Blind Single-Layer Activation Edits Show a Break/Fix Asymmetry in Factual Recall

Zacharie Bugaud
Astera Institute
zacharie.bugaud@gmail.com

Abstract

Can factual errors in language models be repaired by editing a single hidden activation at inference time? We compare *blind* edits, which are not told the correct answer, with oracle edits that receive answer-specific information. On Pythia-6.9B, with corruption replicated on Pythia-1B and GPT-2 XL, we find a strong break/fix asymmetry: single-layer perturbations easily corrupt correct factual recall, flipping 74–100% of initially correct answers, but blind repair is much harder. On EntityConfusion, twelve blind non-gradient interventions from four families fail to repair stable hallucinations in the strict single-layer setting; relaxed multi-layer or multi-head variants improve net accuracy by only +3 percentage points. Blind gradient optimization repairs more errors, but often breaks already-correct answers. In contrast, oracle edits given the correct answer repair many more hallucinations, fixing 68% at the default layer and up to 82% at a better layer. These results suggest that the main barrier is not whether factual recall can be steered, but whether a blind method can identify the right target-specific direction. TriviaQA is a boundary case: blind confidence maximization outperforms the single-token oracle, but the comparison is complicated because evaluation accepts multiple aliases.

1 Introduction

This paper asks whether factual errors can be repaired by editing a model's hidden activations at inference time. The key difficulty is that a repair method may or may not know what answer it is trying to produce. We therefore distinguish *blind* methods, which receive no answer-specific target, from *targeted* and *oracle* methods, which do. This distinction is central: a model may contain a recoverable answer direction, while a blind method may still fail to identify it.

Language models often produce fluent but factually incorrect answers (Ji et al., 2023; Zhang et al., 2023). One approach is to edit the model's internal activations at inference time: adding a direction to a hidden state, suppressing features, or optimizing a perturbation vector (Li et al., 2024; Zou et al., 2023; Burns et al., 2023). Unlike weight editing, these interventions are temporary and can be applied per-example.

We use the following terminology throughout. A method is **blind** if its objective contains no candidate answer, alias, or answer embedding; blind methods may use training-set labels to learn generic directions but never consult the gold answer at test time. We separate blind methods into **blind non-gradient** methods (SAE feature edits, direction steering, probe/ITI steering, activation patching—12 interventions from 4 families, including one random-suppression control, none using per-instance optimization) and **blind gradient** methods (per-instance optimization of confidence, margin, or entropy using only the model's own output distribution; answer-agnostic, though confidence maximization may reinforce the current top token). A **self-targeted** method derives a per-question intervention from the model's own hidden states (e.g., decode-and-steer); it is not blind because it constructs a candidate answer embedding. A **target-informed** method receives question-specific signal about the correct answer at test time (e.g., a contrast prompt known to elicit the gold answer). An **oracle** receives the known correct answer and serves as a controlled reference, not a universal ceiling.

We study single-layer activation edits for factual-recall tasks, primarily on Pythia-6.9B with corruption replicated on Pythia-1B and GPT-2 XL and oracle repair on GPT-2 XL, across two benchmarks. Our main finding is a **break/fix asymmetry**: under the tested perturbation families, perturbing a single layer easily corrupts correct factual recall (74–100% of correct answers flipped), but blind repair is far less effective. Across twelve blind non-

gradient interventions from four families (including one random-suppression control; evaluated on their respective pools; Tables 4, 3), the best net accuracy gain is $+3$ pp on V1, from relaxed multi-layer or multi-head upper-bound variants only (8-layer probe and top-10-head ITI). Among V1-evaluated families, all blind non-gradient single-layer single-direction or single-source methods fix $0/37$ stable hallucinations; on 500-raw, CIS suppression and SAE exchange have near-zero effect ($|\Delta| \leq 1.6$ pp) while direction patching collapses accuracy. Linear probes distinguish correct from hallucinated generations (AUROC .896 on EntityConfusion), but steering along the single-layer probe direction does not repair the outputs.

A natural objection is that non-gradient methods are too constrained. We therefore also study blind per-instance gradient objectives. Unconstrained confidence maximization fixes 39% on EntityConfusion, but the optimized vectors are $3.5\times$ larger than oracle vectors, point in unrelated directions (cosine -0.04; §4.1), and break 48% of correct answers (net: -9 questions, counting fixes and breaks from separate evaluation subsets). When capped to oracle norm, confidence maximization drops to 11%, though norm-capped margin maximization retains 26%; both remain far below the oracle's 68%.

To isolate why repair fails, we conduct oracle ablations (Table 5): a single Adam step achieves the same aggregate fix rate as 50 steps; optimizing toward a wrong or shuffled answer fixes $\leq 3\%$; and Gaussian noise (σ=0.25) drops the fix rate by more than half (from 68% to 30%). These results are consistent with a *direction-selection bottleneck* on EntityConfusion: high-rate, low-collateral repair appears to require target-conditioned directional information that blind methods fail to identify. The oracle gradient is one effective direction, but prompt-contrast results show it is not necessarily unique (§4.1).

The asymmetry varies by error type. EntityConfusion errors are often *latent-recall* errors (median correct-answer rank 16); TriviaQA errors appear predominantly *far-from-surface* under the canonical first token (median rank 1,275; alias proximity unverified). On latent-recall errors, confidence maximization is audit-set net-negative. On TriviaQA, the tested blind non-gradient steering methods still fail ($\approx 1\%$) but blind gradient optimization fixes 62% while breaking 42% of correct answers (audit-set net $+10$; above the single-token oracle; §4.2).

Self-targeted and target-informed methods partially succeed: on EntityConfusion, a gradient oracle fixes 68%, decode-and-steer fixes 8% in-domain (0% on TriviaQA), and prompt-contrast fixes 90% per question (n=21; in-domain only). CIS yields only a weak signal (AUROC 0.58) that vanishes under Top-K SAEs; CIS-guided suppression has zero net accuracy effect (§5).

Contributions.

1. **The break/fix asymmetry.** Corruption (74–100%, replicated across three models using different strong perturbation types) is consistently easier than blind non-gradient repair ($\leq +3$ pp net on Pythia-6.9B, the only model where blind repair was evaluated); blind gradient methods show mixed results with substantial collateral on EntityConfusion (Tables 3, 5).

2. **A target-specific direction-selection bottleneck on EntityConfusion.** Oracle ablations are consistent with blind methods failing to identify an effective per-question direction for high-rate, low-collateral repair (Table 5).

3. **Boundary conditions on repair.** Oracle-gradient edits fix 68% at L_{30} (82% at L_{24}) on Pythia-6.9B (58% on GPT-2 XL); per-question prompt-contrast fixes 90% (n=21). On TriviaQA, blind confidence maximization outperforms the single-token oracle (62% vs. 22%), possibly due to multi-alias evaluation (§4.2).

4. **CIS negative result.** CIS from dense L_1 SAEs ($L_0 \approx 5{,}511$ active features) yields a weak hallucination signal (AUROC 0.58) that vanishes under Top-K SAEs, and CIS-guided suppression has zero net accuracy effect (§5).

Table 2 maps each claim to its evaluation pool, sample size, and result.

2 Background

Factual errors in LMs. Language models routinely produce text that contradicts established knowledge (Ji et al., 2023; Zhang et al., 2023). Prior work detects (Kadavath et al., 2022; Manakul et al., 2023), mitigates (Ouyang et al., 2022; Lewis et al., 2020; Li et al., 2024), and benchmarks (Lin et al., 2022; Min et al., 2023; Li et al., 2023) such errors, but largely treats them as a behavioral phenomenon without probing internal mechanisms.

Regime	Signal	Repair result	Scope / note
Corruption	Perturbation[‡]	74–100%	3 models
Blind non-grad	No target ans.	0%/+3 pp[†]	tested 4 fam.
Blind gradient	No target ans.	11%[*]	norm-ctrl
Self-targeted	Latent→steer	8%	0% Triv.
Prompt contrast	Δh	90% per-Q	n=21
Oracle (EntConf)	Correct grad.	68%/58%[§]	2 models
Oracle (Triv)	Single-token cor.	22%	Py-6.9B

[*]Conf. max at oracle $\|v\|$; 39% uncapped but audit-set net −9.
[†]0% for single-dir./source methods; best relaxed variant (8-layer probe): 16% gross; best net: +3 pp.
[‡]SAE feature zeroing (100% on Py-6.9B V1) or random noise (74–96% elsewhere).
[§]Default Py-6.9B L_{30}; 82% at L_{24}; 58% on GPT-2 XL.

Figure 1: **Intervention regimes ordered by target information.** Blind non-gradient methods rarely repair factual errors. On EntityConfusion, repair improves with answer-specific directional information.

Features in superposition. The *superposition hypothesis* (Elhage et al., 2022) posits that neural networks encode more features than dimensions; SAEs (Sharkey et al., 2022; Bricken et al., 2023; Cunningham et al., 2024; Templeton et al., 2024) decompose these into interpretable features. Our starting hypothesis was that feature-level interference (CIS) might predict or repair hallucinations (Figure 2); its failure (§5) motivated the broader study.

Knowledge localization and editing. Factual knowledge is stored in specific components (Geva et al., 2021; Meng et al., 2022); weight-editing methods (ROME (Meng et al., 2022), MEMIT (Meng et al., 2023)) modify factual associations; representation engineering (Zou et al., 2023) steers along linear directions. Our work targets activation-level interventions for factual-error repair.

Representation-based error detection. CCS (Burns et al., 2023) extracts truth directions from contrast pairs; linear probes reach high AUROCs for hallucination detection (Kadavath et al., 2022; Marks and Tegmark, 2024); ITI (Li et al., 2024) shifts attention heads. Our results complicate this: probes detect hallucinations (AUROC .896), yet steering along the probe direction produces zero repair.

3 Experimental Setup

Models. Our primary model is Pythia-6.9B (32 layers, d=4,096) (Biderman et al., 2023). We extend selected analyses to Pythia-1B (16 layers, d=2,048) and GPT-2 XL (48 layers, d=1,600): corruption on both, oracle repair on GPT-2 XL. For cross-model corruption, noise is calibrated to $1.5\times$ the mean hidden-state norm at the near-final layer.

Table 1: Evaluation settings. The 500-raw pool is the undeduplicated EntityConfusion prompt set; V1 is the 100-question deduplicated subset; V2 extends V1 with 16 additional entity groups (312 Qs total). [*]One borderline item is hallucinated under the gradient evaluation environment but not under the canonical baseline; gradient, oracle, and D&S analyses include it. [†]Break rates use a separate 50-question correct-answer audit subset. [‡]Probes/ITI use 80 hallucinations; oracle/confmax use 50.

Setting	Model	Lyr	Hal. n	Cor. n	Methods
V1 (probes)	Py-6.9B	24	37	63	probe, ITI
V1 (act. patch.)	Py-6.9B	multi[##]	37	63	act. patching
V1 (gradient)	Py-6.9B	30	38[*]	50[†]	oracle, grad
V1 (D&S)	Py-6.9B	24	38[*]	—	D&S
500-raw	Py-6.9B	24	∼222	∼278	CIS, SAE edits
V2	Py-6.9B	30	102	50[†]	robustness
V1	Py-1B	n-f[§]	—	—	corruption
V1[‖]	GPT-2 XL	n-f[§]	50	50	corrupt., oracle
TriviaQA	Py-6.9B	24/30[¶]	80/50[‡]	50	probes, orac., grad

[§]n-f = near-final layer of respective model.
[‖]Same V1 prompts; halluc./correct counts are model-specific.
[##]Same-group: 16 layers; same-entity: all 32 layers.
[¶]Probe/ITI at L_{24}; oracle/confmax at L_{30}.

Benchmark. The EntityConfusion raw pool contains 500 prompts (multiple phrasings per entity-attribute pair) spanning 37 entities in 5 semantic groups (European capitals, physicists, scientists, rivers, historical events); on V1, only 11% of hallucinations (4/37) are same-group substitutions. We define **V1** as the 100-question deduplicated subset (one question per entity-attribute pair; 37 hallucinations, 63 correct); V2 extends V1 with 16 additional entity groups (312 Qs total, 102 halluc.). SAE intervention experiments use the undeduplicated 500-prompt pool ("500-raw" in Table 1); CIS prediction is reported on V1. Different intervention families use different pools; all tables note the pool and baseline.

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 (e.g., model output "The Danube River" matches gold "Danube"). On EntityConfusion, re-evaluating with strict exact match changes no intervention outcome. For TriviaQA, evaluation accepts any of 5–15 aliases per question; this matters for interpreting confidence-maximization results (§4.2).

Default protocol. Unless stated otherwise, results refer to Pythia-6.9B on EntityConfusion V1. Probe, ITI, and activation-patching methods are evaluated on the 37 stable V1 hallucinations; SAE feature edits, CIS suppression, and direction patching use the 500-raw pool (Table 1). Gradient and

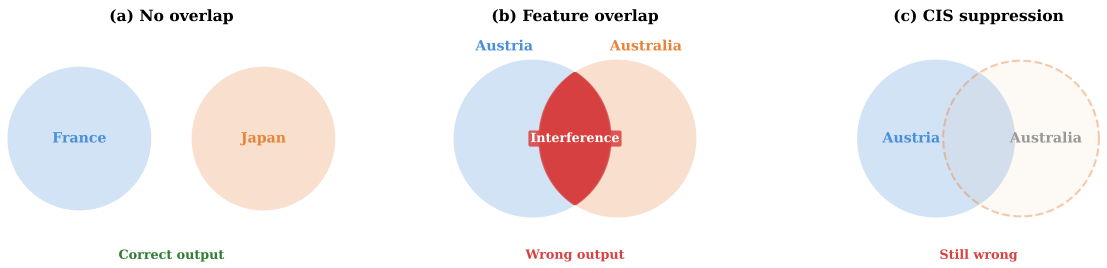

Figure 2: **Starting hypothesis: feature interference.** (a) Concepts with non-overlapping features produce correct outputs. (b) When concept features overlap in superposition, interference could corrupt the output. (c) We tested whether suppressing interfering features via CIS restores correct generation;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

oracle methods use the 38-item gradient pool (one borderline item is hallucinated only in the gradient evaluation environment). For gradient/oracle methods, correct-answer break rates are measured on a separate 50-question correct-answer audit subset; probe/ITI net changes on V1 are computed over the full 100-question pool. Single-layer interventions are our default scope; the best blind non-gradient results (+3 pp) are tied between the 8-layer probe and top-10-head ITI variants and are reported as upper bounds.

Intervention framework. We organize activation-level interventions into six regimes 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 non-gradient*: SAE feature edits, direction steering, probe/ITI steering, activation patching (12 interventions from 4 families, including one random-suppression control). (3) *Blind per-instance optimization*: gradient descent on the model's own output distribution, maximizing top-token probability (confidence), the gap between top-two tokens (margin), or minimizing output entropy, for 50 Adam steps without correct-answer access. (4) *Self-targeted*: decode-and-steer (derives direction from the model's own decoded embedding). (5) *Target-informed*: prompt-contrast. (6) *Oracle*: gradient optimization toward the known correct answer (controlled reference). All gradient-based interventions target the residual stream at a single layer (default: L_{30}; SAE/probe methods use L_{24}). For steering-based non-gradient methods, the intervention is $\alpha \cdot \hat{v}$ where $\hat{v}$ is the learned direction; SAE feature edits and activation patching modify the hidden state differently (Eqs. 3, Table 4). For gradient methods, v is optimized directly. The vector is added at the final sequence position, reapplied at each autoregressive step; gradient objectives are optimized on the first generated token. Multi-layer variants are tested where noted. We report bootstrap 95% CIs (n=2,000) for aggregate statistics, one-sided Clopper–Pearson intervals for binomial fix/break rates, and permutation tests (n=10,000). Because several intervention families were explored with small hallucination pools, best-hyperparameter results should be read as optimistic upper bounds rather than held-out deployment estimates.

Blind gradient objectives. Let $p_v(y \mid x)$ and $z_v(y \mid x)$ denote the first-token probability and logit under additive intervention v. Entropy minimization minimizes $H(p_v)$. Confidence maximization maximizes $\max_y \log p_v(y \mid x)$; the maximizing token is reidentified at each step. Margin maximization maximizes $z_v(y_{(1)}) - z_v(y_{(2)})$; which tokens are top-two is recomputed per step. None of these objectives contains the gold answer. All gradient-based interventions use Adam for the specified number of steps and are evaluated by the final generated string, not by the optimized first-token objective alone. The oracle optimizes $\log p_v(y^* \mid x)$ where y^* is the first token of the known correct answer.

Noise ablation protocol. The noisy-oracle experiment (Table 5) adds Gaussian noise to the one-step oracle direction: $v' = v_{\text{oracle}} + \epsilon$, where each coordinate $\epsilon_j \sim \mathcal{N}(0,\, \sigma^2 \cdot \|v_{\text{oracle}}\|^2)$. Because the noise is i.i.d. across d=4,096 coordinates, the expected noise norm is $\sigma\sqrt{d}\,\|v_{\text{oracle}}\| \approx 64\,\sigma\,\|v_{\text{oracle}}\|$; even σ=0.10 adds noise ∼6× the oracle norm. The perturbed vector is not renormalized, so these rows conflate directional and magnitude perturbation; we do not draw angular-precision conclusions from this experiment alone. Results are averaged over 5 noise seeds; the core direction-selection evidence comes from the target-identity, norm-capping, and cosine-alignment controls (§4.1).

Table 2: **Evidence map.** Each claim is tied to one pool, sample size, and result. V1 has 37 stable hallucinations; oracle and gradient analyses include one borderline question (n=38).

Claim	Pool	n	Result	Ref.
Break/fix asymmetry				
Corrupt breaks	V1,V2,Tr[††]	varies	74–100%[§]	§4.1
Blind n-g (dir/src)	V1	37	0/37	T4
Blind n-g (relaxed)	V1	37/63	+3 pp net	T3
Blind n-g fails	Triv	80	≈1%	§4.2
Direction-selection bottleneck				
Oracle gap	V1	38	68%	T5
1-step=50-step	V1	38	68%=68%	T5
Wrong/shuf fails	V1	38	≤3%	T5
Noise σ=0.25[◇]	V1	38	30%	T5
Capped conf. max	V1	38	11%	T5
Alignment	V1	38	−0.04	§4.1
Orac L_{24}	V1	38	82%	§4.1
Orac replic	GPT2	50	58%	§4.1
Oracle limits				
Orac replic	Triv	50	22%	§4.2
Self-targeted / target-informed				
D&S in-dom	V1	38	8%	T4
D&S out-dom	Triv	80	0%	§4.2
Prompt-contr	V1	21	90%	§4.1
Weight-editing baseline				
ROME edit	V1[¶¶¶]	59	≤2%	§6
CIS negative result				
CIS predict	V1	100	AUC .58	§5
CIS suppress	500-raw	∼500	Δ=0	T4

[§] SAE feature zeroing on Py-6.9B V1; random noise elsewhere.
[◇] Combined direction/magnitude stress test (not renormalized).
[††] V1 tested on 3 models (Py-1B 74%, GPT-2 XL 96%); n varies by pool.
[¶¶¶] n=59 reflects ROME's entity-pair testing; not the V1 hallucination count.

4 The Break/Fix Asymmetry

Table 4 organizes the blind methods into four families: (a) SAE feature editing (CIS suppression and exchange; §5); (b) direction patching; (c) hidden-state steering (probe steering, ITI); (d) activation patching; plus self-targeted decode-and-steer. Table 3 summarizes the core result: within every benchmark shown, corruption flips 74–100% while blind non-gradient methods yield at most +3 pp net on V1 (from relaxed multi-layer/multi-head variants only). Three tiers emerge: (i) single-direction/source steering methods fix 0/37 on V1 (n=38 for gradient analyses); (ii) multi-layer or multi-head variants reach +3 pp net; (iii) blind gradient methods fix more but with collateral (Table 5). Our default scope is single-layer interventions.

4.1 Intervention Results

Blind non-gradient methods do not yield reliable repair. Table 4 reports twelve interventions from four families (including one random-suppression control). CIS suppression produces zero accuracy change ($p > 0.99$); four SAE exchange variants yield at most 1.4 pp; direction patching collapses accuracy to 15.4%. Probe steering has no effect at layer 24; multi-layer steering fixes 6 but breaks 3 (net +3.0 pp). ITI fixes 4 while breaking 1 (+3.0 pp net); activation patching fixes none. The multi-layer (8-layer) probe and multi-head (top-10) ITI cap the net gain at +3 pp.

Each method underwent systematic hyperparameter search; we cannot rule out that a future method with a novel inductive bias could succeed.

Table 3: **Within-benchmark comparison (not same-sample): blind non-gradient methods do not yield more than small gains on any benchmark tested.** Blind non-gradient methods achieve at most +3 pp net accuracy on V1; blind gradient methods have higher gross fix rates but substantial collateral (Table 5). Oracle/gradient use n=38 on V1; break rates use a separate 50-question correct subset. D&S fixes 8% on V1, 0% on TriviaQA.

Pool	Model	Corrupt	Blind n-g	Orac.	n_{eval}
V1	Py-6.9B	100%	+3 pp	68%[†]	37[◇]/63
V2	Py-6.9B	84%	+2 pp[§§]	78%	varies[#]
V1	Py-1B	74%	n.r.[⋆]	n.r.[⋆]	
V1	GPT-2 XL	96%	n.r.[⋆]	58%	50
Triv.	Py-6.9B	82%	≈1% gross[¶]	22%[‡]	80/50[¶¶]

Blind n-g = blind non-gradient (net pp where measured; gross rate on Triv.); blind gradient results in Table 5.
[⋆] n.r. = not run for this model.
[†] Oracle uses n=38 (one item fluctuates across runs).
[¶] ITI fixes 1/80 (1% gross); probe steering 0/80; probe corruption flips 5/50 correct at α=16.
[‡] Blind conf. max fixes 62% on Triv., above oracle; see §4.2.
[¶¶] 80 for probe/ITI; 50 for oracle/confmax.
[#] Oracle: 50-question subset of 102 V2 hallucinations; blind n-g: full V2 pool; corruption: separate correct-answer subset.
[§§] 8-layer probe on V2 (n=102 halluc.); method and α matched to V1 best.
[◇] 37 stable halluc./63 correct; oracle/gradient use 38 (one borderline item; see [†]); +3 pp is over the full 100-Q pool.

Corruption is easy across models (perturbation types differ). We test corruption by (1) zeroing entity-specific SAE features on Pythia-6.9B (100% flipped) and (2) norm-calibrated random noise (1.5× mean hidden-state norm) for cross-model and cross-pool tests. The perturbation type differs (SAE feature destruction vs. random noise), so cross-model rates are not directly comparable; the shared conclusion is that correct recall is fragile under strong single-layer perturbations. Random noise flips 74% on Pythia-1B, 96% on GPT-2 XL, 84% on V2, and 82% on TriviaQA.

Oracle ablation isolates the direction-selection bottleneck. To understand why blind methods fail, we compare them against a gradient oracle that optimizes a steering vector toward the known correct answer at layer 30. The oracle fixes 26/38 hallucinations (68%, CI [51%, 82%]), far above all

Table 4: Activation-level repair methods on EntityConfusion. Each section header specifies the evaluation pool and corresponding baseline accuracy. SAE and direction-patching methods use the full 500-prompt pool (baseline 55.6%, which includes duplicate phrasings); probe, ITI, and patching methods use the deduplicated V1 pool (100 Qs, baseline 63.0%, 37 stable hallucinations). Decode-and-steer uses n=38 hallucinations (one borderline question included). Methods fixing 0/37 on V1 have a one-sided 95% Clopper–Pearson upper bound of 7.8%.

Method	Details	Acc	Δ
SAE feat. edit *(500-Q pool, base 55.6%)*			
CIS suppression	m=30	55.6	+0.0
Random suppr.	m=30	54.0	−1.6
Exchange (4 var.)	SAE feats	54–56	$\lvert\Delta\rvert\leq 1.4$
Dir. steering *(500-Q pool, base 55.6%)*			
Dir. patch	α=4.0	15.4	−40.2
Hidden steering *(V1, base 63.0%)*			
Probe (L_{24})	$\alpha\in[.5, 16]$	63.0	+0.0
Probe (8 layers)	α=4.0	66.0	+3.0
ITI top-10	α=4.0	66.0	+3.0
Act. patching *(V1, base 63.0%)*			
Same-group	16 layers	—	0/37 fixed, 0 broken
Same-entity	all layers	—	0/10 fixed
Decode-&-steer *(V1, n=38 halluc)*			
Answer decoder	L_{24}, best α	—	3/38
Leave-one-out	same	—	3/38

blind methods on EntityConfusion. With our optimizer and hyperparameters, joint multi-layer optimization (all 32 layers) did not outperform single-layer L_{30} (68%), despite the search space containing better single-layer solutions (L_{24}: 82%); we treat this as an optimization failure, not evidence that multi-layer edits lack additional capacity. L_{30} is the default for comparability with blind gradient experiments. Table 5 disentangles three factors:

Budget. A single Adam step at L_{30} fixes 26/38 (68%), the same aggregate rate as 50 steps; additional steps increase norm from 32 to 53 without changing which questions are fixed.

Target identity. The wrong-answer oracle changes all 38 outputs but fixes 0/38. The shuffled-target oracle, which optimizes toward a randomly reassigned correct answer, fixes only 1/38 (3%), even below the random-direction rate (11%), possibly because answer-specific gradients toward another entity actively push mass away from the correct answer. The oracle requires the correct answer for the specific question at hand.

Direction-and-magnitude stress test. Adding Gaussian noise at σ=0.25 (expected noise norm $\sim$16$\times$ the oracle vector) drops the fix rate to 30%; at σ=1.0, to 6%. Because noise dominates the signal at $\sigma \geq 0.25$, this confirms sensitivity to very

Table 5: Oracle ablation (n=38, Pythia-6.9B, L_{30}). Noisy oracle: Gaussian noise with per-coordinate standard deviation $\sigma \cdot \|v_{\text{oracle}}\|$ added to the 1-step direction (see §3); the perturbed vector is not renormalized, so these rows conflate directional and magnitude perturbation. Budget-parity: same 50-step Adam with blind objectives. Norm-capped: $\|\mathbf{v}\| \leq 53$ (50-step oracle mean). Break counts are measured on a separate 50-question correct subset. Fractional counts (noisy oracle, random direction) are averages over 5 seeds.

Condition	Fixed	Rate	Broken
Correct-answer oracle			
1 step	26/38	68%	—
50 steps	26/38	68%	6/50
Noisy oracle (1-step + 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%	—
Rand. dir. @ oracle $\|v\|$	4.2/38	11%	—
Rand. dir. @ confmax $\|v\|$	2.6/38	7%	—
Budget-parity (50 steps, blind, unconstrained)			
Entropy minimization	3/38	8%	—
Confidence maximization	15/38	39%	24/50
Margin maximization	15/38	39%	24/50
Norm-capped ($\|\mathbf{v}\| \leq 53$)			
Conf. max (norm-capped)	4/38	11%	29/50
Margin max (norm-capped)	10/38	26%	n.m.[¶]

[¶] n.m. = not measured.

large unrenormalized perturbations (σ=0.10 leaves repair intact despite $\sim$6$\times$ noise norm) but does not isolate directional precision from magnitude effects.

Interpretation. These ablations are consistent with a target-information bottleneck: on EntityConfusion, high-rate low-collateral repair appears to require question-specific target information. The strongest evidence comes from target-identity controls (wrong-answer and shuffled-target oracles fix $\leq$3%) and norm-capped confidence maximization collapsing to the random-direction rate. Confidence maximization does not recover the oracle-gradient direction (cosine −0.04), but prompt-contrast achieves 90% per-question repair with similarly low oracle cosine (.04); the bottleneck is therefore target-specific directional information, not alignment with one privileged oracle vector.

Norm-constrained analysis reveals magnitude dependence. Unconstrained confidence maximization fixes 39%, but the resulting vectors are 3.5$\times$ larger than oracle vectors ($\|v\| = 183$ vs. 53). Capping at the oracle norm reduces the fix rate to 11% (CI [3%, 25%]). Norm-capped margin maximization retains 26%, above random (11%), indicating some directional signal; however, col-

lateral damage makes the net effect strongly negative for confidence maximization (-25 questions). Gross repair is strongly magnitude-dependent, but collateral damage is not monotonic in norm: norm-capped confidence maximization breaks *more* correct answers ($29/50$) than uncapped ($24/50$); we do not fully understand this inversion. Random directions at the unconstrained norm fix only 7%, well below 39%, so the optimizer finds a weakly informative but non-oracle-aligned direction requiring outsized magnitude.

The optimized directions show no positive alignment with oracle directions (cosine -0.04, permutation $p > 0.99$); overlap in fixed questions ($9/38$) is no larger than chance. Collateral damage compounds: confidence maximization breaks $24/50$ (48%) of the audit subset, net -9 questions; norm-capped, -25 questions. The oracle breaks only $6/50$ (12%), net $+20$ questions.

Replication across pools and models. The magnitude-dependence pattern replicates on V2 (norm-capped: 10%, net -4 questions); GPT-2 XL replicates the oracle/wrong-target contrast (58% vs. 2%), with the 1-step rate (34%) below 50-step. On TriviaQA, the 1-step and 50-step rates match at 22%.

Self-targeted and target-informed methods partially succeed but do not generalize. Prompt-contrast and oracle methods exceed the blind non-gradient ceiling; decode-and-steer (8% gross) modestly exceeds the single-source zero-repair results but not the multi-layer upper bound.

Prompt-contrast uses the hidden-state difference between a prompt eliciting the correct answer and one eliciting the wrong answer. Per-question prompt-contrast fixes $19/21$ questions (90%, CI $[70\%, 99\%]$), but the global mean direction fixes only $12/38$ (32%) while breaking $12/50$ (net ≈ 0). Prompt-contrast directions are no more aligned with oracle directions than random vectors (cosine .04, permutation $p > 0.5$), showing that effective repair directions need not coincide with the oracle gradient.

Decode-and-steer trains a Ridge regression from layer-24 hidden states to the correct-answer embedding, then steers toward it. This fixes $3/38$ in-domain (8%), 0% on TriviaQA; the decoded embeddings point toward the correct answer $27/38$ of the time but steering rarely flips the output.

Oracle repair at the default L_{30} on V1 fixes 68% on Pythia-6.9B and 58% on GPT-2 XL; L_{24}: 82%, and $0.75\times$ magnitude fixes 76% (above full magnitude), so the result depends on layer and magnitude. Decode-and-steer fails on TriviaQA (0%), and the global prompt-contrast direction loses most per-question benefit (32% vs. 90%).

Oracle subspace geometry. The repair subspace is low-rank: projecting each oracle vector onto the top-10 PCs preserves 66% of the fix rate; nearest-neighbor oracle vectors (k=3) achieve 58% (a geometric diagnostic requiring gold-answer vectors). The oracle overshoots: $0.75\times$ magnitude fixes 76% and $0.5\times$ fixes 74%, both above full-magnitude 68%; L_{24} fixes 82%, suggesting the optimal layer differs from the analysis layer.

4.2 Naturalistic Validation: TriviaQA

To test whether the break/fix asymmetry extends beyond curated entity groups, we evaluate Pythia-6.9B on 300 TriviaQA (Joshi et al., 2017) questions (66% hallucination rate). We evaluate probe steering and ITI on 80 hallucinations, oracle and confidence maximization on 50.

Probe steering and ITI. A linear probe distinguishes correct from hallucinated answers at AUROC $.716 \pm .072$ (5-fold CV), above chance but below the EntityConfusion probe at .896. Despite this discriminative signal, steering with the probe direction fixes $0/80$ hallucinations (0% across all α values). ITI fixes $1/80$ (1%, only at α=16). Corruption is limited at moderate strengths ($5/50$ correct answers flip at α=16, vs. 82% from random noise), indicating the probe captures a discriminative but non-causal signal.

Decode-and-steer. A Ridge decoder recovers the correct answer closer than wrong for $89/197$ (45%, vs. 71% on EntityConfusion). Steering fixes $0/80$ (0%). Median correct-answer rank is 1,275 under the canonical first token; because evaluation accepts multiple aliases, surface proximity of the matched alias is unverified.

Single-token oracle. Oracle repair fixes $11/50$ (22%, CI $[12\%, 36\%]$), well below EntityConfusion's 68%. The 1-step=50-step pattern replicates at 22%. Wrong-answer oracle fixes only $2/50$ (4%). Oracle success tracks latent knowledge availability: 24% when the correct answer ranks in the top-100 logits vs. 13% when ranked beyond 100.

Confidence maximization and the two-regime distinction. Unconstrained confidence maximization fixes $31/50$ (62%); norm-capped, it retains 56%, a much smaller drop than on EntityConfusion, where the rate collapses from 39% to

11%. With 21/50 correct answers broken (on a balanced 50-question audit subset), the audit-set net is +10, weakly positive, unlike the net-negative on EntityConfusion.

Surprisingly, blind confidence maximization fixes more TriviaQA hallucinations than the correct-answer oracle (62% vs. 22%), possibly because the oracle targets one first token while evaluation accepts 5–15 aliases; the underlying mechanism is unclear. Additionally, many correct answers have median output rank 1,275, making the oracle gradient signal weak.

These results point to two distinct error regimes. Latent-recall errors, where the correct answer is near the output surface and the model may once have produced it, resist blind non-gradient repair. TriviaQA errors appear far from surface under the canonical first token (median rank 1,275), but alias proximity is unverified; confidence maximization may succeed by producing any accepted surface form. TriviaQA is therefore not a clean replication of the EntityConfusion direction-selection bottleneck; it is a boundary case where the evaluation protocol and error regime differ.

5 The CIS Hypothesis: A Negative Result

This section reports a negative result on our original hypothesis that feature overlap in SAE representations (CIS) would predict and enable repair of factual errors. Its failure led to the broader intervention study above.

SAE features and CIS. We train L_1 and Top-K sparse autoencoders (Bricken et al., 2023; Gao et al., 2024) on Pythia-6.9B residual stream at layer 24 (Appendix A). L_1 SAEs (16,384 features, λ=5×10^{-2}, 50M Pile tokens) yield $L_0 \approx$ 5,511 active features per input; Top-K SAEs enforce exact sparsity ($k \in \{32, 64, 128\}$, 32,768 features). For each entity c, concept features $\mathcal{F}_c$ are identified via specificity:

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

selecting the top-50 features per concept (α_s=10). The Concept Interference Score measures shared activation energy, where $z_i(x)$ denotes the activation of SAE feature i on input x:

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

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

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

CIS identifies the most-overlapping competing concept, but suppression targets features unique to that competitor to avoid destroying target-concept information; this is a conservative downstream intervention, not a direct ablation of the overlapping features in the CIS numerator.

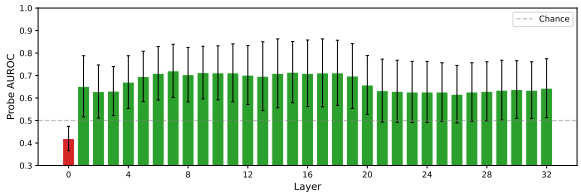

Figure 3: **Per-layer hallucination decodability.** Probe AUROC peaks around layer 24. Single-layer probe steering has no repair effect; an 8-layer variant yields +3 pp net.

CIS as predictor. CIS with L_1 SAEs yields AUROC 0.58 (CI [.533, .633]), well below entropy (0.77) and a linear probe (.896 ± .119, 5-fold CV, permutation $p < 0.001$). Critically, CIS with Top-K SAEs, which enforce exact sparsity and reduce dense co-activation, is non-predictive (AUROC 0.42–0.50 across $k \in \{32, 64, 128\}$, consistent with chance). Post-hoc sparsification of the same L_1 SAE reproduces this drop, consistent with polysemantic co-activation rather than genuine interference. For the linear probe, entity-grouped CV yields .784 ± .123, a 28% drop in above-chance AUROC, indicating roughly a quarter of the signal is entity-specific. The hallucination signal is decodable above chance at every non-embedding layer (Figure 3): probe AUROC peaks around layer 24 but steering along the single-layer L_{24} probe direction has zero net accuracy effect across all α values tested. CIS-guided suppression produces zero accuracy change ($p > 0.99$; Table 4); four SAE exchange variants yield $|\Delta| \leq 1.4$ pp.

Why CIS correlates. L_1 features overlap heavily (Jaccard = 0.269, 35.9% of pairs > 0.3); Top-K features have near-zero overlap (0.032). **Post-hoc sparsification implicates polysemanticity:** retaining only the top-k activations from the same L_1 SAE drops CIS AUROC from .606 (slightly higher than the headline .583 because the post-hoc experiment recomputes per-input CIS with uniform aggregation) at full $L_0 \approx$ 5,511 to .500 at k=32 and .492 at k=64 (chance level), matching the native Top-K result. Encoding through the SAE and de-

coding back preserves probe performance (AUROC $= .896$, $r = .9996$), confirming the failure lies in CIS, not SAE information loss. Group-specific LDA directions achieve near-perfect within-group separation (AUROC $\geq .934$) but show no cross-group alignment (mean cosine $.005$); CIS's single global statistic cannot capture this structure.

Latent-recall errors, not same-group substitution. Only 11% of wrong answers match a same-group entity (4/37); most errors are generic hallucinations. Yet correct knowledge often exists latently: a linear decoder recovers the correct answer from hallucinated states $27/38$ of the time, and the correct answer's median rank in output logits is only 16. Error is distributed across layers (Figure 4); 44% of hallucinations were answered correctly at earlier checkpoints. EntityConfusion's median correct-answer rank is 16 (knowledge is nearly surfaced), vs. $1{,}275$ on TriviaQA (far from surface under the canonical token; alias proximity unverified).

6 Discussion and Conclusion

Scope of claims. Our blind-repair experiments are primarily on Pythia-6.9B across EntityConfusion and TriviaQA; we replicate corruption on Pythia-1B and GPT-2 XL, and oracle repair on GPT-2 XL. These are relatively small, pre-instruction-tuned models. In informal tests, non-activation baselines (LoRA fine-tuning: $32/37$; retrieval-augmented generation: $36/37$) succeed at different budgets, confirming the barrier is specific to blind single-layer activation editing. ROME (Meng et al., 2022) also fails ($1/59$; the larger denominator reflects ROME's requirement to test each entity pairing, not the V1 hallucination count), likely because EntityConfusion's multi-entity structure does not match ROME's single-fact paradigm.

What holds up and what does not. The break/fix asymmetry for blind non-gradient methods is the most robust finding: across twelve interventions from four families evaluated on their respective Pythia-6.9B pools, the best net gain is $+3$ pp on V1 (from relaxed multi-layer/multi-head upper-bound variants only), corruption flips 74–100% across three models (different perturbation types), and blind non-gradient failure holds within every evaluation pool where tested (Table 3; on TriviaQA, only probe steering and ITI were evaluated). On EntityConfusion, blind gradient repair is magnitude-dependent: norm-capped confidence maximization drops to 11%, with no oracle alignment (cosine -0.04) and worse collateral than uncapped, though norm-capped margin maximization retains 26%, above random.

The TriviaQA results complicate the narrative. Blind confidence maximization outperforms the single-token oracle there (62% vs. 22%), possibly because the oracle targets a single token while evaluation accepts multiple aliases; the direction-selection bottleneck interpretation is well-supported on EntityConfusion but should not be extended to TriviaQA without further analysis.

Why corruption is easy and repair is hard. For already-correct examples, any sufficiently large perturbation overwhelms the correct answer's small logit advantage (median gap of 5.0 nats); for hallucinated examples, repair must overcome the wrong answer's logit advantage and select a target-specific correction direction. At the oracle's mean perturbation norm (53), random directions fix 11%. Feature overlap in L_1 SAEs yields a weak hallucination signal (AUROC 0.58) that vanishes under controlled sparsity (Top-K: AUROC 0.42–0.50, chance-level); the tested CIS-guided competitor-feature suppression has zero net accuracy effect, ruling out this specific intervention on this benchmark (other feature-level interventions are not tested).

Detection versus repair. Linear probes achieve AUROCs of $.896$ (EntityConfusion) and $.716$ (TriviaQA) at layer 24, yet single-layer probe steering produces zero repair on both benchmarks; an 8-layer variant yields $+3$ pp net on EntityConfusion. This is the central practical lesson: knowing *that* a model is wrong is not the same as knowing how to fix it.

Implications. On EntityConfusion, reliable low-collateral activation-level repair appears to require target-specific answer information; blind inference-time edits were unreliable or high-collateral in our setting (on TriviaQA, blind confidence maximization fares better, though evaluation asymmetries complicate interpretation). The oracle ablation shows that on EntityConfusion one gradient step suffices when aimed correctly, so geometric structure for repair exists for latent-recall errors, but exploiting it without target information remains unsolved. Whether these patterns hold for instruction-tuned models is the most pressing open question.

Conclusion. We document a break/fix asymmetry: single-layer corruption is easy (74–100%

across three models) but blind non-gradient repair consistently fails ($\leq +3$ pp net on EntityConfusion V1; $\approx 1\%$ gross on TriviaQA). Blind gradient methods fix more but with substantial collateral on EntityConfusion (on TriviaQA: 62% fixed, audit-set net $+10$). Oracle ablations on EntityConfusion implicate a direction-selection bottleneck for latent-recall errors.

Limitations

Our experiments use relatively small, pre-instruction-tuned models (Pythia-6.9B, Pythia-1B, GPT-2 XL); it remains unknown whether the break/fix asymmetry holds for RLHF-aligned models. Different intervention families use different evaluation pools (Table 1), though blind-repair failure holds within each pool (Table 3). Oracle and blind gradient methods use layer 30 while SAE/CIS uses layer 24; oracle at L_{24} achieves 82% vs. 68% at L_{30}, so layer matters but qualitative conclusions hold. Many results rely on small subsets (n=37–102; prompt-contrast: n=21); we report Clopper–Pearson CIs for binomial rates and bootstrap CIs for aggregates. The cosine similarity -0.04 between oracle and confmax vectors is near zero (permutation $p > 0.99$ for positive alignment). On TriviaQA, blind confidence maximization outperforms the single-token oracle (62% vs. 22%), possibly because evaluation accepts 5–15 aliases while the oracle targets a single token; this limits the single-token oracle as a universal ceiling but it remains informative on EntityConfusion. The noisy-oracle experiment is not a pure angular-noise test (the vector is not renormalized); we rely on target identity, norm-capping, and cosine alignment for the core direction-selection argument. We cannot rule out that a future blind method with a novel inductive bias could succeed.

Reproducibility Statement

Code for SAEs, CIS, and intervention experiments will be released. We use Pythia-6.9B (Biderman et al., 2023) and the Pile (Gao et al., 2020); SAE hyperparameters and the EntityConfusion construction are in Appendices A–B.

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 Hallahan, 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*.

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 SAE Training Details

L_1 SAEs: 16,384 features, $\lambda{=}5{\times}10^{-2}$, 50M Pile tokens, $L_0 \approx 5{,}511$. Top-K SAEs: 32,768 features, $k \in \{32, 64, 128\}$. Both trained with Adam (lr $= 3{\times}10^{-4}$) on layer 24. See code release for full details.

B EntityConfusion Dataset

500 prompts across 37 entities in 5 semantic groups (European capitals, physicists, scientists, rivers, historical events) form the raw pool. V1 retains one question per entity-attribute pair (100 Qs, 37 hallucinations); V2 extends V1 with 16 additional entity groups and includes all V1 prompts. All CV splits by question; entity-grouped CV reported separately.

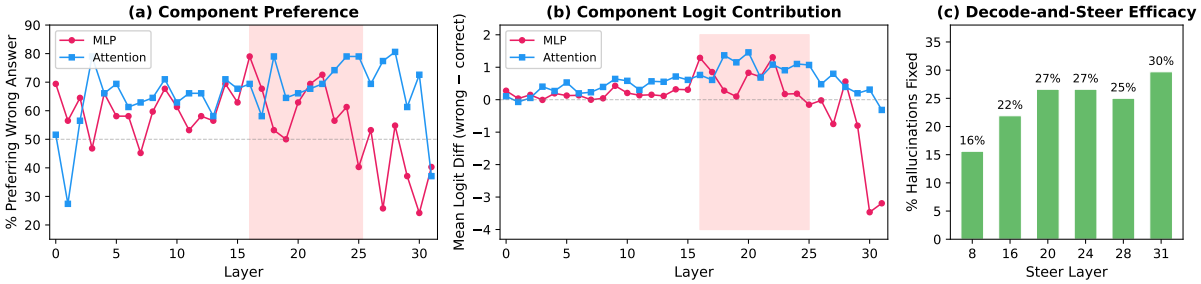

Figure 4: **Readout mechanism.** (a) Per-layer component preference for the wrong answer. (b) Mean logit contribution (wrong − correct) by MLP and attention sub-layers. Mid-layer components push toward the hallucinated answer; late-layer MLPs attempt correction. (c) Decode-and-steer fix rate by steering layer (evaluated on V2; the main-text D&S numbers use V1). Multiple layers contribute, suggesting that single-layer interventions may only partially address a distributed computation.