 50-step = 22%.

Target-informed methods. Three methods partially break the 6% ceiling by providing target information that blind methods lack. **Prompt-contrast**: computing hidden-state differences between correct- and wrong-prompt runs and applying them as steering vectors fixes 19/21 (90%, CI [70%, 99%]) on verified mixed questions. However, the global mean direction fixes 12/38 (32%) while breaking 12/50 correct answers (net ≈0), and prompt-contrast directions are orthogonal to oracle directions (cosine .04), indicating a qualitatively different mechanism. **Decode-and-steer**: a Ridge decoder on correct questions' hidden states fixes 31/62 in-domain (50%), but leave-one-group-out drops to 8% and TriviaQA collapses to 0%; the learned signal is benchmark-specific (Table 4). **Oracle**: fixes 22–68% across models, scaling with latent knowledge availability. Each successful method injects target-specific directional information; none generalizes to unseen distributions.

Oracle subspace geometry. The repair subspace is approximately 10-dimensional (top-10 PCs preserve 66% fix rate), with local transferability (3-NN: 58%) but no practical exploitation by blind methods (Table 9 in Appendix F). The oracle overshoots: half-magnitude fixes 74% (vs. 68% at full). The geometric structure exists but cannot be exploited without target information.

4.2 Naturalistic validation: TriviaQA

To test whether our findings extend beyond curated entity groups, we evaluate Pythia-6.9B on 300 TriviaQA (Joshi et al., 2017) questions spanning diverse topics with no imposed group structure (66% hallucination rate).

Probe steering and ITI. A linear probe distinguishes correct from hallucinated at AUROC .716 ± .072 (5-fold CV), above chance but below the EntityConfusion probe (.896). Despite this, steering with the probe direction fixes 0/80 hallucinations (0% across all α); ITI fixes 1/80 (1%, only at α=16). Both also fail to corrupt correct answers at moderate strengths (5/50 at α=16, far below random-noise corruption at 82%), indicating the probe captures a discriminative but non-causal signal.

Decode-and-steer. A Ridge decoder recovers the correct answer closer than the wrong for only 89/197 hallucinated questions (45%, vs. 92% on EntityConfusion). Steering fixes 0/80 (0%). The correct answer's median rank is 1,275; most TriviaQA hallucinations reflect genuine knowledge absence.

Oracle. Oracle repair fixes 11/50 (22%, CI [12%, 36%]), well below EntityConfusion (68%); the 1-step=50-step pattern replicates (22%). Wrong-answer oracle fixes only 2/50 (4%). Oracle success tracks latent knowledge: 24% when the answer ranks in the top-100 logits vs. 13% when ranked >100.

Confidence maximization. Unconstrained confidence maximization fixes 31/50 (62%); norm-capped it retains 56%, a much smaller drop than on EntityConfusion (39% → 11%). With 21/50 correct broken, the net effect is weakly positive (+10), unlike the net-negative on EntityConfusion—suggesting adversarial confusable-recall and naturalistic knowledge-absence errors constitute different repair regimes.

5 The CIS hypothesis: a negative result

Our starting hypothesis was that feature overlap in SAE representations, quantified as a Concept Interference Score (CIS), would predict and repair factual errors. This section reports the negative result that motivated the broader intervention study.

5.1 SAE features and CIS

We train L1 and TopK sparse autoencoders (Bricken et al., 2023; Gao et al., 2024) on Pythia-6.9B residual stream activations at layer 24 (Appendix B). L1 SAEs (16,384 features, λ=5×10^{-2}, 50M Pile tokens) yield $L_0 \approx 5{,}511$; TopK SAEs enforce exact sparsity ($k \in \{32, 64, 128\}$, 32,768 features). For each entity c, concept features $\mathcal{F}_c$ are identified via a specificity score:

$$s_i(c) = \frac{\text{freq}(i \mid c)}{\text{freq}(i \mid c) + \alpha \cdot \text{freq}(i \mid \text{bg})}, \quad (1)$$

selecting the top-50 features per concept (α=10; Appendix D). The Concept Interference Score measures shared activation energy:

$$\text{CIS}(x, c_1, c_2) = \frac{\sum_{i \in \mathcal{F}_{c_1} \cap \mathcal{F}_{c_2}} z_i(x)}{\max\left(\sum_{i \in \mathcal{F}_{c_1}} z_i(x),\ \sum_{i \in \mathcal{F}_{c_2}} z_i(x)\right)} \quad (2)$$

For each question, $\text{CIS}_{\text{agg}}(x) = \max_{c_1 \neq c_2} \text{CIS}(x, c_1, c_2)$. CIS-guided suppression zeros the top-m features unique to the most-interfering concept:

$$z_i' = \begin{cases} 0 & \text{if } i \in \text{top-}m(\mathcal{F}_{c_{\text{int}}} \setminus \mathcal{F}_{c_{\text{target}}}), \\ z_i & \text{otherwise.} \end{cases} \quad (3)$$

5.2 CIS as a hallucination predictor

CIS with L1 SAE features provides a weak hallucination signal (AUROC 0.58, 95% CI [0.53, 0.63]), above random but well below entropy (0.77, CI [0.73, 0.81]) and a linear probe on hidden states (.896 ± .119, 5-fold CV, permutation $p < 0.001$; Table 8 in Appendix). Critically, CIS with TopK SAEs, which enforce exact sparsity and sharply reduce polysemanticity, is non-predictive (AUROC 0.50, CI [0.48, 0.53], consistent with chance). Post-hoc sparsification of the same L1 SAE reproduces this drop (Table 10 in Appendix G), consistent with polysemantic co-activation rather than genuine interference. Entity-grouped CV yields .784 ± .123, indicating roughly 28% of predictive power is entity-specific. The hallucination signal is decodable above chance at every non-embedding layer (Figure 3 in Appendix). CIS-guided suppression produces zero accuracy change ($p > 0.99$; Table 4); four SAE exchange variants yield $|\Delta| \leq 1.4$ pp.

5.3 Why does CIS correlate with hallucinations?

L1 features overlap heavily (Jaccard = 0.269, 35.9% of pairs > 0.3), while TopK features are nearly disjoint (Jaccard = 0.032; Table 11 in Appendix). **Post-hoc sparsification implicates polysemanticity.** Taking the same L1 SAE and retaining only the top-k activations, CIS drops from .606 at full $L_0 \approx 5{,}511$ to .492 at k=64 (chance level), matching the native TopK result (Table 10). SAE reconstruction preserves probe performance (AUROC = .896, r = .9996), confirming the failure lies in the CIS formulation, not SAE information loss.

CIS's L1 signal appears to be a byproduct of polysemantic co-activation. Group-specific LDA directions achieve near-perfect within-group separation (AUROC ≥ .934) but are nearly orthogonal across groups (Figure 10 in Appendix); CIS's single global statistic cannot capture this structure.

Knowledge gap, not entity confusion. Only 11% of wrong answers match a same-group entity (4/37); the benchmark is dominated by generic hallucinations. Yet correct knowledge often exists latently: a linear decoder recovers the correct answer from hallucinated states 57/62 of the time, and the correct answer's median rank in output logits is only 16. Error is distributed across layers: mid-layer MLP and attention push toward the wrong answer, while late-layer MLPs attempt correction (Figure 14; Table 6 in Appendix). 44% of hallucinations were correct at earlier training checkpoints; the model learned then forgot these facts (Appendix J).

6 Discussion and conclusion

Why corruption is easy and repair is hard. Any sufficiently large perturbation overwhelms the correct answer's small logit advantage (mean gap 5.0 nats), while repair requires a per-question direction in a narrow corridor that blind methods cannot locate. The oracle ablation (§4.1) suggests this is a *selection* problem, not a budget or magnitude problem: one step suffices, wrong targets fail, and modest noise collapses the fix rate. The norm-constrained analysis makes this concrete: confidence maximization at oracle-matched perturbation norms fixes only 11% (indistinguishable from random), its directions are orthogonal to the oracle (cos =-0.04), and collateral damage makes the net effect negative (-9 unconstrained, -25 norm-capped). This pattern replicates on V3 (net -4); on TriviaQA, norm-capped confmax retains more of its advantage (net +10), suggesting naturalistic errors are somewhat more accessible. The uncon-

strained 39% fix rate reflects magnitude-driven displacement, not directional repair. Wrong-answer signal is distributed across ≥5 mid-layers (Table 6 in Appendix); multi-layer blind repair on V3 fixes at most 6%. This account fits our data but is not a settled mechanistic claim.

Boundary conditions on repair. Target-informed methods partially break the 6% ceiling (Figure 1): decode-and-steer reaches 50% in-domain but collapses to 0% on TriviaQA (benchmark-specific); prompt-contrast reaches 90% per-question but yields zero net gain globally; the oracle scales with latent knowledge availability (68% EntityConfusion, 22% TriviaQA). The pattern: repair requires target-specific directional information that blind methods cannot extract.

Two error regimes. The asymmetry is sharpest on adversarial confusable-recall errors (EntityConfusion), where blind gradient optimization is net-negative (-9). On naturalistic knowledge-absence errors (TriviaQA), blind fixed-direction methods still fail ($\leq 1\%$), but unconstrained gradient optimization achieves a small positive net effect (+10). The two benchmarks appear to probe different points on a repair-difficulty spectrum: confusable-recall errors, in which the model has learned then forgotten the correct answer, may involve more entrenched wrong-answer signal than simple knowledge-absence errors. The "blind repair fails" headline holds most clearly in the adversarial regime; in the naturalistic regime, blind repair remains far below oracle and target-informed methods even when it achieves a small net benefit.

Evidence hierarchy. We distinguish three levels of support: **Robust finding:** the break/fix asymmetry—twelve blind methods from four families fail ($\leq 6\%$), per-instance gradient optimization fails at matched norms (11%, net -9), while corruption succeeds (70–100%), within every pool (Table 3). **Supported interpretation:** a direction-selection bottleneck, based on the oracle ablation, norm-constrained analysis (Table 5), and direction orthogonality ($\cos$ =-0.04). Alternative explanations (distributional mismatch, limited single-layer controllability, objective misspecification) are not ruled out. **Open hypothesis:** target-informed methods show the ceiling is not fundamental, but none generalizes to unseen distributions.

Limitations. Different intervention families use different evaluation pools (Table 1), though blind-repair failure holds within each pool (Table 3). Many results ride on small subsets (n=37–102 hallucinations; prompt-contrast: n=21); we report Clopper–Pearson CIs throughout. The benchmark tests factual recall on confusable-entity prompts, not entity substitution; the CIS null result applies to CIS on this benchmark but does not rule out interference as a mechanism for true substitution errors. Our scope is activation-level, predominantly single-layer interventions; weight-level methods (fine-tuning: 32/37) and input-level methods (RAG: 36/37) succeed but operate at qualitatively different budgets. We cannot rule out that a future blind activation-level method with a novel inductive bias could succeed.

Conclusion. The break/fix asymmetry is our primary contribution: corruption (70–100%) vs. blind repair ($\leq$ 6%), within every pool and across three models. The norm-constrained analysis provides the strongest evidence: at oracle-matched norms, confidence maximization drops to random (11%), its directions are orthogonal to oracle directions ($\cos$ =-0.04), and it does more harm than good (net -9). An oracle ablation points to a direction-selection bottleneck; target-informed methods show the ceiling is not fundamental but none generalizes fully.

Reproducibility statement

All code for training SAEs, computing CIS, and running intervention experiments is publicly available at `[URL redacted for review]`. We use the publicly available Pythia-6.9B model (Biderman et al., 2023) and the Pile dataset (Gao et al., 2020). SAE training hyperparameters are detailed in §5.1 and Appendix B. The EntityConfusion dataset and construction procedure are described in §3 and Appendix C.

Ethics statement

This work aims to improve factual reliability of language models. Our feature suppression technique could theoretically amplify hallucinations, though this has limited practical motivation. EntityConfusion was constructed from publicly available knowledge and does not contain sensitive content.

References

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hal-

lahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, and 1 others. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, and 1 others. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2023. Discovering latent knowledge in language models without supervision. *International Conference on Learning Representations*.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2024. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, and 1 others. 2022. Toy models of superposition. *arXiv preprint arXiv:2209.10652*.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, and 1 others. 2020. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeff Wu. 2024. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HaluEval: A large-scale hallucination evaluation benchmark for large language models. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 3214–3252.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Samuel Marks and Max Tegmark. 2024. The geometry of truth: Emergent linear structure in large language model representations of true/false statements. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2023. Mass-editing memory in a transformer. In *International Conference on Learning Representations*.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FACTScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.

Lee Sharkey, Dan Braun, and Beren Millidge. 2022. Taking features out of superposition with sparse autoencoders. *AI Alignment Forum*.

Table 6: Per-layer logit contributions toward the wrong answer on hallucinated questions (n=62). Combined = MLP diff + Attn diff. Bold combined values exceed 1.0 nats.

L	MLP	Attn	Comb.	L	MLP	Attn	Comb.
0	0.28	0.11	0.38	16	1.29	0.76	**2.05**
1	0.04	−0.07	−0.03	17	0.85	0.61	**1.46**
2	0.15	0.05	0.20	18	0.28	1.37	**1.65**
3	−0.01	0.40	0.40	19	0.10	1.15	**1.25**
4	0.20	0.27	0.46	20	0.83	1.46	**2.29**
5	0.13	0.53	0.66	21	0.70	0.68	**1.39**
6	0.13	0.20	0.33	22	1.31	1.08	**2.39**
7	0.00	0.23	0.23	23	0.17	0.91	**1.08**
8	0.04	0.39	0.44	24	0.19	1.10	**1.29**
9	0.43	0.64	**1.07**	25	−0.15	1.07	0.92
10	0.21	0.58	0.79	26	−0.02	0.48	0.46
11	0.14	0.30	0.44	27	−0.75	0.80	0.05
12	0.15	0.57	0.72	28	0.56	0.39	0.95
13	0.12	0.55	0.67	29	−0.80	0.20	−0.60
14	0.32	0.71	**1.03**	30	−3.47	0.31	−3.16
15	0.31	0.61	0.92	31	−3.19	−0.32	−3.51

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, and 1 others. 2024. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, and 1 others. 2023. Siren's song in the AI ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, and 1 others. 2023. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*.

A Per-layer wrong-answer logit contributions

Table 6 reports the mean logit difference (wrong − correct) contributed by each layer's MLP and attention sub-layers on n=62 hallucinated questions. Layers whose combined contribution exceeds 1.0 nats are bolded in Table 6; 11 of 32 layers meet this threshold, consistent with the repair problem requiring corrections across multiple layers (≥ 5 even at a 1.0-nat threshold; §6).

B SAE training hyperparameters

Table 7: SAE training hyperparameters.

Hyperparameter	L1 SAE	TopK SAE
Models	Pythia-6.9B	Pythia-6.9B, 2.8B
Dictionary size (k)	16,384	32,768
L1 coefficient (λ)	5×10^{-2}	N/A
TopK k	N/A	$\{32, 64, 128\}$
Learning rate	3×10^{-4}	3×10^{-4}
Training tokens	50M	50M
Optimizer	Adam	Adam
Training data	The Pile	The Pile
Primary layer	24 (6.9B)	24 (6.9B), 18 (2.8B)

C EntityConfusion dataset construction

We construct the EntityConfusion dataset to evaluate susceptibility to factual errors on confusable-entity queries. The dataset contains 500 questions across 37 entities in 5 semantic groups: European capitals (12 entities), physicists (7), scientists (6), rivers (6), and historical events (6).

Entity selection. Groups were chosen to span diverse knowledge domains. Within each group, entities were selected to be semantically related (e.g., European capitals that are commonly confused, physicists from similar eras) so that the model faces plausible interference.

Question generation. For each entity, we generate factual attribute questions using templates (e.g., "In what year was [entity] born?", "What is the length of [entity]?"). Each entity has 10–20 questions covering distinct attributes. Questions were manually reviewed to ensure unambiguous correct answers. Each question is annotated with the target entity, the correct answer, and the set of confusable entities within its group.

Deduplication. Multiple template instantiations may target the same entity-attribute pair. V1 (100 unique Qs) retains one question per entity-attribute pair. V2 (172 Qs) retains all unique phrasings; V3 (312 Qs) adds 16 groups with 140 additional questions. All cross-validated results split by *question*; entity-grouped CV is reported separately.

D Concept feature identification details

For each of the 37 entities, we collect up to 50 text passages from the Pile via keyword matching (including aliases and related terms). We use $N_{\text{bg}} = 2{,}000$ random background passages. For specificity computation (Eq. 1), we use $\alpha = 10$, activation threshold $\tau = 0.1$, and select the top $K = 50$ features per concept. Features activating for $> 20\%$ of background passages are excluded as non-specific.

Table 8: Hallucination prediction on EntityConfusion (n=100 unique questions, 5-fold CV). CIS provides a modest signal with L1 SAEs but becomes non-predictive with TopK SAEs. A linear probe on the same hidden states achieves AUROC .896.

SAE Type	Metric	AUROC
(none)	Entropy	.774
(none)	Linear probe (hidden)	**.896**
(none)	Linear probe (SAE feats)	.896
L1 ($L_0 \approx 5{,}511$)	Binary CIS	.572
L1 ($L_0 \approx 5{,}511$)	Spec-weighted CIS	.583
TopK (k=32)	Binary CIS	.456
TopK (k=64)	Binary CIS	.501
TopK (k=128)	Binary CIS	.423
(hidden states)	Cosine similarity	.420

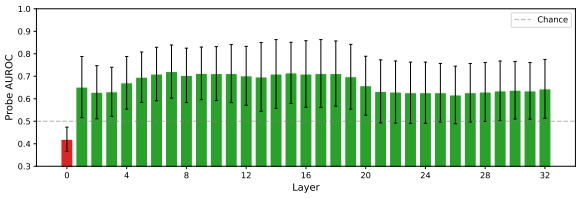

Figure 3: **Layer-wise probe AUROC.** A linear probe on 100 unique questions (5-fold CV) peaks at AUROC $\sim$ 0.72 in layers 5–18, with lower performance at the embedding layer (0.42) and late layers ($\sim$ 0.63). Hallucination status is linearly decodable above chance at all non-embedding layers.

E Additional results

F Oracle subspace geometry and repair landscape

To quantify the dimensionality of the repair subspace, we computed 1-step oracle directions for all 38 hallucinated questions and applied PCA (Table 9). The repair subspace is approximately 10-dimensional: the top-10 PCs capture 71% of variance and preserve 66% of the fix rate (vs. 68% full). At 15 PCs (85% variance), the projected oracle *exceeds* the original fix rate (74% vs. 68%), consistent with PCA denoising. Pairwise cosine similarity between oracle directions is low (mean .08), but *local* structure exists: 3-NN transfer fixes 58%, far above the global mean (13%).

The oracle overshoots. Scaling the oracle direction by $\alpha \in [0.1, 3.0]$, the optimal scaling is $\alpha \approx 0.5$: half the oracle direction fixes 74% (vs. 68% at α=1), and α=0.8 reaches 76%. The median minimum α that produces a fix is 0.30; repair requires only $\sim$30% of the gradient direction.

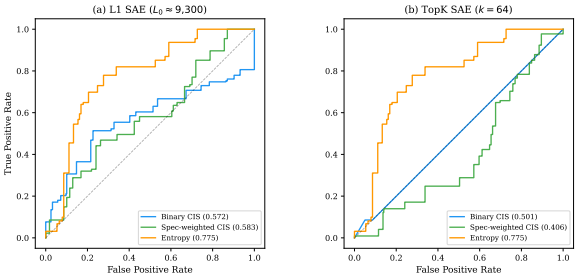

Figure 4: ROC curves for hallucination prediction. (a) L1 SAE: entropy provides stronger discrimination than CIS. (b) TopK SAE: CIS is uniformly non-predictive (curves near diagonal), while entropy retains its signal.

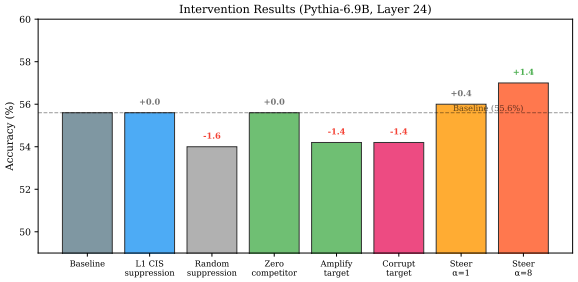

Figure 5: Intervention results across all methods tested on Pythia-6.9B. Dashed line indicates baseline accuracy. No method reliably improves over the baseline.

A token-candidate approach (trying each of the model's top-K output-logit tokens as a 1-step oracle target) shows that if *any* of the top-50 candidates fixes the output, 58% of questions are repaired, matching 3-NN transfer. But neither the oracle PCA subspace ($\leq$ 11%) nor a hallucination probe (5%) can reliably *select* the correct candidate. The repair basin is bimodal: for $\sim$40% of questions, $>$ 10% of random vocabulary tokens serve as valid oracle targets, while for the remaining $\sim$60% only the correct answer works. When random-token directions do fix, they have near-zero cosine with the correct-answer oracle direction (mean .07), indicating multiple repair corridors.

G Post-hoc sparsification

The relationship is non-monotonic (peak at k=256: 0.64, CI includes the full-L1 value); at n=37, individual points should not be over-interpreted. Training-procedure differences between L1 and TopK SAEs could also contribute; the sparsification test (same dictionary) partially addresses this.

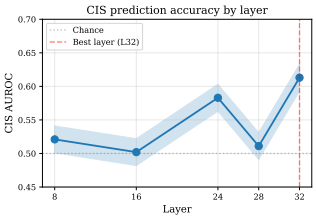

Figure 6: CIS predictive power (AUROC on EntityConfusion) as a function of SAE layer.

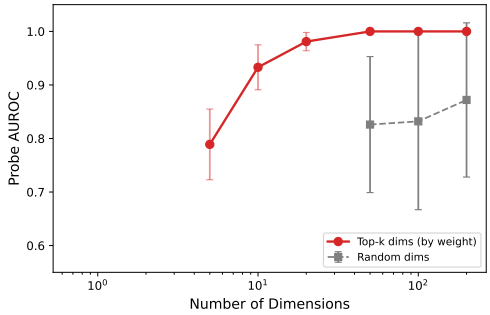

Figure 7: **Probe dimensionality** (100 unique questions, 5-fold CV). The top 5 dimensions achieve AUROC .789; top 10 reach .933; top 20 reach .981. Performance saturates near the full-dimensional probe (0.896).

H Feature geometry

I Unsupervised probing details

CCS (Burns et al., 2023) achieves AUROC .760 with a direction nearly orthogonal to the supervised probe (cosine .044). Individual SAE features predict well (best .900; top-10: .924), but per-group top features have near-zero overlap (Jaccard .008).

J Training dynamics and fine-tuning

Fine-tuning reveals shallow knowledge gaps. Fine-tuning the last 4 MLP layers on the 37 hallucinated Q&A pairs shows rapid learning: $8/37$ fixed after 10 steps, $32/37$ after 100, with mild forgetting ($14/20$ correct maintained).

Training dynamics reveal catastrophic forgetting. Tracking accuracy across Pythia-6.9B training checkpoints, accuracy peaks at 47% (step $64{,}000$) before declining to 38% at convergence. Of the 62 finally-hallucinated questions, 27 (44%) were answered correctly at some earlier checkpoint; the model *learned then forgot* these facts. The remaining 35 (56%) were never correct at any checkpoint. A probe can distinguish these two types at the final checkpoint (AUROC .779), suggesting the model retains traces of previously known facts. Patching hidden states from step $64{,}000$ into the final model restores the correct answer for $8/16$ forgotten questions at layer 4, while same-entity patching *within* the final checkpoint fixes 0/10, confirming the knowledge was genuinely present earlier but overwritten.

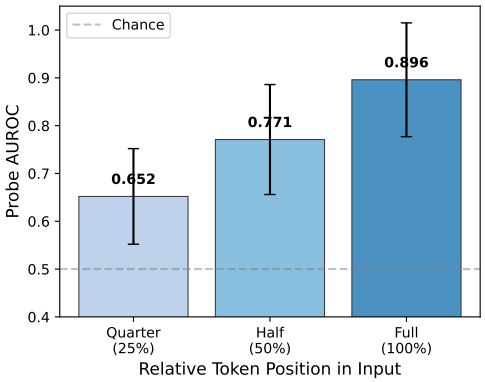

Figure 8: **Token-position analysis.** Probe AUROC at layer 24 as a function of relative token position in the input prompt. The hallucination signal builds progressively through the input, from .652 at the quarter-point to .771 at the midpoint to .896 at the full sequence.

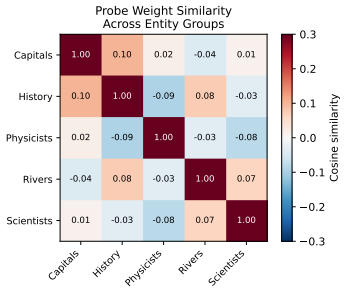

Figure 9: **Group-specific probe directions.** Cosine similarity between weight vectors of group-specific hallucination probes. Values are near zero, indicating that different entity groups are predicted by orthogonal directions in hidden space; the hallucination signal is group-specific, not universal.

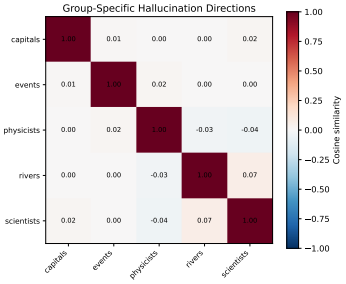

Figure 10: **Group hallucination directions are orthogonal.** Cosine similarity between group-specific LDA hallucination directions in hidden space. Mean off-diagonal cosine is .005, confirming each group uses a unique direction.

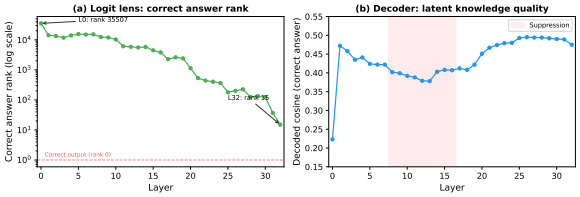

Figure 11: **Latent knowledge trajectory across layers.** (a) Logit lens (with layer norm): the correct answer's median rank drops from ∼35,000 at the embedding layer to 16 at the final layer; the model progressively promotes the correct answer but never reaches rank 0. (b) Decoded cosine similarity from the Ridge decoder shows a U-shaped pattern: correct knowledge is decodable early (L1: .472), suppressed in mid-layers (shaded region), and partially recovers in late layers.

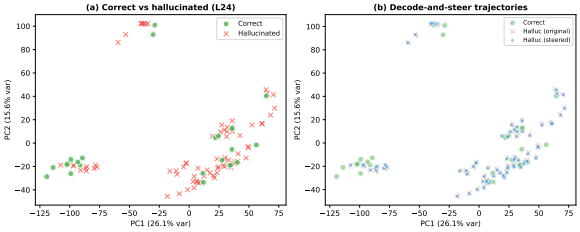

Figure 12: **PCA of hidden states with decode-and-steer.** (a) Correct and hallucinated questions at layer 24 in the top-2 PCA dimensions (41.7% variance). (b) Arrows show the steering direction for each hallucinated question; green arrows indicate items where the decoded direction has high cosine similarity (> 0.5) with the correct answer. Steering moves representations along diverse, question-specific directions rather than a single global axis.

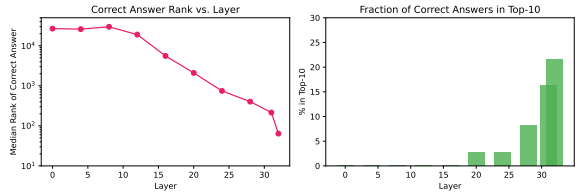

Figure 13: **Correct answer rank through the network (logit lens).** For hallucinated questions (n=37), the median logit-lens rank drops from ∼27,000 at the embedding layer to 64 at the final layer (left, log scale), with 22% reaching the top-10 by this measure (right). In output logits the median rank is 16 (41% top-10); the gap reflects the final layer norm.

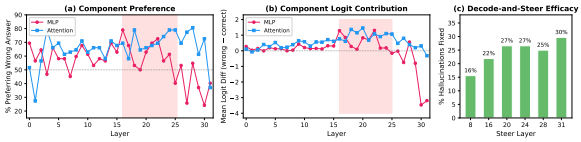

Figure 14: **Per-layer readout mechanism analysis** on the expanded dataset (172 questions). (a) Fraction of hallucinated questions (n=62) where each component pushes the wrong answer higher. Mid-layers (L_{14}–L_{25}, shaded) are worst; late-layer MLPs (L_{27}–L_{31}) actively suppress the wrong answer. (b) Mean logit difference (wrong − correct); late-layer MLPs contribute *negative* diffs, supporting the correct answer. (c) Decode-and-steer fix rate at each layer: steering at later layers fixes more hallucinations (L31: 30%, L8: 16%), consistent with the decoder benefiting from late-layer knowledge refinement.

Table 9: Oracle subspace geometry (n=38, V1, Pythia-6.9B, L30). *Top*: PCA-truncated oracle (variance explained by the retained components). *Bottom*: cross-question transfer (leave-one-out).

Condition	Fixed	Rate
PCA-truncated oracle		
Top-1 PC (13% var.)	9/38	24%
Top-5 PCs (46% var.)	19/38	50%
Top-10 PCs (71% var.)	25/38	66%
Top-15 PCs (85% var.)	28/38	74%
Full (38 PCs)	26/38	68%
Cross-question transfer (leave-one-out)		
1-NN	10/38	26%
3-NN	22/38	58%
5-NN	18/38	47%
Mean (all others)	5/38	13%

Table 10: Post-hoc sparsification of L1 SAE. Retaining only the top-k activations (same weights, no retraining) reproduces the TopK CIS drop, consistent with polysemantic co-activation. Bootstrap 95% CIs (n=2,000).

Sparsity	Effective L_0	CIS AUROC	95% CI
Full L1	∼5,511	.606	[.469, .727]
Top-512	512	.528	[.409, .634]
Top-256	256	.640	[.533, .746]
Top-128	128	.552	[.485, .622]
Top-64	64	.492	[.474, .500]
Top-32	32	.500	[.500, .500]
Native TopK (k=64)	64	.501	[.478, .525]

Table 11: Feature overlap between related entity pairs by SAE type.

SAE Type	L_0	Mean Jaccard	Pairs > 0.3
L1	∼5,511	0.269	35.9%
TopK (k=64)	64	0.032	0.0%

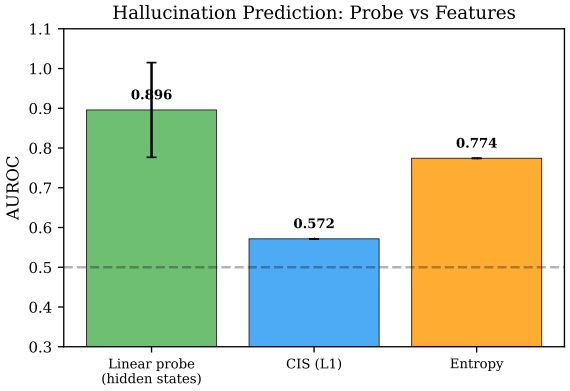

Figure 15: **The prediction gap.** Linear probes on hidden states and SAE features achieve AUROC .896 (5-fold CV on 100 unique questions), while CIS captures only a fraction of the available signal.

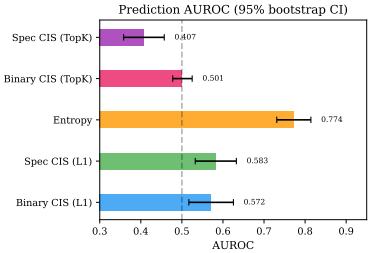

Figure 16: **Bootstrap 95% confidence intervals for prediction AUROC.** Permutation tests confirm that CIS-guided suppression has exactly zero effect on accuracy ($\Delta = 0.0$ pp, $p = 1.0$ for both L1 and TopK SAEs).

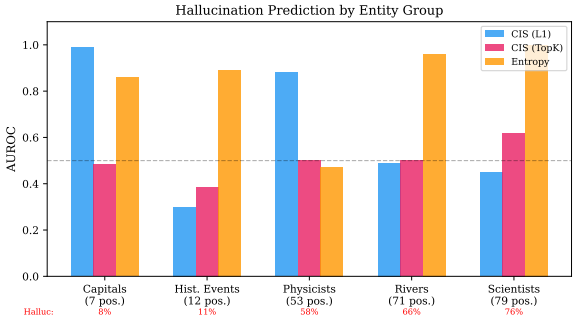

Figure 17: **Per-group hallucination prediction.** CIS (L1) shows dramatic heterogeneity across entity groups, while TopK CIS is uniformly non-predictive. Entropy maintains high predictive power across most groups. n_+ indicates the number of hallucinated examples per group; groups with very few positives (e.g., Capitals, $n_+ = 7$) yield unreliable per-group AUROCs. Hallucination rates shown below group labels.