ABSORPTION AWARE EDITING: PREDICTING AND CLOSING THE LEAKAGE CHANNELS OF SAE-BASED CONCEPT ERASURE

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse autoencoders (SAEs) provide interpretable latent features that can be used to edit concepts in language models without retraining. The usefulness of such edits depends on concept-associated latents providing reliable control over the corresponding information in the model representation. However, ablating these latents often leaves substantial concept information behind, and existing methods provide little guidance about when these edits will fail or how to repair them. We show that the extent of feature absorption, a known SAE pathology, quantitatively predicts the resulting concept leakage, and surviving information is distributed across absorbing SAE latents and the reconstruction residual. Based on this decomposition, we propose a training-free, absorption-aware editing method that removes the target concept direction from both. Across multiple SAE training methods and model families, the proposed method substantially reduces concept leakage with minimal effects on unrelated concepts and model utility. Together, these results turn a systematic source of SAE editing failure into one that can be diagnosed, predicted, and corrected.

1 INTRODUCTION

Targeted model editing seeks to modify how a model represents or uses a particular concept while preserving its behavior elsewhere. Sparse autoencoders (SAEs) offer an appealing interface for such intervention, as they decompose language model activations into sparse, interpretable latents (Huben et al., 2024; Gao et al., 2025), allowing a practitioner to identify latents associated with a target concept and intervene on them directly. A common strategy is therefore to ablate the concept-associated latents, with the expectation that removing these latents will remove the corresponding concept from the model representation (Marks et al., 2025; Karvonen et al., 2024).

However, SAE-based editing can be incomplete. Concept information can remain detectable after the latents selected to represent that concept are ablated (Karvonen et al., 2024; 2025), and recent work shows that suppressed behavior can be recovered through information retained in the SAE reconstruction residual (Cui et al., 2026). Thus, ablating the identified concept-associated latents does not guarantee complete removal of the concept from the representation. Feature absorption, a known SAE pathology (Chanin et al., 2025), provides one possible explanation for this failure. It occurs when, on some inputs, information normally associated with a target latent is instead carried by other SAE latents. Ablating the target latent then leaves this information untouched.

Motivated by this failure mode, we study whether feature absorption can explain and predict failures of SAE-based concept erasure. We find that the extent of feature absorption predicts post-edit concept leakage and the surviving concept information can reside both in other SAE latents and in the reconstruction residual. Based on these observations, we propose a training-free, absorption-aware editing method. Specifically, our method removes the component aligned with the target concept direction from the contributions of absorbing latents and from the reconstruction residual. Across multiple SAE training methods and model families, our method substantially improves the completeness of SAE-based concept erasure while largely preserving non-target concepts and overall language model utility. Together, these results turn a systematic source of SAE editing failure into one that can be diagnosed, predicted, and corrected.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2 Related Work

Our work builds on three closely related lines of research: concept erasure, SAE-based editing, and feature absorption.

Concept erasure. Concept erasure aims to remove information about a target concept from a model's internal representation while preserving information unrelated to that concept. Existing approaches include representation-level methods that remove concept-predictive directions, such as iterative nullspace projection (Ravfogel et al., 2020), mean-difference projection, and closed-form affine erasure (Belrose et al., 2023), as well as methods that modify model weights to suppress targeted knowledge or behavior (Arditi et al., 2024; Limisiewicz et al., 2024; Guo et al., 2025). A growing line of work instead performs concept-level interventions through sparse autoencoders (SAEs), using learned latent features as editing handles. Our work focuses on this SAE-based setting and asks when ablating the latents selected to represent a concept actually removes that concept from the representation.

SAE-based editing. An SAE decomposes a model activation into sparse latent activations over learned decoder directions together with a reconstruction residual (Olah, 2023; Huben et al., 2024; Gao et al., 2025). SAE-based editing identifies latents associated with a target concept and intervenes on their contributions, for example by ablating, clamping, or rescaling their activations. Such interventions have been used for behavior steering (Templeton et al., 2026; Chalnev et al., 2024; Ferrando et al., 2025; Zhao et al., 2025; Arad et al., 2025), removal of spurious signals (Marks et al., 2025), and concept or knowledge removal (Farrell et al., 2025; Khoriaty et al., 2025; Muhamed et al., 2025; Ashuach et al., 2026; Gur-Arieh et al., 2025; Wang et al., 2025; Cywiński & Deja, 2025; Cassano et al., 2026).

The reliability of these edits depends on whether the selected latents capture the target information completely and selectively. Existing intervention-based evaluations assess whether manipulating concept-associated latents changes the intended concept or downstream behavior without excessive collateral effects (Karvonen et al., 2024; 2025). Recent work shows that SAE interventions can be incomplete or have unintended effects, including recoverability through information outside the targeted latents (Cui et al., 2026) and predictable steering side effects (Duan, 2026). These works establish that SAE interventions can fail, but leave open when and why concept erasure remains incomplete at the level of individual instances.

Feature absorption. Feature absorption provides a specific mechanism through which SAE-based concept erasure can fail. It occurs when information about a broader parent concept is carried by a more specific child latent while the corresponding parent latent remains inactive. Ablating only the parent latent can therefore leave the parent concept detectable (Chanin et al., 2025). Subsequent work extends feature absorption beyond a binary label by characterizing graded parent–child dependence (Grandien et al., 2026). Other approaches modify SAE training or architecture to better capture hierarchical structure (Bussmann et al., 2025; Muchane et al., 2025). Prior work primarily treats absorption as a property of SAE representation quality. We instead study its downstream consequence for concept erasure.

3 Background and Preliminaries

Let $\mathbf{x} \in \mathbb{R}^d$ denote a model activation at a fixed layer. A sparse autoencoder (SAE) represents $\mathbf{x}$ as $\mathbf{x} = \sum_j a_j \mathbf{d}_j + \mathbf{b}_{\text{dec}} + \mathbf{r}$, where a_j is the activation of SAE latent j, $\mathbf{d}_j$ is its decoder direction, $\mathbf{b}_{\text{dec}}$ is the decoder bias, and $\mathbf{r}$ is the reconstruction residual (Huben et al., 2024).

To study how a particular concept is represented by these SAE latents, we first need a readout of that concept in the original model representation. For a target concept L, a linear probe on $\mathbf{x}$ measures whether information about L is present, and $\hat{\mathbf{w}}_L$ denotes the corresponding unit probe direction. We keep this probe fixed throughout editing and evaluation. After editing, we use its prediction to determine whether the target concept remains detectable, and leakage is the fraction of target examples for which this occurs. SAE-based editing acts on a set of SAE latents P associated with the target concept L and ablates their contributions, i.e., $\mathbf{x}_{\text{base}} = \mathbf{x} - \sum_{p \in P} a_p \mathbf{d}_p$, where P can be obtained by using k-sparse probing over SAE latent activations $\mathbf{a}(\mathbf{x})$ (Chanin et al., 2025; Karvonen et al., 2025).

Feature absorption asks whether these parent latents consistently carry the target concept across inputs. It occurs when information associated with a broader feature is instead represented by more specific SAE latents on some inputs, while the corresponding parent latents are inactive (Chanin et al., 2025). This creates a failure mode for SAE-based editing: ablating the parent latents can leave its information intact when that information has been absorbed elsewhere. The absorption evaluation attributes such concept-related signal to non-parent latents at the token level. For an input token with activation $\mathbf{x}$, let $C(\mathbf{x})$ denote the absorbing latents attributed to that token. When the target concept is carried by absorbing latents while the parent latents are inactive, the token is fully absorbed. The partial absorption is quantified as

$$\phi(\mathbf{x}) = \frac{\sum_{c \in C(\mathbf{x})} a_c \langle \mathbf{d}_c, \hat{\mathbf{w}}_L \rangle}{\sum_{p \in P} a_p \langle \mathbf{d}_p, \hat{\mathbf{w}}_L \rangle + \sum_{c \in C(\mathbf{x})} a_c \langle \mathbf{d}_c, \hat{\mathbf{w}}_L \rangle}. \tag{1}$$

For a fully absorbed token, all parent latents are inactive, so $a_p = 0$ for every $p \in P$ and $\phi(\mathbf{x}) = 1$.

4 Feature Absorption as a Source of Editing Failure

Feature absorption allows concept information to be carried outside the parent latents targeted by standard SAE editing. We first study whether this structure systematically explains the failure of standard SAE editing.

The token-specific set $C(\mathbf{x})$ in Eq. 1 contains the non-parent latents attributed to absorption for a particular token. To describe the broader set of non-parent latents that may carry target-aligned signal, we define

$$C = \{j \notin P : \cos(\mathbf{d}_j, \hat{\mathbf{w}}_L) \geq \tau\},$$

which contains non-parent latents whose decoder directions are sufficiently aligned with the target concept direction. Unlike $C(\mathbf{x})$, C is fixed for a given concept and SAE.

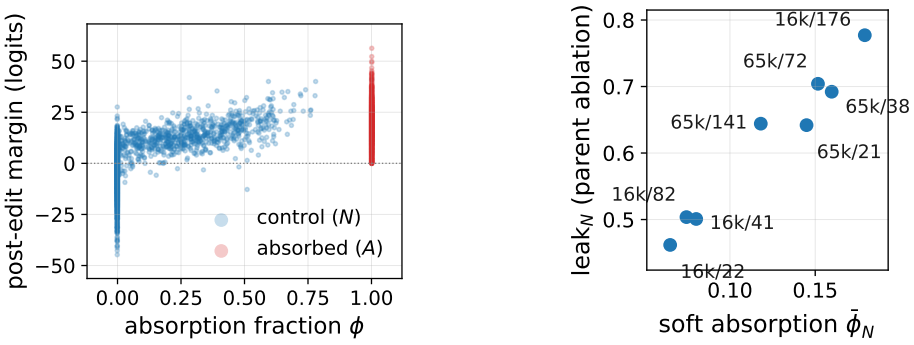
Under the linear-probe readout, the component of the model representation along the target concept direction decomposes as

$$\langle \mathbf{x}, \hat{\mathbf{w}}_L \rangle = \underbrace{\sum_{p \in P} a_p \langle \mathbf{d}_p, \hat{\mathbf{w}}_L \rangle}_{\text{parent latents}} + \underbrace{\sum_{c \in C} a_c \langle \mathbf{d}_c, \hat{\mathbf{w}}_L \rangle}_{\text{candidate latents}} + \underbrace{\sum_{j \notin P \cup C} a_j \langle \mathbf{d}_j, \hat{\mathbf{w}}_L \rangle}_{\text{other latents}} + \underbrace{\langle \mathbf{b}_{\text{dec}}, \hat{\mathbf{w}}_L \rangle}_{\text{decoder bias}} + \underbrace{\langle \mathbf{r}, \hat{\mathbf{w}}_L \rangle}_{\text{residual}}. \tag{2}$$

The standard SAE edit removes only the contribution from the parent latents. For a fully absorbed token, $a_p = 0$ for every $p \in P$, so parent ablation leaves the activation unchanged, i.e., $\mathbf{x}_{\text{base}} = \mathbf{x}$. For a partially absorbed token, a larger $\phi(\mathbf{x})$ indicates that more of the contribution along the target concept direction is carried by absorbing latents. We therefore examine, at the token level, how the pre-edit absorption fraction $\phi(\mathbf{x})$ is associated with the amount of target-concept signal that remains after editing, as measured by the post-edit probe margin.

We use the first-letter task from the SAEBench absorption evaluation (Karvonen et al., 2025) as our primary diagnostic setting. Each target concept L corresponds to one of the 26 initial-letter classes, and concept information is read from the activation of the word token using the 26-way linear probe described in Section 3. Our primary analysis uses the Gemma-2-2B layer-12 Gemma Scope SAE with 16k latents and $\ell_0 = 82$ (Lieberum et al., 2024). Figure 1a shows that the post-edit probe margin is strongly correlated with the absorption fraction across all evaluated tokens (Spearman's $\rho = 0.80$). The correlation remains substantial when we restrict the analysis to not fully absorbed tokens ($\rho = 0.66$). Consistent with the definition of full absorption, parent ablation leaves the target concept detectable on 99.95% of fully absorbed tokens, compared with 53.7% of not fully absorbed tokens. Therefore, tokens with a larger absorption fraction retain more target-concept signal after parent ablation.

We next ask whether this relationship also reflects systematic differences in editing reliability across SAEs. Let A denote tokens labeled fully absorbed for a given concept, and N denote tokens that are not fully absorbed of the same concept. Examples in N may therefore have $\phi(\mathbf{x}) > 0$. Let π_A denote the fraction of tokens that are fully absorbed and $\bar{\phi}_N = \frac{1}{|N|} \sum_{\mathbf{x} \in N} \phi(\mathbf{x})$ denote the mean absorption fraction among not fully absorbed tokens, and let leak_N denote the fraction of these tokens for which the target concept remains detectable after parent ablation. Figure 1b shows that leak_N is strongly correlated with $\bar{\phi}_N$ across the eight Gemma Scope SAEs (Spearman's $\rho = 0.93$).



(a) Token level: ρ=0.80

(b) SAE level: ρ=0.93

Figure 1: Feature absorption is correlated with failure of standard parent ablation at both the token and SAE levels. (a) On the Gemma-2B 16k/L_0=82 SAE, each point represents one evaluated token activation. The post-edit probe margin is correlated with the absorption fraction $\phi(\mathbf{x})$ ($\rho = 0.80$ over all evaluated tokens; $\rho = 0.66$ over not fully absorbed tokens only). (b) Across eight Gemma Scope SAEs, leak_N is correlated with the mean absorption fraction $\bar{\phi}_N$ among not fully absorbed tokens ($\rho = 0.93$). Point labels are shown as A/B, where A denotes the number of SAE latents and B the average number of active latents per token.

We further test whether this relationship generalizes across both concepts and SAEs. Here, each sample is one (SAE, L) pair, giving 132 samples in total. We fit a ridge regressor to predict leak_N and evaluate it using leave-one-SAE-out prediction. A baseline model using SAE and concept-localization features achieves a Spearman correlation of 0.72 with leak_N on held-out SAEs (details are available in Appendix A.4). Adding $(\pi_A, \bar{\phi}_N)$ as predictors increases this to 0.93, while using only $(\pi_A, \bar{\phi}_N)$ achieves 0.85. These results show that the relationship between absorption and editing failure extends beyond individual tokens, and absorption statistics can predict relative editing failure across concepts in held-out SAEs. Motivated by these findings, we propose an absorption-aware operator for editing.

5 Absorption-Aware Operator

Motivated by the decomposition in Eq. 2, we propose an operator which targets two channels through which absorption-related concept signal can survive parent ablation, specifically, the candidate non-parent latents C and the reconstruction residual.

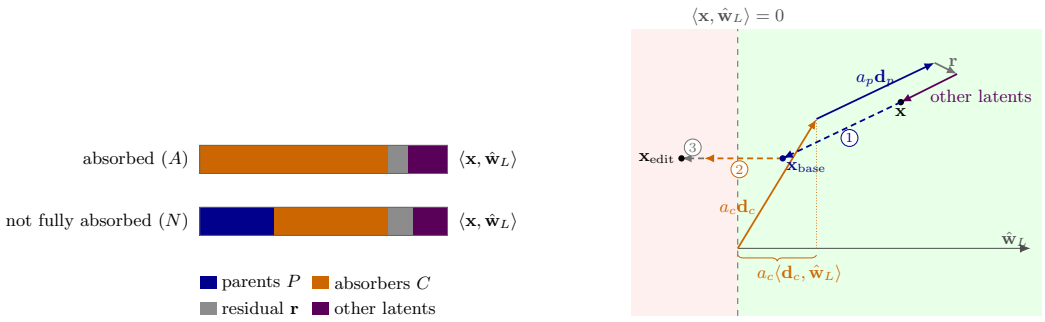
Given a concept L with probe direction $\hat{\mathbf{w}}_L$ and parent set P, let $\mathbf{a} = \text{enc}(\mathbf{x})$ be the SAE activations of an unedited token. We correct the dictionary and residual channels as

$$\mathbf{x}_{\text{edit}} = \mathbf{x}_{\text{base}} - \sum_{c \in C} a_c \langle \mathbf{d}_c, \hat{\mathbf{w}}_L \rangle \hat{\mathbf{w}}_L - \langle \mathbf{r}, \hat{\mathbf{w}}_L \rangle \hat{\mathbf{w}}_L. \tag{3}$$

Figure 2 illustrates the target-direction decomposition and the corresponding absorption-aware edit. The first correction term in Eq. 3 is the *dictionary projection*. For each $c \in C$, we subtract only the component of its contribution $a_c \mathbf{d}_c$ along $\hat{\mathbf{w}}_L$, while preserving the component orthogonal to $\hat{\mathbf{w}}_L$. Because the correction scales with a_c, only candidate latents that are active on the current token contribute to the edit, and the operator does not require a token-specific decision about whether the token is absorbed.

The second correction term is the *residual projection*. The reconstruction residual $\mathbf{r} = \mathbf{x} - \hat{\mathbf{x}}$ can also retain information along the target concept direction. We therefore remove its projection onto $\hat{\mathbf{w}}_L$, closing the channel the dictionary term cannot reach.

Combining these two parts, our operator differs from projecting $\mathbf{x}_{\text{base}}$ as a whole onto the orthogonal complement of $\hat{\mathbf{w}}_L$. It removes target-aligned contributions only from the selected non-parent latents and the residual, while leaving the remaining latent contributions unchanged. The operator requires



(a) Target-direction decomposition

(b) Absorption-aware operator

Figure 2: Illustration of the absorption-aware operator. (a) $\langle \mathbf{x}, \hat{\mathbf{w}}_L \rangle$ is decomposed across parent latents P, absorbers C, the residual r and other latents. Parent ablation removes only the parent contribution, leaving absorption-related signal elsewhere. (b) Starting from $\mathbf{x}_{\text{base}}$ after parent ablation, our operator additionally removes the components of selected non-parent latent contributions and the residual along $\hat{\mathbf{w}}_L$, leading to $\mathbf{x}_{\text{edit}}$.

no retraining of the language model or SAE and uses only the fixed concept probe, the parent set, and the SAE decoder.

6 Experiments

We first evaluate the absorption-aware operator against standard SAE editing and existing concept-erasure baselines on the first-letter task (Section 6.1). We then analyze design choices to understand where the improvement comes from (Section 6.2). Next, we evaluate the operator across different SAE training methods, model scales, and SAE families (Section 6.3). Finally, we evaluate whether the operator generalizes beyond orthographic concepts to syntactic and semantic concepts (Section 6.4).

6.1 Editing Performance on the First-Letter Task

We use the same SAEBench first-letter setting as in Section 4 (Chanin et al., 2025; Karvonen et al., 2025), with Gemma-2-2B layer 12 and the Gemma Scope 16k-latent SAE with $\ell_0 = 82$ (Lieberum et al., 2024). The SAEBench absorption evaluation classifies 32,701 words as fully absorbed or not fully absorbed. We retain the ten letters with at least 100 fully absorbed examples on the primary SAE and at least 50 on every other evaluated SAE, ensuring sufficient examples for stable estimates of leak_A and leak_N. For each retained letter, we use up to 300 fully absorbed and 300 not fully absorbed examples, resulting in 5,165 evaluation examples. Additional dataset details are provided in Appendix A.

We test whether the absorption-aware operator improves concept erasure compared with standard parent ablation and existing concept-erasure baselines. We compare against LEACE (Belrose et al., 2023), mean-difference projection (Haghighatkhah et al., 2022), and SAE latent clamping (Farrell et al., 2025). We report leak_A and leak_N, the fractions of fully absorbed and not fully absorbed examples, respectively, for which the target concept remains detectable after editing. We also report $\Delta\text{acc}_{\bar{L}}$, the drop in probe accuracy on non-target letters, and the relative activation change $\|\Delta\mathbf{x}\|_2/\|\mathbf{x}\|_2$.

Table 1 shows that standard parent ablation leaves the target concept detectable on 99.95% of fully absorbed examples and 53.7% of not fully absorbed examples. Our absorption-aware operator reduces these leakage rates to 3.9% and 1.7%, respectively, while changing the accuracy on non-target letters by only 0.07 percentage points. LEACE and mean-difference projection reduce leakage on fully absorbed examples to 36.0% and 24.5%, respectively, whereas Clamping leaves 99.5% of fully absorbed examples detectable despite reducing leak_N to 1.9%. Under full absorption, the parent latents are inactive by definition, so clamping them does not affect concept information carried elsewhere.

Table 1: Editing performance on the primary first-letter setting (Gemma-2-2B, layer 12, Gemma Scope 16k-latent SAE with $\ell_0 = 82$). Lower is better in all columns.

Edit	$\text{leak}_A \downarrow$	$\text{leak}_N \downarrow$	$\Delta\text{acc}_{\bar{L}} \downarrow$	$\|\Delta\mathbf{x}\|_2/\|\mathbf{x}\|_2 \downarrow$
Parent ablation	99.95%	53.7%	0.00 pp	0.041
LEACE	36.0%	28.2%	0.57 pp	0.088
Mean difference	24.5%	21.4%	5.9 pp	0.123
Clamping ($c = -20$)	99.5%	1.9%	0.07 pp	0.122
Ours (dictionary projection)	22.6%	6.0%	0.03 pp	0.069
Ours (residual projection)	83.4%	46.8%	0.07 pp	0.049
Ours (full operator)	**3.9%**	**1.7%**	0.07 pp	0.075

We further examine the contribution of the dictionary and residual corrections in Eq. 3 by applying each one separately. The dictionary projection reduces leak_A to 22.6% and leak_N to 6.0%, and the residual projection alone yields 83.4% and 46.8%, respectively. Combining the two reduces leakage further to 3.9% and 1.7%. Thus, both corrections contribute to the full operator.

6.2 Analysis of Design Choices

We next examine the design choices behind the dictionary correction in Eq. 3. In particular, we ask whether it is necessary to (i) remove only the target-aligned component of each selected latent, (ii) restrict the correction to the candidate set C, and (iii) use a fixed concept-level candidate set rather than the token-specific attributed set $C(\mathbf{x})$. All comparisons use the same first-letter setting as in Section 6.1.

Table 2: Comparison of alternative operator designs on the first-letter setting. Dictionary projection is included as the reference implementation of the proposed selective latent correction. Lower is better in all columns.

Edit	$\text{leak}_A \downarrow$	$\text{leak}_N \downarrow$	$\Delta\text{acc}_{\bar{L}} \downarrow$	$\|\Delta\mathbf{x}\|_2/\|\mathbf{x}\|_2 \downarrow$
Dictionary projection	22.6%	6.0%	0.03 pp	0.069
Absorber zeroing	**22.2%**	**5.3%**	8.4 pp	0.225
All-latent projection	90.8%	69.6%	1.2 pp	0.048
Token-specific $C(\mathbf{x})$	47.1%	22.3%	**0.00 pp**	0.059

Projection versus latent zeroing. Our operator removes only the component of each selected latent contribution along $\hat{\mathbf{w}}_L$. As an alternative, *absorber zeroing* removes the selected latent contributions entirely. Table 2 shows that the two approaches achieve similar leakage on fully absorbed examples (22.6% for dictionary projection and 22.2% for zeroing), but zeroing substantially increases the drop in non-target accuracy (8.4pp versus 0.03pp) and the relative activation change (0.225 versus 0.069). Thus, projecting only the target-aligned component achieves comparable concept removal while better preserving non-target information.

Selecting concept-aligned latents. We next test whether the candidate set C is necessary by applying the projection to the aggregate contribution of all non-parent latents. This *all-latent projection* leaves 90.8% of fully absorbed examples and 69.6% of not fully absorbed examples detectable. On fully absorbed examples, positively aligned latent contributions sum to $+7.75$ along $\hat{\mathbf{w}}_L$, while negatively aligned contributions sum to -5.89, leaving a net contribution of only 1.86. Including both therefore causes substantial cancellation, motivating the restriction to positively aligned candidate latents.

Fixed versus token-specific candidate sets. Finally, we replace the fixed candidate set C with the token-specific attributed set $C(\mathbf{x})$ from the absorption evaluation. This variant leaves 47.1% of fully absorbed examples and 22.3% of not fully absorbed examples detectable, compared with 22.6% and 6.0% for the fixed candidate set. The token-specific attribution set contains only a small number of latents selected for each example, whereas the fixed set C allows any active non-parent latent sufficiently aligned with the target direction to contribute to the correction. These results suggest that

the token-specific attribution set is too restrictive for editing, even though it is useful for diagnosing absorption.

6.3 Generalization Across SAEs and Model Scales

In this section, we explore whether the absorption-aware operator generalizes beyond the primary Gemma Scope SAE used in Sections 6.1 and 6.2. We consider two sources of variation: SAE training method and model scale. All experiments in this section use the first-letter task.

Across SAE training methods. We use different SAEs while keeping the underlying model and layer fixed at Gemma-2-2B layer 12. Specifically, we evaluate SAEs trained with Standard ReLU (Olah, 2023), TopK (Gao et al., 2025), BatchTopK (Bussmann et al., 2024), Gated (Rajamanoharan et al., 2024a), JumpReLU (Rajamanoharan et al., 2024b), P-Anneal, and Matryoshka, together with the 32k Matryoshka releases from Chanin et al. (2025) and the hierarchical H-SAE of Muchane et al. (2025). Most SAEs have 16k latents. The two Matryoshka releases from Chanin et al. (2025) have 32k latents.

Because the prevalence of full absorption varies substantially across SAEs, we summarize overall post-edit leakage by

$$\mathcal{M} = \pi_A \mathrm{leak}_A + (1 - \pi_A)\mathrm{leak}_N,$$

where π_A is the fraction of evaluation examples that are fully absorbed.

Table 3 shows that the operator substantially reduces leakage across all evaluated SAE training methods. Under parent ablation, $\mathcal{M}$ ranges from 0.357 to 0.835, whereas after our edit it ranges from 0.002 to 0.071. Notably, the Matryoshka variants exhibit substantially lower full-absorption rates than most other SAEs, yet parent ablation still leaves substantial overall leakage. Our operator further reduces this remaining leakage, suggesting that SAE designs associated with lower absorption and edit-time correction provide complementary benefits.

Table 3: Generalization across SAE training methods on the first-letter task. All SAEs are for Gemma-2-2B layer 12. We report the full-absorption rate π_A and leakage under standard parent ablation and our operator with $\tau = 0.025$. Lower is better for all leakage metrics.

		parent ablation			ours		
training method	π_A	leak_A	$\mathrm{leak}_N \downarrow$	$\mathcal{M} \downarrow$	$\mathrm{leak}_A \downarrow$	$\mathrm{leak}_N \downarrow$	$\mathcal{M} \downarrow$
Matryoshka BatchTopK	0.040	0.997	0.330	0.357	0.010	0.010	0.010
Matryoshka 32k (Chanin)	0.081	0.999	0.416	0.463	0.008	0.001	0.002
Matryoshka 32k snap (Chanin)	0.041	0.998	0.439	0.461	0.013	0.005	0.006
P-Anneal	0.196	1.000	0.524	0.617	0.055	0.030	0.035
H-SAE	0.307	1.000	0.505	0.657	0.105	0.056	0.071
JumpReLU	0.342	0.999	0.501	0.671	0.040	0.025	0.030
BatchTopK	0.351	1.000	0.533	0.697	0.047	0.025	0.033
TopK	0.402	1.000	0.527	0.717	0.055	0.038	0.045
Gated	0.393	1.000	0.634	0.778	0.041	0.028	0.033
Standard (ReLU)	0.272	1.000	0.774	0.835	0.071	0.051	0.056

Across model scales and SAE families. To evaluate whether the operator transfers across model scales and SAE families, we additionally run experiments using Gemma-3-4B and Gemma-3-27B. For the two models, we use the corresponding Gemma Scope 2 SAEs at layers 17 and 31, respectively, selecting the 16k-latent SAE at medium sparsity in each case. This setting changes both the underlying language model and the SAE family. Note that the Gemma Scope 2 SAEs have a different distribution of decoder–probe cosine similarities from the Gemma Scope SAE. We therefore use $\tau = 0.005$ for both Gemma-3 models. Appendix C shows a threshold analysis.

As shown in Table 4, parent ablation leaves substantial overall leakage at both settings, with $\mathcal{M} = 0.723$ for Gemma-3-4B and $\mathcal{M} = 0.800$ for Gemma-3-27B. Our operator reduces these values to 0.006 and 0.025, respectively. It also achieves lower overall leakage than LEACE at both settings (0.006 vs. 0.041 at 4B and 0.025 vs. 0.060 at 27B). Clamping remains ineffective on fully absorbed

Table 4: Generalization to larger models and a different SAE family on the first-letter task. We evaluate Gemma-3-4B (layer 17) and Gemma-3-27B (layer 31) using the corresponding Gemma Scope 2 16k-latent SAEs at medium sparsity. Results compare standard parent ablation, our operator ($\tau = 0.005$), and the baselines from Table 1. Lower is better in all columns.

	Gemma-3-4B				Gemma-3-27B			
edit	$\text{leak}_A \downarrow$	$\text{leak}_N \downarrow$	$\mathcal{M} \downarrow$	$\Delta\text{acc}_{\bar{L}} \downarrow$	$\text{leak}_A \downarrow$	$\text{leak}_N \downarrow$	$\mathcal{M} \downarrow$	$\Delta\text{acc}_{\bar{L}} \downarrow$
Parent ablation	1.000	0.706	0.723	0.00	0.999	0.783	0.800	0.00
LEACE	0.052	0.040	0.041	0.00	0.103	0.057	0.060	0.00
Mean difference	0.279	0.356	0.352	98.1	0.407	0.500	0.492	91.6
Clamp ($c = -20$)	1.000	0.598	0.622	0.00	0.999	0.740	0.760	0.00
Ours	**0.009**	**0.005**	**0.006**	0.07	**0.055**	**0.023**	**0.025**	0.00

examples, with $\text{leak}_A = 1.000$ and 0.999, while mean-difference projection causes large drops in non-target accuracy (98.1pp and 91.6pp). These results show that the operator remains effective when both the underlying language model and the SAE family change. Additional discussion about Gemma-3-4B and Gemma-3-27B, together with sensitivity analyses for τ, are reported in Appendix C.

6.4 Generalization across Concept Types

Sections 6.1–6.3 focus on orthographic concepts defined by the first letter of a word. In this section, we study whether the absorption-aware operator generalizes to concepts defined by syntax and semantics. We consider part-of-speech tags and two semantic categories, country names and US state names, and compare against the same editing baselines as in Section 6.1. Additional setup details are provided in Appendix A.

Part of speech. We consider syntactic concepts using the development split of Universal Dependencies English-EWT (Silveira et al., 2014). We fit a 14-way token-level logistic probe for universal part-of-speech tags and apply the same absorption evaluation as in the first-letter setting. We use the Gemma-2-2B layer-12 Gemma Scope 16k-latent SAE with $\ell_0 = 82$. We evaluate the nine tags for which both the fully absorbed and not fully absorbed sets contain at least 50 examples, ensuring sufficiently stable estimates of leak_A and leak_N. As shown in Table 5, standard parent ablation leaves substantial overall leakage, with a mean $\mathcal{M}$ of 0.780 across the nine tags. Our operator reduces the mean $\mathcal{M}$ to 0.002, reaching zero on eight of the nine tags and 0.014 on VERB, with a mean drop of only 0.10 percentage points in non-target probe accuracy. For comparison, LEACE, mean-difference projection, and clamping achieve mean $\mathcal{M}$ values of 0.032, 0.742, and 0.529, respectively. Overall, our operator provides the strongest balance between concept erasure and preservation of non-target information.

Table 5: We report overall leakage M and the drop in probe accuracy on non-target tags, $\Delta\text{acc}_{\bar{L}}$. Lower is better in all columns. Our operator achieves near-zero leakage with minimal loss of non-target accuracy.

	$\mathcal{M} \downarrow$					$\Delta\text{acc}_{\bar{L}}$ (pp) $\downarrow$			
Tag	Parent	**Ours**	LEACE	Mean diff	Clamp	**Ours**	LEACE	Mean diff	Clamp
DET	0.222	**0.000**	0.009	0.828	0.165	0.11	0.40	16.0	**0.00**
AUX	0.527	**0.000**	0.008	0.900	0.217	0.14	0.22	34.7	**0.01**
PUNCT	0.762	**0.000**	0.006	0.926	0.617	0.05	0.27	13.5	**0.00**
ADP	0.886	**0.000**	0.049	0.998	0.776	**0.12**	0.33	56.7	0.18
PRON	0.828	**0.000**	0.028	0.293	0.391	0.08	0.20	4.0	**0.00**
NUM	0.941	**0.000**	**0.000**	0.920	0.711	0.07	0.09	65.5	**0.00**
ADJ	0.932	**0.000**	0.052	0.907	0.629	0.08	0.29	34.3	**0.04**
VERB	0.942	**0.014**	0.104	0.864	0.400	0.07	0.24	30.3	**0.06**
PROPN	0.983	**0.000**	0.033	0.040	0.851	0.16	0.24	5.4	**0.00**
mean	0.780	**0.002**	0.032	0.742	0.529	0.10	0.25	28.9	**0.03**

Semantic concepts. We consider two semantic categories, country names and US-state names, using occurrences extracted from natural text corpora. We evaluate them on the Gemma Scope 65k-latent SAE with $\ell_0 = 21$, the only evaluated setting in which both concepts have at least 50 fully absorbed examples.

As shown in Table 6, standard parent ablation leaves substantial overall leakage, with $\mathcal{M} = 0.399$ for country names and $\mathcal{M} = 0.337$ for US state names. Our operator eliminates this leakage for both concepts ($\mathcal{M} = 0$), while changing non-target probe accuracy by at most 0.06 percentage points. LEACE leaves $\mathcal{M} = 0.476$ and 0.005, mean-difference projection leaves $\mathcal{M} = 0.001$ and 0.047, and clamping leaves all fully absorbed examples detectable for both concepts ($\mathrm{leak}_A = 1$). Additional analyses show that our semantic edits incur little model-level utility loss and are substantially more resistant to residual-space recovery than LEACE and clamping (Appendix B). Together with the part-of-speech results, these experiments show that our operator generalizes beyond first-letter concepts to both syntactic and semantic concepts.

Table 6: Editing performance on the two semantic concepts using the Gemma Scope 65k-latent SAE with $\ell_0 = 21$. We report leakage on fully absorbed and not fully absorbed examples, overall leakage $\mathcal{M}$, the change in non-target probe accuracy $\Delta\mathrm{acc}_{\bar{L}}$, and relative activation change. Lower is better in all columns. Our operator eliminates leakage for both concepts with minimal change in non-target accuracy.

concept	edit	$\mathrm{leak}_A \downarrow$	$\mathrm{leak}_N \downarrow$	$\mathcal{M} \downarrow$	$\Delta\mathrm{acc}_{\bar{L}} \downarrow$	$\lVert\Delta x\rVert/\lVert x\rVert \downarrow$
country	Parent ablation	1.000	0.367	0.399	-0.04	**0.15**
	Ours	0.000	**0.000**	**0.000**	$\mathbf{-0.06}$	0.20
	LEACE	0.229	0.489	0.476	$+0.28$	0.17
	Mean difference	0.000	0.001	0.001	$+0.08$	0.46
	Clamp ($c = -20$)	1.000	0.029	0.079	$+0.41$	0.26
US state	Parent ablation	1.000	0.271	0.337	0.00	**0.17**
	Ours	0.000	**0.000**	**0.000**	$+0.02$	0.24
	LEACE	0.000	0.005	0.005	$\mathbf{-0.01}$	0.20
	Mean difference	0.032	0.048	0.047	$+4.95$	0.28
	Clamp ($c = -20$)	1.000	0.012	0.102	$+0.06$	0.28

7 Conclusion

We show in this paper that feature absorption provides a systematic explanation for the failure of standard SAE-based editing. When concept information is carried outside the parent latents selected for editing, ablating those latents leaves predictable leakage. Motivated by this structure, we introduce an absorption-aware operator that removes target-aligned components from candidate non-parent latents and the reconstruction residual. Across diverse SAE training methods, three model scales, and orthographic, syntactic, and semantic concepts, our operator substantially improves the completeness of concept erasure while largely preserving non-target information and model utility. In future work, the operator could be explored as a lightweight post-hoc intervention for settings in which a target concept must be removed without retraining, such as targeted unlearning or suppression of unwanted behaviors.

AI use statement

In this work, we used generative AI tools to edit the manuscript for readability and to polish the writing. We have not used generative AI tools to generate synthetic data, develop the theoretical framework, formulate or prove mathematical claims, propose or refine hypotheses, design or evaluate the methodology or experiments, implement methods, interpret results, or clean or reformat datasets. Translation and qualitative data analysis are not applicable to this work. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Dana Arad, Aaron Mueller, and Yonatan Belinkov. SAEs are good for steering – if you select the right features. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 10241–10259, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.519. URL https://aclanthology.org/2025.emnlp-main.519/.

Andy Arditi, Oscar Balcells Obeso, Aaquib Syed, Daniel Paleka, Nina Rimsky, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=pH3XAQME6c.

Tomer Ashuach, Dana Arad, Aaron Mueller, Martin Tutek, and Yonatan Belinkov. CRISP: Persistent concept unlearning via sparse autoencoders. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1806–1825, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.82. URL https://aclanthology.org/2026.acl-long.82/.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=awIpKpwTwF.

Bart Bussmann, Patrick Leask, and Neel Nanda. Batchtopk sparse autoencoders. In *NeurIPS 2024 Workshop on Scientific Methods for Understanding Deep Learning*, 2024. URL https://openreview.net/forum?id=d4dpOCqybL.

Bart Bussmann, Noa Nabeshima, Adam Karvonen, and Neel Nanda. Learning multi-level features with matryoshka sparse autoencoders. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=m25T5rAy43.

Enrico Cassano, Riccardo Renzulli, Marco Nurisso, Mirko Zaffaroni, Alan Perotti, and Marco Grangetto. Saemnesia: Erasing concepts in diffusion models with supervised sparse autoencoders, 2026. URL https://arxiv.org/abs/2509.21379.

Sviatoslav Chalnev, Matthew Siu, and Arthur Conmy. Improving steering vectors by targeting sparse autoencoder features, 2024. URL https://arxiv.org/abs/2411.02193.

David Chanin, James Wilken-Smith, Tomáš Dulka, Hardik Bhatnagar, Satvik Golechha, and Joseph Isaac Bloom. A is for absorption: Studying feature splitting and absorption in sparse autoencoders. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=R73ybUciQF.

Mingyue Cui, Linghui Shen, and Xingyi Yang. Sae interventions are unreliable: Post-intervention recovery of suppressed behavior, 2026. URL https://arxiv.org/abs/2606.18322.

Bartosz Cywiński and Kamil Deja. SAeuron: Interpretable concept unlearning in diffusion models with sparse autoencoders. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=6N0GxaKdX9.

Evan Duan. Pre-intervention prediction of sparse autoencoder steering side effects, 2026. URL https://arxiv.org/abs/2606.08365.

Eoin Farrell, Yeu-Tong Lau, and Arthur Conmy. Applying sparse autoencoders to unlearn knowledge in language models, 2025. URL https://openreview.net/forum?id=ZtvRqm6oBu.

Javier Ferrando, Oscar Balcells Obeso, Senthooran Rajamanoharan, and Neel Nanda. Do i know this entity? knowledge awareness and hallucinations in language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=WCRQFlji2q.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The pile: An 800gb dataset of diverse text for language modeling, 2020. URL https://arxiv.org/abs/2101.00027.

Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=tcsZt9ZNKD.

Nils Grandien, David Steinmann, Felix Friedrich, and Kristian Kersting. Do sparse autoencoders learn meaningful concept hierarchies?, 2026. URL https://arxiv.org/abs/2606.22994.

Phillip Huang Guo, Aaquib Syed, Abhay Sheshadri, Aidan Ewart, and Gintare Karolina Dziugaite. Mechanistic unlearning: Robust knowledge unlearning and editing via mechanistic localization. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=92oBV5HAGl.

Yoav Gur-Arieh, Clara Haya Suslik, Yihuai Hong, Fazl Barez, and Mor Geva. Precise in-parameter concept erasure in large language models. *ArXiv*, abs/2505.22586, 2025. URL https://api.semanticscholar.org/CorpusID:278959968.

Pantea Haghighatkhah, Antske Fokkens, Pia Sommerauer, Bettina Speckmann, and Kevin Verbeek. Better hit the nail on the head than beat around the bush: Removing protected attributes with a single projection. In *Conference on Empirical Methods in Natural Language Processing*, 2022. URL https://api.semanticscholar.org/CorpusID:254408718.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=F76bwRSLeK.

Adam Karvonen, Can Rager, Samuel Marks, and Neel Nanda. Evaluating sparse autoencoders on targeted concept erasure tasks, 2024. URL https://arxiv.org/abs/2411.18895.

Adam Karvonen, Can Rager, Johnny Lin, Curt Tigges, Joseph Isaac Bloom, David Chanin, Yeu-Tong Lau, Eoin Farrell, Callum Stuart McDougall, Kola Ayonrinde, Demian Till, Matthew Wearden, Arthur Conmy, Samuel Marks, and Neel Nanda. SAEBench: A comprehensive benchmark for sparse autoencoders in language model interpretability. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=qrU3yNfX0d.

Matthew Khoriaty, Andrii Shportko, Gallay Ambroise Mercier, and Zach Wood-Doughty. Don't forget it! conditional sparse autoencoder clamping works for unlearning. *ArXiv*, abs/2503.11127, 2025. URL https://api.semanticscholar.org/CorpusID:277043381.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2, 2024. URL https://arxiv.org/abs/2408.05147.

Tomasz Limisiewicz, David Mareček, and Tomáš Musil. Debiasing algorithm through model adaptation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=XIZEFyVGC9.

Samuel Marks, Can Rager, Eric J Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=I4e82CIDxv.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Byj72udxe.

Mark Muchane, Sean Richardson, Kiho Park, and Victor Veitch. Incorporating hierarchical semantics in sparse autoencoder architectures. *arXiv preprint arXiv:2506.01197*, 2025.

Aashiq Muhamed, Jacopo Bonato, Mona T. Diab, and Virginia Smith. SAEs can improve unlearning: Dynamic sparse autoencoder guardrails for precision unlearning in LLMs. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=kaPAalWAp3.

Trenton Bricken Adly Templeton Joshua Batson Brian Chen Adam Jermyn Tom Conerly Nicholas L Turner Cem Anil Carson Denison Amanda Askell Robert Lasenby Yifan Wu Shauna Kravec Nicholas Schiefer Tim Maxwell Nicholas Joseph Alex Tamkin Karina Nguyen Brayden McLean Josiah E Burke Tristan Hume Shan Carter Tom Henighan Chris Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Anthropic*, 2023. URL https://www.anthropic.com/research/towards-monosemanticity-decomposing-language-models-with-dictionary-learning.

Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Tom Lieberum, Vikrant Varma, János Kramár, Rohin Shah, and Neel Nanda. Improving dictionary learning with gated sparse autoencoders, 2024a. URL https://arxiv.org/abs/2404.16014.

Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. Jumping ahead: Improving reconstruction fidelity with jumprelu sparse autoencoders, 2024b. URL https://arxiv.org/abs/2407.14435.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7237–7256, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.647. URL https://aclanthology.org/2020.acl-main.647/.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel R. Bowman, Miriam Connor, John Bauer, and Chris Manning. A gold standard dependency corpus for English. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 2897–2904, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). URL https://aclanthology.org/L14-1067/.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, Alex Tamkin, Esin Durmus, Tristan Hume, Francesco Mosconi, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet, 2026. URL https://arxiv.org/abs/2605.29358.

Xu Wang, Zihao Li, Benyou Wang, Yan Hu, and Difan Zou. Model unlearning via sparse autoencoder subspace guided projections. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 26530–26546, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1348. URL https://aclanthology.org/2025.emnlp-main.1348/.

Yu Zhao, Alessio Devoto, Giwon Hong, Xiaotang Du, Aryo Pradipta Gema, Hongru Wang, Xuanli He, Kam-Fai Wong, and Pasquale Minervini. Steering knowledge selection behaviours in LLMs via SAE-based representation engineering. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5117–5136, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN

979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.264. URL https://aclanthology.org/2025.naacl-long.264/.

A EXPERIMENT DETAILS

This appendix provides additional details of the experimental setup, including the absorption evaluation, dataset construction, evaluation metrics, and the cross-SAE prediction analysis.

A.1 ABSORPTION EVALUATION

Every quantity used in analysis and editing operator comes from the SAEBench absorption evaluation (Chanin et al., 2025; Karvonen et al., 2025). For the first-letter task, we use the evaluation outputs provided by SAEBench directly. For the part-of-speech and semantic tasks, we implement the same evaluation procedure on the new datasets and verify that our implementation reproduces the SAEBench outputs on the primary first-letter setting.

The Probe $\hat{\mathbf{w}}_L$ We train a multiclass logistic probe on the model activations $\mathbf{x}$ to predict the concept label. For a target concept L, we take the corresponding probe direction, normalize it to unit length, and denote it by $\hat{\mathbf{w}}_L$.

The Parent Set P We identify a small set of SAE latents that best predict the target concept and use them as the parent set P. We first fit an ℓ_1-regularized logistic probe on the SAE latent activations using 70% of the labeled examples, with a stratified split. The latents are ranked by their positive probe coefficients. We then consider the top-k latents for $k = \{1, \ldots, 10\}$. For each k, we fit an unregularized logistic probe using only these top-k latents and evaluate its F1 score on the held-out 30% of the data. We increase k only when the F1 score improves by at least 0.03. Otherwise, we stop and take the last accepted set as P. The resulting parent sets contain one to three latents for first-letter concepts, two to six for part-of-speech concepts, and one for each semantic concept.

Fully absorbed and not fully absorbed examples. For each SAE latent j, we measure its contribution along the target concept direction as $a_j \langle \mathbf{d}_j, \hat{\mathbf{w}}_L \rangle$. Let $j^\star$ denote the latent with the largest such contribution. Following the SAEBench absorption evaluation, a token is labeled fully absorbed when all of the following hold: (i) all parent latents are inactive, i.e., $a_p < 10^{-8}$ for every $p \in P$; (ii) the latent that contributes the most is aligned with the target concept direction, i.e, $\cos(\mathbf{d}_{j^\star}, \hat{\mathbf{w}}_L) \geq \tau$; (iii) the target concept is detectable in the original model activation, i.e., $\langle \mathbf{x}, \hat{\mathbf{w}}_L \rangle \geq 0$; (iv) the top-contributing latent carries a large enough share of the projection, i.e., $a_{j^\star} \cos_{j^\star} / \langle \mathbf{x}, \hat{\mathbf{w}}_L \rangle \geq 0.4$. We denote the fully absorbed examples by A and the remaining target examples by N.

The absorption fraction $\phi(\mathbf{x})$. The absorption fraction in Eq. 1 uses the token-specific attribution set $C(\mathbf{x})$ defined in Section 3. Before computing the ratio, we apply the same validity checks as in the absorption evaluation. If the parent contribution already accounts for the full target-direction projection, or if the attributed non-parent contribution is too small relative to the total projection, we set $\phi(\mathbf{x}) = 0$. Otherwise, the attributed non-parent contribution is capped by the portion of the target-direction projection not explained by the parents before computing Eq. 1.

A.2 DATASETS

First letter We use the words from the SAEBench spelling task (Chanin et al., 2025; Karvonen et al., 2025). Each word is presented in the template "{word} has the first letter:" and we extract the model activation $\mathbf{x}$ at the word's token. We train a 26-way logistic probe on its training split to predict the first-letter class. For the primary SAE, the absorption evaluation assigns labels to 32,701 words. We include a letter if it has at least 100 fully absorbed examples under the primary SAE and at least 50 under every other evaluated SAE. Ten letters satisfy this criterion, leading to a pool of 18,020 labeled words in total. For each included letter, sample at most 300 fully absorbed examples for A and at most 300 not fully absorbed examples for N. For collateral evaluation, we use not fully absorbed words from the other letter classes, sampling 300 examples per target letter for the primary SAE and 200 for each other SAE. Each example is evaluated with the probe direction corresponding to its own letter class.

Part of Speech We use the Universal Dependencies English-EWT treebank (Silveira et al., 2014), including the first 10,000 sentences from the training split (156,758 labeled tokens) and the whole

development split (24,174 labeled tokens). Sentences are reconstructed from the CoNLL-U forms, and contraction subwords are excluded. In the fourteen universal POS tags that serve as labels, SYM, INTJ, and X, each of which accounts for less than 0.4% of tokens, are excluded. We train a 14-way logistic probe on the training tokens, and use the probe direction corresponding to each target tag as $\hat{\mathbf{w}}_L$. We further exclude ADV and SCONJ because their per-tag probe accuracy is below 0.9. For evaluation, we include a POS tag only if both the fully absorbed set A and the not fully absorbed set N contain at least 50 examples. Nine of the remaining twelve tags satisfy this criterion (Table 7), leading to 16,470 positive tokens in total. For collateral evaluation, we use development-set tokens whose POS tag is different from the edited target tag. We measure the drop in accuracy, $\Delta\mathrm{acc}_{\bar{L}}$, on these non-target tokens before and after editing.

Table 7: Part-of-speech dataset composition on the UD-EWT development split, using 16k-latent SAE with $\ell_0 = 82$. We report the parent-set size $|P|$ and the numbers of fully absorbed (A) and not fully absorbed (N) tokens for each POS tag.

	DET	AUX	PUNCT	ADP	PRON	NUM	ADJ	VERB	PROPN
$\lvert P\rvert$	2	4	5	2	6	3	4	2	5
$\lvert A\rvert$	290	240	1646	1364	577	131	406	174	64
$\lvert N\rvert$	1533	1020	1422	584	1454	243	1320	2418	1584

Semantic concepts We consider two target concepts: country names and US-state names. The country list contains 115 names, and the US-state list contains 36. We construct five additional classes consisting of city names, first names, month names, weekday names, and other capitalized words, resulting in seven classes in total. We collect occurrences from Pile-10k (Gao et al., 2020), excluding the GitHub, DM Mathematics, and arXiv subsets, and from the first 10% of the WikiText-103 training split (Merity et al., 2017). We retain sentences between 20 and 600 characters. Listed names are matched as whole words, and ambiguous names and mentions preceded by a compound prefix are excluded.

The resulting 64,694 occurrences are split by document into 51,758 training and 12,936 test examples. We cap each country or US-state name at 400 occurrences and each contrast class at 8,000 occurrences within each split. For each occurrence, we extract the layer-12 model activation at the final subtoken of the mention and encode it with each evaluated SAE. We train a 7-way logistic probe on the training examples. For each target concept L, we use the corresponding probe direction as $\hat{\mathbf{w}}_L$. We evaluate editing using the Gemma Scope 65k-latent SAE with $\ell_0 = 21$ (Table 8). For collateral evaluation, we use test examples from the other six classes.

Table 8: Semantic-concept dataset composition for the 65k-latent SAE with $\ell_0 = 21$. We report the number of names, parent-set size $|P|$, and the numbers of fully absorbed (A) and not fully absorbed (N) evaluation examples for each target concept.

concept	members	$\lvert P\rvert$	$\lvert A\rvert$	$\lvert N\rvert$
Country names	115	1	227	4186
US-state names	36	1	222	2218

A.3 METRICS

Detection and margin. For an example with activation $\mathbf{x}$ and target concept L, let $z_{L'}(\mathbf{x})$ denote the probe logit for class L'. The concept is *detectable* on $\mathbf{x}$ when the probe predicts L, that is, $z_L(\mathbf{x}) > \max_{L' \neq L} z_{L'}(\mathbf{x})$. We define the probe margin as $m(\mathbf{x}) = z_L(\mathbf{x}) - \max_{L' \neq L} z_{L'}(\mathbf{x})$. So the concept is detectable when $m(\mathbf{x}) > 0$. The post-edit margin in Figure 1a is $m(\mathbf{x}_{\mathrm{base}})$ after parent ablation, in logits.

Leakage. For a concept L, let A and N denote its fully absorbed and not fully absorbed evaluation examples, respectively (Appendix A.2). Only examples on which the concept is detectable before the edit are counted for the leakage rates. leak_A and leak_N are the fractions of examples in A and N, respectively, for which the concept remains detectable after editing: $m(\mathbf{x}_{\mathrm{edit}}) > 0$. Under standard

parent ablation $\mathrm{leak}_A \approx 1$ by construction because the parent latents are inactive on a fully absorbed example, so the edit leaves its activation unchanged.

Absorption statistics. π_A is the fraction of all labeled examples of a concept that are classified as fully absorbed by the absorption evaluation (see Appendix A.1). $\bar{\phi}_N$ is the mean of the absorption fraction $\phi(\mathbf{x})$ among its not fully absorbed examples. Both quantities are computed before editing using all labeled examples of the concept.

Overall leakage. Because the prevalence of full absorption differs across concepts and SAEs, we report the prevalence-weighted leakage: $\mathcal{M} = \pi_A \,\mathrm{leak}_A + (1 - \pi_A)\,\mathrm{leak}_N$.

Collateral effect. The collateral set for a target concept L consists of examples of the other concepts (Appendix A.2). We apply the edit for L to these examples and evaluate the multiclass probe against their original concept labels. We report $\Delta\mathrm{acc}_L$, the probe accuracy before editing minus the accuracy after editing, in percentage points.

Activation change. We report the relative activation change $\|\Delta\mathbf{x}\|_2/\|\mathbf{x}\|_2$ where $\Delta\mathbf{x} = \mathbf{x}_{\mathrm{edit}} - \mathbf{x}$ averaged over the evaluation examples of the edited concept. It measures the magnitude of the intervention independently of the probe.

A.4 Cross SAE Prediction of editing failure

We provide additional details for the cross-SAE prediction experiment in Section 4. The goal is to test whether quantities available before editing can predict the failure of standard parent ablation on held-out SAEs. The experimental design follows the pre-intervention prediction setting of Duan (2026).

SAEs and prediction samples. We use the first-letter dataset and Gemma-2-2B layer-12 Gemma Scope SAEs (Lieberum et al., 2024). We use ten SAE releases with 16k latents ($\ell_0 \in \{22, 41, 82, 176, 445\}$) and 65k latents ($\ell_0 \in \{21, 38, 72, 141, 297\}$). We exclude the 16k SAE with $\ell_0 = 445$ and the 65k SAE with $\ell_0 = 297$ because only two letter concepts satisfy the evaluation criteria for these SAEs. The remaining eight SAEs yield 132 (SAE, L) pairs in total which are used for Figure 1b. For each pair, we use up to 300 fully absorbed and 300 not fully absorbed examples and apply standard parent ablation. We compute leak_A and leak_N, as in Section 6.1, restricting evaluation to examples correctly classified by the concept probe before editing.

Features. All features are computed before editing. The baseline features are the parent count $|P|$, the largest cosine between $\mathbf{d}_p$ and any other decoder direction, averaged over $p \in P$, the mean decoder norm and the mean encoder norm of the parents, the mean cosine between parent and probe $\cos(\mathbf{d}_p, \hat{\mathbf{w}}_L)$ over P, the coverage of candidate set $|C|/\mathrm{width}$, and the mean $\cos(\mathbf{d}_c, \hat{\mathbf{w}}_L)$ over C, the ℓ_0 and width of SAEs. We additionally consider two absorption features: π_A, the fraction of labeled examples that are fully absorbed, and $\bar{\phi}_N$, the mean absorption fraction among not fully absorbed examples. Both are computed over all labeled examples for the target concept.

Prediction and evaluation. We fit ridge regression models with regularizer set to 1. Features are standardized to zero mean and unit variance.

We evaluate prediction using leave-one-SAE-out cross-validation. In each fold, all concept pairs from one SAE are held out, and the predictor is trained on pairs from the remaining seven SAEs. We pool the predictions from all eight held-out SAEs and report the Spearman correlation between predicted and observed editing failure.

We compare three feature sets: the nine baseline features, the baseline features augmented with $(\pi_A, \bar{\phi}_N)$, and $(\pi_A, \bar{\phi}_N)$ alone. Table 9 reports results for leak_N, as discussed in Section 4, together with the corresponding results for $\mathcal{M}$.

Table 9: Leave-one-out Spearman correlation between predicted and observed editing failure under parent ablation.

Features	leak_N	$\mathcal{M}$
baseline features	0.72	0.71
baseline features + $(\pi_A, \bar{\phi}_N)$	**0.93**	**0.95**
$(\pi_A, \bar{\phi}_N)$ only	0.85	0.94

B Model-Level Utility and Robustness to Concept Recovery

We provide two additional evaluations of the absorption-aware edit. First, we examine whether applying the edit throughout the model degrades general language-model utility. Second, motivated by prior work showing that suppressed concept information can be recovered through the SAE reconstruction residual, we test how easily the edited concept can be recovered by a perturbation.

Unless otherwise stated, the first-letter and the position of speech experiments use Gemma-2-2B with the 16k-latent SAE with $\ell_0 = 82$. We report the corresponding analyses for the semantic concepts using the 65k-latent SAE with $\ell_0 = 21$.

Model-level utility. We evaluate whether the edits affect general model behavior when applied at every non-BOS position at layer 12. We measure next-token prediction on WikiText-103 and the Pile using 50,000 tokens from each dataset, and evaluate downstream performance on a 570-question subset of MMLU (Hendrycks et al., 2021). The results in Table. 10 Table. 12 and Table. 11 shows that our method retains better or comparable model utility in terms of KL divergence and a comparable performance in terms of MMLU accuracy and cross entropy compared to original models when compared to outside baselines, showing that our method does not lead to a degradation in the utility of language models.

Table 10: Language-model utility on the first letter task. We report the change in next token cross entropy (ΔCE) and the KL divergence from the unedited model, both in nats/token, on WikiText-103 and the Pile, and the change in accuracy on our 570 question MMLU subset. Lower KL is better. ΔCE and ΔMMLU should stay close to zero.

	WikiText-103		Pile		
method	$\Delta\mathrm{CE} \to 0$	$\mathrm{KL} \downarrow$	$\Delta\mathrm{CE} \to 0$	$\mathrm{KL} \downarrow$	$\Delta\mathrm{MMLU} \to 0$
parent ablation	$+0.0006$	0.0011	$+0.0005$	0.0008	-0.1pp
clamp ($c = -20$)	$+0.0034$	0.0040	$+0.0019$	0.0023	-1.1pp
LEACE	$+0.0033$	0.0037	$+0.0023$	0.0029	$+0.5$pp
mean-difference	$+0.0023$	0.0041	$+0.0024$	0.0034	$+0.5$pp
ours	$+0.0012$	**0.0018**	$+0.0005$	**0.0014**	-0.2pp

Table 11: Language-model utility on the POS task. We report the change in next token cross entropy (ΔCE) and the KL divergence from the unedited model, both in nats/token, on WikiText-103 and the Pile, and the change in accuracy on our 570-question MMLU subset (one question is 0.18pp). Lower KL is better; ΔCE and ΔMMLU should stay close to zero.

	WikiText-103		Pile		
method	$\Delta\mathrm{CE} \to 0$	$\mathrm{KL} \downarrow$	$\Delta\mathrm{CE} \to 0$	$\mathrm{KL} \downarrow$	$\Delta\mathrm{MMLU} \to 0$
parent ablation	$+0.018$	0.019	$+0.025$	0.026	$+0.2$pp
clamp ($c = -20$)	$+0.061$	0.063	$+0.072$	0.072	-0.1pp
LEACE	$+0.022$	0.023	$+0.016$	0.018	$+0.7$pp
mean-difference	$+0.083$	0.091	$+0.090$	0.104	$+0.4$pp
ours	$+0.026$	0.028	$+0.034$	0.034	$+0.4$pp

Residual-space recovery. Following Cui et al. (2026), we test whether the target concept can be recovered after editing by perturbing the edited activation. For each example, we optimize a

Table 12: Language-model utility on the closed set concept task. We report the change in next token cross entropy (ΔCE) and the KL divergence from the unedited model, both in nats/token, on WikiText-103 and the Pile, and the change in accuracy on our 570-question MMLU subset (one question is 0.18pp). Lower KL is better; ΔCE and ΔMMLU should stay close to zero.

		WikiText-103		Pile		
concept	method	ΔCE $\to 0$	KL $\downarrow$	ΔCE $\to 0$	KL $\downarrow$	ΔMMLU $\to 0$
country	parent ablation	-0.0002	0.0005	$+0.0002$	0.0000	0.0pp
	clamp ($c = -20$)	$+0.0014$	0.0015	$+0.0001$	0.0001	0.0pp
	LEACE	$+0.0265$	0.0268	$+0.0169$	0.0184	$+3.3$pp
	mean-difference	$+0.0765$	0.0752	$+0.0360$	0.0404	$+0.4$pp
	ours	$+0.0014$	**0.0019**	-0.0002	**0.0010**	0.0pp
US state	parent ablation	-0.0001	0.0001	$+0.0000$	0.0000	0.0pp
	clamp ($c = -20$)	$+0.0002$	0.0004	$+0.0000$	0.0000	0.0pp
	LEACE	$+0.0076$	0.0067	$+0.0051$	0.0054	$+0.2$pp
	mean-difference	$+0.0087$	0.0110	$+0.0087$	0.0104	$+1.4$pp
	ours	$+0.0000$	**0.0015**	-0.0004	**0.0008**	0.0pp

perturbation δ to increase the target probe margin subject to $\|\delta\|_2 \leq \varepsilon \|\mathbf{x}\|$. We additionally penalize changes in the SAE encoding with λ=10 on $\|\mathrm{enc}(\mathbf{x}+\delta) - \mathrm{enc}(\mathbf{x})\|_1$, so that the perturbation has to travel through the reconstruction residual, the channel prior work identifies as the route of recovery and the one the operator projects out.

Table 13, Table 14 and Table 15 report recovery rates on our three datasets. For smaller perturbation budgets, our method consistently outperforms all baselines on all datasets, showing that our method is consistently robust under small perturbations.

For larger perturbation budgets, our method is still robust, only being worse than mean-difference whose performance is not competitive in main experiment(Table 1) and produces larger model-level utility changes as reported above, also showing that our model is robust under large perturbations.

Table 13: Concept recovery by residual-space perturbation for first letter. *leak*: share of these tokens that is still detected after the edit and before any attack. ε columns: share on which a perturbation of budget ε (relative to $\|\mathbf{x}\|$) recovers it. $\varepsilon_{50\%}$: the smallest budget at which the mean recovery reaches one half. Lower leak and recovery and a larger $\varepsilon_{50\%}$ are better.

		recovery at ε					
method	leak $\downarrow$	ε=0.01 $\downarrow$	0.02 $\downarrow$	0.05 $\downarrow$	0.10 $\downarrow$	0.20 $\downarrow$	$\varepsilon_{50\%}$ $\uparrow$
parent ablation	0.97	0.99	1.00	1.00	1.00	1.00	0.01
clamp ($c = -20$)	0.96	0.99	0.99	1.00	1.00	1.00	0.01
LEACE	0.02	0.19	0.53	0.99	1.00	1.00	0.02
mean-difference	0.10	0.19	0.31	0.67	0.95	1.00	0.05
ours	0.02	**0.17**	**0.48**	0.99	1.00	1.00	0.05

Table 14: Concept recovery by residual-space perturbation for POS. *leak*: share of these tokens on which the tag is still detected after the edit and before any attack. ε columns: share on which a perturbation of budget ε (relative to $\|\mathbf{x}\|$) recovers it. $\varepsilon_{50\%}$: the smallest budget at which the mean recovery reaches one half. Lower leak and recovery and a larger $\varepsilon_{50\%}$ are better.

		recovery at ε					
method	leak $\downarrow$	ε=0.01 $\downarrow$	0.02 $\downarrow$	0.05 $\downarrow$	0.10 $\downarrow$	0.20 $\downarrow$	$\varepsilon_{50\%}$ $\uparrow$
parent ablation	1.00	1.00	1.00	1.00	1.00	1.00	0.01
clamp ($c = -20$)	1.00	1.00	1.00	1.00	1.00	1.00	0.01
LEACE	0.03	0.19	0.46	0.93	1.00	1.00	0.05
mean-difference	0.71	0.79	0.86	0.91	0.92	0.94	0.01
ours	0.00	**0.04**	**0.26**	0.84	1.00	1.00	0.05

Table 15: Concept recovery by residual-space perturbation for closed set semantic concepts. *leak*: share of these tokens that is still detected after the edit and before any attack. ε columns: share on which a perturbation of budget ε (relative to $\|\mathbf{x}\|$) recovers it. $\varepsilon_{50\%}$: the smallest budget at which the mean recovery reaches one half. Lower leak and recovery and a larger $\varepsilon_{50\%}$ are better.

concept	method	leak ↓	recovery at ε ε=0.01 ↓	0.02 ↓	0.05 ↓	0.10 ↓	0.20 ↓	$\varepsilon_{50\%}$ ↑
country	parent ablation	1.00	1.00	1.00	1.00	1.00	1.00	0.01
	clamp ($c = -20$)	1.00	1.00	1.00	1.00	1.00	1.00	0.01
	LEACE	0.20	0.53	0.81	1.00	1.00	1.00	0.01
	mean-difference	0.00	0.00	0.00	0.02	0.17	0.87	0.20
	ours	0.00	0.00	0.00	0.05	0.99	1.00	0.10
US state	parent ablation	1.00	1.00	1.00	1.00	1.00	1.00	0.01
	clamp ($c = -20$)	1.00	1.00	1.00	1.00	1.00	1.00	0.01
	LEACE	0.00	0.02	0.15	0.79	1.00	1.00	0.05
	mean-difference	0.04	0.05	0.05	0.11	0.65	1.00	0.10
	ours	0.00	0.00	0.00	0.01	0.94	1.00	0.10

C SENSITIVITY TO τ

Our proposed operator uses a threshold τ to define the candidate set $C = \{j \notin P : \cos(\mathbf{d}_j, \hat{\mathbf{w}}_L) > \tau\}$. We use $\tau = 0.025$ on Gemma Scope and $\tau = 0.005$ on Gemma Scope 2 in the main experiments. Here, we examine how the choice of τ affects editing performance across the evaluated SAEs.

Figure 3 shows the value of $\mathcal{M}$ as τ varies for the 16k-latent SAE in each model setting. For Gemma Scope on Gemma-2-2B, $\mathcal{M}$ is already low at $\tau = 0.025$ and changes little as the threshold is reduced. In contrast, the same threshold leaves substantially higher leakage for the Gemma Scope 2 SAEs on Gemma-3-4B and Gemma-3-27B. Lowering the threshold admits more candidate latents and reduces $\mathcal{M}$, reaching 0.006 and 0.025, respectively, at $\tau = 0.005$. Non-target accuracy remains stable across the evaluated thresholds.

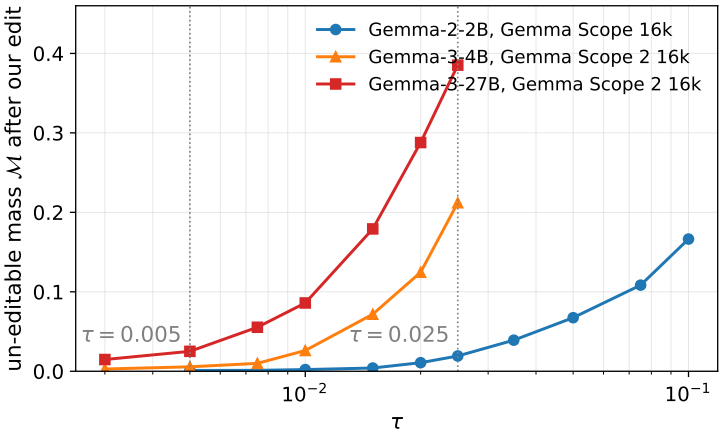


Figure 3: The change in $\mathcal{M}$ with respect to τ for the 16k-latent SAE across different model settings.

Table 16 reports the fraction of SAE latents included in C, averaged over all 26 first-letter concepts. At τ=0.025, the Gemma-2-2B SAE includes 11.6% of its latents on average, whereas the Gemma Scope 2 SAE includes fewer, e.g., 5.5–8.8% at 4B and 2.0–2.8% at 27B. Lowering τ to 0.005 increases this fraction to approximately 33–39% across the evaluated settings.

The final column of Table 16 reports the value of τ required to match the 11.6% candidate-set fraction of the primary 2B SAE. This threshold decreases from 0.026 at 2B to 0.019–0.023 at 4B and

0.014–0.016 at 27B. Thus, the same value of τ can correspond to substantially different candidate-set sizes across SAEs.

Table 16: Fraction of SAE latents included in C, averaged over the 26 first-letter concepts. The final column gives the value of τ that matches the 11.6% candidate-set fraction of the primary Gemma-2-2B SAE.

family	SAE	d	coverage @0.025	coverage @0.005	τ for 11%
Gemma Scope (2B, L12)	16k medium	2304	11.6%	38.8%	0.026
Gemma Scope 2 (4B, L17)	16k medium	2560	5.5%	36.9%	0.019
	65k medium	2560	7.5%	38.4%	0.021
	262k medium	2560	8.8%	38.9%	0.023
Gemma Scope 2 (27B, L31)	16k medium	5376	2.0%	32.7%	0.014
	65k medium	5376	2.4%	33.9%	0.015
	262k medium	5376	2.8%	34.8%	0.016