NEURALYZING THE TRACE: SELECTIVE REPRESENTATION-LEVEL UNLEARNING WITH CONTRASTIVE SPARSE AUTOENCODERS

ABSTRACT

Machine unlearning aims to remove targeted information while preserving a model's other abilities. In realistic settings, such as privacy requests under the EU GDPR, the target may be narrow, for example information associated with a single person. Behavioral forgetting alone may be insufficient, motivating interventions directly on internal representations. However, standard mechanistic-interpretability extractors are poorly selective for such targets. We identify an *energy bias* in reconstruction-based extraction, which favors dominant background structure over low-energy target-specific components. We introduce SCALPEL, a contrastive sparse autoencoder designed to learn more selective forget features. We show theoretically that contrastive training promotes target-selective features and that our selection score controls expected background knowledge perturbation. We validate SCALPEL experimentally on TOFU across Qwen, Llama, and Gemma, where it substantially improves over NMF and standard SAE interventions and is competitive with Gradient Difference and RMU, bridging mechanistic interpretability and fine-grained unlearning.

1 INTRODUCTION

Once a large language model (LLM) has learned a piece of information, it is hard to take back. Yet private, copyrighted, or otherwise sensitive information may need to be removed without retraining from scratch. This is the goal of machine unlearning and relates directly to requirements such as the right to erasure under Article 17 of the EU GDPR (European Parliament and Council of the European Union, 2016).

Most existing methods modify model behavior through weight optimization, representation steering, or inference-time conditioning (Bu et al., 2024; Huu-Tien et al., 2024; Pawelczyk et al., 2023). They can make a model stop answering forget-set queries, but leave a basic question open: was the knowledge removed, or only prevented from being expressed?

If the internal representation survives, the knowledge may resurface under paraphrases, indirect prompts, or attacks. Recent work on parametric knowledge traces shows that behavioral forgetting can leave much of the underlying trace intact (Hong et al., 2025). This motivates identifying target-associated representations and intervening on them directly.

Mechanistic interpretability provides natural tools for this. Information accumulates through the residual stream (Elhage et al., 2021), while PCA, NMF, and sparse autoencoders (SAEs) decompose activations into meaningful directions (Fel et al., 2023; Jourdan et al., 2023; Cunningham et al., 2023; Bricken et al., 2023; Gao et al., 2024). However, existing approaches are better suited to broad, separable concepts (Farrell et al., 2024; Patil, 2024), whereas realistic forget requests may be much narrower: forgetting one author is not removing an entire topic. In this regime, we find that standard PCA- and SAE-based interventions are poorly selective and damage retain knowledge together with forget knowledge.

We identify one cause in the training objective. Reconstruction-based extractors are rewarded according to the *energy* of the structure they capture, so a low-energy target-specific component may receive much less incentive than the high-energy background in which it is embedded (Section 3).

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.

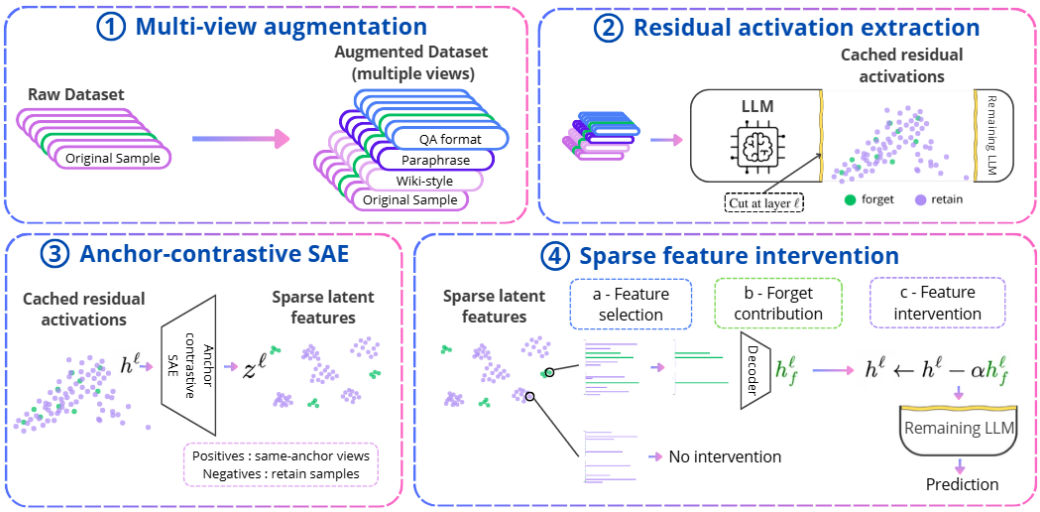


Figure 1: Overview of SCALPEL. (1) We generate multiple views of each forget anchor and (2) cache residual-stream activations. (3) An anchor-contrastive SAE learns sparse features from same-anchor positives and retain negatives. (4) Forget-specific feature contributions are decoded and subtracted when they activate, yielding an input-dependent internal filter.

We address this with SCALPEL, a contrastive sparse autoencoder that learns from multiple views of each forget target. Views of the same target are treated as positives and retain examples as negatives, encouraging sparse features that are stable across reformulations of the target but weak on background data. Selected feature contributions are then subtracted only when they activate, producing an explicit, input-dependent intervention (Figure 1).

Our goal is not certified parametric deletion, but to test whether internal features can be extracted selectively enough to support fine-grained representation-level unlearning.

Contributions. We make four contributions:

- We identify an **energy-driven selectivity bottleneck**: reconstruction objectives can under-represent low-energy forget targets (Section 3).
- We introduce **SCALPEL**, an anchor-contrastive SAE that learns target-selective sparse features and a norm-aware selection criterion (Section 4).
- We formalize how contrastive learning promotes target selectivity and how selection controls expected background perturbation (Section 5).
- Across three model families on TOFU, SCALPEL substantially improves over NMF and standard SAE interventions and is competitive with RMU and Gradient Difference (Section 7).

2 Related Work

General approaches to LLM unlearning. Unlearning removes the influence of a forget set $\mathcal{D}_f$ while preserving behavior on a retain set $\mathcal{D}_r$; an auxiliary set $\mathcal{D}_w$ (e.g. Wiki data) is often used to check general knowledge (Maini et al., 2024; Shi et al., 2024; Dorna et al., 2025). A first family updates the weights with gradient-based objectives. Gradient Ascent raises the loss on forget examples but quickly damages the model (Maini et al., 2024); Gradient Difference adds descent on $\mathcal{D}_r$ to limit the damage, with normalized variants (Maini et al., 2024; Bu et al., 2024); Negative Preference Optimization replaces the unstable ascent term with a smoother preference-based loss (Zhang et al., 2024); other approaches finetune the model toward generic or incorrect answers (Eldan & Russinovich, 2023). Unlearning can also avoid weight changes: in-context unlearning conditions

the model at inference time (Pawelczyk et al., 2023), which changes behavior but not the stored representation. Model editing such as ROME updates parameters tied to one factual association (Meng et al., 2022), though editing replaces an association rather than removing an influence.

A particularly relevant baseline is Representation Misdirection for Unlearning (RMU) (Li et al., 2024; Huu-Tien et al., 2024), which pushes forget activations at a chosen layer toward a random direction while pinning retain activations to those of the frozen original model. RMU shares our core principle act on internal representations, not just outputs but never identifies an interpretable representation of the target: it moves the whole forget activation and modifies the weights. SCALPEL instead extracts a small set of sparse target-specific features and intervenes only along their decoded directions. Comparing against RMU therefore tests whether an interpretable, selective intervention can match a stronger but opaque one.

Behavioral evaluation alone cannot establish that knowledge is gone: an internal trace may remain recoverable through other prompts or light finetuning (Hong et al., 2025; Yuan et al., 2024). Representation-level methods do not guarantee parametric deletion either, but they make the mechanism explicit by causally linking behavioral change to an identified internal intervention.

Mechanistic-interpretability-based unlearning. In transformers, a natural intervention point is the residual stream, through which information flows between layers (Elhage et al., 2021; Dar et al., 2022; Geva et al., 2022; 2023). Concept-decomposition methods express activations through a small set of meaningful directions: PCA extracts high-variance orthogonal directions and NMF factorizes non-negative activations (Jolliffe, 2002; Lee & Seung, 1999); we use NMF as a classical baseline whose components can be scored on forget/retain data and used as intervention directions. Sparse autoencoders offer a more flexible decomposition and have become a standard tool for interpretable features (Cunningham et al., 2023; Bricken et al., 2023; Gao et al., 2024; Bussmann et al., 2024; Rajamanoharan et al., 2024). Given an activation h, an SAE computes a sparse code $z(h) = E_\phi(h)$ and reconstructs $\widehat{h} = b_D + \sum_{q=1}^{Q} z_q(h)\, d_q$, where b_D is the decoder bias and d_q the decoder direction of feature q; the linear decoder associates each feature with an explicit direction in residual space, making feature-level interventions decomposable and inspectable.

A standard SAE unlearning pipeline trains the SAE, identifies forget-related features from activation statistics, and modifies them at inference. A direct intervention writes $\widetilde{h} = h - \alpha \sum_{q \in S_f} z_q(h)\, d_q$ for a selected set S_f and strength α, with clamping or negative scaling as variants (Farrell et al., 2024). An alternative is to edit the sparse code and replace h by the full SAE reconstruction, as in LLM Neurosurgeon (Zhou et al., 2025). That choice, however, additionally injects the SAE's reconstruction error, $\widetilde{h} - h = -\sum_{q \in S_f} z_q(h) d_q + (b_D + Dz(h) - h)$, which in our experiments degraded retain and Wiki performance even before any feature was removed. We therefore always intervene on the original activation. SAE features can also constrain weight updates through decoder-direction subspaces (Wang et al., 2025); this suits broad, globally separable topics, whereas our targets overlap heavily with retain data and do not occupy a clean linear subspace. Existing MI-based unlearning methods generally reuse features learned by generic reconstruction objectives; we instead study how to train the extractor itself for the selectivity required by narrow forget targets.

The main limitation of these pipelines is not finding features that activate on forget examples, but finding features that activate specifically on the target while remaining weak on shared background structure. SCALPEL addresses this selectivity bottleneck by adapting the feature-learning objective itself. The next section explains why reconstruction objectives are poorly aligned with this requirement.

3 Energy Bias in Reconstruction-Based Extraction

One source of the selectivity bottleneck is an objective mismatch: reconstruction favors dominant structure, whereas targeted unlearning requires specificity. Following the additive concept-manifold view of Bhalla et al. (2026), let $h = \sum_{j=1}^{J} m_j$, with $m_j \in \mathcal{M}_j$ and energy $E_j = \mathbb{E}[\|m_j\|_2^2]$. Under centered, mutually uncorrelated components, the reconstruction loss decomposes as

$$\mathcal{L}_{\mathrm{rec}}(\Theta) = \sum_{j=1}^{J} E_j \big(1 - C_j(\Theta)\big),$$

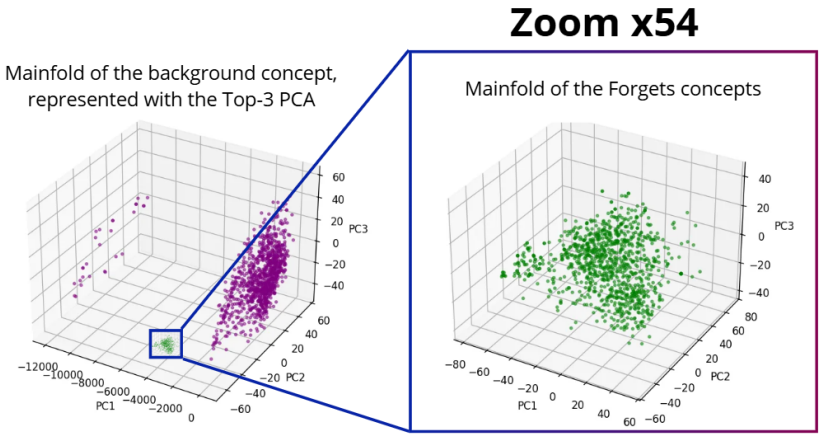


Figure 2: Energy comparison at layer -4 of Qwen. Left: dominant background structure from retain and Wiki activations. Right: forget activations after removing the top-three background directions and refitting PCA on the residuals. The two panels use different bases and should be compared through their scale: the background top-3 PCA energy is $\sim 329\times$ larger.

where $C_j(\Theta)$ is the fraction of concept j's energy captured by the extractor. Improving its capture by ΔC_j therefore yields a gain $E_j \Delta C_j$: reconstruction allocates more incentive to high-energy structure. A low-energy but important forget concept can therefore be blended into broader background features (Appendix A).

We test whether this regime occurs in Qwen's residual stream. We fit PCA on background activations $H_{\text{bg}} = H_r \cup H_w$, remove their top-three directions from forget activations, and fit a second PCA on the residuals. The corresponding top-three energies are $\widehat{E}_{\text{bg}}^{(3)} = 3{,}787{,}695$ and $\widehat{E}_f^{\perp \text{bg},(3)} = 11{,}517$, a $\sim 329\times$ gap (Figure 2). The residual is not a pure forget manifold, but the gap is far larger than under a random split of the same data. This supports the proposed mechanism and motivates an objective that explicitly rewards *selectivity*.

4 SCALPEL: Sparse Contrastive Autoencoders for Precise Representation-Level Unlearning

SCALPEL defines fine-grained forget targets, constructs multiple semantic views, learns target-selective sparse features from cached residual activations, and filters their decoded contribution at inference (Figure 1).

4.1 Target granularity and multi-view construction

Targeted unlearning first needs a precise definition of a forget concept. Let $\mathcal{D}_f, \mathcal{D}_r, \mathcal{D}_w$ be the forget, retain, and Wiki datasets and $\mathcal{D}_{\text{bg}} = \mathcal{D}_r \cup \mathcal{D}_w$. In TOFU, forget and retain authors come from the same generation process and the split is random, not semantic: they share writing style, factual structure, vocabulary, and QA format, and their activations overlap heavily (Appendix D, Figure 5). Treating the whole forget set as one concept would push the extractor toward features of the common TOFU format rather than the knowledge to remove. We instead define one concept per forget author, $G_a = \{x \in \mathcal{D}_f : \text{author}(x) = a\}$, which matches the right-to-be-forgotten setting and gives several related samples per target (sample-level granularity is discussed in Appendix I).

For each target G_a we build a set of augmented views $\mathcal{V}_a = \{x_{a,1}, \ldots, x_{a,m_a}\}$ using paraphrasing, prompt-style rewriting with `Qwen2.5-7B-Instruct`, and backtranslation (Sennrich et al., 2016; NLLB Team et al., 2022), in the spirit of view-based contrastive learning (Chen et al., 2020). Before augmentation we extract the key facts of the target (names, dates, titles, relations) and constrain them to remain unchanged, so the surface form varies while the defining target facts are retained.

The same recipe extends to document-level datasets such as MUSE by treating each chunk as a target; we leave that evaluation to future work.

4.2 RESIDUAL-STREAM ACTIVATION EXTRACTION

We run the finetuned model and cache residual-stream activations $h_t^\ell(x) \in \mathbb{R}^d$ at a fixed layer ℓ, in practice the fourth layer before the output: deep enough to carry high-level factual information, but before the final layers specialize for next-token prediction (Yu et al., 2024; Mabrok, 2026). We keep this layer fixed across all representation-based methods. We do not claim that it is optimal, and a systematic study of layer selection is left for future work. For a dataset $\mathcal{D}_s$ and retained token positions $\mathcal{I}(x)$, we write $\mathcal{H}_s = \{h_t^\ell(x) : x \in \mathcal{D}_s,\, t \in \mathcal{I}(x)\}$ for $s \in \{f, r, w\}$, $\mathcal{H}_{\text{bg}} = \mathcal{H}_r \cup \mathcal{H}_w$, and $\mathcal{H}_{G_a}$ for the tokens of target G_a. Each activation dimension is standardized with statistics from the training cache, $\overline{h} = (h - \mu_H) \oslash \sigma_H$; both the SAE input and its reconstruction live in this standardized space, which we found important for robustness across model families.

4.3 CONTRASTIVE SPARSE AUTOENCODER

Sparse decomposition. All variants share the same base. The encoder maps a standardized activation to a non-negative sparse code $z(\overline{h}) = E_\phi(\overline{h}) \in \mathbb{R}_+^Q$, and a linear decoder reconstructs $\widehat{\overline{h}} = b_D + \sum_{q=1}^Q z_q(\overline{h})\, d_q$. Training combines the usual reconstruction loss $\mathcal{L}_{\text{rec}} = \mathbb{E}_{\overline{h}}[\|\overline{h} - \widehat{\overline{h}}\|_2^2]$ over forget, retain, and Wiki activations with an ℓ_1 penalty $\mathcal{L}_{\text{sparse}} = \mathbb{E}_{\overline{h}}[\|z(\overline{h})\|_1]$.

The raw coefficient z_q alone does not measure a feature's decoded contribution and is sensitive to reciprocal rescaling of the encoder and decoder. We therefore use the scale-invariant norm-aware activation $\widetilde{z}_q(\overline{h}) = z_q(\overline{h})\, \|d_q\|_2 \geq 0$, and represent token $i = (x, t)$ by $r_i = (\widetilde{z}_1(\overline{h}_i), \ldots, \widetilde{z}_Q(\overline{h}_i)) \in \mathbb{R}_+^Q$. The contrastive objective operates directly on these token-level codes through the cosine similarity $\text{sim}(i, j) = \langle r_i, r_j \rangle / (\|r_i\|_2 \|r_j\|_2)$; no token averaging is done before the similarity (sample-level aggregation is used only later, for feature scoring).

Token-level multi-positive InfoNCE. All cached activations are flattened into one pool of tokens, each inheriting the concept identifier and split of its source sample; the pool is shuffled and cut into minibatches each epoch, independently of sample boundaries. For an anchor token i, positives $\mathcal{P}(i)$ are the other in-batch tokens sharing its concept identifier (another token of the same sample, or a token of another view of the same target), and negatives $\mathcal{N}(i)$ are in-batch tokens with a different concept identifier that satisfy the split-specific contrastive mask; depending on the variant, negatives may include retain tokens or Wiki tokens. Anchors with an empty positive or negative set are dropped. With temperature $\tau > 0$ and $\kappa(i, j) = \exp(\text{sim}(i, j)/\tau)$, define the positive and negative masses $A_i^+ = \sum_{j \in \mathcal{P}(i)} \kappa(i, j)$ and $A_i^- = \sum_{j \in \mathcal{N}(i)} \kappa(i, j)$, and the multi-positive InfoNCE loss (van den Oord et al., 2018; Khosla et al., 2020)

$$\ell_{\text{NCE}}(i) = -\log \frac{A_i^+}{A_i^+ + A_i^-}, \qquad \mathcal{L}_{\text{NCE}} = \frac{1}{|\mathcal{I}_{\text{valid}}|} \sum_{i \in \mathcal{I}_{\text{valid}}} \ell_{\text{NCE}}(i).$$

This pulls together token codes of the same target and pushes them away from other concepts and from background data. Crucially, negatives come from retain data rather than from other forget examples, so related forget samples are never pushed apart.

4.4 SCALPEL

SCALPEL. Forget and background samples share a lot of structure, and forcing one latent code to serve both reconstruction and separation can be restrictive. Inspired by contrastive SAEs of (Poupart, 2024), we split the code into two blocks produced by two encoders, $z(\overline{h}) = [z^{\text{com}}(\overline{h}), z^{\text{diff}}(\overline{h})]$, decoded jointly. The common block captures structure shared across samples ($\mathcal{L}_{\text{com}}$); the target-specific block contrasts views of the same forget target against retain and Wiki examples ($\mathcal{L}_{\text{diff}}$). The objective is $\mathcal{L}_{\text{split}} = \mathcal{L}_{\text{rec}} + \lambda \mathcal{L}_{\text{sparse}} + \beta_{\text{com}} \mathcal{L}_{\text{com}} + \beta_{\text{diff}} \mathcal{L}_{\text{diff}}$, and only features of the target-specific block are eligible for intervention.

4.5 FEATURE SELECTION AND RESIDUAL-STREAM INTERVENTION

After training, we score each feature by how active it is on one target relative to the background. With $\mu_q(G_a) = \mathbb{E}_{\overline{h} \in \overline{\mathcal{H}}_{G_a}}[\widetilde{z}_q(\overline{h})]$, and $\mu_q(\mathrm{bg}) = \mathbb{E}_{\overline{h} \in \overline{\mathcal{H}}_{\mathrm{bg}}}[\widetilde{z}_q(\overline{h})]$, the target-selectivity score is

$$s_q(G_a) = \frac{\mu_q(G_a)}{\mu_q(\mathrm{bg}) + \varepsilon}. \tag{1}$$

For each target, we select $S_{G_a} = \{q : s_q(G_a) \geq s_{\min}\}$, with a top-feature fallback if the set is empty. For a forget request containing several targets, the fixed global intervention set is $S_f = \bigcup_{a \in \mathcal{A}_f} S_{G_a}$, where shared features are included only once.

The same set S_f is applied to every input, without access to author identity or forget/retain labels; its effect stays input-dependent, since a selected feature is subtracted only when the encoder activates it. As the SAE operates in standardized space, we compute $\overline{h}^{\ell} = (h^{\ell} - \mu_H) \oslash \sigma_H$ and map the decoded contribution back:

$$\widetilde{h}^{\ell} = h^{\ell} - \alpha \sum_{q \in S_f} z_q(\overline{h}^{\ell}) \big(\sigma_H \odot d_q\big). \tag{2}$$

Modifying the original activation rather than replacing it with the full reconstruction avoids injecting reconstruction error, and each component remains a sparse activation times a decoder direction.

5 THEORETICAL ANALYSIS

We formalize two properties of SCALPEL: contrastive learning favors target-selective sparse representations, and selecting such features limits their expected effect on background activations. Complete assumptions, quantitative statements, and proofs are given in Appendices B and C.

Contrastive learning promotes target-selective features. Write $\mathcal{L}_{\mathrm{SAE}} = \mathcal{L}_{\mathrm{rec}} + \lambda \mathcal{L}_{\mathrm{sparse}}$ and $\mathcal{L}_{\mathrm{SCALPEL}} = \mathcal{L}_{\mathrm{SAE}} + \beta \mathcal{L}_{\mathrm{NCE}}$. Reconstruction preserves information and sparsity concentrates it on few coordinates, but neither explicitly rewards target-background separation. The contrastive term provides this signal.

Let $\widetilde{z}_q(h) = z_q(h)\|d_q\|_2$ and $u_i = \widetilde{z}(h_i)/\|\widetilde{z}(h_i)\|_2$.

Proposition 5.1 (Contrastive regularization favors target selectivity)**.** *Assume non-negative codes, controlled decoder scales, and that sparse target-background separation is compatible with comparable reconstruction and sparsity cost. As the contrastive objective approaches its optimum, same-target representations satisfy* $\langle u_i, u_j \rangle \to 1$*, while their average similarity with background negatives approaches* 0*.*

Thus, views of the same target increasingly share sparse coordinates while their overlap with background codes decreases. Combined with sparsity, this favors features that are active on a target and weak on the background, precisely the structure measured by $s_q(G_a)$ in Eq. equation 1. A quantitative finite-gap version is given in Appendix B.

Selectivity controls background perturbation. In standardized activation space, let $\delta_{G_a}(\overline{h}) = \alpha \sum_{q \in S_{G_a}} z_q(\overline{h}) d_q$.

Proposition 5.2 (Selectivity controls background perturbation)**.** *If every selected feature satisfies* $s_q(G_a) \geq s_{\min}$*, then*

$$\mathbb{E}_{\overline{h} \sim \overline{\mathcal{H}}_{\mathrm{bg}}}\left[\|\delta_{G_a}(\overline{h})\|_2\right] \leq \alpha \sum_{q \in S_{G_a}} \mu_q(\mathrm{bg}) \leq \frac{\alpha}{s_{\min}} \sum_{q \in S_{G_a}} \mu_q(G_a).$$

The bound controls expected representation-space perturbation, not downstream behavior or worst-case effects, and ignores the orientation of d_q, so similarly scored features may have different causal effects (Section 7.3).

Together, these results explain the intended mechanism: reconstruction preserves information, sparsity concentrates it, contrastive learning separates target from background structure, and selectivity limits the expected background footprint of the intervention.

6 EXPERIMENTS

We run one main benchmark and two diagnostics. The benchmark compares representation-level unlearning against classical, mechanistic-interpretability, and optimization-based baselines across three model families. The diagnostics test whether contrastive training improves feature selectivity and whether the selectivity score is associated with intervention quality.

Models and reference checkpoints. We use `Qwen2.5-1.5B-Instruct`, `Llama-3.2-1B-Instruct`, and `Gemma-2-2B-it` (Yang et al., 2024; Grattafiori et al., 2024; Gemma Team, 2024). For each model and seed, we finetune a full model θ_{full} on $\mathcal{D}_f \cup \mathcal{D}_r$ and a retain-only reference θ_{ret} on $\mathcal{D}_r$. All methods start from the same full checkpoint. Augmented views are generated with `Qwen2.5-7B-Instruct`.

Dataset and augmentation. We evaluate on TOFU (Maini et al., 2024), with Wiki data $\mathcal{D}_w$ for general-knowledge preservation. For each forget author, we generate 16 types of semantic views, including QA reformulations, paraphrases, biographies, Wiki-style passages, and multi-hop or adversarial prompts, followed by backtranslation through French, Spanish, and German (Zou et al., 2023; Dahal et al., 2026; Sennrich et al., 2016; NLLB Team et al., 2022). The transformations vary surface form while preserving the defining target facts.

Compared methods. We compare our methods **contrastive SAEs** and **SCALPEL** with **NMF** (Lee & Seung, 1999), a reconstruction-only **standard SAE** (Cunningham et al., 2023; Farrell et al., 2024), **Gradient Difference** (Bu et al., 2024), and **RMU** (Huu-Tien et al., 2024). All representation-based methods use layer $L - 4$ and the same token positions, selection procedure, and intervention whenever applicable.

Protocol. Each method is tuned independently for every model family using Optuna (Akiba et al., 2019); search spaces and budgets are reported in Appendix E. The selected configuration is frozen and rerun over five independent end-to-end seeds, including reference-model finetuning, method training, feature selection, and evaluation with OpenUnlearning (Dorna et al., 2025).

Metrics and trade-off curves. The main trade-off compares TOFU model utility with forget-set ground-truth answer probability (Maini et al., 2024). We additionally report forget, retain, and Wiki metrics including Q-A probability, ROUGE, truth ratio, privacy leakage, extraction strength, and paraphrase-based evaluation (Lin, 2004; Maini et al., 2024; Shi et al., 2024; Carlini et al., 2021; Dorna et al., 2025). For methods with intervention strength α, we trace the trade-off using adaptive bisection in $\log \alpha$ and aggregate the five seed curves on a common grid (Appendix H). Complete multi-metric results are in Appendix F.2.

Diagnostics. On Qwen, we first compare within-target and target-background cosine similarity in the norm-aware sparse codes of the standard SAE, contrastive SAE, and SCALPEL (Section 7.2). We then evaluate 50 individual author-feature pairs across intervention strengths and compare their forget-retain trajectories across selectivity-score bins (Section 7.3).

7 RESULTS

7.1 MAIN UNLEARNING BENCHMARK

Figure 3 shows the main benchmark on Qwen; Llama and Gemma behave consistently and are reported in Appendix F.1. Three patterns stand out. First, NMF and the standard SAE are poorly selective: reductions in forget probability come with substantial loss in model utility, as predicted by the objective mismatch of Section 3 their features fire on forget examples but also capture structure shared with retain and Wiki data.

Second, contrastive feature learning substantially improves the trade-off. The plain contrastive SAEs already outperform the reconstruction-only SAE, while the main SCALPEL variants lie on or close to the Pareto front over the practically relevant range. This supports our central claim that adapting

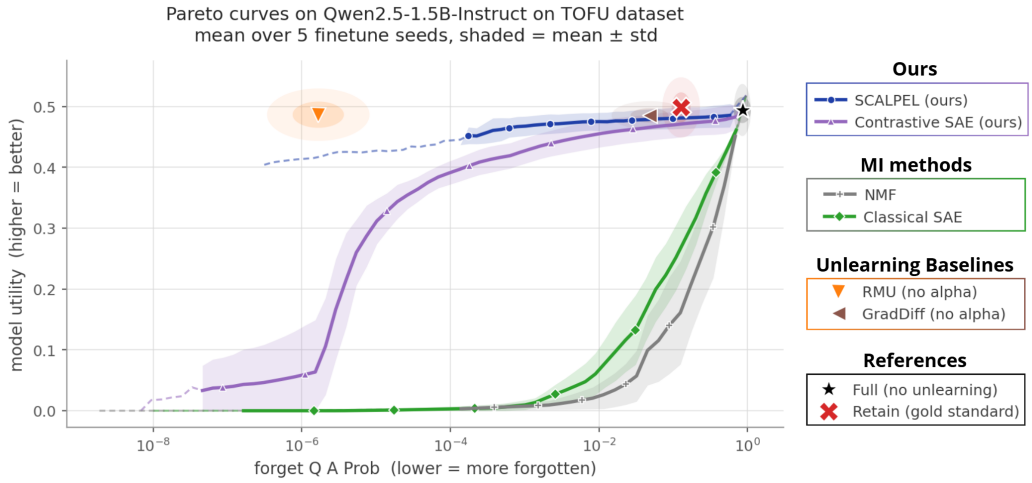


Figure 3: Unlearning trade-off on TOFU forget10 for `Qwen2.5-1.5B-Instruct`, after method-specific hyperparameter optimization and aggregation over five independent end-to-end seeds. Each curve varies the intervention strength: lower forget probability indicates stronger forgetting, while higher model utility indicates better preservation. RMU and GradDiff are shown as single points because they do not expose an inference-time intervention strength analogous to α. Results for Llama and Gemma are reported in Appendix F.1 and show similar trends. Additional metrics are reported in Appendix F.2.

the extraction objective to target selectivity produces more effective representation-level interventions.

Third, RMU and Gradient Difference remain strong baselines, but SCALPEL achieves competitive trade-offs while keeping the base model fixed and expressing the intervention through explicit sparse feature activations and decoder directions. Corresponding experiments on Llama and Gemma are provided in Appendix F.1.

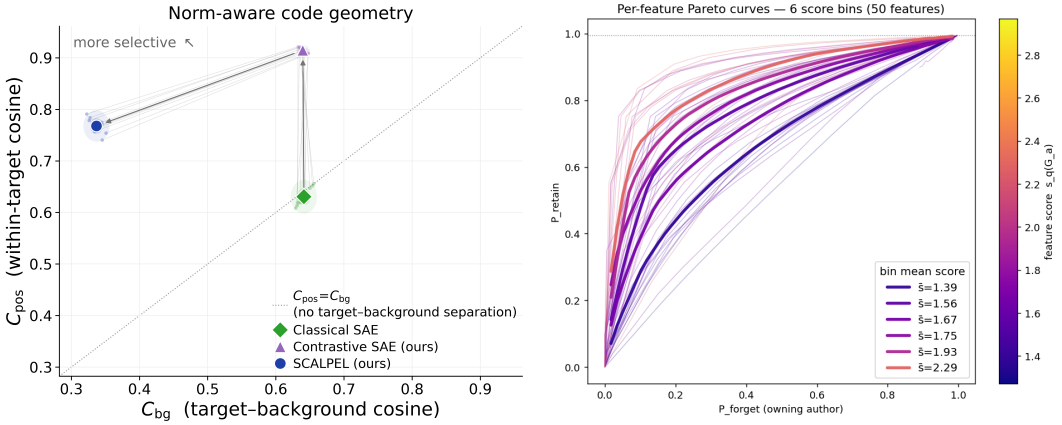
7.2 Contrastive learning increases feature selectivity

Figure 4a tests the geometric prediction of Proposition 5.1. The reconstruction-only SAE lies close to $C_{\text{pos}} = C_{\text{bg}}$, with a separation gap $\Delta C = C_{\text{pos}} - C_{\text{bg}} \approx -0.01$, indicating little target-background separation. Adding the contrastive objective increases within-target similarity and raises the gap to $\Delta C \approx 0.27$. SCALPEL-Split further reduces target-background similarity, reaching $\Delta C \approx 0.43$. The representation therefore moves progressively toward the high-C_{pos}, low-C_{bg} regime favored by the sparse contrastive objective.

This geometric separation is also reflected at the feature level. Appendix G compares the top-5 selectivity scores of the same models across authors and seeds. Mean top-5 selectivity increases from approximately 1.9 for the reconstruction-only SAE to 4.0 for the contrastive SAE, and to 24.5 for SCALPEL-Split. Together, these results support the proposed role of contrastive learning: reconstruction and sparsity produce a compact representation, while the contrastive objective organizes its features around the target concepts and away from background structure.

7.3 The selectivity score predicts unlearning quality

Figure 4b tests whether Eq. equation 1 predicts the quality of individual interventions. Each author-feature pair is evaluated in isolation while varying its intervention strength. In aggregate, higher-score bins preserve more retain probability at fixed forget probability, or achieve stronger forgetting at comparable retain probability. The selectivity score is therefore associated not only with feature selection but also with downstream intervention quality. Together with the contrastive-geometry diagnostic, this connects the learned representation to feature selectivity and, in turn, to intervention quality. The ordering remains imperfect at the level of individual features: the score captures



(a) Norm-aware sparse-code geometry across five Qwen seeds. C_{pos} measures within-target similarity and C_{bg} target-background similarity. More selective representations lie toward the top-left: contrastive learning increases target alignment, while SCALPEL-Split further reduces background overlap.

(b) Single-feature unlearning curves grouped by selectivity-score quantile. Thin curves represent individual author-feature pairs; thick curves show the average within each bin. Higher-score features generally provide better forget-retain trade-offs.

Figure 4: Diagnostics linking representation geometry, feature selectivity, and intervention quality for SCALPEL. Left: contrastive training increases within-target alignment and reduces target-background overlap in the norm-aware sparse code. Right: features with higher selectivity scores generally achieve better single-feature forget-retain trade-offs.

activation magnitude and target-background selectivity, but ignores the orientation of d_q in the local activation geometry. A geometry-aware score would require substantially more expensive local manifold estimation (Appendix I).

8 Conclusion and Limitations

We proposed a first approach for connecting mechanistic interpretability with fine-grained unlearning of specific information. Our study shows that reconstruction-based concept extractors can lack the selectivity required to isolate a narrow forget target from the background knowledge that should be preserved. We identify one mechanism behind this behavior: reconstruction rewards captured energy, while target-specific information may correspond to a low-energy component entangled with much larger background structure; at the studied layer, we measure a $\sim 329\times$ energy gap. SCALPEL addresses this problem by adapting the feature-learning objective rather than the intervention pipeline, contrasting multiple views of each target against background data to learn sparse, target-selective features. Across Qwen, Llama, and Gemma on TOFU, SCALPEL substantially improves over NMF and reconstruction-only SAE interventions and is competitive with RMU and Gradient Difference. Our theoretical analysis and diagnostics further show how contrastive learning promotes target-selective representations and how feature selectivity relates to intervention quality.

Limitations. Our evaluation covers TOFU and three relatively small models; larger models, less controlled datasets, and document-, fact-, or sample-level targets remain untested. The theoretical analysis is local, and the perturbation bound controls representation-space magnitude rather than model behavior. Most importantly, SCALPEL does not erase the parametric trace: it applies a persistent internal filter to selected feature contributions whenever they activate, so our results establish selective behavioral and representation-level unlearning under intervention, not certified deletion. Using the extracted features to guide permanent parameter updates while preserving their selectivity is the natural next step. A fuller discussion is in Appendix I.

AI USE STATEMENT

AI was used at two levels during this project. First, it was used as a research assistant to implement and explore our ideas more quickly, allowing faster iteration and helping us identify promising or unpromising directions early. Second, we wrote the paper ourselves and used AI only to polish the writing style and improve clarity. All research ideas, methodological choices, experiments, and interpretations are our own; AI was used only to accelerate the exploration and writing process.

ETHICS STATEMENT

Machine unlearning may contribute to better control over information stored in machine-learning models, with potential applications to privacy and data governance. However, techniques for selectively modifying model knowledge could also be used to suppress legitimate information or manipulate model behavior. The representation-level interventions studied in this work should therefore be considered research tools for understanding and controlling model representations, rather than guarantees of data deletion or privacy compliance. Our main experiments use the synthetic TOFU benchmark, avoiding direct experimentation on personal information associated with real individuals.

REFERENCES

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

Usha Bhalla, Thomas Fel, Can Rager, Sheridan Feucht, Tal Haklay, Daniel Wurgaft, Siddharth Boppana, Matthew Kowal, Vasudev Shyam, Jack Merullo, Atticus Geiger, and Ekdeep Singh Lubana. Do sparse autoencoders capture concept manifolds?, 2026. URL https://arxiv.org/abs/2604.28119.

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, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E. Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits Thread, 2023. https://transformer-circuits.pub/2023/monosemantic-features.

Zhiqi Bu, Xiaomeng Jin, Bhanukiran Vinzamuri, Anil Ramakrishna, Kai-Wei Chang, Volkan Cevher, and Mingyi Hong. Unlearning as multi-task optimization: A normalized gradient difference approach with an adaptive learning rate, 2024.

Bart Bussmann, Patrick Leask, and Neel Nanda. BatchTopK sparse autoencoders, 2024.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *30th USENIX Security Symposium*, pp. 2633–2650. USENIX Association, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models, 2023.

Chahana Dahal, Ashutosh Balasubramaniam, and Zuobin Xiong. GONE: Structural knowledge unlearning via neighborhood-expanded distribution shaping, 2026.

Guy Dar, Mor Geva, Ankit Gupta, and Jonathan Berant. Analyzing transformers in embedding space, 2022.

V. Dorna et al. Openunlearning: Accelerating llm unlearning via unified benchmarking of methods and metrics, 2025.

Ronen Eldan and Mark Russinovich. Who's harry potter? approximate unlearning in LLMs, 2023.

Nelson Elhage, Neel Nanda, Catherine Olsson, et al. A mathematical framework for transformer circuits. Transformer Circuits Thread, 2021. https://transformer-circuits.pub/2021/framework/index.html.

European Parliament and Council of the European Union. Regulation (eu) 2016/679 of the european parliament and of the council (general data protection regulation). Official Journal of the European Union, L 119, 2016. URL https://eur-lex.europa.eu/eli/reg/2016/679/oj. Article 17: Right to erasure ('right to be forgotten').

Eoin Farrell, Yeu-Tong Lau, and Arthur Conmy. Applying sparse autoencoders to unlearn knowledge in language models, 2024.

Thomas Fel, Victor Boutin, Mazda Moayeri, Rémi Cadène, Louis Bethune, Léo Andéol, Mathieu Chalvidal, and Thomas Serre. A holistic approach to unifying automatic concept extraction and concept importance estimation. In *Advances in Neural Information Processing Systems*, 2023.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders, 2024.

Gemma Team. Gemma 2: Improving open language models at a practical size, 2024.

Mor Geva, Avi Caciularu, Kevin Ro Wang, and Yoav Goldberg. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 30–45. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.3.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12216–12235. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.751.

Aaron Grattafiori et al. The Llama 3 herd of models, 2024.

Yihuai Hong, Lei Yu, Haiqin Yang, Shauli Ravfogel, and Mor Geva. Intrinsic test of unlearning using parametric knowledge traces. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 19513–19535. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.985.

Dang Huu-Tien, Trung-Tin Pham, Hoang Thanh-Tung, and Naoya Inoue. On effects of steering latent representation for large language model unlearning, 2024.

Ian T. Jolliffe. *Principal Component Analysis*. Springer, 2002.

Fanny Jourdan, Agustin Picard, Thomas Fel, Laurent Risser, Jean Michel Loubes, and Nicholas Asher. COCKATIEL: COntinuous Concept ranKed ATtribution with Interpretable ELements for explaining neural net classifiers on NLP tasks. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 5120–5136, 2023.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

Nathaniel Li, Alexander Pan, Anjali Gopal, et al. The WMDP benchmark: Measuring and reducing malicious use with unlearning, 2024.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81. Association for Computational Linguistics, 2004.

Mohamed A. Mabrok. Latent semantic manifolds in large language models, 2026.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C. Lipton, and J. Zico Kolter. TOFU: A task of fictitious unlearning for LLMs, 2024.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

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

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. No language left behind: Scaling human-centered machine translation, 2022.

Kaustubh Patil. Targeted knowledge removal in LLMs using sparse autoencoders. OpenReview, 2024. `https://openreview.net/forum?id=aeQeXlG2Pw`.

Martin Pawelczyk, Seth Neel, and Himabindu Lakkaraju. In-context unlearning: Language models as few shot unlearners, 2023.

Yoann Poupart. Contrastive sparse autoencoders for interpreting planning of chess-playing agents, 2024.

Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. Jumping ahead: Improving reconstruction fidelity with JumpReLU sparse autoencoders, 2024.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 86–96. Association for Computational Linguistics, 2016.

Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A. Smith, and Chiyuan Zhang. MUSE: Machine unlearning six-way evaluation for language models, 2024.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Xu Wang, Zihao Li, Benyou Wang, Yan Hu, and Difan Zou. Model unlearning via sparse autoencoder subspace guided projections. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 26530–26546. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1348.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Lin, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2024.

Lei Yu, Meng Cao, Jackie Chi Kit Cheung, and Yue Dong. Mechanisms of non-factual hallucinations in language models, 2024.

Xiaojian Yuan, Tianyu Pang, Chao Du, Kejiang Chen, Weiming Zhang, and Min Lin. A closer look at machine unlearning for large language models, 2024.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning, 2024.

Dylan Zhou, Kunal Patil, Yifan Sun, Karthik Lakshmanan, Senthooran Rajamanoharan, and Arthur Conmy. LLM Neurosurgeon: Targeted knowledge removal in LLMs using sparse autoencoders. In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*, 2025. https://openreview.net/forum?id=aeQeXlG2Pw.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023.

A Energy Bias of Reconstruction-Based Extraction

This appendix gives the formal statement behind Section 3: reconstruction-only objectives reward an extractor in proportion to the *energy* of what it captures, so a low-energy forget concept is systematically deprioritized relative to high-energy background structure. The result is a motivation, not a theorem about trained SAEs: it isolates the incentive created by the reconstruction loss under an idealized concept decomposition, in the spirit of the additive concept-manifold view of Bhalla et al. (2026). We then describe in full the PCA experiment that measures the corresponding energy gap on real activations (Figure 2).

A.1 Setting

Let $h \in \mathbb{R}^d$ be a random residual-stream activation, and suppose it decomposes into concept-manifold components

$$h \;=\; \sum_{j=1}^{J} m_j, \qquad m_j \in \mathcal{M}_j, \tag{3}$$

where $\mathcal{M}_j$ denotes the j-th concept manifold. Define the energy of concept j as

$$E_j \;=\; \mathbb{E}\big[\|m_j\|_2^2\big] \;>\; 0. \tag{4}$$

Let Θ denote an extractor (for instance an SAE), producing a reconstruction $\hat{h}^{\Theta}$ of h, with reconstruction loss $\mathcal{L}_{\mathrm{rec}}(\Theta) = \mathbb{E}[\|h - \hat{h}^{\Theta}\|_2^2]$.

Assumption A.1 (Component orthogonality)**.** *The concept components are centered and mutually uncorrelated:*

$$\mathbb{E}[m_j] = 0 \quad \textit{for all } j, \qquad \mathbb{E}\big[\langle m_j, m_k\rangle\big] = 0 \quad \textit{for } j \neq k.$$

Under Assumption A.1, the total activation energy splits exactly across concepts, $\mathbb{E}[\|h\|_2^2] = \sum_{j=1}^{J} E_j$, which is what makes E_j interpretable as concept j's share of the signal.

Assumption A.2 (Concept-aligned reconstruction with orthogonal errors)**.** *The reconstruction admits a concept-aligned decomposition $\hat{h}^{\Theta} = \sum_{j=1}^{J} \hat{m}_j^{\Theta}$ (any decoder bias is absorbed into the estimates), and the per-concept reconstruction errors are mutually uncorrelated:*

$$\mathbb{E}\big[\langle m_j - \hat{m}_j^{\Theta},\; m_k - \hat{m}_k^{\Theta}\rangle\big] = 0 \quad \textit{for } j \neq k.$$

Assumption A.2 holds, for example, when the concept manifolds occupy mutually orthogonal subspaces and each estimate $\hat{m}_j^{\Theta}$ and its error lie in the corresponding subspace, or when the errors are independent across concepts. It is an idealization: real concept components overlap, and a trained SAE does not come with a canonical per-concept decomposition of its reconstruction. The assumption serves to isolate the incentive structure of the loss; the experiment of Section A.3 provides the empirical counterpart on real activations.

For each concept, define the *captured energy fraction*

$$C_j(\Theta) \;=\; 1 - \frac{\mathbb{E}\big[\|m_j - \hat{m}_j^{\Theta}\|_2^2\big]}{E_j} \;\leq\; 1, \tag{5}$$

with $C_j(\Theta) \geq 0$ whenever the estimate is at least as accurate as the trivial estimate $\hat{m}_j^{\Theta} = 0$.

A.2 Energy-weighted decomposition of the reconstruction loss

Proposition A.1 (Energy bias of reconstruction objectives)**.** *Under Assumptions A.1-A.2,*

$$\mathcal{L}_{\mathrm{rec}}(\Theta) \;=\; \sum_{j=1}^{J} E_j\big(1 - C_j(\Theta)\big). \tag{6}$$

Equivalently, minimizing the reconstruction loss is equivalent to maximizing the energy-weighted capture

$$G_{\mathrm{rec}}(\Theta) \;=\; \sum_{j=1}^{J} E_j\, C_j(\Theta). \tag{7}$$

Proof. Write $h - \hat{h}^{\Theta} = \sum_{j=1}^{J}(m_j - \hat{m}_j^{\Theta})$ using Eq. equation 3 and Assumption A.2. Expanding the squared norm,

$$\mathcal{L}_{\mathrm{rec}}(\Theta) = \sum_{j=1}^{J} \mathbb{E}\big[\|m_j - \hat{m}_j^{\Theta}\|_2^2\big] + \sum_{j\neq k} \mathbb{E}\big[\langle m_j - \hat{m}_j^{\Theta},\, m_k - \hat{m}_k^{\Theta}\rangle\big].$$

The cross terms vanish by Assumption A.2, and each diagonal term equals $E_j(1 - C_j(\Theta))$ by Eq. equation 5, which gives Eq. equation 6. Since $\mathcal{L}_{\mathrm{rec}}(\Theta) = \sum_j E_j - G_{\mathrm{rec}}(\Theta)$ and $\sum_j E_j$ does not depend on Θ, minimizing the former maximizes the latter. □

Corollary A.1 (Marginal capacity is allocated by energy)**.** *Consider a marginal change of the extractor that improves the capture fractions by $(\Delta C_1, \ldots, \Delta C_J)$. The induced reconstruction gain is*

$$\Delta G_{\mathrm{rec}} \;=\; \sum_{j=1}^{J} E_j\, \Delta C_j.$$

In particular, when a unit of capacity can improve either a background concept $\mathcal{M}_b$ by ΔC_b or a forget concept $\mathcal{M}_f$ by ΔC_f, the reconstruction objective prefers the background concept whenever

$$E_b\, \Delta C_b \;>\; E_f\, \Delta C_f,$$

and, for comparable capture improvements $\Delta C_b \approx \Delta C_f$, the preference is decided by the energies alone.

Remark A.1. Proposition A.1 does not say that the forget concept is unimportant to the model, only that it contributes little to the reconstruction objective compared with broad background structure. PCA is the extreme case: for linear orthogonal projections, maximizing captured energy is exactly the PCA criterion, so high-variance background directions are selected by construction. NMF and reconstruction-only SAEs are softer versions of the same incentive. This is the mechanism addressed by the contrastive objective of SCALPEL: Proposition 5.1 (proved in Appendix B) shows that adding contrastive learning to reconstruction and sparsity favors sparse representations with stronger within-target alignment and lower target-background overlap.

A.3 Measuring the energy gap on real activations

We now describe in full the experiment summarized in Section 3 and Figure 2, which measures the energy gap between background structure and target-specific structure on residual-stream activations extracted at layer -4 of Qwen.

Background PCA. Let $H_{\mathrm{bg}} = H_r \cup H_w$ be the union of retain and Wiki activations and H_f the forget activations. We estimate the background mean and covariance using only the background data,

$$\overline{h}_{\mathrm{bg}} = \frac{1}{|H_{\mathrm{bg}}|} \sum_{h \in H_{\mathrm{bg}}} h, \qquad \Sigma_{\mathrm{bg}} = \frac{1}{|H_{\mathrm{bg}}|} \sum_{h \in H_{\mathrm{bg}}} \big(h - \overline{h}_{\mathrm{bg}}\big)\big(h - \overline{h}_{\mathrm{bg}}\big)^{\top}.$$

Let $\lambda_1^{\mathrm{bg}} \geq \lambda_2^{\mathrm{bg}} \geq \lambda_3^{\mathrm{bg}} \geq \cdots$ be the eigenvalues of Σ_{bg} and $U_{\mathrm{bg},3} \in \mathbb{R}^{d\times 3}$ its first three eigenvectors. The left panel of Figure 2 shows the background activations in these coordinates, $y_{\mathrm{bg}} = U_{\mathrm{bg},3}^{\top}(h - \overline{h}_{\mathrm{bg}})$ for $h \in H_{\mathrm{bg}}$.

Background-orthogonal forget residuals. We remove the dominant background component from every forget activation,

$$\widetilde{h}_f = \big(I - U_{\mathrm{bg},3}U_{\mathrm{bg},3}^\top\big)\big(h_f - \overline{h}_{\mathrm{bg}}\big), \qquad h_f \in H_f.$$

Projecting these residuals back onto $U_{\mathrm{bg},3}$ gives zero coordinates by construction, so visualizing the remaining forget structure requires a second basis: we fit a second PCA directly on $\widetilde{H}_f = \{\widetilde{h}_f : h_f \in H_f\}$. With $\overline{\widetilde{h}}_f$ the mean of the residuals, define

$$\Sigma_f^{\perp\mathrm{bg}} = \frac{1}{|\widetilde{H}_f|}\sum_{\widetilde{h}_f \in \widetilde{H}_f}\big(\widetilde{h}_f - \overline{\widetilde{h}}_f\big)\big(\widetilde{h}_f - \overline{\widetilde{h}}_f\big)^\top,$$

with eigenvalues $\lambda_i^{f\perp\mathrm{bg}}$. The right panel of Figure 2 shows the residual forget activations in the top-3 eigenvectors of $\Sigma_f^{\perp\mathrm{bg}}$. The two panels therefore use two different PCA bases and must be compared through their scale and explained energy, not their absolute coordinates.

Energy comparison. We measure the scale of each displayed manifold by the sum of its first three PCA eigenvalues,

$$\widehat{E}_{\mathrm{bg}}^{(3)} = \sum_{i=1}^{3}\lambda_i^{\mathrm{bg}}, \qquad \widehat{E}_f^{\perp\mathrm{bg},(3)} = \sum_{i=1}^{3}\lambda_i^{f\perp\mathrm{bg}},$$

and obtain

$$\widehat{E}_{\mathrm{bg}}^{(3)} = 3{,}787{,}695, \qquad \widehat{E}_f^{\perp\mathrm{bg},(3)} = 11{,}517, \qquad \frac{\widehat{E}_{\mathrm{bg}}^{(3)}}{\widehat{E}_f^{\perp\mathrm{bg},(3)}} \approx 329.$$

What the experiment does and does not show. The experiment does not claim that the complete forget manifold carries exactly $329\times$ less energy: it measures the scale of the forget structure that *remains* after removing the three dominant background directions, which is the component a targeted intervention must isolate. The size of the gap nevertheless supports the energy-bias argument of Proposition A.1: the target-specific component is orders of magnitude smaller than the background structure that a reconstruction-based extractor is primarily rewarded for capturing.

B Sparse Contrastive Learning Promotes Target-Selective Features

We give a geometric account of why the combination of reconstruction, sparsity, and contrastive learning used by SCALPEL favors target-selective sparse features. The result characterizes the geometry preferred by the objective when such a representation is compatible with reconstruction; it is not a global convergence result for SAE training.

B.1 Setting

For a standardized residual activation h_i, the SAE produces non-negative activations $z(h_i) \in \mathbb{R}_+^Q$ and reconstructs

$$\widehat{h}_i = b_D + \sum_{q=1}^{Q} z_q(h_i)d_q.$$

The contrastive loss operates on the norm-aware code

$$\widetilde{z}_q(h_i) = z_q(h_i)\|d_q\|_2, \qquad \widetilde{z}_i = \big(\widetilde{z}_1(h_i), \ldots, \widetilde{z}_Q(h_i)\big).$$

For every non-zero code, define

$$\rho_i = \|\widetilde{z}_i\|_2, \qquad u_i = \frac{\widetilde{z}_i}{\rho_i}.$$

Thus

$$u_i \in \mathbb{R}_+^Q, \qquad \|u_i\|_2 = 1,$$

and the cosine similarity used by the contrastive objective is simply

$$\mathrm{sim}(i,j) = \langle u_i, u_j \rangle.$$

We assume that decoder scales are controlled:

$$0 < d_{\min} \le \|d_q\|_2 \le d_{\max} < \infty. \tag{A1}$$

This removes the reciprocal rescaling ambiguity between encoder activations and decoder directions.

For anchor i, let $P(i)$ and $N(i)$ be its positive and negative sets, with $p_i = |P(i)|$ and $n_i = |N(i)|$. In SCALPEL, positives belong to the same target concept and negatives are drawn from the background. The multi-positive InfoNCE loss is

$$\ell_i = -\log \frac{\sum_{j \in P(i)} e^{\langle u_i, u_j \rangle / \tau}}{\sum_{j \in P(i)} e^{\langle u_i, u_j \rangle / \tau} + \sum_{j \in N(i)} e^{\langle u_i, u_j \rangle / \tau}}.$$

B.2 Reconstruction prevents the trivial representation

The contrastive direction u_i is undefined when $\widetilde{z}_i = 0$. Reconstruction prevents the all-zero solution from being useful.

Lemma B.1 (Non-triviality from reconstruction)**.** *If $z(h) = 0$ almost surely, the best possible reconstruction is the constant decoder bias*

$$b_D^\star = \mathbb{E}[h],$$

with loss

$$L_{\text{const}} = \mathbb{E}\left[\|h - \mathbb{E}[h]\|_2^2\right].$$

Consequently, any SAE satisfying

$$L_{\text{rec}} < L_{\text{const}}$$

must use non-zero latent codes on a set of positive probability.

Proof. If $z(h) = 0$, then $\widehat{h} = b_D$ for every input. The minimizer of $\mathbb{E}\|h - b_D\|_2^2$ is $b_D = \mathbb{E}[h]$, giving the stated loss. Any reconstruction below this value therefore requires non-zero latent activations. □

Reconstruction therefore supplies a non-trivial representation, while the remaining terms determine how this representation is organized.

B.3 Sparsity concentrates the norm-aware code

For a normalized code u, let $u_{(1)} \ge \cdots \ge u_{(Q)}$ denote its coordinates in decreasing order.

Lemma B.2 (Sparse energy concentration)**.** *For every $K \ge 1$,*

$$\sum_{q > K} u_{(q)}^2 \le \frac{\|u\|_1^2}{K},$$

and hence

$$\sum_{q=1}^{K} u_{(q)}^2 \ge 1 - \frac{\|u\|_1^2}{K}.$$

Moreover,

$$\|u\|_1 \ge \|u\|_2 = 1,$$

with equality if and only if u is one-sparse.

Proof. Since the coordinates are sorted,

$$u_{(K+1)} \le \frac{\|u\|_1}{K}.$$

Therefore

$$\sum_{q>K} u_{(q)}^2 \le u_{(K+1)} \sum_{q>K} u_{(q)} \le \frac{\|u\|_1^2}{K}.$$

The second result follows from $\|u\|_2^2 = 1$. Finally, $\|u\|_1 \ge \|u\|_2$, with equality for a non-negative vector only when at most one coordinate is non-zero. □

Under Assumption (A1), sparsifying z also sparsifies $\widetilde{z}$, since multiplication by bounded positive decoder norms does not change its support. Thus the ℓ_1 term favors concentrating the representation on a small number of coordinates, while reconstruction prevents complete collapse.

B.4 Contrastive learning aligns targets and separates background

We now characterize the geometry preferred by the contrastive term.

Proposition B.1 (Optimal contrastive geometry)**.** *For every valid anchor i,*

$$\ell_i \ge \log\left(1 + \frac{n_i}{p_i} e^{-1/\tau}\right).$$

Equality holds if and only if

$$\langle u_i, u_j \rangle = 1 \quad \forall j \in P(i),$$

and

$$\langle u_i, u_j \rangle = 0 \quad \forall j \in N(i).$$

Therefore, whenever this geometry is compatible with reconstruction, the contrastive objective favors identical normalized codes within a target and disjoint supports between that target and its background negatives.

Proof. Because u_i, u_j are non-negative unit vectors,

$$0 \le \langle u_i, u_j \rangle \le 1.$$

Define

$$A_i^+ = \sum_{j \in P(i)} e^{\langle u_i, u_j \rangle / \tau}, \qquad A_i^- = \sum_{j \in N(i)} e^{\langle u_i, u_j \rangle / \tau}.$$

Then

$$A_i^+ \le p_i e^{1/\tau}, \qquad A_i^- \ge n_i.$$

Since

$$\ell_i = \log\left(1 + \frac{A_i^-}{A_i^+}\right),$$

we obtain

$$\ell_i \ge \log\left(1 + \frac{n_i}{p_i} e^{-1/\tau}\right).$$

Equality requires simultaneously

$$\langle u_i, u_j \rangle = 1 \quad (j \in P(i))$$

and

$$\langle u_i, u_j \rangle = 0 \quad (j \in N(i)).$$

For unit vectors, the first condition implies $u_i = u_j$. Because all coordinates are non-negative, the second condition implies

$$\operatorname{supp}(u_i) \cap \operatorname{supp}(u_j) = \varnothing.$$

Thus same-target examples share a normalized representation, whereas target and background negatives use disjoint coordinates. □

Combined with Lemma B.2, this gives the preferred geometry of the joint objective: reconstruction keeps the representation non-trivial, sparsity concentrates it onto few coordinates, and the contrastive term encourages those coordinates to be shared by examples of the same target while remaining inactive on background examples.

Importantly, the proposition only separates pairs defined as negatives. If two forget targets are not contrasted against each other, they may share coordinates. This is consistent with the shared features observed empirically in SCALPEL.

B.5 Near-optimal contrastive loss implies near-optimal geometry

The previous result describes the exact optimum. We can also quantify how approaching this optimum forces the representations toward the same geometry.

Let

$$R_i = \frac{A_i^-}{A_i^+}, \qquad R_i^\star = \frac{n_i}{p_i} e^{-1/\tau}.$$

Suppose

$$R_i \le (1+\delta) R_i^\star, \qquad \delta \ge 0. \tag{A2}$$

Proposition B.2 (Approximate contrastive separation)**.** *Under Eq. (A2),*

$$\frac{1}{n_i} \sum_{j \in N(i)} \langle u_i, u_j \rangle \le \tau\delta,$$

and

$$\frac{1}{p_i} \sum_{j \in P(i)} (1 - \langle u_i, u_j \rangle) \le \frac{\delta}{(1+\delta)(1-e^{-1/\tau})}.$$

Consequently,

$$\frac{1}{p_i} \sum_{j \in P(i)} \|u_i - u_j\|_2^2 \le \frac{2\delta}{(1+\delta)(1-e^{-1/\tau})}.$$

Hence, as the contrastive loss approaches its optimum, positive codes collapse toward one another while their average similarity with background negatives approaches zero.

Proof. Observe that

$$\frac{R_i}{R_i^\star} = \underbrace{\frac{A_i^-}{n_i}}_{\ge 1} \underbrace{\frac{p_i e^{1/\tau}}{A_i^+}}_{\ge 1}.$$

If their product is at most $1+\delta$, each factor is itself at most $1+\delta$.

For the negatives,

$$\frac{1}{n_i} \sum_{j \in N(i)} e^{\langle u_i, u_j \rangle / \tau} \le 1 + \delta.$$

Using $e^x \ge 1 + x$ gives

$$\frac{1}{n_i} \sum_{j \in N(i)} \langle u_i, u_j \rangle \le \tau\delta.$$

For the positives,

$$\frac{1}{p_i} \sum_{j \in P(i)} e^{-(1 - \langle u_i, u_j \rangle)/\tau} \ge \frac{1}{1+\delta}.$$

For $x \in [0,1]$, concavity of $1 - e^{-x/\tau}$ gives

$$1 - e^{-x/\tau} \ge (1 - e^{-1/\tau}) x.$$

Therefore,

$$(1 - e^{-1/\tau})\frac{1}{p_i}\sum_{j \in P(i)} (1 - \langle u_i, u_j \rangle) \leq 1 - \frac{1}{1+\delta} = \frac{\delta}{1+\delta},$$

which proves the second inequality. Finally,

$$\|u_i - u_j\|_2^2 = 2\,(1 - \langle u_i, u_j \rangle)\,,$$

giving the last result. □

Since

$$\ell_i = \log(1 + R_i),$$

condition (A2) is equivalently a bound on the excess contrastive loss: $\delta \to 0$ whenever $\ell_i \to \ell_i^\star$.

B.6 Connection to the selectivity score

Recall the SCALPEL score

$$s_q(G_a) = \frac{\mu_q(G_a)}{\mu_q(\mathrm{bg}) + \epsilon},$$

where

$$\mu_q(G_a) = \mathbb{E}_{i \in G_a}[\widetilde{z}_q(h_i)], \qquad \mu_q(\mathrm{bg}) = \mathbb{E}_{j \in \mathrm{bg}}[\widetilde{z}_q(h_j)].$$

At the ideal sparse contrastive geometry, suppose target G_a uses coordinate q. Reconstruction ensures a non-zero code scale, sparsity concentrates the target representation on q, and Proposition B.1 gives

$$\widetilde{z}_q(h_i) > 0 \quad \text{for } i \in G_a, \qquad \widetilde{z}_q(h_j) = 0 \quad \text{for } j \in \mathrm{bg}.$$

Hence

$$\mu_q(G_a) > 0, \qquad \mu_q(\mathrm{bg}) = 0,$$

and therefore

$$s_q(G_a) = \frac{\mu_q(G_a)}{\epsilon}.$$

Thus the geometry favored by the joint objective is exactly the geometry rewarded by the selectivity criterion: the sparsity term concentrates target information onto few coordinates, while the contrastive term increases agreement within the target and suppresses overlap of those coordinates with background examples. Proposition B.2 further shows that this behavior emerges continuously as the contrastive loss approaches its optimum.

C Selectivity and Intervention Safety

This appendix proves Proposition 5.2 of the main text: the selectivity score used by SCALPEL controls the expected magnitude of the perturbation applied to background activations. Together with the sparse contrastive analysis of Appendix B, it completes the training-to-intervention argument: contrastive learning favors sparse representations with stronger within-target alignment and lower target-background overlap, while selecting high-score features limits their expected footprint on retain and Wiki representations under intervention.

C.1 Setting

We use the notation of Sections 4.3 and 4.5. For a standardized activation $\overline{h}$, the encoder produces a non-negative sparse code $z(\overline{h}) \in \mathbb{R}_+^Q$, so that

$$z_q(\overline{h}) \;\geq\; 0 \qquad \text{for all } q,\ \overline{h}, \tag{8}$$

and $d_q \in \mathbb{R}^d$ is the decoder direction of feature q. For an author-level target G_a, with $\overline{\mathcal{H}}_{\mathrm{bg}} = \overline{\mathcal{H}}_r \cup \overline{\mathcal{H}}_w$ the standardized background activations, recall the mean norm-aware activations and the selectivity score (Eq. equation 1 of the main text),

$$\mu_q(G_a) = \mathbb{E}_{\overline{h} \sim \overline{\mathcal{H}}_{G_a}}\big[z_q(\overline{h})\,\|d_q\|_2\big], \qquad \mu_q(\mathrm{bg}) = \mathbb{E}_{\overline{h} \sim \overline{\mathcal{H}}_{\mathrm{bg}}}\big[z_q(\overline{h})\,\|d_q\|_2\big], \qquad s_q(G_a) = \frac{\mu_q(G_a)}{\mu_q(\mathrm{bg}) + \varepsilon},$$

with $\varepsilon > 0$. SCALPEL selects $S_{G_a} = \{q : s_q(G_a) \geq s_{\min}\}$ for a threshold $s_{\min} > 0$, and the intervention applies the perturbation

$$\delta_{G_a}(\overline{h}) \;=\; \alpha \sum_{q \in S_{G_a}} z_q(\overline{h})\, d_q, \qquad \alpha \geq 0. \tag{9}$$

C.2 The perturbation bound

Proposition C.1 (Selectivity controls background perturbation; Proposition 5.2 of the main text)**.** *If every selected feature satisfies $s_q(G_a) \geq s_{\min}$, then*

$$\mathbb{E}_{\overline{h} \sim \overline{\mathcal{H}}_{\mathrm{bg}}}\big[\|\delta_{G_a}(\overline{h})\|_2\big] \;\leq\; \alpha \sum_{q \in S_{G_a}} \mu_q(\mathrm{bg}) \;\leq\; \frac{\alpha}{s_{\min}} \sum_{q \in S_{G_a}} \mu_q(G_a) - \alpha\,\varepsilon\,\big|S_{G_a}\big| \;\leq\; \frac{\alpha}{s_{\min}} \sum_{q \in S_{G_a}} \mu_q(G_a). \tag{10}$$

Proof. By the triangle inequality and the non-negativity of the codes (Eq. equation 8), for every $\overline{h}$,

$$\|\delta_{G_a}(\overline{h})\|_2 \;\leq\; \alpha \sum_{q \in S_{G_a}} z_q(\overline{h})\, \|d_q\|_2.$$

Taking the expectation over $\overline{h} \sim \overline{\mathcal{H}}_{\mathrm{bg}}$ and using linearity gives the first inequality of Eq. equation 10. For the second, each selected feature satisfies $s_q(G_a) = \mu_q(G_a)/(\mu_q(\mathrm{bg}) + \varepsilon) \geq s_{\min}$, hence

$$\mu_q(\mathrm{bg}) \;\leq\; \frac{\mu_q(G_a)}{s_{\min}} - \varepsilon.$$

Summing over $q \in S_{G_a}$ and multiplying by α yields the second inequality; dropping the non-negative term $\alpha\varepsilon|S_{G_a}|$ yields the third. □

C.3 Discussion

Remark C.1 (Interpretation and monotonicity). The quantity $\sum_{q \in S_{G_a}} \mu_q(G_a)$ is the total norm-aware activation mass of the selected features on the target: it is the natural scale of the intervention, since the same triangle-inequality argument applied to target activations gives $\mathbb{E}_{\overline{h} \sim \overline{\mathcal{H}}_{G_a}}[\|\delta_{G_a}(\overline{h})\|_2] \leq \alpha \sum_{q \in S_{G_a}} \mu_q(G_a)$. Proposition C.1 states that the corresponding bound on *background* data is smaller by the factor $1/s_{\min}$: relative to the activation mass they carry on the target, high-selectivity features can only perturb background activations weakly. The bound is monotone in the threshold, raising $s_{\min}$ tightens it, and the intermediate form in Eq. equation 10 shows that the constant ε in the score buys an additional slack of $\alpha\varepsilon$ per selected feature.

Remark C.2 (Standardized versus original residual space). The bound is stated in standardized activation space, where the SAE operates (Section 4.2), while the intervention of Eq. equation 2 is applied to the residual stream. If the decoded contribution is mapped back to the original space through the fixed standardization, the perturbation becomes $\alpha \sum_{q \in S_{G_a}} z_q(\overline{h})\,(\sigma_H \odot d_q)$, and Proposition C.1 holds verbatim with $\|d_q\|_2$ replaced by $\|\sigma_H \odot d_q\|_2 \leq \|\sigma_H\|_\infty \|d_q\|_2$ in the definition of μ_q; in particular the same bound holds up to the fixed factor $\|\sigma_H\|_\infty$. The selectivity ratio $s_q(G_a)$, and hence the $1/s_{\min}$ gain, is unchanged by any per-feature rescaling of $\|d_q\|_2$, since the rescaling appears in both the numerator and the denominator of the score.

Remark C.3 (What the bound does and does not guarantee). Proposition C.1 is a *representation-level safety bound*, and its scope should be read precisely. First, it controls the *expected* perturbation magnitude over background activations, not the worst case: individual background tokens on which a selected feature fires strongly can still be perturbed substantially. Second, it aggregates over the selected set, so the bound grows with $|S_{G_a}|$; selecting few, highly selective features is favored on both counts. Third, and most importantly, it says nothing about model *behavior*: relating a perturbation at layer ℓ to a change in the output distribution would require additional assumptions on the downstream layers. Fourth, the score measures activation magnitude and decoder norm but not the *orientation* of d_q relative to the local geometry of the forget and background manifolds; two features with similar scores can therefore produce different behavioral effects, a variability that Section 7.3 measures directly and that a geometry-aware score could reduce at substantially higher cost (Appendix I).

Remark C.4 (Connection to the training analysis). Appendix B shows that the contrastive objective favors sparse representations with stronger within-target alignment and lower target-background overlap, thereby promoting features with higher selectivity scores $s_q(G_a)$. Proposition C.1 completes the argument: selecting features above a threshold $s_{\min}$ limits their expected perturbation on background representations relative to their activation on the target.

D UMAP : MAINFOLD OF THE DATASETS

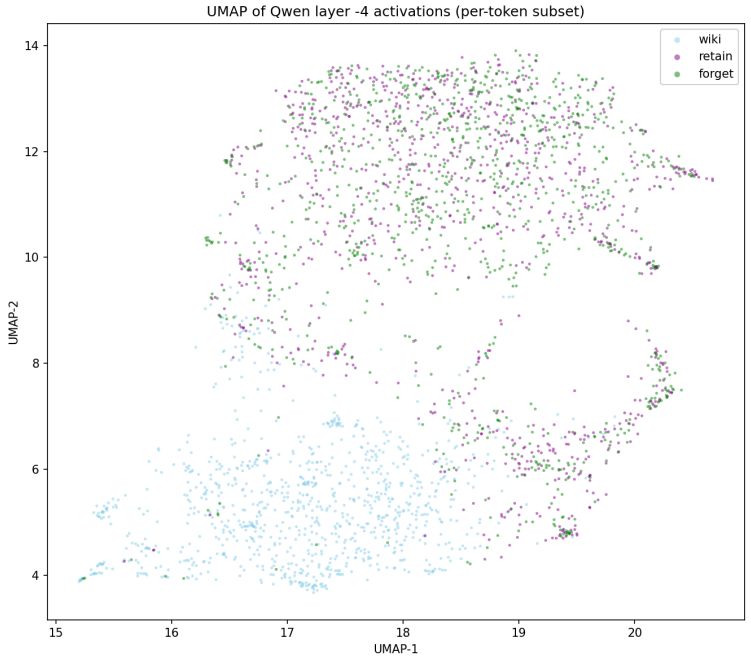


Figure 5: UMAP projection (McInnes et al., 2018) of residual-stream activations from the forget, retain, and Wiki sets. Forget and retain examples strongly overlap because they come from the same generation process and differ only through the random assignment of authors to the splits: the full forget set is not a globally separable concept in activation space.

E HYPERPARAMETER OPTIMIZATION PROTOCOL AND SEARCH SPACES

This appendix specifies the complete model-selection protocol. It has two levels. **Level 1** is the hyperparameter search: one independent Optuna study (Akiba et al., 2019) per (model $\times$ method) pair , 3 models $\times$ (5 SAE-family methods + 2 trainer methods) = 21 studies , all run with a 5 seeds (0-4). A study never samples over methods: each method is pinned to its own study and therefore receives its own full trial budget. **Level 2** freezes the winning configuration of each study and reruns the complete pipeline over five independent seeds; its essentials are summarized in Section E.5 below, and the reported results come exclusively from Level 2.

E.1 SELECTION OBJECTIVE

All 21 studies maximize the same scalar, the *retain-forget margin*

$$\text{margin} \;=\; \tfrac{1}{2}\big(P_{\text{retain}} + P_{\text{holdout}}\big) \;-\; P_{\text{forget}}, \tag{11}$$

where P_{forget}, P_{retain}, and P_{holdout} are teacher-forced ground-truth answer probabilities on OpenUnlearning's own probe sets (`forget05` and `holdout05`; probe batch size 32; parity with the OpenUnlearning data verified by a dedicated check). Model selection never uses generation-based

or test metrics, so the selection signal is decoupled from the reported evaluation. An earlier constrained objective (smallest α such that P_{forget} falls below a threshold) is kept in the codebase for diagnostics only; all reported studies use the margin of Eq. equation 11.

Operating point within a trial (SAE track). For representation-based methods, the free knob inside a trial is the intervention strength α. Each trial trains the extractor, selects features, then evaluates a coarse grid $\alpha \in \{0, 2, 4, 8, 16, 32\}$ in a single prefix-shared forward pass, followed by golden-section refinement (ratio 0.618) on the bracket around the coarse argmax, with a budget of at most 8 refinement evaluations and a relative resolution of 0.05. The trial value is the best margin over all evaluated α. This fast, margin-maximizing search is used only for model selection; it is distinct from the log-space bisection of Appendix H, which traces the full trade-off curves reported in the benchmark.

Operating point within a trial (trainer track). GradDiff and RMU have no α; the free knob is the training checkpoint. Every checkpoint (stride 1) is scored with the same margin, and the trial value is the maximum over *non-degenerate* checkpoints. A degeneracy guard rejects checkpoints whose forget probability falls below $0.1\times$ the retain-reference threshold as collapsed rather than unlearned (the threshold is computed from a retain-only finetuned reference); if all checkpoints are degenerate, the trial receives a penalty of -10^9.

E.2 Search spaces

Table 1 gives the SAE-family search space and Table 2 the trainer search space; the distributions below are exactly those recorded in the study databases.

Table 1: SAE-family search space (per trial). "Trained SAEs" means all methods except NMF; "contrastive family" means all contrastive variants.

Hyperparameter	Range / choices	Sampling	Applies to
Latent dimension Q	$\{64, 512, 1024, 2048, 4096, 8192\}$	categorical	all
Selection threshold $s_{\min}$	$[1.0,\ 100.0]$	uniform	all
Batch size	$\{64, 128, 256, 512\}$	categorical	trained SAEs
Training epochs	$[5,\ 40]$	int. uniform	trained SAEs
Encoder depth	$\{1, 2\}$	categorical	trained SAEs
Sparsity coefficient λ	$[10^{-4},\ 10^{-1}]$	log-uniform	trained SAEs
Temperature τ	$\{0.1, 0.5\}$	categorical	contrastive family
Contrastive weight β	$[10^{-4},\ 1.0]$	log-uniform	contrastive SAE
Common-block fraction	$\{0.25, 0.5, 0.75\}$	categorical	Split
β_{com}, β_{diff}	$[10^{-4},\ 1.0]$ each	log-uniform	Split
NMF max iterations	$[50,\ 500]$	int. uniform	NMF

Fixed (not searched), SAE family: learning rate 10^{-3}; extraction layer -4 (fourth from last); Wiki reconstruction weight $\lambda_{\text{wiki}} = 1.0$, so the training loss is $\text{MSE}(\text{TOFU}) + \lambda_{\text{wiki}}\,\text{MSE}(\text{Wiki}) + \lambda\,\mathcal{L}_{\text{sparse}}$ + contrastive terms; author-level contrastive grouping (Section 4.1); per-dimension z-score input normalization computed once from the TOFU token cache and frozen across trials; feature selection by the per-author norm-aware forget/background ratio of Eq. equation 1 with top-1 fallback, using at most $20{,}000$ tokens per side.

Fixed, trainers: per-device train batch size 4; RMU retain loss `EMBED_DIFF`; checkpoints saved per epoch (model weights only); OpenUnlearning's generation-based evaluation disabled during training. The epoch grid was raised from the default $\{2, 4, 6, 8, 10\}$ to $\{10, \ldots, 30\}$ because the retain utility of GradDiff and RMU is non-monotonic in training and recovers late.

E.3 Search algorithm, budget, and pruning

All studies use Optuna's TPE sampler with seed 0, direction *maximize*. SAE-family studies additionally use a median pruner (3 startup trials): the pruning signal reported before refinement is the best margin on the coarse α grid, so a pruned trial skips the golden-section refinement but its record is still written. Trainer studies report the same intermediate value but never prune, since

Table 2: Trainer search space (GradDiff and RMU). The names `gamma` and `alpha` are the trainer implementation's loss weights and are unrelated to the paper's intervention strength α and norm exponent γ.

Hyperparameter	Range / choices	Sampling	Method
Learning rate	$[10^{-6},\ 10^{-4}]$	log-uniform	both
Training epochs	$\{10, 15, 20, 25, 30\}$	categorical	both
Forget-loss weight (`gamma`)	$[0.1,\ 5.0]$	log-uniform	both
Retain-loss weight (`alpha`)	$[0.1,\ 5.0]$	log-uniform	both
Retain loss type	$\{\mathrm{NLL}, \mathrm{KL}\}$	categorical	GradDiff
Steering coefficient	$[0.5,\ 40.0]$	log-uniform	RMU

training has already happened by scoring time. The budget is 25 trials per SAE-family study and 20 per trainer study; the budget is a total target, and resubmitted jobs resume the persistent study database until it is reached, which makes walltime-limited cluster jobs safe. For cost control, the base model and the $\sim$25 GB activation cache are loaded once per study rather than once per trial. A built-in uniform-random searcher over the same spaces exists as a fallback, but all reported results use Optuna/TPE.

E.4 Study outcomes and winning configurations

Table 3 reports, for every study, the number of completed and pruned trials (pruned trials consume budget) and the best margin found, as of the reported snapshot; resumable studies below their trial target are marked by their counts.

Table 3: Per-study outcomes: completed trials / pruned trials / best margin (Eq. equation 11).

Method	Qwen2.5	Llama3.2	Gemma2
Contrastive SAE	7 / 18 / 0.528	16 / 9 / 0.532	7 / 18 / 0.518
SCALPEL-Split	17 / 8 / 0.526	11 / 14 / 0.533	9 / 15 / 0.517
Standard SAE	15 / 10 / 0.073	15 / 8 / 0.088	12 / 13 / 0.313
NMF	20 / 5 / 0.020	18 / 7 / $\approx$0	11 / 14 / 0.020
GradDiff	20 / 0 / 0.509	20 / 0 / 0.517	20 / 0 / 0.527
RMU	20 / 0 / 0.531	20 / 0 / 0.526	20 / 0 / 0.524

The selection margins already anticipate the benchmark: reconstruction-only extraction (standard SAE, NMF) cannot reach a useful margin on any model, while the contrastive variants match the strong trainer baselines at selection time.

Each study's winning configuration is copied verbatim into the experiment configuration by a single generator script; no configuration is hand-edited. As an example, the winning Qwen contrastive-SAE configuration is: latent dimension 1024, encoder depth 2, 19 epochs, batch size 256, sparsity coefficient 2.34×10^{-3}, $\tau = 0.1$, contrastive weight 2.95×10^{-2}, selection threshold 12.7.

One documented post-hoc override. For GradDiff only, the winning epoch count is replaced by a single epoch with sub-epoch checkpointing (a checkpoint every 20 steps, $\sim$40 checkpoints, degeneracy floor 0.1): at the selected learning rate GradDiff collapses the forget set inside the first epoch, so per-epoch checkpoints only ever contain already-collapsed models, and sub-epoch checkpointing is required to resolve the transition. This is the only manual deviation from the search winners, and it is applied identically across models and seeds.

E.5 From frozen configurations to the multi-seed benchmark

With hyperparameters frozen, Level 2 runs five independent end-to-end pipelines per model (finetuning seeds 0-4; unlearning and SAE seeds fixed to 0), so the reported error bars cover the full pipeline, not just the unlearning stage. For each seed: (i) the full model and the retain-only reference are finetuned from the base checkpoint (learning rate 5×10^{-5}, 5 base epochs with a balanced

$\{-1, 0, +1\}$ epoch jitter dealt round-robin and shuffled deterministically, so no jitter level is missed over five seeds; full and retain share the seed's epoch count to remain comparable); (ii) the forget side uses all $13{,}000$ augmented forget rows, and the retain side a seed-deterministic random sample of $64{,}000$ rows from the augmented retain pool, while probe and evaluation data are identical across seeds; (iii) a divergence gate verifies that the full and retain models actually separate on the forget set before the seed's branch runs; (iv) the activation cache (layer -4) and the retain-reference threshold are re-extracted from that seed's own models; (v) every method retrains with its frozen configuration and re-selects its operating point against that seed's threshold , SAE-family methods rerun the α search, trainers re-select the checkpoint; and (vi) each winner receives a full OpenUnlearning TOFU evaluation and one row in the results registry. The whole pipeline is emitted as a dependency-ordered SLURM DAG, resumable through per-stage sentinels. Aggregation follows Appendix H: per-method mean $\pm$ population standard deviation across seeds; where a cell has fewer than five completed seeds at the reported snapshot, we state the per-cell n alongside the error bars.

E.6 Compute footprint

Each study runs on a single H100 with 64-128 GB of RAM and 6 CPUs; SAE-family studies take 4-24 h of walltime (resumed until the trial target), trainer studies 12-24 h. Benchmark stages run on smaller MIG slices where they fit (SAE training and α search on a `2g.20gb` slice, raised to `3g.40gb` for Gemma-2 due to its hidden size of 2304 and 256k vocabulary). Gemma-2 uses eager attention throughout, as the SDPA soft-capping implementation produced numerical instabilities.

F Benchmarks for Qwen, llama and Gemma

F.1 Model Utility vs forget prob

F.2 Trade-offs Across Multiple Unlearning Metrics

G Feature Selectivity Across Extractors

Figure 12 complements the representation-level geometry of Figure 4a. The contrastive SAE produces more selective top-ranked features than the reconstruction-only SAE, while SCALPEL-Split yields the strongest selectivity. This supports the interpretation that contrastive learning does not merely reorganize the latent geometry globally, but translates this separation into individual features that are more specific to the forget targets.

H Adaptive Intervention-Strength Search

Most methods expose a scalar parameter that controls the strength of unlearning; we denote it generically by α, although its precise meaning depends on the method (intervention strength for representation-based methods, update scale for parameter-based ones). A small value leaves the model close to θ_{full}, while a very large value produces near-complete forgetting together with substantial degradation. A uniform grid over α is inefficient: the practically relevant transition often occupies a narrow interval, and α can span several orders of magnitude. We therefore trace each trade-off curve with an adaptive bisection in $\log \alpha$.

Procedure.

1. **Endpoints.** Evaluate two endpoints α_{low} and α_{high}, chosen to bracket the transition from the nearly unchanged full model to a strongly degraded model. Each evaluation runs the OpenUnlearning pipeline and returns the pair $\big(P_f(\alpha), P_r(\alpha)\big)$ of forget and retain ground-truth probabilities. All evaluated points are cached, so an interrupted search resumes without recomputation.
2. **Interval selection.** At each iteration, sort the evaluated points by α and consider every adjacent pair (α_i, α_{i+1}). Prioritize intervals whose forget-probability range overlaps the focus region $\mathcal{B}_f = [0.05,\ 0.70]$, where the practically relevant part of the trade-off lies;

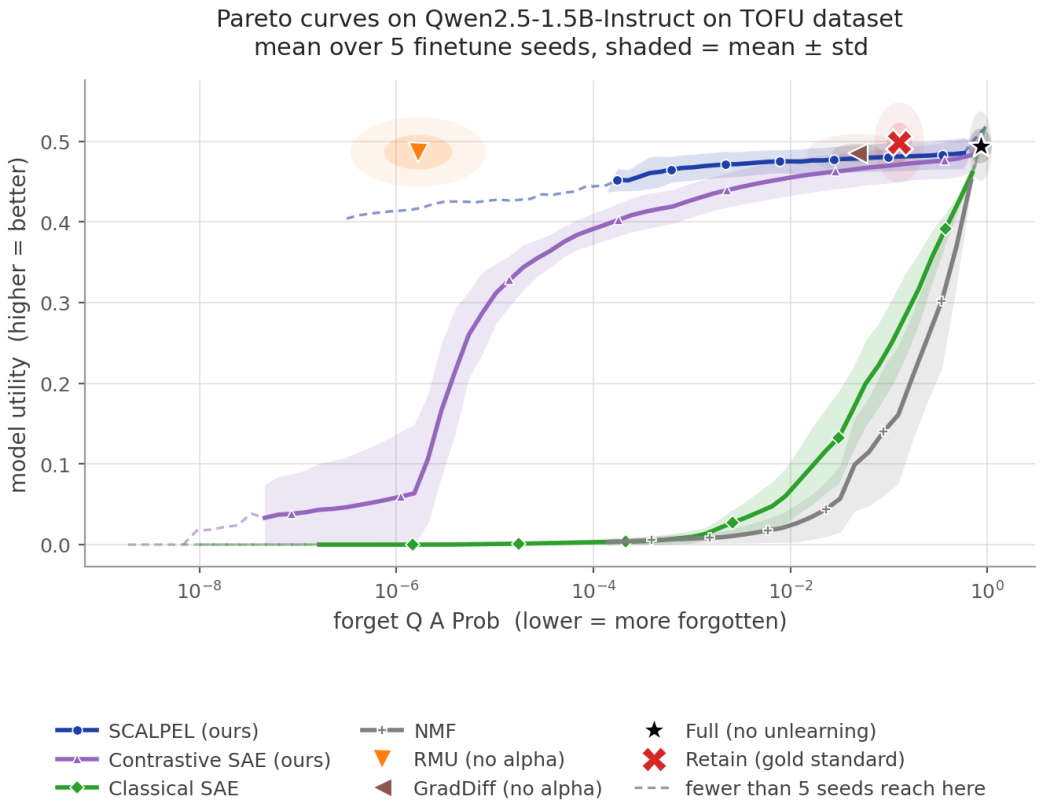


Figure 6: Unlearning trade-off on TOFU forget10 for `Qwen2.5-1.5B-Instruct`, after method-specific hyperparameter optimization and aggregation over five independent end-to-end seeds. Each curve varies the intervention strength: lower forget probability indicates stronger forgetting, while higher model utility indicates better preservation. Results for Llama and Gemma are reported in Appendix F.1.

among intervals with the same overlap, select the one with the largest gap in forget probability.

3. **Refinement.** Evaluate the geometric midpoint $\alpha_{\text{new}} = \sqrt{\alpha_i\,\alpha_{i+1}}$, i.e. the midpoint of the interval in $\log \alpha$ space.
4. **Stopping.** Repeat until the prescribed budget of evaluation points is reached or no interval contains a meaningful unresolved gap.

The procedure keeps the two endpoint evaluations and concentrates the remaining budget in the steep region of the curve, producing a dense estimate of the Pareto front without a large uniform sweep.

Aggregation across seeds. After evaluation, the points of each seed are sorted by α. To aggregate the five seeds, each seed-specific curve is interpolated onto a common forget-probability grid over the seeds' shared support; the corresponding retain values are then averaged across seeds, and error bars are computed from the five independent runs. This yields comparable mean curves and uncertainty estimates even though the adaptive search evaluates different α values in different seeds.

I Extended Limitations and Future Work

This appendix expands the limitations summarized in Section 8.

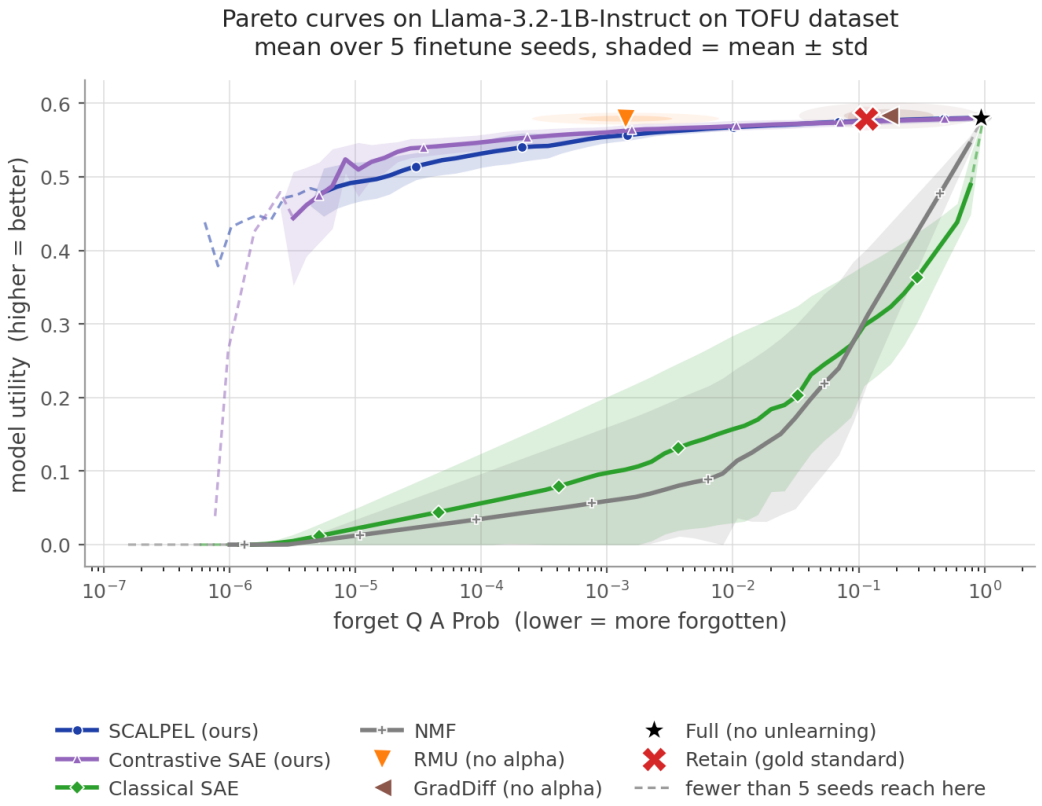


Figure 7: Unlearning trade-off on TOFU forget10 for `Llama-3.2-1B-Instruct`, after method-specific hyperparameter optimization and aggregation over five independent end-to-end seeds. Each curve varies the intervention strength: lower forget probability indicates stronger forgetting, while higher model utility indicates better preservation.

Scope of the evaluation. Our experiments cover TOFU and three relatively small models. Although Qwen, Llama, and Gemma represent distinct families, further work is needed to establish whether the same behavior holds for larger models and less controlled datasets. We have not yet completed the evaluation on MUSE, which would require concept groups at the document or chunk level rather than at the author level, and on WMDP-Bio. Our theory also predicts where the advantage should shrink: the energy-bias argument (Appendix A) implies that SCALPEL helps most when target concepts carry low energy in representation space, so the gap with existing methods may narrow on datasets where targets already contribute strongly to the activations.

Concept granularity. The author-level definition of a concept fits the structure of TOFU and the right-to-be-forgotten setting, where each person is naturally a separate target. Other applications may need a different granularity, a document, a fact, or a single sample. The anchor-based formulation supports these settings in principle through suitable data augmentation (each anchor's views providing the positives), but their stability and computational cost remain to be studied. Relatedly, we have not yet measured how the method behaves as the fraction of data to unlearn grows; because the intervention acts locally on selected directions, we expect robustness to this parameter, but it must be verified.

Computational cost and design choices. The method depends on several design choices: the intervention layer, the SAE architecture and latent dimension, the contrastive weight and temperature, the augmentation procedure, the selection threshold, and the strength α. We reduce this dependence through per-method hyperparameter optimization and multi-seed evaluation, but the full protocol is expensive: it requires finetuning full and retain-only references, caching token-level activations,

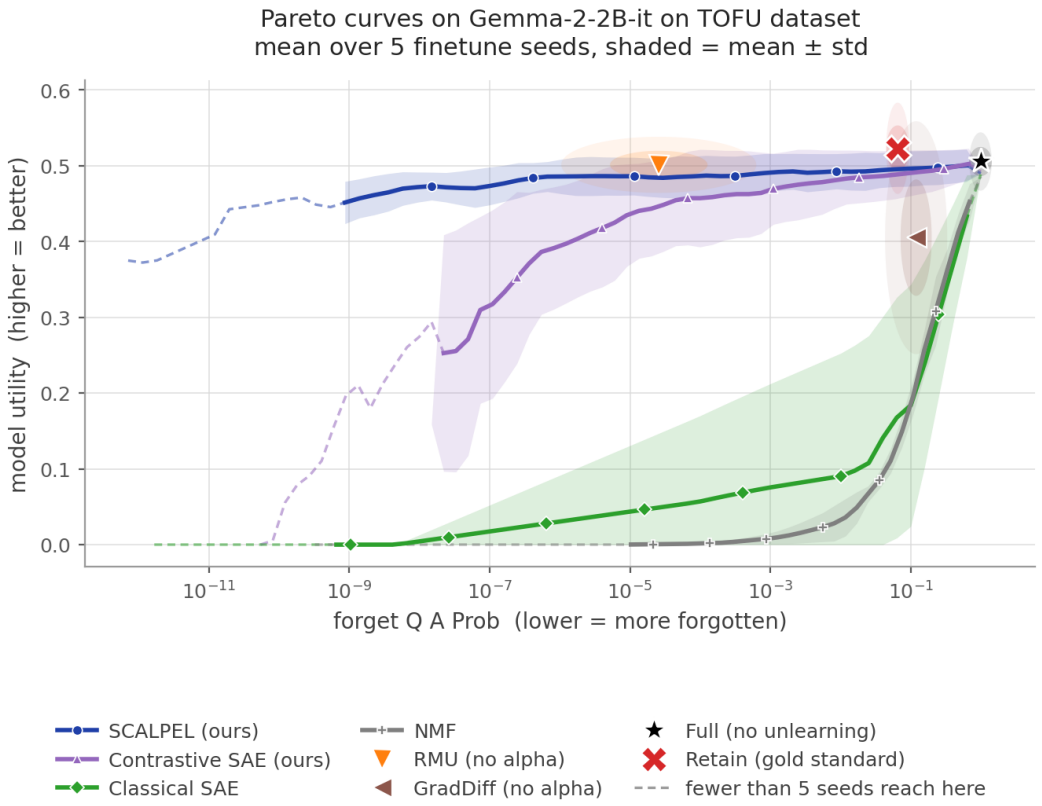


Figure 8: Unlearning trade-off on TOFU forget10 for `Gemma-2-2B-it`, after method-specific hyperparameter optimization and aggregation over five independent end-to-end seeds. Each curve varies the intervention strength: lower forget probability indicates stronger forgetting, while higher model utility indicates better preservation.

training several sparse dictionaries, and evaluating many intervention strengths. The quality of the augmented data is itself a limitation, we rely on a fixed set of paraphrasing and transformation procedures, and richer augmentation could improve the stability of the extracted features and their coverage of the target concept.

Scope of the theory. The theoretical results are deliberately local and rely on stated idealizations. The energy-bias result assumes centered, uncorrelated concept components and orthogonal per-concept reconstruction errors (Appendix A); the contrastive-selectivity analysis assumes non-negative codes, controlled decoder scales, and that sparse target-background separation is compatible with comparable reconstruction and sparsity cost (Appendix B); and the perturbation bound controls only the expected representation-level footprint of the intervention, not downstream behavior (Appendix C). These assumptions make the mechanisms analyzable but do not fully capture the geometry of language-model representations, and none of the results guarantees complete forgetting or behavioral preservation.

Score geometry. The score $s_q(G_a)$ uses activation magnitudes and decoder norms but ignores the orientation of d_q relative to the local forget and background manifolds, which likely explains why features with similar scores can produce different unlearning effects (Section 7.3). A geometry-aware score could compare each decoder direction with local principal directions of the two manifolds, and a geometry-aware intervention could move the activation along, or back toward, the retain manifold instead of applying a straight subtraction, potentially approaching the behavior of a model trained only on retained data. Both require repeated local manifold estimation (e.g. local

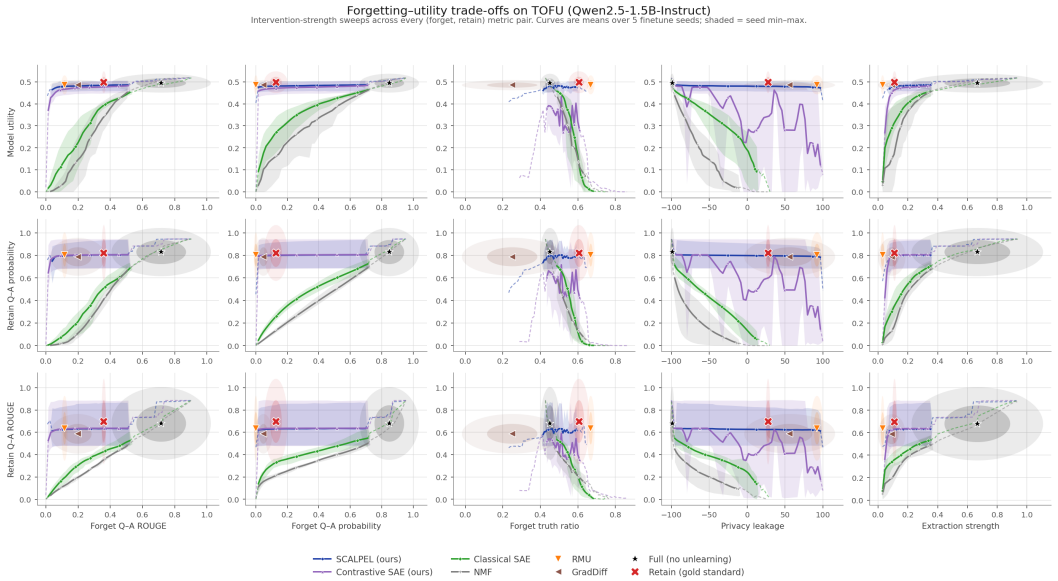


Figure 9: Multi-metric forgetting-utility trade-offs on TOFU forget10 for `Qwen2.5-1.5B-Instruct`, after method-specific hyperparameter optimization. Curves show the mean over five finetuning seeds, and shaded regions show the seed-wise minimum-maximum range. Columns correspond to different forget-side metrics, while rows report preservation through model utility, retain Q-A probability, and retain Q-A ROUGE, for which higher values are better. Lower forget Q-A ROUGE and forget Q-A probability indicate stronger forgetting. Forget truth ratio and extraction strength are interpreted relative to the retain-only reference, while privacy leakage is best when close to zero. The black star and red cross denote the full and retain-only reference models, respectively; RMU and GradDiff are shown at their selected operating points.

PCA), making selection and intervention substantially more expensive; we view them as promising extensions rather than free improvements.

Shared features across targets. Several forget authors select the same latent features. This sharing is useful, it means forgetting all ten authors needs far fewer than fifty directions, reducing memory and inference cost, but it may also couple targets: removing a shared feature for one author could affect another author represented by the same coordinate. Further analysis is needed to determine whether these features capture genuinely shared forget structure or introduce unwanted interference.

Inference overhead and sequential unlearning. SCALPEL adds a small inference-time overhead: the SAE encoder and the selected decoder directions must remain available so the intervention can be applied at the chosen layer. Sequential unlearning is an open problem: after unlearning one group, removing a second currently requires either retraining SCALPEL on the combined forget set or stacking a second intervention, which increases inference cost with every new request. Making the method scale to repeated unlearning requests is an important direction.

Context-aware extraction. The encoder processes token representations individually. These already carry contextual information from earlier transformer layers, but the extraction mechanism does not explicitly model the full sample context; a context-aware encoder could better isolate associations that only become identifiable at the sentence or document level.

Behavioral versus parametric removal. Finally, two limitations concern what "unlearning" means here. First, existing metrics mainly measure output behavior and cannot reliably determine whether the target information is still stored internally or recoverable through alternative prompts or attacks; better evaluations of true removal versus hiding are needed. Second, SCALPEL modifies activations at inference and does not erase the parametric trace: the original information may persist at other layers or become accessible if the intervention is disabled, bypassed, or challenged. Our

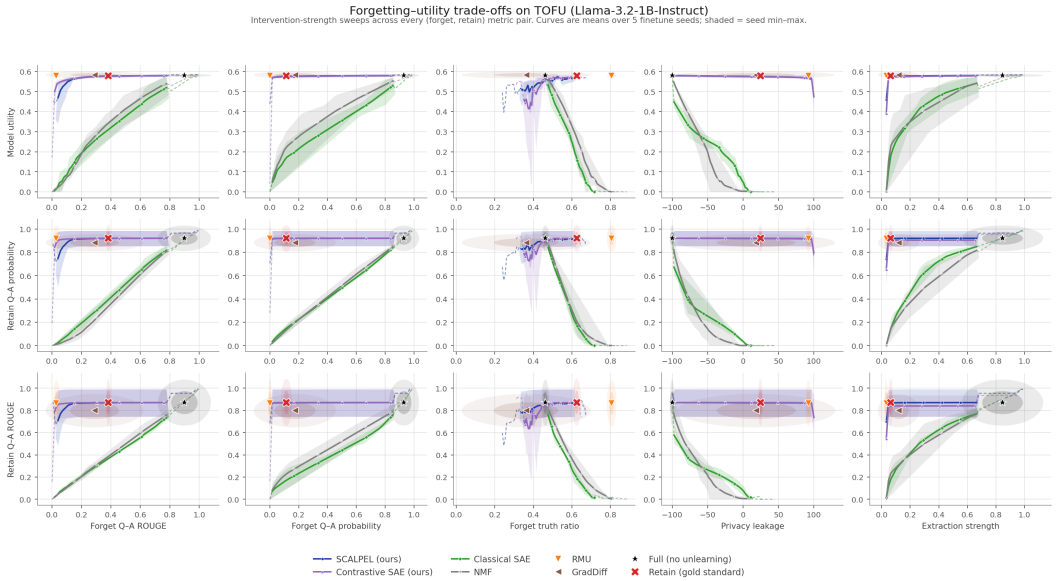


Figure 10: Multi-metric forgetting-utility trade-offs on TOFU forget10 for `Llama-3.2-1B-Instruct`, after method-specific hyperparameter optimization. Curves show the mean over five finetuning seeds, and shaded regions show the seed-wise minimum-maximum range. Columns correspond to different forget-side metrics, while rows report preservation through model utility, retain Q-A probability, and retain Q-A ROUGE, for which higher values are better. Lower forget Q-A ROUGE and forget Q-A probability indicate stronger forgetting. Forget truth ratio and extraction strength are interpreted relative to the retain-only reference, while privacy leakage is best when close to zero. The black star and red cross denote the full and retain-only reference models, respectively; RMU and GradDiff are shown at their selected operating points.

results therefore establish selective behavioral and representation-level unlearning *under the proposed intervention*, not certified deletion from the model. A natural next step is to use the extracted features to guide permanent parameter updates while preserving the selectivity of the representation-level approach; we view the present work as a first step toward bridging mechanistic interpretability and machine unlearning.

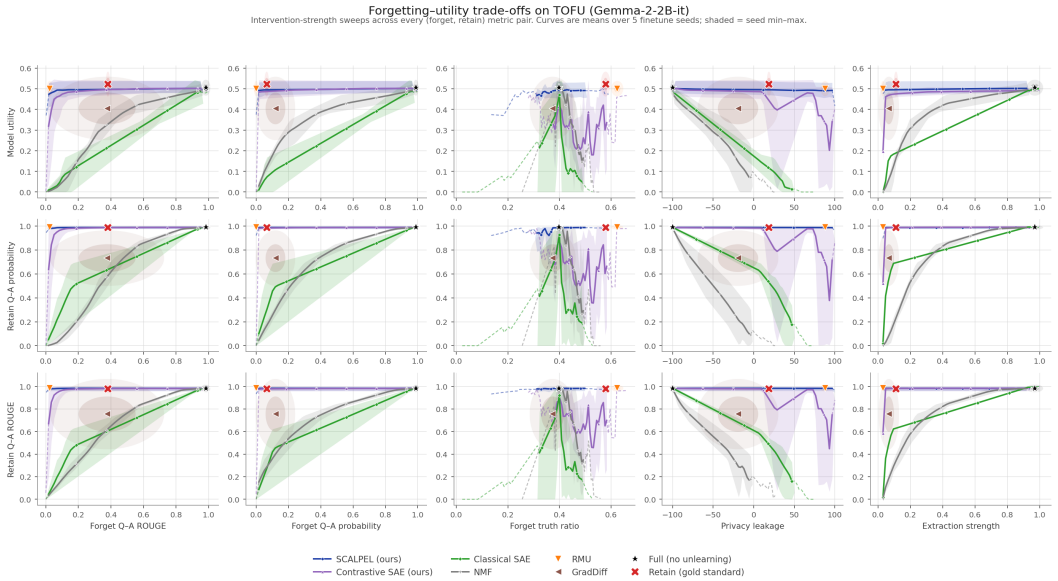


Figure 11: Multi-metric forgetting-utility trade-offs on TOFU forget10 for `Gemma-2-2B-it`, after method-specific hyperparameter optimization. Curves show the mean over five finetuning seeds, and shaded regions show the seed-wise minimum-maximum range. Columns correspond to different forget-side metrics, while rows report preservation through model utility, retain Q-A probability, and retain Q-A ROUGE, for which higher values are better. Lower forget Q-A ROUGE and forget Q-A probability indicate stronger forgetting. Forget truth ratio and extraction strength are interpreted relative to the retain-only reference, while privacy leakage is best when close to zero. The black star and red cross denote the full and retain-only reference models, respectively; RMU and GradDiff are shown at their selected operating points.

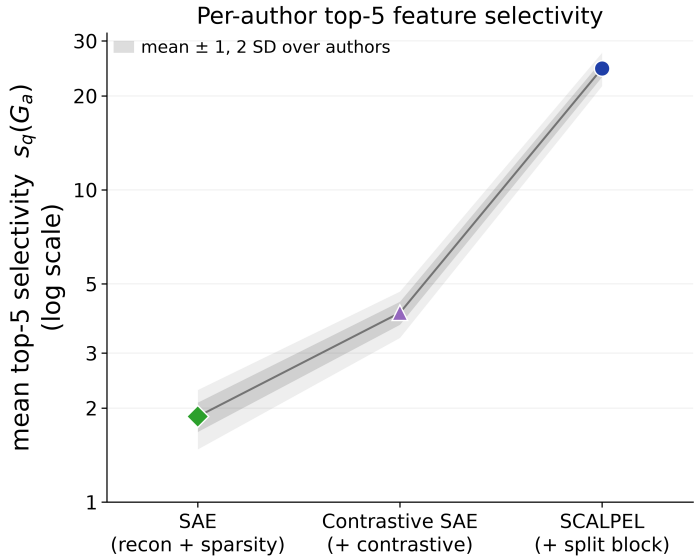


Figure 12: Per-author top-5 feature selectivity for the reconstruction-only SAE, contrastive SAE, and SCALPEL-Split, evaluated across the five Qwen benchmark seeds. Points correspond to the mean top-5 selectivity score $s_q(G_a)$ for each author, while the shaded region summarizes variation across authors. Adding contrastive learning increases feature selectivity over the standard SAE, and the split architecture further amplifies this effect.