THE MODEL ALREADY HAS A DICTIONARY: CUMULATIVE SVD FEATURES AS A TRAINING-FREE ALTERNATIVE TO SPARSE AUTOENCODERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse autoencoders (SAEs) have become the prevailing paradigm for learning interpretable feature dictionaries from neural activations, but they require large activation corpora, introduce reconstruction errors, and vary across training seeds; consequently, public SAE releases remain largely confined to a few model families and activation sites. We ask how much of this feature dictionary can instead be recovered directly from the model's intrinsic weights? For each Transformer residual stream state, we construct an inherently overcomplete, model-intrinsic dictionary in a training-free manner by unifying two principles: accumulating historical module outputs across all upstream layers, and applying singular value decomposition (SVD) to attention head OV circuits $\mathbf{W}_{OV}$ and MLP down projections $\mathbf{W}_{\text{down}}$ to extract compact, intra-module orthogonal read and write channels. The resulting dictionary reconstructs the accumulative residual activations exactly without training corpora or optimization. Across Gemma 3 (1B-27B) and Qwen3-8B, SAE features remain slightly easier to describe from activating contexts, while SVD attains full coarse topic coverage and captures complementary fine-grained semantics. Moreover, cumulative SVD features excel at sparse concept prediction and enable more selective attribute disentanglement alongside stronger factual relation steering under multi-feature composition. Extraction uses no corpus tokens and is about seven orders of magnitude more efficient than training a matched SAE pair. These findings position weight-derived feature dictionaries as a strong training-free baseline and a complementary source of concepts for SAE-based interpretability.

1 INTRODUCTION

A central quest in mechanistic interpretability is to isolate a feature dictionary that decomposes entangled neural activations into distinct computational primitives. Under the linear representation hypothesis, semantic concepts and computational variables are posited to occupy linear directions in activation space (Elhage et al., 2022; Park et al., 2024; Garg et al., 2026). To uncover these directions from polysemantic activations, sparse autoencoders (SAEs) have become the prevailing paradigm (Bricken et al., 2023; Templeton et al., 2024; Huben et al., 2024). Yet this approach rests on an implicit premise that discovering a feature dictionary requires auxiliary training over massive activation corpora. While learned features are often easy to describe from activating contexts, auxiliary training introduces severe representational, optimization, and computational costs.

Prioritizing sparsity compromises representational fidelity. The resulting approximation error can undermine causal control, rendering interventions prone to unintended side effects (Makelov et al., 2025; Karvonen et al., 2025). Moreover, optimization instability driven by stochastic initialization, feature splitting, and absorption causes separate training runs to discover divergent dictionaries even at identical loss, obscuring whether features reflect true mechanisms or training artifacts (Chanin et al., 2025; Paulo & Belrose, 2026). Worse still, auxiliary dictionary learning demands immense compute, with open suites like Gemma Scope consuming over 20% of GPT-3's pretraining compute (Lieberum et al., 2024; Brown et al., 2020). These raise a fundamental question: *how much of such a feature dictionary can be read from the model's weights, without learning an external dictionary?*

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.

We show in this paper that the answer is positive. Our starting point is that the residual stream is additive. Within a Transformer, all attention head OV circuits $\mathbf{W}_{OV}$ (Elhage et al., 2021) and MLP down projections $\mathbf{W}_{\text{down}}$ serve as the linear operators that write directly into the residual stream.

Singular Value Decomposition (SVD) yields an exact rank-one decomposition for any such operator,

$$\mathbf{W} = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top, \qquad \mathbf{W}\mathbf{x} = \sum_{i=1}^{r} \underbrace{(\sigma_i \mathbf{v}_i^\top \mathbf{x})}_{\text{read / activation}} \underbrace{\mathbf{u}_i}_{\text{write / feature}} . \tag{1}$$

The right singular vector $\mathbf{v}_i$ reads an input pattern with scalar gain σ_i, while the left singular vector $\mathbf{u}_i$ defines the feature write direction in the output space.

For each residual state, we define a cumulative dictionary unifying two core operations: (i) **cross-depth module accumulation** keeps the write of every upstream head and MLP as a separate source rather than collapsing them into the summed residual and (ii) **weight SVD** decomposes each of these sources into intra-module orthogonal singular vectors, achieving exact linear reconstruction while offering a more compact basis than neurons. Overcompleteness arises from stacking sources across depth rather than within any single module, which provides the capacity to partially disentangle features in superposition without external optimization (Elhage et al., 2022).

In fact, our construction builds on an established evidence (Xue & Andrzejak, 2026) that the singular vectors of Transformer weights are interpretable, where late MLP singular subspaces map onto coherent output semantics such as sentiment through Logit Lens (nostalgebraist, 2020). Yet, vocabulary-level readout of isolated components leaves the foundational premise of dictionary learning untested: can singular vectors genuinely disentangle fine-grained semantic primitives, sparsely decode abstract latent concepts, and causally steer intermediate computational mechanisms?

We go a step further to demonstrate that they do in ways that complement SAEs. We evaluate intrinsic SVD features across three pillars: *Descriptively*, although SAE features remain slightly easier to describe from activating contexts, SVD achieves full coarse topic coverage while capturing complementary fine-grained semantics. *Predictively*, cumulative SVD features excel at sparse concept prediction, both with individual features and feature compositions. *Causally*, interventions show that the cumulative residual SVD dictionary supports more selective attribute disentanglement and stronger factual relation steering under multi-feature composition. Together, these findings position intrinsic SVD features as a training-free alternative to SAEs and a complementary source of concepts.

We summarize our main contributions as follows: (1) **Model-intrinsic dictionary.** We formulate residual-stream dictionaries grounded in two complementary principles: *cross-depth module accumulation*, which gathers upstream attention and MLP outputs to naturally yield overcompleteness, and *weight SVD*, which provides exact linear reconstruction with intra-module orthogonality, yielding a more compact basis than neurons. (2) **Descriptive complementarity.** SVD features achieve full coarse topic coverage, while capturing fine-grained semantic concepts complementary to SAEs. (3) **Predictive utility.** Cumulative SVD features excel at sparse concept prediction, both with individual features and feature compositions. (4) **Causal control.** Interventions on intermediate representations show that the cumulative residual SVD dictionary supports more selective attribute disentanglement alongside stronger factual relation steering under multi-feature composition. (5) **Zero training.** Requiring neither activation corpora nor auxiliary training, weight SVD provides a deterministic, reproducible, and compute-efficient feature dictionary foundation.

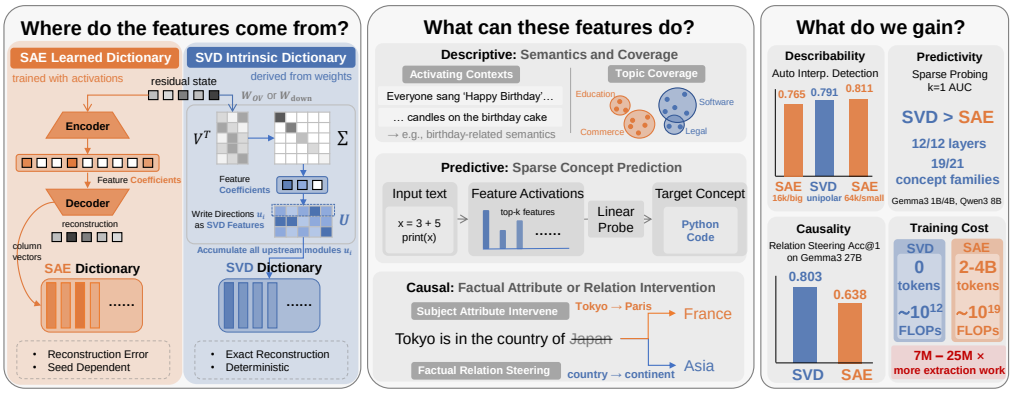


Figure 1: Training-free SVD: construction, semantics, prediction, and causal control versus SAEs.

2 From Learned to Model-Intrinsic Dictionaries via Weight SVD

2.1 Learned SAE Feature Dictionaries

Let d_{model} denote the residual stream width. Under the linear representation hypothesis (Elhage et al., 2022), a hidden state $\mathbf{h} \in \mathbb{R}^{d_{\text{model}}}$ can be approximated by a sparse combination of feature directions. A sparse autoencoder (SAE) learns a dictionary of m features $\mathbf{F}^{\text{SAE}} \in \mathbb{R}^{d_{\text{model}} \times m}$ and sparse coefficients $\mathbf{a}^{\text{SAE}} \in \mathbb{R}^m$ (Bricken et al., 2023) as

$$\mathbf{a}^{\text{SAE}} = \phi^{\text{SAE}}(\mathbf{W}_{\text{enc}}\mathbf{h} + \mathbf{b}_{\text{enc}}), \qquad \hat{\mathbf{h}} = \mathbf{F}^{\text{SAE}}\mathbf{a}^{\text{SAE}} + \mathbf{b}_{\text{dec}}, \tag{2}$$

where the decoder $\mathbf{F}^{\text{SAE}}$ holds the learned feature directions, and ϕ^{SAE} typically implements TopK (Bussmann et al., 2024) or JumpReLU (Rajamanoharan et al., 2024) to induce sparsity.

2.2 Model-Intrinsic SVD Feature Dictionaries

Instead of training external models to discover features post hoc, we extract feature dictionaries directly from Transformer linear weight operators. For any operator $\mathbf{W} \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ of rank r, singular value decomposition decomposes it as

$$\mathbf{W} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top, \tag{3}$$

where orthonormal matrices $\mathbf{U} \in \mathbb{R}^{d_{\text{out}} \times r}$ and $\mathbf{V} \in \mathbb{R}^{d_{\text{in}} \times r}$ contain the left and right singular vectors, and $\mathbf{\Sigma} = \text{diag}(\sigma_1, \ldots, \sigma_r)$ holds the singular values. For a local input $\mathbf{x} \in \mathbb{R}^{d_{\text{in}}}$ to $\mathbf{W}$, this yields r rank one channels whose feature directions, activations, and realized contributions are

$$\mathbf{f}_i^{\text{SVD}} = \mathbf{u}_i \in \mathbb{R}^{d_{\text{out}}}, \qquad a_i^{\text{SVD}}(\mathbf{x}) = \sigma_i \mathbf{v}_i^\top \mathbf{x} \in \mathbb{R}, \qquad \mathbf{c}_i(\mathbf{x}) = a_i^{\text{SVD}}(\mathbf{x})\mathbf{f}_i^{\text{SVD}}. \tag{4}$$

Here $\mathbf{v}_i$ reads from the input, σ_i scales the response, and $\mathbf{u}_i$ writes to the output, exactly recovering the linear map as $\sum_{i=1}^{r} \mathbf{c}_i(\mathbf{x}) = \mathbf{W}\mathbf{x}$ without reconstruction error.

Attention heads. For each head $p = 1, \ldots, H$, we apply SVD to its OV circuit $\mathbf{W}_{OV}^{(p)} = \mathbf{W}_O^{(p)}\mathbf{W}_V^{(p)}$, the head's full residual writing linear operator (Elhage et al., 2021; Olsson et al., 2022). Its left singular vectors $\mathbf{u}_i^{(p)}$ thus define feature directions in residual activation space. Given contextual head input $\mathbf{x}_{\text{attn}}^{(p)} \in \mathbb{R}^{d_{\text{model}}}$ aggregated by head p, channel i has activation $\sigma_i^{(p)}\mathbf{v}_i^{(p)\top}\mathbf{x}_{\text{attn}}^{(p)}$.

MLPs. For a gated MLP with gate function ϕ, intermediate width d_{ff}, and state $\mathbf{z}(\mathbf{x}) = \phi(\mathbf{W}_{\text{gate}}\mathbf{x}) \odot (\mathbf{W}_{\text{up}}\mathbf{x}) \in \mathbb{R}^{d_{\text{ff}}}$, we apply SVD to the down projection $\mathbf{W}_{\text{down}} \in \mathbb{R}^{d_{\text{model}} \times d_{\text{ff}}}$, the linear operator that directly writes into the residual stream (Xue & Andrzejak, 2026). Given $\mathbf{z}(\mathbf{x})$, channel i has activation $\sigma_i \mathbf{v}_i^\top \mathbf{z}(\mathbf{x})$ along $\mathbf{u}_i$. Thus across both attention and MLP blocks, these left singular vectors directly define feature write directions in the residual stream.

Residual stream states. Before layer ℓ, the residual state $\mathbf{h}_\ell(\mathbf{x}) \in \mathbb{R}^{d_{\text{model}}}$ accumulates the initial token embedding $\mathbf{e}(\mathbf{x}) \in \mathbb{R}^{d_{\text{model}}}$ and the outputs written by all preceding modules,

$$\mathbf{h}_\ell(\mathbf{x}) = \mathbf{e}(\mathbf{x}) + \sum_{j=0}^{\ell-1} \left[\sum_{p=1}^{H} \mathbf{W}_{OV}^{(j,p)} \mathbf{x}_{\text{attn}}^{(j,p)} + \mathbf{W}_{\text{down}}^{(j)} \mathbf{z}^{(j)} \right]. \tag{5}$$

Collecting the singular write vectors across all upstream attention heads and MLP modules of layer ℓ forms an inherently overcomplete cumulative output dictionary,

$$\mathbf{F}_\ell^{\text{SVD}} = \left[\mathbf{U}_{OV}^{(0,1)}, \ldots, \mathbf{U}_{OV}^{(\ell-1,H)}, \mathbf{U}_{\text{down}}^{(0)}, \ldots, \mathbf{U}_{\text{down}}^{(\ell-1)} \right], \tag{6}$$

with corresponding input dependent feature activations $\mathbf{a}_\ell^{\text{SVD}}(\mathbf{x})$ gathering the channel responses $\mathbf{\Sigma}_{OV}^{(j,p)}\mathbf{V}_{OV}^{(j,p)\top}\mathbf{x}_{\text{attn}}^{(j,p)}$ and $\mathbf{\Sigma}_{\text{down}}^{(j)}\mathbf{V}_{\text{down}}^{(j)\top}\mathbf{z}^{(j)}$ across all preceding modules. The residual state then decomposes exactly as

$$\mathbf{h}_\ell(\mathbf{x}) = \mathbf{e}(\mathbf{x}) + \mathbf{F}_\ell^{\text{SVD}}\mathbf{a}_\ell^{\text{SVD}}(\mathbf{x}). \tag{7}$$

3 SVD FEATURES SPAN DESCRIBABLE AND COMPLEMENTARY SEMANTICS

While SVD provides exact reconstruction, it does not guarantee that individual features are interpretable. In this section, we systematically evaluate the descriptive semantics of SVD features across three progressive levels: individual feature describability (§3.1), semantic coverage and density (§3.2), and concept complementarity with learned SAEs (§3.3).

3.1 INDIVIDUAL SVD FEATURES CARRY DESCRIBABLE ACTIVATING SEMANTICS

We first test whether individual SVD features correspond to describable semantics. Following SAEBench AutoInterp Detection (Karvonen et al., 2025), we evaluate whether descriptions generated from activating contexts reliably predict feature activation on held-out text. We evaluate attention and MLP outputs at four layers in each of Gemma 3 1B, 4B, and 27B, following the main layer selection in Gemma Scope 2 (McDougall et al., 2025) across approximately 25%, 50%, 65%, and 85% of model depth, and at two layers of 50% and 65% relative depth in Qwen3-8B-Base, yielding 28 model $\times$ layer $\times$ module evaluation sites. Each singular channel yields two signed directions, which we treat as independent features. We define SVD feature firing using an activation threshold calibrated to match the median firing frequency of active SAE latents at the same site. We report *raw SVD* as the average Detection accuracy across both polarities, and define *unipolar SVD* as an oracle upper bound that selects the more describable polarity of each channel (Appendix C.1).

Detection accuracy. We observe clear singular polarity asymmetry across all 28 sites: among 27,859 channels evaluated, 59.4% exhibit an accuracy gap exceeding 0.10 between their two polarities (Appendix Figure 8), showing that individual channels typically provide one clearly describable direction. Across all 28 sites pooling Gemma and Qwen, *raw SVD* achieves an average Detection accuracy of 0.708. Under the *unipolar SVD* oracle upper bound, accuracy reaches 0.791, surpassing compact 16k/big SAEs at 0.765 and approaching top-performing 65k/small SAEs at 0.811 (Figure 2a; Appendix C.4). While trained SAEs maintain an advantage in deeper MLPs, SVD channels achieve comparable describability in attention heads and earlier layers (Appendix C.3, Figure 9). Thus, even without dictionary training, the activating semantics of intrinsic SVD channels carry meaningful interpretability, with individual channels often providing a comparatively interpretable direction.

3.2 SVD DICTIONARIES PROVIDE BROAD AND DENSE SEMANTIC COVERAGE

Beyond individual feature describability, we evaluate dictionary coverage across diverse semantic domains. Using a semantic topic map of 73 topics clustered from the Pile (Gao et al., 2020) independently of dictionary activations (Appendix D.1), we benchmark cumulative SVD and SAE dictionaries across four residual layers in Gemma 3 1B, 4B, and Qwen3 8B, following the same Gemma Scope 2 main layer selection as in §3.1, yielding 12 evaluation layers. We compare SVD against the widest publicly available residual SAEs: 262k features for Gemma and 64k for Qwen, which afford SAEs an expansive candidate pool to detect and cover concepts (Appendix Table 4).

Semantic breadth. We first evaluate *semantic breadth*, where a topic is covered if at least one feature concentrates its activating contexts within that topic significantly above a matched permutation null (Appendix D.1). Across all 12 layers, both SVD and SAE dictionaries cover all 73 topics, confirming that SVD features span the full range of coarse domains rather than clustering in narrow areas.

Coverage density. To evaluate how efficiently a compact dictionary covers semantic space, we measure *coverage density* using the coverage score G under a fixed feature budget B (Appendix D.1). Grounded in activating contexts, G computes the average similarity between text samples broadly drawn across the semantic space and the nearest activating semantic prototypes of dictionary features. Higher G indicates denser coverage with fewer semantic blind spots. At low feature budgets, compact SVD dictionaries achieve substantially higher coverage density than equal-sized SAE dictionaries (Figure 2b; Appendix D.2, Table 6). At $B = 16$, SVD improves coverage by $\Delta G = +0.0718$ over SAEs, a $+46.0\%$ relative gain, outperforming SAEs across all low feature budgets. Beyond pure dictionaries, SVD features also exhibit *conditional semantic novelty* by providing non-redundant reach that complements learned SAEs. When augmenting a fixed SAE core under an identical budget, substituting SVD features improves coverage over adding further SAE features by $\Delta G = +0.0253$, an 11.4% relative improvement across feature budgets $B \in \{16, 32\}$ and extension ratios $q \in \{0.05, 0.10, 0.20\}$ (Figure 2c; Appendix Table 7). At $B = 64$, these gains moderate as both

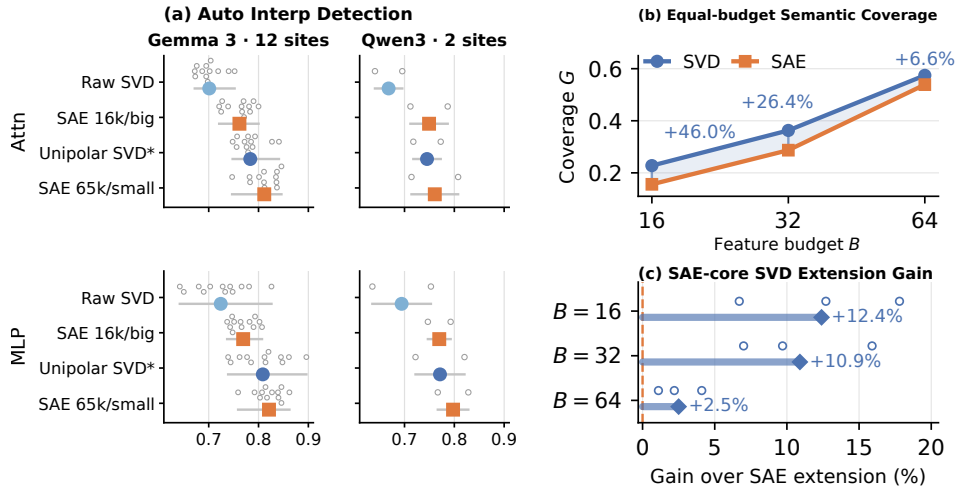


Figure 2: **Describability and semantic coverage.** (a) AutoInterp Detection (hollow dots: sites; filled: means; **unipolar SVD*: oracle upper bound selecting one polarity per channel). (b) Pure-dictionary coverage G. (c) Extension gain ΔG over a fixed SAE core (hollow dots: ratio $q \in \{0.05, 0.10, 0.20\}$).

representations approach broader coverage. Together, these results establish intrinsic SVD features as a compact, complementary basis that efficiently spans semantic space and enriches learned SAEs.

3.3 SVD and SAE Capture Complementary Semantic Concepts

Beyond aggregate coverage gains, we next examine which specific concepts drive the complementarity between SVD and SAE across two granularities: (1) at the *topic level*, we evaluate relative preferences across the 73-topic semantic map (topic naming in Appendix D.1) to test whether SVD and SAE consistently favor different domains across layers and models; and (2) at the *feature level*, we identify specific SVD concepts that nearby SAE features fail to semantically substitute for on held-out text.

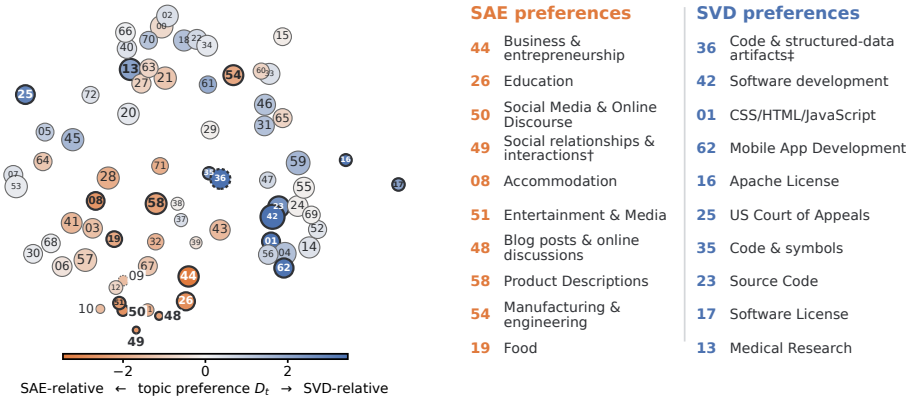


Figure 3: **SVD–SAE preferences across 73 topics.** Position: semantic proximity; area: frequency; color: relative allocation D_t (blue: SVD, orange: SAE); outlines/key: top-10 preferences in each direction (†: small leaf node; ‡: descriptive reference label; Appendix Table 8).

Topic preferences. We evaluate relative feature allocation internal to each dictionary rather than absolute coverage (Figure 3; calculation and complete map in Appendix D.3). The two dictionaries specialize in clearly different domains. SVD features concentrate heavily on formal, structured domains, allocating substantially more capacity to software engineering, source code, legal proceedings, and open-source licenses. In contrast, SAEs prioritize social interaction, online discourse, education, and commerce. Crucially, these specializations are highly consistent across architectures and depths. The median Spearman rank correlation across all 66 within- and cross-model layer pairs between Gemma and Qwen is 0.805 (Appendix D.3, Table 9). Furthermore, the 15 topics with the strongest

preferences consistently favor the same dictionary across all 12 evaluation layers without exception, demonstrating that SVD and SAE develop fundamentally distinct, invariant conceptual emphases.

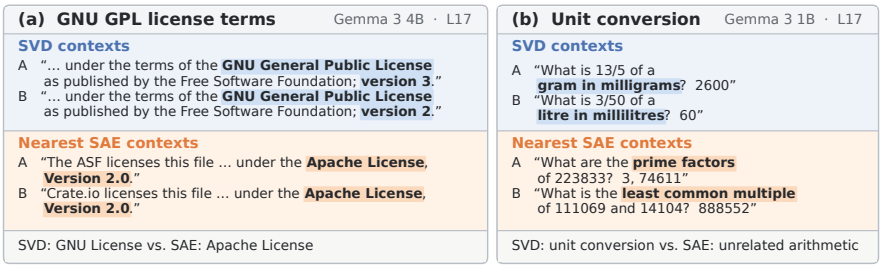


Figure 4: **SVD-specific concepts vs. nearest SAE neighbors.** Representative cases.

Concept specificity. At the individual feature level, we identify specific SVD concepts that cannot be semantically substituted by their nearest SAE neighbors (Appendix D.3). Across the 12 evaluation layers, we find 1,519 instances where a coherent SVD feature isolates specific activating semantics that none of its three nearest SAE neighbors can reproduce. Figure 4 illustrates two representative cases where nearest SAE neighbors fail to substitute for SVD concepts, either by blurring into a broader category or diverging into a different domain. In software, an SVD feature isolates GNU GPL text, whereas its nearest SAE neighbor responds broadly to general licenses, including Apache. In mathematics, an SVD feature tracks unit conversions, while its nearest SAE neighbor mixes unrelated arithmetic such as prime factorization and least common multiples. Additional cases, with longer activating texts and judge reasons, appear in Appendix D.3, Table 12.

Importantly, these advantages are not an artifact of candidate inventory size: both the low-budget coverage gains (§3.2) and SVD-specific concepts persist across all seven evaluation sites where the full SVD feature pool is smaller than the SAE pool (Appendix Table 11). Together, these findings establish intrinsic SVD as a complementary semantic basis, which encodes verifiable semantics in individual features, achieves denser coverage in compact dictionaries, and preserves structured concepts that learned SAEs blur into broader categories or overlook.

4 SVD Features Excel at Sparse Concept Prediction

Beyond describability, can intrinsic singular features linearly predict target concepts as effectively as learned SAEs? We evaluate prediction performance across sparse feature budgets k in §4.1 and show in §4.2 that both dictionaries specialize across different concept families(Gurnee et al., 2023).

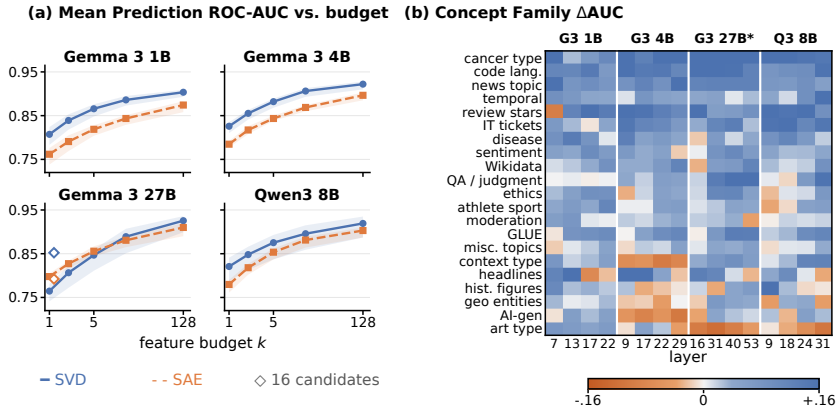


Figure 5: **Sparse prediction.** (a) Mean ROC-AUC vs. feature budget k. (b) Family-mean SVD–SAE ΔAUC ($k = 1$). Diamonds in (a) and * in (b): 16-candidate validation selection for Gemma 3 27B.

4.1 SVD Provides a More Predictive Basis under Sparse Budgets

Following standard SAE-Probes conditions (Kantamneni et al., 2025; Karvonen et al., 2025), we test whether top-k features selected by mean difference can linearly predict 113 binary concepts

(Appendix E). Consistent with the layer selection in §3, we evaluate four matched residual layers in Gemma 3 1B, 4B, 27B, and Qwen3 8B across probe budgets $k \in \{1, 2, 5, 16, 128\}$, treating $k = 1$ as the primary test of single feature predictivity and larger k for compositional scaling (complete model–layer curves in Appendix Figure 12). For Gemma, we benchmark cumulative residual SVD against the top-performing 65k/L_0-big residual SAE, selected by our Gemma 3 4B width and sparsity sweep (Appendix Figure 13a). For Qwen, we evaluate the public 64k SAE.

SVD features consistently achieve higher linear concept predictivity than learned SAEs. At $k = 1$, SVD outperforms SAEs across all 12 sites in Gemma 3 1B, 4B, and Qwen3 8B (Figure 5a). This advantage widens with depth, where $\Delta\mathrm{AUC}$ more than doubles from $+0.026$ to $+0.055$ across layers in Qwen3 8B. This lead persists across budgets $k \in \{2, 5, 16, 128\}$, winning up to 100 of 113 tasks at $k = 5$ in Gemma 3 4B. In Gemma 3 27B, while naive mean-difference ranking favors SAE at small k, selecting from just 16 candidates via validation restores SVD's substantial advantage across all four layers (Figure 5a, open diamonds; Appendix E).

4.2 Broad SVD Predictive Advantage with Specialized SAE Niches

Beyond aggregate averages, predictive differences $\Delta\mathrm{AUC} = \mathrm{SVD} - \mathrm{SAE}$ across 21 task families reveal domain-specific strengths (Figure 5b; Appendix Table 13). SVD has higher cross-site mean AUC in 19 families, with a mean margin of $+0.192$ on *cancer type* classification. *Art type* favors SAE in 10 of 16 sites and has a cross-site mean margin of -0.020; *AI-generated text* also has a small mean SAE advantage of -0.003. The SVD advantage is largest and most consistent on *cancer type*, *code language*, and *news topic* tasks, while *art type* retains a model-dependent SAE advantage.

These results demonstrate that core conceptual directions are already explicit in the model's weight geometry. Intrinsic SVD thus forms a powerful training-free predictive basis, while situating learned SAEs within specialized, complementary stylistic niches.

5 SVD Features Enable Targeted Causal Control

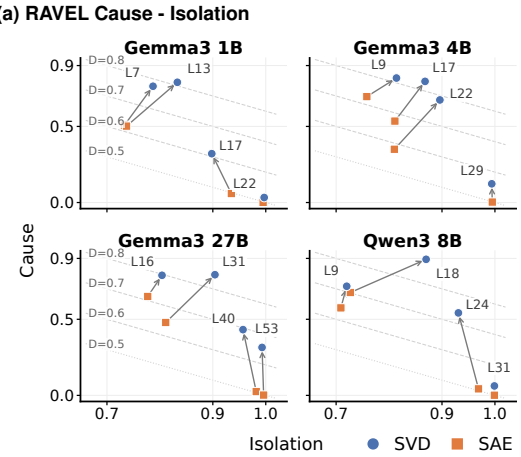
Beyond describability and predictivity, can intrinsic singular features exert targeted, effective causal control over model computation (Makelov et al., 2025)? We evaluate selective attribute disentanglement on subject entities in §5.1 and demonstrate factual relation steering in §5.2.

5.1 SVD Features Enable More Selective Attribute Disentanglement

An effective concept representation must causally govern target model behavior with high selectivity, altering intended attributes while leaving unrelated semantics intact (Canby et al., 2025). Following the mask-learning feature interchange protocol of Chaudhary & Geiger (2024) and SAEBench (Karvonen et al., 2025), we evaluate attribute disentanglement across the official SAEBench subset and the complete RAVEL benchmark (Huang et al., 2024), measuring causal efficacy on the target attribute via *Cause*, non-target preservation via *Isolation*, and their joint mean Disentangle score D. Consistent with previous sections, we benchmark four residual layers across depths in Gemma 3 1B, 4B, 27B, and Qwen3-8B against top-performing SAEs across 16 paired sites (Appendix Figure 14).

The cumulative residual SVD dictionary achieves consistently stronger and more selective attribute disentanglement than learned SAEs, as shown in the *Cause–Isolation* space of Figure 6a. SVD achieves higher Disentangle across all 16 evaluation sites, driven by simultaneous gains in *Cause* rather than passive interventions that merely preserve *Isolation*. At Gemma 3 4B layer 22, SVD raises *Cause* from 0.350 to 0.674 and *Isolation* from 0.810 to 0.896, boosting D by $+0.205$. This advantage extends to the complete RAVEL benchmark across all five entity types (Appendix F.4, Table 17).

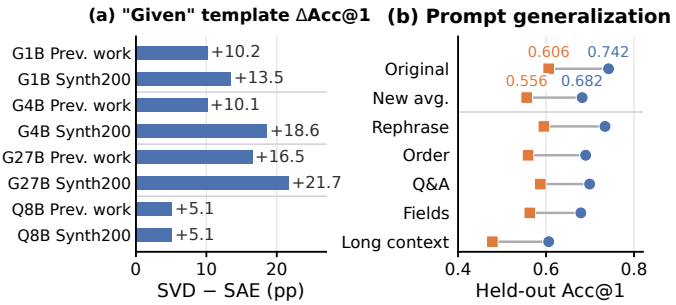
Beyond SAEs, the cumulative SVD dictionary outperforms all other unsupervised dictionaries across Gemma 3 4B layers (Figure 6b; Appendix F.3), averaging 0.748 against 0.620 for residual neurons and SAEs, 0.563 for residual activation PCA, and 0.719 for cumulative module output neurons, which accumulate upstream attention and MLP module outputs in the neuron basis. Even compared with MDAS, which optimizes its projection basis under task supervision, cumulative SVD matches its 0.748 mean. These comparisons decouple the impact of cross-depth accumulation from basis orientation. The marked advantage of cumulative methods over local residual representations confirms that maintaining source-specific write coordinates across depth preserves accessible attribute



(b) Other Baselines on Gemma3 4B

Method	L9	L17	L22	L29	Mean
SVD†	**0.816**	**0.832**	**0.785**	0.559	**0.748**
Res. SAE	0.727	0.673	0.580	0.499	0.620
Res. neurons	0.730	0.663	0.588	0.499	0.620
Res. PCA	0.627	0.593	0.534	0.500	0.563
Module neurons†	0.799	0.796	0.753	0.528	0.719
MDAS	0.804	0.827	0.768	**0.587**	0.746

Figure 6: **RAVEL causal interchange.** (a) Cause–Isolation across 16 evaluation sites; (b) Gemma 3 4B baselines (D). † accumulates upstream module writes; MDAS is task-supervised basis.



(c) Other Baselines G4B

Method	Acc@1	Acc@5
SVD†	**71.65**	**88.29**
SAE	61.50	77.40
Resid. neurons	65.81	81.03
Resid. PCA	58.03	75.07
SVD (one attn+MLP)	62.91	81.93
Module neurons†	71.63	83.83
Weight columns†	70.37	86.42

Figure 7: **Relation intervention.** (a) Template "*Given [subject]...*" ΔAcc@1; (b) Prompt generalization Acc@1; (c) Gemma 3 4B baselines (Acc@1/5). † accumulates upstream module writes.

pathways significantly better than residual streams. Concurrently, SVD consistently surpasses cumulative module output neurons across all four layers, indicating that intra-module singular decomposition yields a low-crosstalk, high-leverage basis that mitigates collinearity within the module output neurons, offering a more effective coordinate frame for selective causal intervention.

5.2 SVD Feature Compositions Drive Stronger Relation Intervention

Beyond isolated attributes, factual knowledge recall requires executing relational mappings from subjects to entities. While prior work steers relational mechanisms through coarse, monolithic residual overwrites (Wang et al., 2024), whether intrinsic SVD or learned SAE features can assemble into fine-grained, composable and effective relation concepts remains unknown.

To identify relation concepts, we rank decomposed features via Integrated Gradients along a corruption-to-recovery path (Sundararajan et al., 2017; Lundstrom et al., 2022). Selecting top-k candidates, we grid-search composition size k and steering strength on randomized relation prompts, evaluating on held-out test subjects. For instance, steering an injected random relation on subject Japan toward *capital of* should recover Tokyo. Beyond 80 published relations (Hernandez et al., 2024; Christ et al., 2025), we introduce Synth200, a benchmark of 200 relations spanning 12 diverse semantic domains such as science, technology, and biology to systematically evaluate relational steering across broader knowledge categories (Appendix G).

Intrinsically, cumulative SVD feature compositions steer factual relations more accurately than learned SAEs, raising macro Acc@1 across eight model-dataset groups on the standard template *Given [subject], the [relation] of this [type] is* from 0.606 to 0.742 ($+0.136$ margin, 95% CI $[+0.117, +0.155]$) in Figure 7a. This advantage generalizes across prompt variations: on ten unseen templates across five structural families (paraphrases, inverted clauses, questions, structured fields,

extended contexts), frozen SVD vectors maintain a decisive lead in Figure 7b, reaching Acc@1 of 0.682 vs. 0.556 for SAE ($+0.125$ margin), confirming that SVD steers abstract relational mappings rather than surface cues.

Benchmarking baselines on published relations on Gemma 3 4B highlights the dual benefits of cross-depth accumulation and SVD basis orientation (Figure 7c; Appendix G.3). Cross-depth accumulation proves essential. Restricting SVD to one upstream attention plus MLP block yields only 62.91% Acc@1, and local residual baselines plateau at 58.03% for activation PCA and 65.81% for residual neurons. In contrast, accumulating all upstream module writes elevates performance past 71%. Under this shared cumulative formulation, SVD achieves the highest mean Acc@1 of 71.65% and decisively leads on Acc@5 at 88.29%, surpassing cumulative module neurons at 83.83% by 4.46 points and raw weight columns (W_O, W_{down}) at 86.42% by 1.87 points. These gains demonstrate that intra-module singular vectors yield a cleaner, more readily identifiable basis for causal steering.

6 Eliminating Auxiliary Training and Seed Instability

Intrinsic SVD eliminates the data and compute burdens of auxiliary training. For Qwen3-8B-Base layers 18 and 24, extracting SVD features for one attention together with one MLP module requires zero tokens and 2.63×10^{12} FLOPs. In contrast, training matched SAE pairs under the Gemma Scope JumpReLU recipe (Appendix B) consumes 2.0–4.0 billion tokens and 1.92×10^{19}–6.77×10^{19} FLOPs, yielding an SAE-to-SVD work ratio of 7.28×10^{6}–2.57×10^{7} (Table 23, Appendix H).

Furthermore, weight SVD yields deterministic dictionaries, whereas learned SAEs vary substantially across training seeds. Across three Qwen3-8B-Base layer-18 seeds, exact permutation alignment yields mean cosine similarities of only 0.689 for attention and 0.851 for MLP outputs. Even after matching, roughly 41% of attention and 17% of MLP features fall below 0.7 cosine similarity, showing that many learned directions fail to replicate (Appendix H.2).

7 Related Work

Training-based dictionary learning. Under the linear representation hypothesis (Elhage et al., 2022; Park et al., 2024), sparse autoencoders (SAEs) and transcoders are trained on massive activations to uncover interpretable features (Bricken et al., 2023; Templeton et al., 2024; Huben et al., 2024; Gao et al., 2025; Dunefsky et al., 2024). Yet auxiliary training incurs fundamental trade-offs: reconstruction error harms intervention fidelity (Makelov et al., 2025; Karvonen et al., 2025), random seeds yield divergent dictionaries (Paulo & Belrose, 2026), and feature absorption masks fine-grained concepts (Chanin et al., 2025). While methods like subspace-aware SAEs (Dalili & Mahdavi, 2026) attempt improvements, scaling full suites remains computationally prohibitive.

Training-free interpretability and weight SVD. Beyond training, other works study internal circuits, communication channels, and low-rank subspaces (Elhage et al., 2021; Olsson et al., 2022; Meng et al., 2022; Wang et al., 2024), alongside native neuron bases and activation signs (Arora et al., 2026; Nalagatla, 2026). Weight SVD in particular probes localized mechanisms: tracing head communication (Merullo et al., 2024), projecting late MLP subspaces to vocabulary logits (Xue & Andrzejak, 2026), isolating task-masked circuits (Ahmad et al., 2025), and extracting vision semantic OV directions (Gentile et al., 2026). Yet these studies treat singular vectors as localized probes of isolated components or circuits, leaving their potential as a feature dictionary foundation unexplored.

8 Conclusion

Transformer weights intrinsically contain an interpretable feature dictionary. By accumulating singular write directions across upstream attention and MLP operators, weight SVD yields a deterministic, training-free dictionary that exceeds or complements learned SAEs across descriptive, predictive, and causal capabilities, establishing a principled foundation for intrinsic mechanistic interpretability.

AI USE STATEMENT

Generative AI assisted with experimental evaluation, benchmark generation, and manuscript editing. Llama 3.1 70B Instruct generated and scored feature descriptions, audited semantic clusters, and judged SVD–SAE concept overlap (Appendices C, D.1, and D.3). For the Synth200 benchmark, `gpt-5.6-sol` proposed relations and `gpt-5.6-terra` generated subject–object pairs (Appendix G). AI writing assistants were used for text drafting and polishing. The authors validated all model outputs and take full responsibility for the paper's contents.

ETHICS STATEMENT

This work uses publicly available open-weight models and benchmarks, involving no human subjects, sensitive personal data, or private information. While representation steering could theoretically be leveraged for behavioral manipulation, our interventions are purely diagnostic and aimed at evaluating representation selectivity and mechanistic transparency.

REPRODUCIBILITY STATEMENT

Mathematical formulations and extraction algorithms are detailed in Section 2 and Appendix A. Model and SAE checkpoints, training settings, and hyperparameters are specified in Appendix B. Appendices C–G document all datasets, prompts, splits, and intervention protocols, with compute FLOPs and seed stability analyzed in Appendix H. Code, evaluation scripts, and precomputed dictionary artifacts are provided in the supplementary material and will be open-sourced upon publication.

REFERENCES

Areeb Ahmad, Abhinav Joshi, and Ashutosh Modi. Beyond components: Singular vector-based interpretability of transformer circuits. In *Advances in Neural Information Processing Systems*, volume 38, pp. 167772–167811, 2025. doi: 10.52202/085713-5593. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/f51afa708f3d4b3cae8dc9a8a9bbea87-Abstract-Conference.html.

Aryaman Arora, Zhengxuan Wu, Jacob Steinhardt, and Sarah Schwettmann. Language model circuits are sparse in the neuron basis, 2026. URL https://arxiv.org/abs/2601.22594.

Trenton Bricken, Adly Templeton, Joshua Batson, et al. Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits Thread, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features/.

Tom B. Brown et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Bart Bussmann, Patrick Leask, and Neel Nanda. BatchTopK sparse autoencoders, 2024. URL https://arxiv.org/abs/2412.06410.

Marc E. Canby, Adam Davies, Chirag Rastogi, and Julia Hockenmaier. How reliable are causal probing interventions? In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 857–878. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.ijcnlp-long.47. URL https://aclanthology.org/2025.ijcnlp-long.47/.

David Chanin, James Wilken-Smith, Tomáš Dulka, Hardik Bhatnagar, Satvik Golechha, and Joseph Bloom. A is for absorption: Studying feature splitting and absorption in sparse autoencoders. In *Advances in Neural Information Processing Systems*, volume 38, pp. 82318–82355, 2025. doi: 10.52202/085713-2755. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/764ff7477b8e24dbe01531f6791e8bdf-Abstract-Conference.html.

Maheep Chaudhary and Atticus Geiger. Evaluating open-source sparse autoencoders on disentangling factual knowledge in GPT-2 Small, 2024. URL https://arxiv.org/abs/2409.04478.

Miranda Anna Christ, Adrián Csiszárik, Gergely Becsó, and Dániel Varga. The structure of relation decoding linear operators in large language models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 43187–43215, 2025. doi: 10.52202/085713-1438. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/3dada6990a8e8af682ed2c9803793e38-Abstract-Conference.html.

Seyed Arshan Dalili and Mehrdad Mahdavi. Subspace-aware sparse autoencoders for effective mechanistic interpretability, 2026. URL https://arxiv.org/abs/2606.06333.

Boyi Deng, Xu Wang, Yaoning Wang, Yu Wan, Yubo Ma, Baosong Yang, Haoran Wei, Jialong Tang, Huan Lin, Ruize Gao, Tianhao Li, Qian Cao, Xuancheng Ren, Xiaodong Deng, An Yang, Fei Huang, Dayiheng Liu, and Jingren Zhou. Qwen-scope: Turning sparse features into development tools for large language models, 2026. URL https://arxiv.org/abs/2605.11887.

Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. Transcoders find interpretable LLM feature circuits. In *Advances in Neural Information Processing Systems*, volume 37, pp. 24375–24410, 2024. doi: 10.52202/079017-0768. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/2b8f4db0464cc5b6e9d5e6bea4b9f308-Abstract-Conference.html.

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

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition, 2022. URL https://transformer-circuits.pub/2022/toy_model/.

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

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *International Conference on Learning Representations*, pp. 26721–26754, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/42ef3308c230942d223c411adf182c88-Abstract-Conference.html.

Nikhil Garg, Jon Kleinberg, and Kenny Peng. How many features can a language model store under the linear representation hypothesis? In *Proceedings of Thirty Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 5358–5376. PMLR, 2026. URL https://proceedings.mlr.press/v336/garg26a.html.

Gemma Team et al. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Francesco Gentile, Nicola Dall'Asen, Francesco Tonini, Massimiliano Mancini, Lorenzo Vaquero, and Elisa Ricci. From weights to concepts: Data-free interpretability of CLIP via singular vector decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2895–2906, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Gentile_From_Weights_to_Concepts_Data-Free_Interpretability_of_CLIP_via_Singular_CVPR_2026_paper.html.

Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. Finding neurons in a haystack: Case studies with sparse probing. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=JYs1R9IMJr.

XinYang He, Wei Wang, Bing Zhao, Xuan Ren, WenBo Li, WeiXu Qiao, Hu Wei, and Lin Qu. Discovering millions of interpretable features with sparse autoencoders, 2026. URL https://arxiv.org/abs/2606.26620.

Evan Hernandez, Arnab Sen Sharma, Tal Haklay, Kevin Meng, Martin Wattenberg, Jacob Andreas, Yonatan Belinkov, and David Bau. Linearity of relation decoding in transformer language models. In *International Conference on Learning Representations*, pp. 10504–10526, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/2b6d7285306fe896fca9b2e9a02ff6ec-Abstract-Conference.html.

Jing Huang, Zhengxuan Wu, Christopher Potts, Mor Geva, and Atticus Geiger. RAVEL: Evaluating interpretability methods on disentangling language model representations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8669–8687. Association for Computational Linguistics, 2024. URL https://aclanthology.org/2024.acl-long.470/.

Robert Huben, Hoagy Cunningham, Logan Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, pp. 7827–7845, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/1fa1ab11f4bd5f94b2ec20e794dbfa3b-Abstract-Conference.html.

Subhash Kantamneni, Joshua Engels, Senthooran Rajamanoharan, Max Tegmark, and Neel Nanda. Are sparse autoencoders useful? a case study in sparse probing. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 29018–29049. PMLR, 2025. URL https://proceedings.mlr.press/v267/kantamneni25a.html.

Adam Karvonen, Can Rager, Johnny Lin, Curt Tigges, Joseph Isaac Bloom, David Chanin, Yeu-Tong Lau, Eoin Farrell, Callum Stuart Mcdougall, Kola Ayonrinde, Demian Till, Matthew Wearden, Arthur Conmy, Samuel Marks, and Neel Nanda. SAEBench: A comprehensive benchmark for sparse autoencoders in language model interpretability. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 29223–29264. PMLR, 2025. URL https://proceedings.mlr.press/v267/karvonen25a.html.

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

Daniel Lundstrom, Tianjian Huang, and Meisam Razaviyayn. A rigorous study of integrated gradients method and extensions to internal neuron attributions, 2022. URL https://arxiv.org/abs/2202.11912.

Aleksandar Makelov, Georg Lange, and Neel Nanda. Towards principled evaluations of sparse autoencoders for interpretability and control. In *International Conference on Learning Representations*, pp. 33588–33636, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/53356aebeea8ffd40a8ac3bb66243162-Abstract-Conference.html.

Callum McDougall, Arthur Conmy, János Krámár, Tom Lieberum, Senthooran Rajamanoharan, and Neel Nanda. Gemma Scope 2 - technical paper, 2025. Technical paper.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372, 2022. doi: 10.52202/068431-1262. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. Talking heads: Understanding inter-layer communication in transformer language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 61372–61418, 2024. doi: 10.52202/079017-1962. URL https://papers.nips.cc/paper_files/paper/2024/hash/70e5444e5f331f7f5431f302110b97af-Abstract-Conference.html.

Varun Reddy Nalagatla. The signs were always there: Training-free concept detection and steering in raw transformer dimensions, 2026. URL https://arxiv.org/abs/2606.12629.

nostalgebraist. interpreting GPT: the logit lens. LessWrong, 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

Catherine Olsson, Nelson Elhage, Neel Nanda, et al. In-context learning and induction heads. Transformer Circuits Thread, 2022. URL https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 39643–39666. PMLR, 2024. URL https://proceedings.mlr.press/v235/park24c.html.

Gonçalo Paulo and Nora Belrose. Sparse autoencoders trained on the same data learn different features. In *International Conference on Learning Representations*, pp. 35748–35760, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/3c1fe56b043848b211030c202764c6a7-Abstract-Conference.html.

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

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328. PMLR, 2017. URL https://proceedings.mlr.press/v70/sundararajan17a.html.

Adly Templeton, Tom Conerly, Jonathan Marcus, et al. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. Transformer Circuits Thread, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/.

Sinie van der Ben, Neele Roch, Anna Hedström, and Mennatallah El-Assady. Building fast, evaluating slow: Pipeline choices dominate autointerpretability score variance, 2026. URL https://arxiv.org/abs/2607.19386.

Zijian Wang, Britney Whyte, and Chang Xu. Locating and extracting relational concepts in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 4818–4832. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.287. URL https://aclanthology.org/2024.findings-acl.287/.

Min Xue and Artur Andrzejak. SVD as a fast interpretability method for transformers. In *Forty-Third International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/66016.

An Yang et al. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

A Architecture-Specific Implementation of the SVD Dictionary

The SVD dictionary (§2) factorizes the linear projection weights of the transformer text decoder. All evaluated Gemma 3 and Qwen3 checkpoints feature bias-free linear projections.

Attention OV decomposition under GQA. For grouped-query attention (GQA), H query heads share H_{KV} key-value heads. Each query head $p \in \{0, \ldots, H-1\}$ maps to value head $g(p) = \lfloor p/(H/H_{\mathrm{KV}}) \rfloor$. We factorize the per-head OV operator $\mathbf{W}_{OV}^{(p)} = \mathbf{W}_{O}^{(p)} \mathbf{W}_{V}^{(g(p))} = \mathbf{U}_{OV}^{(p)} \mathbf{\Sigma}_{OV}^{(p)} \mathbf{V}_{OV}^{(p)\top}$, where $\mathbf{W}_{O}^{(p)}$ slices the output projection columns for head p and $\mathbf{W}_{V}^{(g(p))}$ slices the value projection rows for head $g(p)$. During forward execution, dynamic attention weights, causal masks, and positional encodings modulate the contextual value vector $\mathbf{v}^{(p)}$ captured at `self_attn.o_proj.input`; singular channel activations are obtained by projecting through the head's output slice onto the left singular vectors: $\mathbf{a}^{(p)} = \mathbf{U}_{OV}^{(p)\top} \mathbf{W}_{O}^{(p)} \mathbf{v}^{(p)}$.

MLP down-projection decomposition. For gated MLPs, SVD factorizes the down-projection weight $\mathbf{W}_{\mathrm{down}} = \mathbf{U}_{\mathrm{down}} \mathbf{\Sigma}_{\mathrm{down}} \mathbf{V}_{\mathrm{down}}^{\top}$. Channel activations are obtained by projecting module writes onto $\mathbf{U}_{\mathrm{down}}$, hooked at `mlp.down_proj.output` for Qwen3 and `post_feedforward_layernorm.input` for Gemma 3. Bases are computed in float32 retaining full output rank. Each singular vector is canonically oriented such that its maximum-magnitude coordinate is positive, with positive and negative polarities both available as distinct semantic directions.

Module normalization and residual writes. Qwen3 modules write directly to the residual stream. In Gemma 3, post-attention and post-feedforward RMSNorm layers precede residual accumulation. For residual-stream causal interventions (§5, Appendix F), the effective write basis incorporates the input-dependent RMS factor $s(\mathbf{x}) = 1/\,\mathrm{RMS}(\mathbf{x})$ and learned diagonal scaling $(\mathbf{1} + \mathbf{w}_{\mathrm{norm}})$, exactly recovering the residual additions. For local module-site interpretability (§3, Appendix C), activations are evaluated directly at the pre-norm module hooks.

B Models and SAE Weights

B.1 Language Models and Public SAE Weights

We evaluate pretrained base checkpoints of Gemma 3 (1B, 4B, 27B) (Gemma Team et al., 2025) and Qwen3-8B-Base (Yang et al., 2025), analyzing the text decoders for multimodal variants. Evaluated layers span approximately 25%, 50%, 65%, and 85% of decoder depth (Table 1), matching the primary evaluation layers established in Gemma Scope 2 (McDougall et al., 2025).

Base model	Decoder layers	Evaluated layers (zero-indexed)
Gemma 3 1B PT	26	7, 13, 17, 22
Gemma 3 4B PT	34	9, 17, 22, 29
Gemma 3 27B PT	62	16, 31, 40, 53
Qwen3-8B-Base	36	9, 18, 24, 31

Table 1: **Representative decoder layers.** Gemma layers follow Gemma Scope 2; Qwen layers cover corresponding fractions of decoder depth.

Public Gemma SAEs are taken from the Gemma Scope 2 JumpReLU release (McDougall et al., 2025) across `attn_out`, `mlp_out`, and `resid_post` sites. For Qwen residual-stream benchmarks, we use the public Qwen-Scope TopK SAE (Deng et al., 2026) (65,536 latents, TopK 100). For Qwen attention and MLP module evaluations, we train matched JumpReLU SAEs as detailed below. Checkpoints are sourced from Hugging Face: Gemma 3 base models (`google/gemma-3-{1b,4b,27b}-pt`), Gemma Scope 2 (`google/gemma-scope-2-{1b,4b,27b}-pt`), Qwen3-8B-Base (`Qwen/Qwen3-8B-Base`), and Qwen-Scope (`Qwen/SAE-Res-Qwen3-8B-Base-W64K-L0_100`).

B.2 Qwen Attention and MLP SAE Training

To evaluate module-level representations in Qwen3-8B-Base where public module SAEs are unavailable, we train JumpReLU SAEs following the Gemma Scope 2 recipe (McDougall et al., 2025).

Activation sites and configurations. At layers 18 and 24, we train separate SAEs on the concatenated attention-head output (`self_attn.o_proj.input`) and the MLP output (`mlp.output`). At each site, we train two configurations: a wide, sparse setting (65k width, target $L_0^* = 20$, 4B tokens) and a compact setting (16k width, target $L_0^* = 150$, 2B tokens), yielding eight checkpoints in total (Table 2).

Layer	Configuration	Width	Target L_0^*	Training Tokens
18	65k/small	65,536	20	4,000,000,000
18	16k/big	16,384	150	2,000,001,024
24	65k/small	65,536	20	4,000,000,000
24	16k/big	16,384	150	2,000,001,024

Table 2: **Trained Qwen attention and MLP SAEs.** Each row represents two independently trained models, one per activation site. Configuration labels denote target sparsity L_0^*.

Dataset and tokenization. Following recent SAE training on Qwen3 models (He et al., 2026), training data is sourced from FineWeb-Edu (`sample-10BT`), tokenized with the Qwen3-8B tokenizer and packed into 2,048-token sequences without boundary padding. The frozen base model runs in bfloat16 through the target layer; activations are stored in float16 shards of 262,144 tokens. The training set is strictly disjoint from all evaluation corpora.

Architecture and objective. Activations $\mathbf{x}$ are scaled by a fixed scalar such that $\mathbb{E}\|\mathbf{x}'\|_2^2 = 1$. The JumpReLU SAE computes pre-activations $\mathbf{u} = (\mathbf{x}' - \mathbf{b}_{\text{dec}})\mathbf{W}_{\text{enc}} + \mathbf{b}_{\text{enc}}$, sparse latents $\mathbf{z}_i = \mathbf{u}_i \mathbf{1}[\mathbf{u}_i > \theta_i]$, and reconstruction $\hat{\mathbf{x}}' = \mathbf{z}\mathbf{W}_{\text{dec}} + \mathbf{b}_{\text{dec}}$. The per-token objective is

$$\mathcal{L} = \mathbb{E}\|\mathbf{x}' - \hat{\mathbf{x}}'\|_2^2 + \lambda_t \, \mathbb{E}\left[\frac{2}{L_0^*}\left(\|\mathbf{z}\|_0 - L_0^*\right)^2\right], \tag{8}$$

where target sparsity $L_0^* \in \{20, 150\}$, and penalty coefficient λ_t warms up linearly from 0 to 10^{-3} over the first 50,000 steps. Thresholds $\boldsymbol{\theta}$ start at 0.001 with a 0.001-bandwidth straight-through estimator. Decoder dictionary columns are constrained to unit norm with parallel gradient components removed during optimization.

Optimization and checkpoint export. Training uses a global batch of 4,096 tokens, Adam ($(\beta_1, \beta_2) = (0, 0.999)$, $\epsilon = 10^{-8}$), and a learning rate of 7×10^{-5} after a 1,000-step cosine warmup. At export, activation normalization is folded into the SAE parameters, and checkpoints are stored as standard `params.safetensors` files containing encoder and decoder weights, biases, and thresholds.

C AutoInterp Detection of Feature Describability

C.1 Evaluation Method

Detection protocol. We adapt SAEBench AutoInterp Detection (Karvonen et al., 2025) to evaluate whether natural-language explanations reliably predict feature activation on held-out text. For each feature, an explainer model inspects 15 21-token contexts (10 top-activating and 5 activation-weighted) with active tokens highlighted, generating a concise description. A separate judge model evaluates this description on 14 shuffled held-out contexts (2 top positives, 2 activation-weighted positives, and 10 fully inactive negatives). Feature score is binary classification accuracy on these 14 contexts (always-negative baseline is $10/14 \approx 0.714$). Both explainer and judge use quantized `Meta-Llama-3.1-70B-Instruct-AWQ-INT4` at temperature zero.

Evaluation sites and sampling. We benchmark 28 model $\times$ layer $\times$ module sites: 24 sites across Gemma 3 1B, 4B, and 27B (evaluating attention and MLP outputs at the four depth-spanning layers in Table 1), and 4 sites in Qwen3-8B-Base (attention and MLP outputs at layers 18 and 24, using matched JumpReLU SAEs from Appendix B). Attention hooks are placed at the pre-`o_proj` concatenated head output (`self_attn.o_proj.input`); MLP hooks are placed at module outputs before post-layer normalization. Per site, 1,000 SVD axes and 1,000 SAE latents are sampled on calibration shard

`01` of the Pile (2M tokens). SAE latents are sampled across activation frequency deciles, while SVD axes are sampled uniformly across heads and singular rank bins. Scoring is evaluated on disjoint Pile shard `00` (2M tokens).

SVD calibration and polarity handling. Because linear singular vectors are unconstrained and signed, each axis provides two candidate feature directions ($\pm$). For each direction, we calibrate an activation threshold $t_\pm$ on calibration shard `01` such that its firing frequency matches the median firing frequency of active latents from the canonical 65k/small SAE at that site. Positive contexts require activation above $t_\pm$, and negative windows contain no activations exceeding the floor cutoff. To maintain consistent feature identities across benchmarks, calibrated SVD thresholds are held fixed when comparing with alternative SAE widths. *Raw SVD* reports the mean accuracy across both signed directions, while *unipolar SVD* selects the higher-scoring polarity of each channel, capturing the empirical polarity asymmetry of singular features (§3.1).

Measurement stability. Following the stability protocol of van der Ben et al. (2026), we evaluate sensitivity to explanation context resampling at Gemma 3 4B layer 9 across five seeds (42, 123, 456, 789, 2026). Across resampling draws, feature-level binary ROC AUC achieves high intraclass correlation ($\mathrm{ICC}(1,1) \in [0.753, 0.839]$), mean top-$k$ Jaccard lift exceeds the $2\times$ reliability line across all $k \in \{10, \ldots, 50\}$ (reaching $66.01\times$–$134.19\times$ at $k = 10$), and aggregate site accuracies vary by at most 0.0048. These metrics confirm that AutoInterp Detection reliably captures intrinsic feature interpretability.

C.2 Polarity Asymmetry

For each SVD axis with completed Detection scores s_+ and s_-, we define its polarity gap as $|s_+ - s_-|$. Across all 28 evaluation sites, 27,859 of 28,000 sampled axes (99.5%) completed evaluation on both directions.

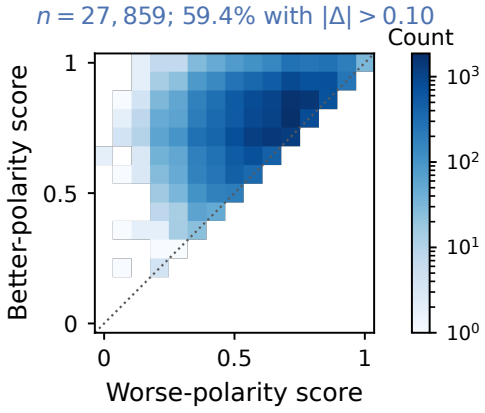


Figure 8: **Polarity asymmetry in AutoInterp Detection.** Distribution of singular channels by their better- and worse-scoring polarities across 27,859 completed channels in Gemma 3 and Qwen3-8B. A substantial accuracy gap (> 0.10, corresponding to ≥ 2 context predictions) occurs for 59.4% of channels.

As shown in Figure 8 and Table 3, singular channels exhibit pronounced polarity asymmetry: 59.4% of paired axes display an accuracy gap exceeding 0.10 (median gap $2/14 = 0.143$). Across all sites, the mean score of the more interpretable polarity reaches 0.791, compared to 0.626 for the opposite polarity (+0.165 gap). This asymmetry is consistent across both architectures (59.8% in Gemma 3, 56.9% in Qwen3-8B), demonstrating that intrinsic singular directions predominantly encode semantic concepts along one canonical orientation.

Model	Layer	Attn. paired	Attn. gap > 0.10	MLP paired	MLP gap > 0.10
Gemma 3 1B	7	996	62.9%	1,000	56.5%
	13	996	61.5%	1,000	64.5%
	17	996	60.8%	1,000	60.1%
	22	1,000	59.7%	1,000	67.1%
Gemma 3 4B	9	996	60.1%	1,000	57.3%
	17	998	58.3%	999	58.8%
	22	997	60.5%	999	57.2%
	29	1,000	59.2%	1,000	65.7%
Gemma 3 27B	16	886	61.1%	1,000	56.4%
	31	999	57.9%	1,000	53.9%
	40	998	55.1%	1,000	61.6%
	53	1,000	60.6%	1,000	58.6%
Qwen3 8B	18	1,000	57.4%	1,000	63.7%
	24	999	55.7%	1,000	50.9%

Table 3: **Polarity gaps across all AutoInterp sites.** Paired counts include SVD axes with completed Detection scores on both polarities; percentages indicate the fraction of axes with gap > 0.10.

C.3 Detection Across Depths and Module Types

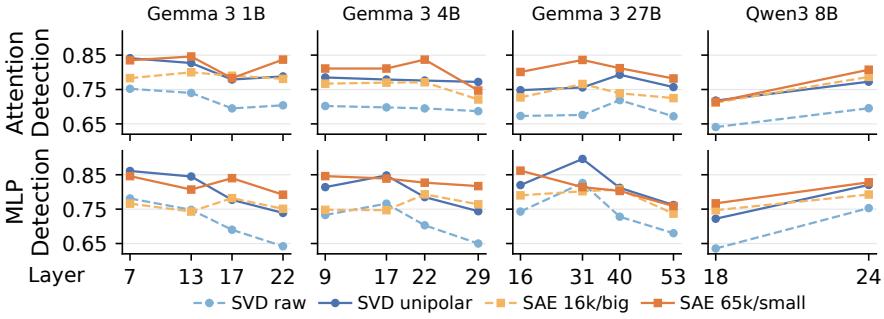


Figure 9: **AutoInterp Detection across depths and module types** in Gemma 3 and Qwen3-8B-Base. Points indicate site means; unipolar SVD captures the higher-scoring polarity of each channel.

Figure 9 details Detection accuracy across models, depths, and module types. Consistent with findings in §3.1, unipolar SVD demonstrates strong describability across layers, matching or closely approaching the canonical 65k/small SAE in attention heads and early-to-mid MLP modules, while consistently exceeding the compact 16k/big SAE across diverse sites. The advantage of trained SAEs is primarily concentrated in deeper MLP layers, whereas attention heads exhibit comparable describability across all evaluated depths.

C.4 SAE Width and Sparsity Selection

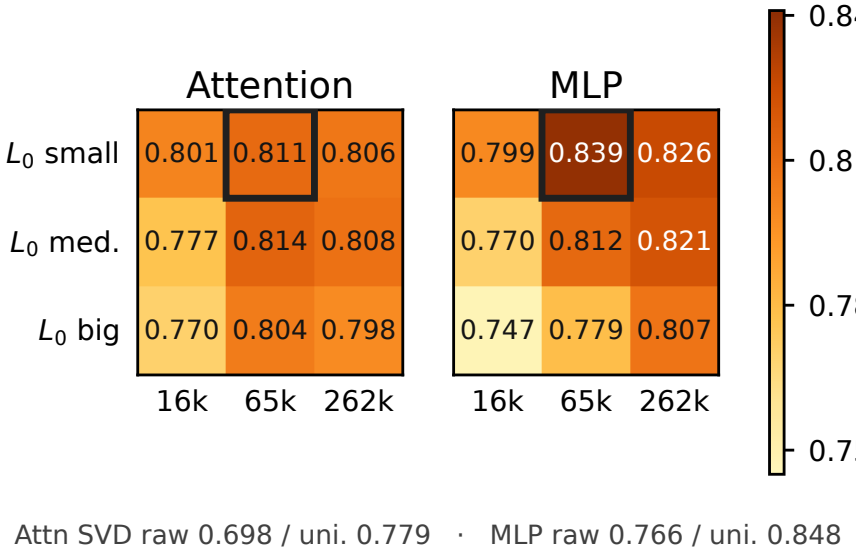


Figure 10: **Gemma 3 4B layer 17 SAE width–L_0 scan.** Cells display Detection accuracy across widths and target sparsities. The outlined 65k/small configuration establishes the strongest overall baseline across module sites.

To identify the strongest SAE comparison baselines, we conducted a 9-checkpoint width and sparsity scan at Gemma 3 4B layer 17 across widths $\{16\text{k}, 65\text{k}, 262\text{k}\}$ and sparsity levels $\{L_0\text{-small, medium, big}\}$ (Figure 10). The 65k/small configuration achieved the strongest overall performance (0.811 on attention, 0.839 on MLP), while maintaining high feature completion rates ($> 95\%$). We therefore established 65k/small as the canonical SAE benchmark across all Gemma sites. In addition, we evaluate 16k/big as a dense, compact SAE baseline across all 28 sites. Across all sites, equal-weighted mean Detection accuracy is 0.765 for 16k/big, 0.791 for unipolar SVD, and 0.811 for 65k/small.

D Semantic Coverage and Complementarity

D.1 Topic Map and Evaluation Method

Residual SAE baselines. At each of the 12 residual model–layer sites, we evaluate against the widest publicly available residual SAE checkpoints: Gemma Scope 2 width 262k (L_0-small) for Gemma 3 1B and 4B, and Qwen-Scope width 64k (TopK 100) for Qwen3-8B (Appendix B). Using these expansive candidate pools ensures a strong learned baseline for concept discovery. In all fine-coverage evaluations, dictionaries are compared under identical, fixed feature budgets B, ensuring controlled comparisons across representations.

Table 4 summarizes the total candidate feature directions prior to cross-fold estimability filtering. For layer L, the cumulative SVD dictionary aggregates left singular vectors across all upstream blocks $0, \ldots, L$ (contributing 2,176 axes per block for Gemma 3 1B, 4,608 for Gemma 3 4B, and 8,192 for Qwen3-8B), with each axis evaluated along both positive and negative directions.

Model	Layer	SVD signed directions	SAE latents
Gemma 3 1B	7	34,816	262,144
	13	60,928	262,144
	17	78,336	262,144
	22	100,096	262,144
Gemma 3 4B	9	92,160	262,144
	17	165,888	262,144
	22	211,968	262,144
	29	276,480	262,144
Qwen3 8B	9	163,840	65,536
	18	311,296	65,536
	24	409,600	65,536
	31	524,288	65,536

Table 4: **Full residual feature pools at the 12 semantic-coverage sites.** SVD counts both polarities of each upstream singular axis as separate candidate features; SAE counts one feature per latent. Counts reflect candidate feature directions prior to cross-fold estimability filtering.

Notably, the candidate SVD pool is substantially smaller than the SAE pool across seven sites (all Gemma 3 1B layers and Gemma 3 4B layers 9, 17, and 22), providing a natural test of whether SVD advantages depend on inventory scale (Table 11).

Construction of the 73-topic map. The 73-topic semantic map is constructed from 225,000 64-token windows sampled from Pile Uncopyrighted, disjoint from the 45M-token activation corpus. Windows are embedded via Qwen3-Embedding-4B. A directly fitted $K = 64$ spherical k-means partition serves as the refinement root based on granularity diagnostics on validation contexts, minimizing label redundancy while preserving broad domain coverage. Heterogeneous clusters are hierarchically split into two or three sub-clusters under geometric and membership constraints (at least 500 fit and 100 validation windows per child, document-bootstrap centroid cosine ≥ 0.80, pairwise child cosine < 0.995, maximum depth 4). The final frozen partition contains 73 disjoint terminal clusters spanning all validation windows: 71 passed the strict automated LLM semantic audit, and the remaining two were manually inspected by the authors to provide descriptive reference labels (marked ‡ in Table 8; $71/73$ strict audit coverage, $73/73$ full descriptive coverage). All 73 clusters are strictly retained in all downstream metric denominators without post-hoc exclusions. Cluster memberships and centroids were frozen prior to all downstream dictionary analyses.

Topic naming and audit. For each node and across three audit seeds, 20 validation contexts (10 nearest the cluster centroid and 10 randomly sampled non-medoid windows) were evaluated by an independent Llama 3.1 70B Instruct auditor. Contexts were randomized, with cluster IDs, hierarchical structure, and model representations completely masked. The auditor returned a coherence judgment, a concise label, a descriptor type, and a brief rationale. A blinded consensus adjudicator synthesized the three audit runs, requiring at least two coherent runs referencing the same dominant concept to establish a canonical topic label. Table 5 displays the operative audit and adjudication prompt templates. Subsequent representation-blind display-name adjudication distinguished genuinely different clusters with similar labels without altering cluster memberships or centroids. For the two terminal clusters that did not reach automated LLM consensus (‡ in Table 8), the authors manually inspected sampled validation passages to provide concise descriptive labels for human readability (Topics 09 and 36), while strictly retaining both in all downstream 73-topic metric denominators.

Table 5: **Operative prompt templates for blinded semantic topic audit and consensus adjudication.** Evaluated via Llama 3.1 70B Instruct with cluster IDs, hierarchical structure, and representation sources fully masked. Placeholders {contexts} and {audits} supply the 20 numbered passages and the three independent audit outputs, respectively.

Blinded Cluster Auditor (Standard E)	**Blinded Label Adjudicator (Consensus)**
Role: Blinded cluster auditor. **Prompt:** Read all 20 contexts. Decide whether most share one short, stable, dominant theme (content domain, genre, symbolic format, or discourse function). Generic prose is not enough. **Output schema:** JSON object with fields: • "coherent": *bool* • "short_label": *str* • "descriptor_type": *str* • "rationale": *str* **Rules:** • coherent=true only if one label covers most passages. • descriptor_type: semantic, genre, syntax-format, or discourse-function. • Label must be short and specific (avoid generic or mixed labels). • If incoherent, return best dominant label with coherent=false.	**Role:** Blinded label adjudicator. **Prompt:** You see three independent short labels for the same cluster (with coherence vote, descriptor type, and rationale). You do not see original contexts, cluster IDs, SVD, or SAE. **Task:** Decide whether at least two coherent labels point to the same dominant concept. **Output schema:** JSON object with fields: • "same_dominant_concept": *bool* • "canonical_label": *str* • "descriptor_type": *str* • "laundry_list": *bool* • "rationale": *str* **Rules:** • If < 2 runs coherent, same_dominant_concept=false. • canonical_label must be short, specific, and non-generic. • Set laundry_list=true if cluster only enumerates unrelated domains.

Coarse semantic breadth. To evaluate coarse semantic breadth, a separate 45M-token Pile activation corpus is partitioned into document-disjoint folds A and B, retaining up to 40 top-activating contexts per feature per fold. A feature is estimable if it activates on at least 20 distinct documents per fold. Contexts are assigned to their nearest topic, yielding document-balanced topic allocation $m(f, t, h)$ with $\sum_t m(f, t, h) = 1$. For representation r and topic t, cross-fold association strength is

$$M_{r,t} = \max_{f \in r} \min\{m(f, t, A), m(f, t, B)\}. \tag{9}$$

Comparing against 1,000 label-permutation null replicates gives the one-sided permutation probability $\widehat{p}_{r,t}$ and overall breadth:

$$\widehat{p}_{r,t} = \frac{1 + \sum_{b=1}^{1000} \mathbf{1}[M_{r,t}^{(b)} \geq M_{r,t}]}{1001}, \qquad C_{\text{breadth},r} = \frac{1}{73} \sum_{t=1}^{73} \mathbf{1}[\widehat{p}_{r,t} \leq 0.01]. \tag{10}$$

Both SVD and SAE dictionaries achieve full breadth ($73/73$ topics, $C_{\text{breadth}} = 1.0$) across all 12 evaluated layers.

Held-out coverage score. Fine coverage assesses how densely a budget-constrained dictionary spans semantic space. For features meeting the significance criterion ($\widehat{p} \leq 0.01$), we form normalized prototype embeddings in the fit fold and select a dictionary D of budget B via greedy facility-location optimization. On held-out evaluation contexts $X_{t,h}$ with associated prototypes $P_t(D)$, coverage score G is computed as

$$G_h(D) = \frac{1}{|T|} \sum_{t \in T} \frac{1}{|X_{t,h}|} \sum_{x \in X_{t,h}} \max_{p \in P_t(D)} K_\sigma(x, p), \tag{11}$$

where $K_\sigma(x, p) = \exp\left[-\arccos(\text{clip}(x^\top p, -1, 1))^2/(2\sigma^2)\right]$ with calibrated bandwidth $\sigma = 1.0967$. We evaluate bidirectional cross-validation ($A \to B$ and $B \to A$).

D.2 Breadth and Coverage Density

Table 6 reports pure-dictionary coverage scores G across all 12 residual sites. Across all sites and both evaluation directions, SVD achieves substantially higher coverage density than matched SAE dictionaries at low feature budgets: mean SVD/SAE scores are 0.2278 / 0.1560 at $B = 16$ ($+46.0\%$ relative gain, $\Delta G = +0.0718$) and 0.3629 / 0.2871 at $B = 32$ ($+26.4\%$ relative gain, $\Delta G = +0.0758$). SVD outperforms SAE in 100% (24/24) of directional evaluations at $B = 16$ and $B = 32$, and maintains an advantage at $B = 64$ (0.5744 vs. 0.5388, $+6.6\%$).

Model	Layer	$B = 16$	$B = 32$	$B = 64$
Gemma 3 1B	7	0.2800 / 0.1385	0.4227 / 0.2695	0.5771 / 0.5221
	13	0.2685 / 0.1732	0.4080 / 0.3046	0.5778 / 0.5569
	17	0.2574 / 0.1478	0.3919 / 0.2791	0.5788 / 0.5327
	22	0.2382 / 0.1381	0.3709 / 0.2687	0.5792 / 0.5217
Gemma 3 4B	9	0.2585 / 0.1622	0.4040 / 0.2933	0.5793 / 0.5451
	17	0.2378 / 0.2109	0.3722 / 0.3431	0.5788 / 0.5840
	22	0.2237 / 0.1595	0.3559 / 0.2905	0.5799 / 0.5432
	29	0.2180 / 0.1377	0.3497 / 0.2686	0.5794 / 0.5217
Qwen3 8B	9	0.2183 / 0.1488	0.3514 / 0.2796	0.5807 / 0.5324
	18	0.1876 / 0.1478	0.3196 / 0.2787	0.5707 / 0.5311
	24	0.1745 / 0.1625	0.3058 / 0.2936	0.5566 / 0.5456
	31	0.1710 / 0.1453	0.3026 / 0.2762	0.5540 / 0.5288
Mean		0.2278 / 0.1560	0.3629 / 0.2871	0.5744 / 0.5388

Table 6: **Held-out pure-dictionary coverage by residual site.** Each entry reports SVD G / SAE G, averaged over $A \to B$ and $B \to A$ with equal site weighting.

Replacement budget	ΔG	Relative gain
$B = 16$	$+0.0193$	$+12.4\%$
$B = 32$	$+0.0312$	$+10.9\%$
$B = 64$	$+0.0133$	$+2.5\%$

Table 7: **Coverage gain from SAE-core extension.** SVD features augment a fixed SAE core at replacement fractions $q \in \{0.05, 0.10, 0.20\}$, averaged over all 12 residual sites and both directions.

Table 7 evaluates the conditional novelty of SVD features when augmenting a fixed SAE dictionary core. Substituting a fraction q of SAE features with SVD features consistently improves semantic reach, yielding an average relative improvement of $+11.4\%$ across low budgets ($B \in \{16, 32\}$).

D.3 Topic Preferences and Specific Concepts

Relative topic preference. For each representation $r \in \{\mathrm{SVD}, \mathrm{SAE}\}$ and site s, let $n_{r,t,s}$ count estimable features significantly associated with topic t ($\widehat{p} \le 0.01$). Smoothed and log-centered allocation rates are defined as

$$\widetilde{\rho}_{r,t,s} = \frac{n_{r,t,s} + 0.5}{N_{r,s} + 1}, \qquad P_{r,t,s} = \log_2 \widetilde{\rho}_{r,t,s} - \frac{1}{73} \sum_{u=1}^{73} \log_2 \widetilde{\rho}_{r,u,s}. \tag{12}$$

Relative topic preference is given by the median contrast across the 12 evaluation sites:

$$D_t = \mathrm{median}_s \, P_{\mathrm{SVD},t,s} - \mathrm{median}_s \, P_{\mathrm{SAE},t,s}, \tag{13}$$

where positive D_t indicates SVD preference (blue) and negative D_t indicates SAE preference (orange). Across all 66 pairwise site comparisons, topic preference profiles exhibit high rank correlation (median Spearman $\rho = 0.805$). The 15 topics with the strongest preferences retain identical signs across all 12 evaluated sites.

Complete topic map and cross-layer consistency. Figure 11 displays the complete 73-topic semantic preference map, with full descriptions, contrasts D_t, and validation sizes listed in Table 8. Strongest SVD preferences concentrate on structured, formal domains: code and structured data ($D_{36} = +4.230$), software development ($D_{42} = +3.683$), and web development ($D_{01} = +3.564$). Strongest SAE preferences concentrate on commerce ($D_{44} = -3.605$), education ($D_{26} = -2.888$), and social discourse ($D_{50} = -2.753$).

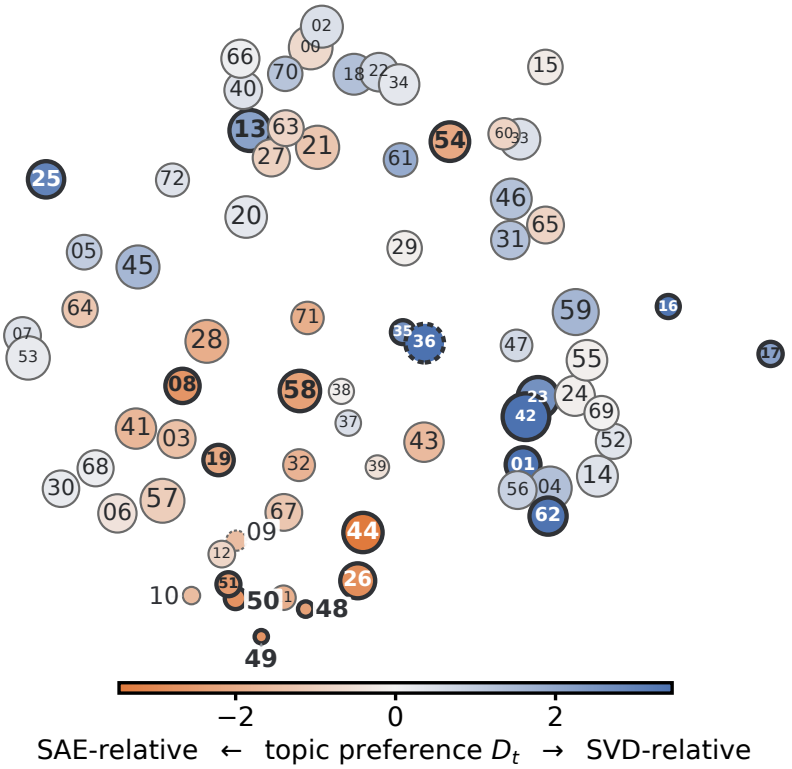


Figure 11: **Complete numbered 73-topic semantic map.** Centroid cosine distances determine 2D layout; validation frequency determines circle area; color reflects relative preference D_t (blue: SVD-favored, orange: SAE-favored). Full topic names and statistics are detailed in Table 8.

Table 8: **Complete 73-topic map key.** Topic IDs match Figure 11; $D_t > 0$ favors SVD and $D_t < 0$ favors SAE. Circle area in the map follows topic map validation prevalence. † marks the small audited leaf; ‡ marks author-inspected descriptive reference labels for the two clusters where automated LLM audit did not reach consensus.

ID	Topic display name	D_t	Validation windows
00	Physiology & pharmacology	-0.706	925
01	CSS/HTML/JavaScript	+3.564	588
02	Cell biology	+0.466	853
03	Events	-1.425	698
04	Frontend Development	+1.350	905
05	Geography	+1.014	582
06	Sports news and discussion	-0.452	712
07	Sports biographies and statistics	+0.411	655
08	Accommodation	-2.669	570
09	Informal narratives & personal anecdotes‡	-1.525	198
10	Parenting & family	-1.548	147
11	Personal Reflections	-1.909	291
12	Personal Blogs	-0.785	345
13	Medical Research	+2.143	814
14	SQL & database queries	+0.412	811
15	Semiconductor manufacturing	-0.215	577
16	Apache License	+3.338	257
17	Software License	+2.205	284
18	Molecular biology	+1.329	809
19	Food	-2.084	447
20	Healthcare	+0.277	838
21	Surgery	-1.253	906

ID	Topic display name	D_t	Validation windows
22	Chemistry	+0.667	721
23	Source Code	+2.596	811
24	Programming semantics & object behavior	-0.108	777
25	US Court of Appeals	+2.986	639
26	Education	-2.888	593
27	Neurology & psychiatry	-0.921	674
28	Economy & finance	-1.953	889
29	Scientific Research	-0.112	575
30	Music	+0.205	646
31	Math exercises & calculations	+1.304	713
32	Video Games	-1.794	491
33	Mechanical Devices	+0.473	805
34	Chemistry & Materials Science	+0.291	792
35	Code & symbols	+2.654	277
36	Code & structured-data artifacts‡	+4.230	724
37	Website policies & terms	+0.533	316
38	Business	-0.085	307
39	Online Community Forum	-0.505	267
40	Microbiology & infectious diseases	+0.411	688
41	Crime & law enforcement	-1.692	789
42	Software development	+3.683	1,055
43	Technology & Computing	-1.599	754
44	Business & entrepreneurship	-3.605	749
45	Legal proceedings	+1.567	897
46	Math theory & derivations	+1.290	808
47	Graphics	+0.657	485
48	Blog posts & online discussions	-2.350	115
49	Social relationships & interactions†	-2.677	86
50	Social Media & Online Discourse	-2.753	224
51	Entertainment & Media	-2.478	279
52	File I/O	+0.373	598
53	Biographies	+0.192	912
54	Manufacturing & engineering	-2.109	721
55	Data processing & I/O programming	-0.122	799
56	Backend Development	+0.867	695
57	Politics & governance	-1.039	934
58	Product Descriptions	-2.336	793
59	C/C++ source code	+1.595	1,025
60	Optical imaging and display	-0.863	478
61	Physics & mathematics	+1.902	547
62	Mobile App Development	+3.388	670
63	Medical treatment	-0.823	623
64	Historic Buildings & Locations	-1.224	605
65	Telecommunications	-0.854	658
66	Cancer Research	+0.087	705
67	Christianity	-1.250	655
68	Film	+0.159	632
69	Regular Expressions	-0.038	567
70	Cardiovascular medicine	+1.219	574
71	Vehicles	-1.934	510
72	Biology & Zoology	+0.429	528

Table 9 summarizes cross-layer preference correlations. Within the same model, layers exhibit remarkable consistency (median $\rho = 0.931$); across different models, substantial alignment persists (median $\rho = 0.741$). Pairwise correlation details across all 66 site combinations are provided in Table 10.

Site pairs	Pairs	Median Spearman ρ	Range
All model–layer sites	66	0.805	0.411–0.982
Layers within one model	18	0.931	0.852–0.982
Sites from different models	48	0.741	0.411–0.953

Table 9: **Cross-layer consistency of topic preferences.** Spearman rank correlations compare 73-dimensional preference vectors $D_{t,s}$ across residual sites.

Site	Gemma 3 1B				Gemma 3 4B				Qwen3 8B			
	L7	L13	L17	L22	L9	L17	L22	L29	L9	L18	L24	L31
G1 L7	1	—	—	—	—	—	—	—	—	—	—	—
G1 L13	0.904	1	—	—	—	—	—	—	—	—	—	—
G1 L17	0.915	0.964	1	—	—	—	—	—	—	—	—	—
G1 L22	0.929	0.894	0.956	1	—	—	—	—	—	—	—	—
G4 L9	0.863	0.770	0.759	0.801	1	—	—	—	—	—	—	—
G4 L17	0.867	0.830	0.809	0.827	0.933	1	—	—	—	—	—	—
G4 L22	0.920	0.911	0.940	0.930	0.874	0.903	1	—	—	—	—	—
G4 L29	0.891	0.852	0.912	0.953	0.852	0.867	0.951	1	—	—	—	—
Q8 L9	0.728	0.721	0.776	0.749	0.587	0.620	0.773	0.744	1	—	—	—
Q8 L18	0.562	0.619	0.686	0.627	0.411	0.464	0.667	0.615	0.940	1	—	—
Q8 L24	0.614	0.678	0.734	0.663	0.435	0.487	0.704	0.641	0.938	0.982	1	—
Q8 L31	0.733	0.738	0.817	0.801	0.580	0.603	0.811	0.799	0.941	0.921	0.934	1

Table 10: **Pairwise topic-preference consistency at all 12 residual sites.** Entries are Spearman correlations across the 73 topic-wise $D_{t,s}$ contrasts. G1/G4 abbreviate Gemma 3 1B/4B PT; Q8 abbreviates Qwen3-8B-Base. Only the lower triangle is shown; the diagonal is one. Values are rounded to three decimals.

SAE substitute test. To identify specific SVD concepts that cannot be substituted by learned SAE features, we first filter SVD directions for stable semantic concentration across document-disjoint folds A and B. Let $m(f, u, h)$ denote the document-balanced fraction of feature f's top contexts assigned to topic u in fold h, and c_t be topic t's normalized centroid. With centroid cosine distance $d(t, u) = 1 - \cos(c_t, c_u)$ and calibrated topic bandwidth $\tau_0 = \text{median}_t \min_{u \neq t} d(t, u) = 0.205076$, kernelized semantic concentration (KSC) measures contextual mass around the optimal topic center:

$$\text{KSC}(f, h) = \max_t \sum_{u=1}^{73} m(f, u, h) \exp\left[-\frac{d(t, u)}{\tau_0}\right]. \tag{14}$$

To ensure sharp local specificity alongside neighborhood concentration, we additionally evaluate nearest-topic purity $\text{NTP}(f, h) = \max_t\{m(f, t, h) + m(f, n(t), h)\}$, where $n(t)$ is topic t's nearest neighboring centroid. Candidates are selected via two tiers: a strict tier ($\text{KSC}, \text{NTP} \geq 0.70$, ≥ 20 documents/fold, projected prototype cosine ≥ 0.80) and a topic-expansion tier ($\text{KSC}, \text{NTP} \geq 0.55$, ≥ 10 documents/fold, projected cosine ≥ 0.75), with both tiers requiring original-dimensional cross-fold prototype stability ≥ 0.85.

For each selected SVD feature, candidate SAE substitutes are retrieved from the full estimable SAE pool via prototype cosine, reranking the top 128 in full embedding space to obtain the three closest independent latents. An independent, blinded Llama 3.1 70B Instruct judge inspects activating contexts from document-disjoint folds across three runs. An SVD feature is marked coherent if at least four of its six contexts isolate a specific concept, and each SAE latent is classified as a `full`, `partial`, `none`, or `unclear` substitute. A feature is verified as having no SAE substitute when consensus confirms SVD coherence and assigns ≥ 2 votes of `partial` or `none` to each SAE neighbor. Across 2,875 judged feature–target-layer pairs, 1,519 instances (52.8%, spanning 925 unique signed SVD source features) have no full substitute among their nearest SAE latents.

Model and layers	Pool ratio	ΔG_{16}	ΔG_{32}	ΔG_{ext}	No full SAE substitute / judged pairs
Gemma 3 1B, 7/13/17/22	0.13–0.38	+0.1116	+0.1179	+0.0327	417/813
Gemma 3 4B, 9/17/22	0.35–0.81	+0.0625	+0.0684	+0.0249	352/657
Gemma 3 4B, 29	1.05	+0.0803	+0.0811	+0.0263	101/245
Qwen3 8B, 9/18/24/31	2.50–8.00	+0.0367	+0.0378	+0.0179	649/1,160

Table 11: **Semantic coverage and concept specificity stratified by feature pool size.** Pool ratio indicates SVD candidate directions divided by SAE latents (Table 4). ΔG denotes held-out pure-dictionary coverage gains; ΔG_{ext} denotes SAE-core extension gains; the final column reports the ratio of judged SVD features without full SAE substitute.

Crucially, as shown in Table 11, these advantages are not driven by candidate pool size. Across all seven sites where the full SVD candidate pool is smaller than the SAE pool (Gemma 3 1B and Gemma 3 4B layers 9, 17, 22), SVD maintains decisive pure-dictionary gains ($\Delta G_{16} \in [+0.0625, +0.1116]$), consistent core-extension gains ($\Delta G_{\text{ext}} \in [+0.0249, +0.0327]$), and high proportions of unique concepts (769 of 1,470 judged pairs lack an SAE substitute). This confirms that intrinsic SVD features provide structurally distinct semantic reach that complements learned dictionaries.

Table 12 presents eight representative case studies illustrating concepts cleanly resolved by SVD features where nearest SAE latents blur into broad categories or drift into adjacent topics.

Table 12: **Additional SVD-specific cases and judge explanations.** A/B are document-disjoint folds. Each text is a longer excerpt of one activating window in that fold; ellipses mark omitted window text. The closest SAE is shown; the three majority labels refer to three distinct retrieved SAE latents. Judge reasons are quoted from each of the three generation runs. The full input comprises six SVD windows and four windows for each SAE latent. Source layers, attention heads, and SVD axes are zero-based; the axis sign gives the activating polarity, and the arrow identifies the target residual layer. Raw artifact keys are retained in the accompanying case-index file. The GNU GPL and unit-conversion cases appear in Figure 4.

Fold	SVD activating text	Closest SAE activating text
Case 1: 3D scene rendering code Gemma 3 1B: L6 attn head 1, SVD axis $42^{-} \rightarrow$ residual L13 SAE 1/2/3: partial / none / none		
A	… const instancedMesh = new THREE.Mesh(instancedGeometry.geometry, material); instancedMesh.frustumCulled = false; instancedMesh.castShadow = true; …	… Slug does all the work in a fragment shader. If so, this would make its closest relative the Will Dobbie vector texture approach [1]. I considered the Dobbie …
B	… Seems odd I would have to do this to render a static 3D scene. /* globals THREE, window, document */ var scene = new THREE.Scene(); var camera = new THREE.PerspectiveCamera(75, window.innerWidth	ctx.drawImage(this, 300, 0, 200, 200, 0, 0, 400, 400); } var ctx = document.getElementById('output').getContext('2d'); ctx. webkitImageSmoothingEnabled = ctx. mozImage
Judge reasons: *Run 1:* "SVD contexts share 3D graphics rendering details, but SAE contexts have different concepts" *Run 2:* "SVD contexts share 3D rendering details, but SAE contexts lack a common concept" *Run 3:* "SVD contexts share 3D graphics rendering, but SAE features lack a common concept"		

Table 12: Additional SVD-specific cases and judge explanations (continued).

Fold	SVD activating text	Closest SAE activating text
Case 2: MDA-MB-231 breast-cancer cell line Qwen3 8B: L9 MLP, SVD axis 1472^{+} → residual L18 SAE 1/2/3: none / none / none		
A	expression in MCF-7 and MDA-MB-231 breast cancer cells promotes cell cycle progression, while silencing PRX3 impairs cell …	… Human melanoma cell line LiBr was transfected with recombinant plasmids encoding *mda-7* gene (pCI-Neo-*mda-7*) or with negative control plasmid (pCI-Neo). Two …
B	Chloroquine inhibits tumor-related Kv10.1 channel and decreases migration of MDA-MB-231 breast cancer cells in vitro. Chloroquine (CQ) is an old antimalarial drug …	… carcinoma cell line Huh7, a cell line that supports HCV pseudovirus entry and productive infection …
Judge reasons: *Run 1:* "SVD contexts share a specific cell line, but SAE contexts do not match" *Run 2:* "SVD contexts share a specific cell line, but SAE contexts do not share a common concept with SVD or among themselves." *Run 3:* "SVD contexts share a specific cell line, but SAE contexts do not share a common concept with SVD contexts"		
Case 3: Angular template syntax Qwen3 8B: L3 attn head 17, SVD axis 64^{-} → residual L31 SAE 1/2/3: none / partial / none		
A	… <mat-checkbox [(ngModel)]="isChecked" (change)="checkCheckBoxvalue(isChecked)"> Check me!</mat-checkbox> Add variable timeout on top timeout; checkCheckBoxvalue(event: …	… <PSButton type="button" class="update-qty float-sm-right" :class="classObject" :
B	… Main component template: <evt-bibliography-item *ngFor="let biblCit of biblCits"	… import Tab from '../tabs/Tab.vue' export default { name: 'ThemeSelectTab',
Judge reasons: *Run 1:* "SVD contexts share Angular syntax, but SAE features do not consistently match" *Run 2:* "SVD contexts share Angular template syntax, but SAE contexts do not share a common concept" *Run 3:* "Angular template syntax appears in most contexts"		
Case 4: Dithiothreitol (DTT) Qwen3 8B: L5 MLP, SVD axis 2549^{-} → residual L9 SAE 1/2/3: none / none / none		
A	NaCl, 400 mM; dithiothreitol (DTT), 1 mM; IGEPAL, 1%; glycerol, 20% (Carlo Erba, Italy), plus the protease inhibitor cocktail (Roche, …	… Certain bis-isothiuronium compounds having an intermediate ureylene or ether moiety between isothiuronium moieties are also known as activator-stabilizers. These …
B	enzymes and DNA ligase from Biolabs, and polyethyleneimine, dithiothreitol (DTT), ATP, dATP, and the poorly hydrolyzable ATP analog ATPγS from Sigma. DEAE-, …	… quinoline, isoquinoline, benzothiophene, benzofuran, benzoxazole, benzothiazole, benzimidazole, 5,6-benzoquinoline, 7,8-benzoquinoline, 1,2-benzisoxazole
Judge reasons: *Run 1:* "SVD contexts share a repeated chemical compound" *Run 2:* "SVD contexts consistently mention dithiothreitol (DTT) as a repeated distinguishing detail, but none of the SAE features share this concept." *Run 3:* "SVD contexts share a specific chemical compound, but SAE contexts do not share a specific concept"		

Table 12: Additional SVD-specific cases and judge explanations (continued).

Fold	SVD activating text	Closest SAE activating text
Case 5: Ineffective assistance of counsel Qwen3 8B: L13 attn head 4, SVD axis 71^{-} → residual L18 SAE 1/2/3: none / none / none		
A	… To prove he has received ineffective assistance of counsel, '(t)he defendant must show both that counsel's performance was deficient and that this deficiency prejudiced the …	deprived of the effective assistance of counsel and his pleas were not made knowingly and intelligently because his trial counsel misled him as to the range of punishment …
B	petition for a writ of habeas corpus, alleging that he had received ineffective assistance of counsel due to Crone's failure to brief fully and to argue specific instances of …	appellant was physically and mentally fit at the time he waived counsel, venue and indictment, pleaded guilty and was sentenced by Judge Kennamer, and, further, that at the …
Judge reasons: *Run 1:* "SVD contexts consistently mention ineffective assistance of counsel, but SAE contexts do not share a common concept" *Run 2:* "SVD contexts consistently mention ineffective assistance of counsel, but SAE contexts do not share a common concept" *Run 3:* "SVD contexts consistently mention ineffective assistance of counsel, but SAE contexts do not share a common concept"		
Case 6: Ephesians 2 Gemma 3 4B: L3 attn head 3, SVD axis 225^{-} → residual L22 SAE 1/2/3: none / none / none		
A	… The author would have you believe 'debt' was nailed to the cross but not the law. Ephesians 2: 15 tells us, whatever was abolished, that act united Jew and Gentile. I ask	In Isaiah 61:1-3, the wake-up called is clear, "The Spirit of the Lord God is upon Me, because the Lord has anointed Me to preach good tidings to the poor; He has …
B	… Ephesians 2:8 - "For by grace are ye saved THROUGH FAITH: and that NOT OF YOURSELVES....." No it is not of ourselves, our faith, our own personal human faith. We are …	prophecy, only God can inspire real prophecy (Isaiah 46:9–10). In light of all of the forgery and occultism associated with Rome, why would the God of the Bible …
Judge reasons: *Run 1:* "SVD contexts share a specific Bible verse, but SAE contexts do not match" *Run 2:* "SVD contexts consistently reference Ephesians 2, but SAE contexts do not share a common concept" *Run 3:* "SVD contexts consistently reference Ephesians 2, but SAE contexts do not share a common concept"		
Case 7: JUnit assertions Gemma 3 1B: L7 attn head 2, SVD axis 167^{-} → residual L7 SAE 1/2/3: none / none / none		
A	… Assert.assertEquals; import static org.junit. Assert.assertFalse; import static org.junit.Assert. assertNotSame; import static org.junit.Assert. assertSame; import …	package com.jack.mall.model; import io. swagger.annotations.ApiModelProperty; import java.io.Serializable; public class PmsProductAttributeValue implements Serializable { private …
B	… import static org.junit.Assert.assertThat; public class GelfUDPSenderTest extends TestCase { @Test public void …	package com.novoda.noplayer.internal. exoplayer.forwarder; import android.view. Surface; import com.google.android.exoplayer2. ExoPlaybackException; import …
Judge reasons: *Run 1:* "SVD contexts share Junit assertions, but SAE features do not" *Run 2:* "SVD contexts share JUnit assertion methods, but SAE features lack a common concept" *Run 3:* "SVD contexts consistently use JUnit assertions, but SAE features do not share a common concept"		

Table 12: Additional SVD-specific cases and judge explanations (continued).

Fold	SVD activating text	Closest SAE activating text
Case 8: Arthroplasty Qwen3 8B: L12 attn head 18, SVD axis $18^{-} \rightarrow$ residual L18 SAE 1/2/3: none / none / none		
A	arthroplasty by the authors between 1982 and 1993 was performed. Sixty-two total knee arthroplasties with late infections (greater than 6 months after their procedure) were …	soft tissue reconstruction was achieved with vertical rectus abdominus myocutaneous pedicled flap. This video is available in the "Related Videos" section of the Full-Text …
B	arthroplasty. A prospective analysis of 33 patients undergoing primary knee arthroplasty performed by 3 surgeons in one institution. The first 12 patients (3 bilateral and 9 …	… The average Knee Society Score was 82.3 +/ - 6.4. Considering its less morbidity, excellent cosmesis, no need for second operation to remove the implant, excellent cost …
Judge reasons: *Run 1:* "SVD contexts share arthroplasty, but SAE contexts do not match" *Run 2:* "SVD contexts share arthroplasty, but SAE contexts do not" *Run 3:* "SVD contexts share arthroplasty as a common concept, but SAE contexts do not match"		

E Sparse Concept Prediction

E.1 Evaluation Protocol and Tasks

Task collection and family grouping. Following standard SAE-Probes conditions (Kantamneni et al., 2025; Karvonen et al., 2025), we evaluate linear concept predictivity across 113 native binary classification tasks under class-balanced conditions. The benchmark spans diverse conceptual domains: historical facts, geographic entities, Wikidata attributes, news topics and headlines, programming languages, arithmetic, GLUE linguistic judgments, sentiment ratings, content moderation, AI-generated text, ethics, and biomedical categories (diseases, cancer types). For domain-level analysis (Figure 5b and Table 13), the 113 tasks are grouped into 21 coherent families. Each task receives equal weight in aggregate AUC calculations, and family scores represent the unweighted mean over constituent tasks.

Sampling, splits, and feature extraction. Each task uses up to 5,000 examples, randomly partitioned into class-balanced training and held-out test splits (seed 42), with training set size capped at 1,024 examples. Input sequences are left-truncated to 1,024 tokens. Activations are extracted from the final non-padding token at the target residual stream layer, then encoded via the frozen residual SAE or projected onto the cumulative SVD dictionary axes (Appendix A). SVD and SAE dictionaries are evaluated on identical texts, labels, splits, and layers.

Coordinate ranking and sparse probe fitting. For each dictionary coordinate j, activations z_{ij} are normalized by their mean absolute non-zero activation m_j for ranking:

$$s_j = \left|\mathbb{E}[z_{ij}/m_j \mid y_i = 1] - \mathbb{E}[z_{ij}/m_j \mid y_i = 0]\right|, \qquad I_k = \mathrm{TopK}_j(s_j). \tag{15}$$

An L_1-regularized logistic regression probe is fitted on the top-k unnormalized activations, where $k \in \{1, 2, 5, 16, 128\}$ specifies the probe capacity. Regularization strength is tuned via validation AUC over 10 logarithmically spaced values in $[10^{-5}, 10^{5}]$, and the selected model is refit on the full training split. We report held-out test ROC–AUC averaged equally across the 113 tasks. The $k = 1$ setting evaluates single-feature concept isolation, while larger k tests compositional linearly accessible information.

Evaluated models and layers. We evaluate 16 model–layer sites across Gemma 3 1B, 4B, 27B PT, and Qwen3-8B-Base at four depth-spanning residual layers (Table 1). Gemma sites are benchmarked against the canonical Gemma Scope 2 residual JumpReLU 65k/L_0-big checkpoints, and Qwen3 sites are benchmarked against the public Qwen-Scope TopK 64k ($L_0 = 100$) residual SAE (Appendix B).

E.2 Predictive Performance Across Scales and Concept Families

Comprehensive predictive advantage. Figure 12 displays test ROC–AUC across all 16 model–layer sites for probe budgets $k \in \{1, 2, 5, 16, 128\}$. SVD provides a consistently superior predictive basis across model families and depths. Across all eight sites in Gemma 3 1B and 4B, SVD outperforms SAE by $+0.029$ to $+0.054$ AUC at $k = 5$. In Qwen3-8B, the SVD advantage steadily expands with network depth, rising from near-parity at layer 9 to $+0.019$, $+0.031$, and $+0.037$ AUC gains at layers 18, 24, and 31. At broader readout budgets ($k = 128$), SVD surpasses SAE across all 16 evaluated sites.

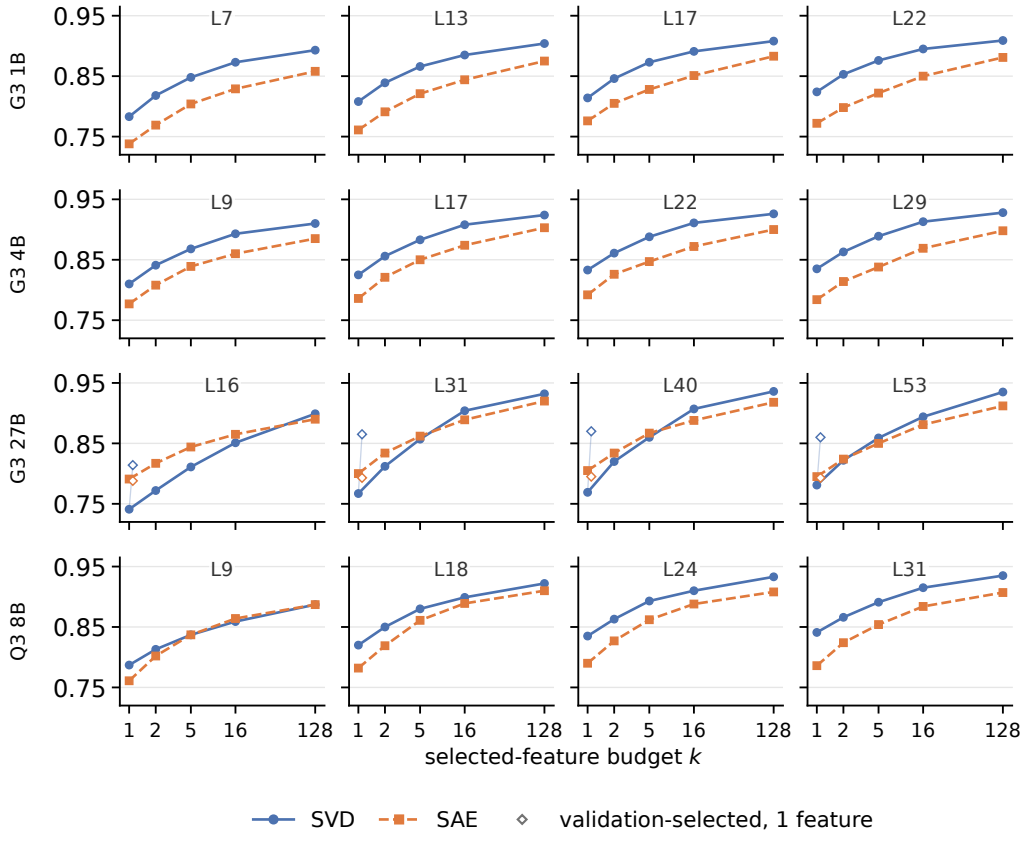


Figure 12: **Sparse prediction across all 16 model–layer sites.** Panels display mean held-out ROC–AUC over 113 tasks across feature budgets $k \in \{1, 2, 5, 16, 128\}$ across depth-spanning layers. Solid lines use standard mean-difference selection. Hollow diamonds in Gemma 3 27B show single-feature performance under 16-candidate validation selection.

Task family specialization. Table 13 details single-feature ($k = 1$) predictive performance across the 21 concept families. SVD achieves higher cross-site mean AUC in 19 of the 21 families, winning all 16 sites on cancer type ($+0.1917$ mean ΔAUC), code language ($+0.1318$), news topic ($+0.1059$), and temporal concepts ($+0.0572$). In contrast, SAE advantages are confined to specialized stylistic niches, primarily art type (-0.0195 mean ΔAUC, where SVD leads in Gemma 3 1B while SAE leads in 27B and Qwen3) and AI-generated text (-0.0027). This confirms that intrinsic SVD features provide a broadly stronger linear representation of core factual and semantic concepts, while learned SAEs occupy specialized complementary niches.

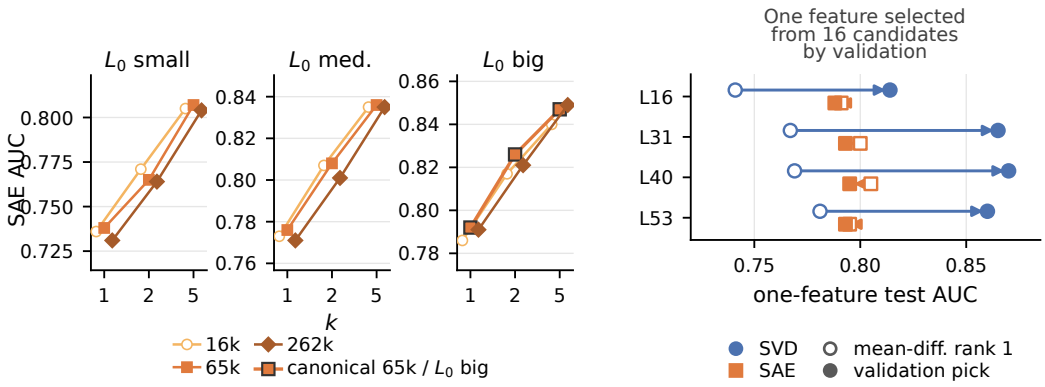
Task family	Tasks	Mean ΔAUC	SVD/SAE sites
cancer type	3	+0.1917	16/0
code lang.	3	+0.1318	16/0
review stars	1	+0.1197	15/1
news topic	6	+0.1059	16/0
IT tickets	3	+0.0926	15/1
GLUE	8	+0.0706	13/3
disease	3	+0.0626	15/1
temporal	4	+0.0572	16/0
QA / judgment	6	+0.0571	14/2
headlines	5	+0.0543	11/5
sentiment	4	+0.0511	15/1
Wikidata	9	+0.0493	15/1
ethics	5	+0.0441	14/2
athlete sport	3	+0.0358	14/2
hist. figures	3	+0.0270	10/6
context type	3	+0.0254	12/4
moderation	6	+0.0212	14/2
misc. topics	21	+0.0203	13/3
geo entities	12	+0.0074	10/6
AI-gen	2	-0.0027	9/7
art type	3	-0.0195	6/10

Table 13: **One-feature prediction by task family.** Mean ΔAUC is SVD minus SAE, averaged across the 16 model–layer sites after averaging tasks within each family. SVD/SAE sites count which dictionary has the higher family AUC at each site. Gemma 3 27B uses validation selection of one feature from 16 candidates; the other 12 sites use the standard mean-difference rank-1 feature.

E.3 SAE Configuration Sweep and Coordinate Selection Dynamics

SAE configuration sweep. To establish a rigorous SAE baseline, we conducted an exhaustive 9-checkpoint sweep over widths $\{16\text{k}, 65\text{k}, 262\text{k}\}$ and target sparsities $\{L_0\text{-small, medium, big}\}$ at Gemma 3 4B layer 22 (Figure 13a). The 65k/L_0-big configuration attained the highest performance among all SAE variants at $k = 1$ and $k = 2$ (0.792 and 0.826 AUC), and matched the widest 262k/L_0-big checkpoint at $k = 5$ (0.847 vs. 0.849). SVD decisively outperformed every tested SAE configuration at this site across all budgets, achieving 0.833, 0.861, and 0.888 AUC at $k \in \{1, 2, 5\}$.

Coordinate selection in large models. In large-scale models such as Gemma 3 27B, the expanded dictionary inventory can make single-feature identification sensitive to simple unnormalized 1D ranking. We examined whether a symmetric validation selection rule resolves this: evaluating the top 16 candidates ranked by mean difference and selecting the single feature with highest validation AUC. When applied identically to both dictionaries, SVD establishes a decisive advantage across all four layers of Gemma 3 27B (Figure 13b): achieving SVD/SAE test AUCs of $0.814/0.788$ at layer 16 ($+0.026$), $0.865/0.793$ at layer 31 ($+0.072$), $0.870/0.795$ at layer 40 ($+0.075$), and $0.860/0.793$ at layer 53 ($+0.067$). Each final probe still utilizes exactly one feature, demonstrating that highly predictive individual singular axes are present throughout the model's depth.



(a) Gemma-3 4B L22 SAE width–L_0 sweep. The outlined point is the canonical 65k/L_0-big checkpoint.

(b) Gemma-3 27B one-feature selection. Hollow markers denote standard rank-1; filled markers denote validation selection.

Figure 13: **SAE sweep and coordinate selection dynamics. Left:** The canonical 65k/L_0-big SAE establishes the strongest overall SAE baseline; SVD outperforms all nine SAE configurations at this layer across budgets. **Right:** Under symmetric validation selection among 16 candidate coordinates, SVD decisively outperforms SAE at all four layers of Gemma 3 27B ($+0.026$ to $+0.075$ AUC gain).

F RAVEL Attribute Disentanglement

F.1 Evaluation Protocol and Interchange Setup

Attribute benchmarks and entity collections. Following Huang et al. (2024), Chaudhary & Geiger (2024), and the SAEBench RAVEL evaluator (Karvonen et al., 2025), we evaluate causal attribute disentanglement via feature-space interchange. We assess two benchmark suites: (1) *Official SAEBench selection*: covers cities (target attributes: Country, Continent, Language) and Nobel Prize winners (Country of Birth, Field, Gender); (2) *Expanded coverage*: spans five distinct entity categories—cities, Nobel Prize winners, occupations, physical objects, and verbs—using all valid local attributes. To guarantee that interventions test facts within the base model's knowledge, prompts are filtered to retain entities whose unmodified greedy completion matches ground truth, selecting up to 500 top entities and 90 templates per attribute. For each target attribute A, we sample up to 5,000 *Cause* pairs (where base and source entities have different A values) and up to 5,000 *Isolation* pairs (where the base prompt queries another attribute of the same entity type). Using seed 42, pairs are stratified into 70% training, 15% validation, and 15% held-out test splits. SVD and SAE dictionaries are evaluated on identical prompts, pairs, splits, and test examples.

Feature interchange and mask optimization. Interventions hook the `resid_post` state after decoder layer L at the final token of the entity span. The base model and dictionary representations remain strictly frozen. For each attribute A, mask logits q_A are learned over dictionary coordinates. During training, a continuous soft mask $m_A = \sigma(q_A/T)$ mixes source and base coordinates as $\tilde{z} = (1 - m_A) \odot z_b + m_A \odot z_s$, with temperature T annealed from 1.0 to 10^{-4}. We optimize cross-entropy on the first continuation token (targeting the source A value for Cause pairs and the base non-target value for Isolation pairs) using Adam (learning rate 10^{-3}, 2 epochs), selecting the mask checkpoint with lowest validation loss. At test time, the mask is binarized ($M_A(i) = \mathbf{1}[q_A(i) > 0]$) and evaluated via greedy six-token generation scored with RAVEL's exact-match evaluation rules.

For an SAE, the patched state is $h_{\text{patch}} = D((1 - M_A) \odot z_b + M_A \odot z_s) + [h_b - D(z_b)]$. For cumulative SVD, coordinates z concatenate signed projections across all upstream attention-head and MLP writes through block L, and the patched state is $h_{\text{patch}} = h_b + D_{\text{SVD}}(M_A \odot (z_s - z_b))$, where D_{SVD} accounts for Gemma 3 post-module RMSNorm scalings (Appendix A). In both representations, unmasked coordinates preserve the base state exactly.

Causal disentanglement metrics. For each target attribute, *Cause* measures the proportion of held-out Cause continuations matching the source attribute value, while *Isolation* measures the

proportion of Isolation continuations preserving the base non-target value. Overall performance is quantified by the joint Disentangle score:

$$D = \frac{\text{Cause} + \text{Isolation}}{2}. \tag{16}$$

Scores are averaged equally across attributes within each entity type, then averaged across entity types. Paired evaluation on identical splits ensures controlled comparisons across representations.

Evaluated models and layers. We evaluate 16 model–layer residual sites across Gemma 3 1B, 4B, 27B PT, and Qwen3-8B-Base at four depth-spanning layers (Table 1). Gemma sites are benchmarked against the canonical Gemma Scope 2 65k/L_0-big residual SAEs (identified via the sweep in Appendix F.5), and Qwen3 sites are benchmarked against the public Qwen-Scope 64k/TopK-100 residual SAE (Appendix B).

F.2 Official Six-Attribute Results Across Depths and Models

Table 14: **SAEBench six-attribute RAVEL selection.** Held-out Cause (C), Isolation (I), and Disentangle (D) for SAE and SVD at each paired residual site. Scores average attributes within city and Nobel Prize winner, then the two entity types equally. Gemma uses 65k/L_0-big SAEs; Qwen uses its public 64k/TopK-100 SAE. All four 27B layers are paired.

Model	Layer	SAE C	SVD C	SAE I	SVD I	SAE D	SVD D
G3 4B	L9	0.696	0.818	0.758	0.814	0.727	0.816
G3 4B	L17	0.535	0.796	0.811	0.868	0.673	0.832
G3 4B	L22	0.350	0.674	0.810	0.896	0.580	0.785
G3 4B	L29	0.003	0.123	0.995	0.994	0.499	0.559
G3 1B	L7	0.508	0.764	0.733	0.787	0.621	0.775
G3 1B	L13	0.502	0.790	0.737	0.833	0.620	0.812
G3 1B	L17	0.059	0.321	0.935	0.898	0.497	0.610
G3 1B	L22	0.001	0.033	0.995	0.997	0.498	0.515
Qwen3 8B	L9	0.575	0.717	0.709	0.720	0.642	0.719
Qwen3 8B	L18	0.676	0.894	0.727	0.870	0.702	0.882
Qwen3 8B	L24	0.043	0.542	0.969	0.931	0.506	0.737
Qwen3 8B	L31	0.001	0.062	0.999	0.999	0.500	0.531
G3 27B	L16	0.650	0.789	0.777	0.804	0.714	0.796
G3 27B	L31	0.479	0.793	0.811	0.904	0.645	0.849
G3 27B	L40	0.025	0.432	0.982	0.957	0.504	0.695
G3 27B	L53	0.002	0.315	0.996	0.993	0.499	0.654

Table 14 presents held-out Cause, Isolation, and Disentangle scores across all 16 residual sites on the official SAEBench benchmark. Cumulative SVD achieves higher Disentangle (D) and higher Cause (C) than matched residual SAEs across all 16 evaluated sites. Averaged across depths, mean SAE/SVD Disentangle is $0.559/0.678$ in Gemma 3 1B ($+0.119$), $0.620/0.748$ in Gemma 3 4B ($+0.128$), $0.591/0.748$ in Gemma 3 27B ($+0.157$), and $0.588/0.717$ in Qwen3-8B ($+0.129$). Across all 16 sites, cumulative SVD achieves a mean Disentangle score of 0.723 versus 0.589 for SAE ($+0.134$), driven by a substantial gain in Cause (0.554 vs. 0.319) alongside superior Isolation (0.892 vs. 0.859).

Crucially, the SVD advantage expands at deeper residual layers where learned SAEs exhibit catastrophic causal failure. At Gemma 3 4B layer 29, Qwen3-8B layer 31, and Gemma 3 27B layer 53, SAE Cause drops to near zero (≤ 0.003), indicating that learned SAE latents fail to steer model predictions, leaving Isolation near 1.0 merely as a passive default. In contrast, cumulative SVD maintains active causal control across all depths, achieving substantial Cause gains (e.g., reaching 0.674 vs. 0.350 at Gemma 3 4B L22, and 0.432 vs. 0.025 at Gemma 3 27B L40).

F.3 Representation Baselines on Gemma-3 4B

Table 15 compares six representation baselines on Gemma 3 4B across layers 9, 17, 22, and 29 on the official six attributes, with full Cause and Isolation decompositions in Table 16. We evaluate:

1. **Residual neurons**: Interchanges coordinates directly in the 2,560-dimensional native residual basis, $h_{\text{int}} = h_b + m_A \odot (h_s - h_b)$.

Table 15: **Gemma-3 4B RAVEL baselines.** Held-out Disentangle on the official six attributes; layers and entity subsets are equally weighted. MDAS is a validation-selected, task-supervised reference. Bold marks the best score at each layer.

Method	L9	L17	L22	L29	Mean
Cumulative SVD	**0.816**	**0.832**	**0.785**	0.559	0.748
Residual SAE	0.727	0.673	0.580	0.499	0.620
Residual neurons	0.730	0.663	0.588	0.499	0.620
Residual PCA	0.627	0.593	0.534	0.500	0.563
Cumulative module output neurons	0.799	0.796	0.753	0.528	0.719
MDAS (rank 128)	0.804	0.827	0.768	**0.587**	0.746

Table 16: **Components of the Gemma 3 4B six-method RAVEL comparison.** Each cell gives held-out Cause/Isolation on the official six attributes, with city and Nobel Prize winner equally weighted after averaging their three attributes. The corresponding Disentangle scores are in Table 15.

Method	L9 C/I	L17 C/I	L22 C/I	L29 C/I
SVD	0.818/0.814	0.796/0.868	0.674/0.896	0.123/0.994
Res. SAE	0.696/0.758	0.535/0.811	0.350/0.810	0.003/0.995
Res. neurons	0.710/0.750	0.563/0.763	0.347/0.829	0.001/0.997
Res. PCA	0.591/0.663	0.502/0.684	0.302/0.766	0.001/0.999
Module neurons	0.803/0.795	0.743/0.849	0.619/0.888	0.059/0.997
MDAS	0.784/0.824	0.800/0.853	0.647/0.890	0.179/0.995

2. **Residual PCA**: Standardizes entity-token training activations by mean μ and standard deviation s, computing a full-rank orthogonal basis U from the correlation matrix: $h_{\text{int}} = h_b + s \odot U[m_A \odot U^\top((h_s - h_b) \oslash s)]$.

3. **Cumulative module output neurons**: Retains raw attention and MLP writes from upstream blocks $0, \ldots, L$ in the native residual basis, accumulating $2(L+1) \times 2560$ coordinates.

4. **Residual SAE**: Gemma Scope 2 65k/L_0-big residual SAE.

5. **Cumulative SVD**: Decomposes attention and MLP outputs into intra-module singular vectors, retaining head-specific singular axes.

6. **MDAS**: Directly learns an attribute-specific orthogonal projector $QQ^\top$ ($Q \in \mathbb{R}^{2560\times 128}$, $Q^\top Q = I$) optimized under task supervision via counterfactual cross-entropy.

These comparisons provide three fundamental insights:

- **Decoupling cross-depth accumulation from basis orientation**: Cumulative module output neurons (0.719 mean D) substantially outperform local residual representations (residual neurons 0.620, residual SAE 0.620, and residual PCA 0.563). This demonstrates that preserving source-specific write coordinates across depth prevents attribute representations from collapsing into an entangled residual bottleneck.

- **Causal efficacy of singular vector decomposition**: Cumulative SVD consistently surpasses cumulative module output neurons across all four layers ($+0.017, +0.036, +0.031, +0.031$, four-layer mean 0.748 vs. 0.719). This demonstrates that intra-module singular value decomposition diagonalizes write covariance, resolving collinearity across attention heads and providing an optimal orthogonal coordinate frame for selective intervention.

- **Parity with task-supervised subspaces**: Training-free, unsupervised cumulative SVD matches or exceeds the performance of fully supervised MDAS (0.748 vs. 0.746), winning at layers 9, 17, and 22. This confirms that intrinsic SVD coordinates naturally align with true causal attribute mechanisms without requiring counterfactual training.

F.4 Expanded Five-Entity Benchmark

Table 17: **Expanded five-entity RAVEL coverage.** Held-out Cause (C), Isolation (I), and Disentangle (D) for SAE and SVD at each paired residual site. Scores average common evaluable attributes within each of city, Nobel Prize winner, occupation, physical object, and verb, then average the five entity types equally. SAE checkpoints match Table 14; all four 27B layers are paired.

Model	Layer	SAE C	SVD C	SAE I	SVD I	SAE D	SVD D
G3 4B	L9	0.460	0.583	0.636	0.666	0.548	0.624
G3 4B	L17	0.304	0.531	0.701	0.724	0.502	0.627
G3 4B	L22	0.210	0.367	0.781	0.825	0.496	0.596
G3 4B	L29	0.009	0.089	0.990	0.986	0.499	0.537
G3 1B	L7	0.371	0.494	0.653	0.658	0.512	0.576
G3 1B	L13	0.295	0.487	0.679	0.692	0.487	0.590
G3 1B	L17	0.058	0.196	0.933	0.896	0.496	0.546
G3 1B	L22	0.012	0.052	0.987	0.991	0.499	0.522
Qwen3 8B	L9	0.465	0.546	0.594	0.625	0.529	0.586
Qwen3 8B	L18	0.418	0.630	0.633	0.735	0.525	0.683
Qwen3 8B	L24	0.178	0.404	0.821	0.860	0.499	0.632
Qwen3 8B	L31	0.018	0.067	0.986	0.989	0.502	0.528
G3 27B	L16	0.461	0.570	0.664	0.674	0.563	0.622
G3 27B	L31	0.293	0.608	0.756	0.789	0.525	0.699
G3 27B	L40	0.097	0.368	0.903	0.907	0.500	0.638
G3 27B	L53	0.008	0.178	0.992	0.980	0.500	0.579

Table 17 reports results on the expanded RAVEL benchmark across five entity categories: cities, Nobel Prize winners, occupations, physical objects, and verbs. Cumulative SVD consistently outperforms residual SAEs across all 16 model–layer sites in both Cause and Disentangle. Four-layer mean Disentangle improves from 0.498 to 0.558 in Gemma 3 1B ($+0.060$), 0.511 to 0.596 in Gemma 3 4B ($+0.085$), 0.522 to 0.634 in Gemma 3 27B ($+0.112$), and 0.514 to 0.607 in Qwen3-8B ($+0.093$). Across all 16 sites, SVD achieves a mean Disentangle score of 0.599 versus 0.511 for SAE, driven by an average Cause gain from 0.229 to 0.386 ($+0.157$) while maintaining higher Isolation (0.812 vs. 0.794). This confirms that SVD's causal superiority generalizes across diverse syntactic, lexical, and entity domains.

F.5 SAE Width and Sparsity Sweep

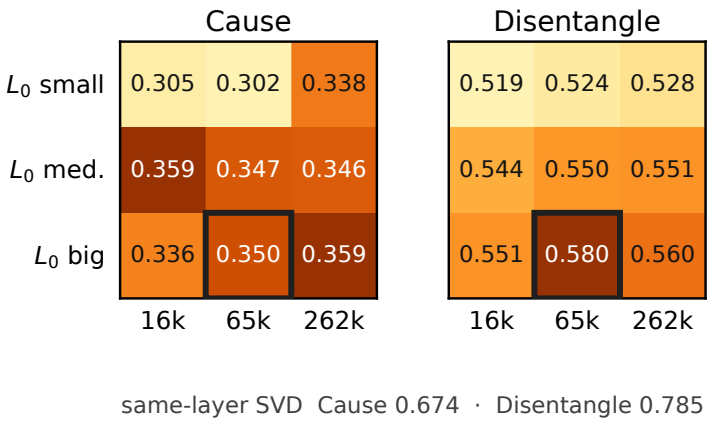


Figure 14: **Gemma 3 4B layer-22 SAE sweep on the official six RAVEL attributes.** Heatmaps display held-out Cause and Disentangle for nine residual SAE configurations: widths $\{16\text{k}, 65\text{k}, 262\text{k}\}$ crossed with target $L_0 \in \{20, 60, 150\}$. The outlined 65k/L_0-big cell achieves the highest SAE Disentangle (0.580). Same-layer SVD achieves Cause 0.674 and Disentangle 0.785.

To ensure a competitive baseline, we evaluated all nine combinations of widths $\{16\text{k}, 65\text{k}, 262\text{k}\}$ and target sparsities $\{L_0\text{-small, medium, big}\}$ at Gemma 3 4B layer 22 under identical task pairs and evaluation protocols (Figure 14). The 65k/L_0-big checkpoint achieved the highest Disentangle score (0.580) among all nine SAE variants, outperforming the wider 262k/L_0-big model (0.560). In expanded coverage, 65k/L_0-big likewise achieved the highest SAE Disentangle (0.496). We therefore established 65k/L_0-big as the canonical SAE configuration across Gemma sites. SVD dramatically outperforms every SAE configuration in the sweep, achieving Cause of 0.674 and Disentangle of 0.785 at this layer.

G Factual Relation Concept Discovery and Intervention

G.1 Relation Datasets and Causal Steering Setup

Relation benchmarks and Synth200 construction. To evaluate relational knowledge representation, we benchmark 80 published relations from Hernandez et al. (2024) and Christ et al. (2025), alongside Synth200, a newly constructed benchmark expanding relation coverage across 12 semantic domains: geography, biology, science, technology, linguistics, culture, history, functional objects, materials, mathematics, social institutions, and procedural sequences. Synth200 was generated via a multi-stage protocol using `gpt-5.6-sol` and `gpt-5.6-terra` with structured constraints ensuring stable subject–object types, non-overlapping schema identifiers, and clean prompt templates. Across all 200 accepted relation files, Synth200 comprises 32,422 unique subject–object pairs (25–260 pairs per relation).

Subject filtering and split protocol. To ensure that causal interventions evaluate knowledge already stored within the model rather than inducing hallucinated facts, we apply a strict clean-recall filter: retaining only records where the target object's first token is among the base model's top-5 next-token predictions on clean relation prompts. Relations with fewer than four valid records are omitted. Evaluated subjects are partitioned into disjoint training and held-out test splits. For each evaluation subject s, interventions are tested across 100 relation-replacement prompts $T_{\text{random}}(s, q)$, where an unrelated relation q is injected while evaluating whether steering recovers the original relation's target object.

IG feature discovery and stability scoring. To identify features causally mediating the relation, we apply Integrated Gradients (IG) along a 64-step path from a Gaussian-corrupted embedding baseline $E^{(0)}$ to the clean prompt embedding E:

$$\bar{A}_j = \frac{1}{|S_{\text{train}}|} \sum_{s \in S_{\text{train}}} \sum_{m=1}^{64} \langle g_j(E(t_m)) - g_j(E(t_{m-1})), \nabla_{g_j} \log p_\theta(y \mid E(t_m)) \rangle, \tag{17}$$

where g_j denotes the candidate feature write at the final prompt token. To favor consistent mechanisms over noisy activations, features are ranked by a stability-adjusted score:

$$S_j = \frac{\bar{A}_j}{1 + V_j / \max(\widetilde{V}, 10^{-8})}, \tag{18}$$

where V_j is the candidate's write variance across training subjects and $\widetilde{V}$ is the median variance.

Vector composition and causal injection. For a clean training record s_{rec}, the top-k ranked SVD features are composed into a relation steering vector $v^{\text{SVD}}_{r,L,k} = \sum_{j=1}^{k} c_{\pi_j}(T_r(s_{\text{rec}}))u_{\pi_j}$ (incorporating Gemma 3 post-module RMSNorm scalings, Appendix A). The SAE counterpart similarly sums active latent decoder writes $v^{\text{SAE}}_{r,L,k} = \sum_{j=1}^{k} z_{\pi_j}(T_r(s_{\text{rec}}))d_{\pi_j}$. At the final prompt token of a relation-replacement query, the normalized vector is injected into residual layer L with steering strength α:

$$h'_L = h_L + \alpha \|h_L\|_2 \frac{v_{r,L,k}}{\|v_{r,L,k}\|_2}. \tag{19}$$

A symmetric two-stage search (coarse grid over 420 configurations, followed by refine search over 2,400 configurations) optimizes (L, k, α) purely on training subjects to maximize Acc@1 and Acc@5, respectively. The selected configuration is then frozen and evaluated on held-out test subjects.

Table 18: **Original Given-template relation intervention.** Held-out scores after choosing separate configurations on training subjects for Acc@1 and Acc@5. Each row macro-averages n paired relations equally; Δ is SVD minus SAE. The overall row weights all 794 relation–model–collection units equally.

Model	Relations	n	Acc@1			Acc@5		
			SVD	SAE	Δ	SVD	SAE	Δ
Gemma 3 1B	Prev. Work	58	0.741	0.639	+0.102	0.861	0.791	+0.070
	Synth200	99	0.605	0.470	+0.135	0.753	0.633	+0.120
Gemma 3 4B	Prev. Work	68	0.716	0.615	+0.101	0.883	0.774	+0.109
	Synth200	142	0.686	0.500	+0.186	0.851	0.679	+0.172
Gemma 3 27B	Prev. Work	69	0.803	0.638	+0.165	0.950	0.823	+0.127
	Synth200	154	0.806	0.589	+0.217	0.937	0.768	+0.169
Qwen3 8B	Prev. Work	66	0.834	0.782	+0.051	0.934	0.904	+0.030
	Synth200	138	0.766	0.715	+0.051	0.916	0.886	+0.029
Overall	Paired macro	794	0.742	0.606	+0.136	0.886	0.774	+0.112

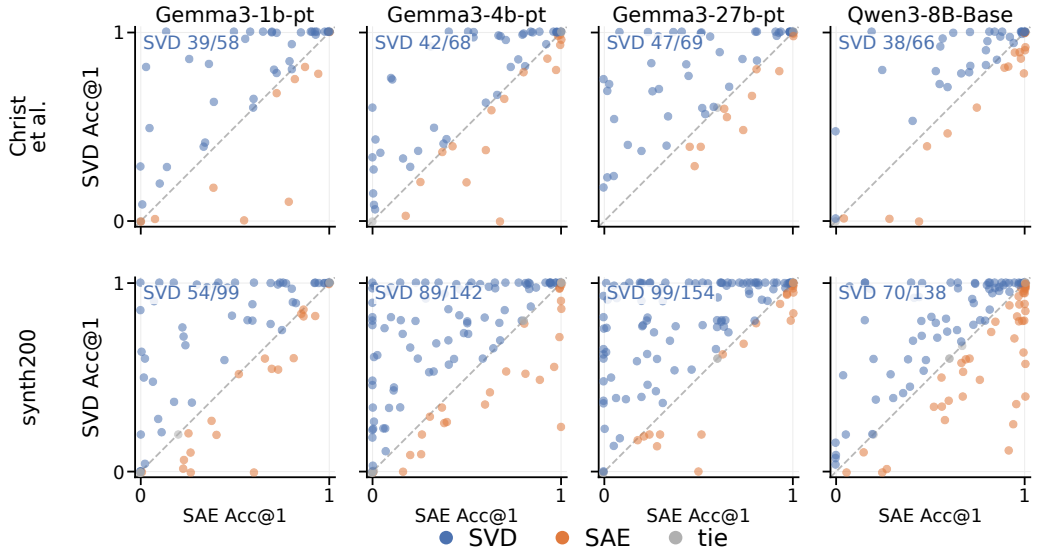


Figure 15: **Per-relation original-template Acc@1.** Scatter plots compare held-out Acc@1 for each of the 794 paired relation–model units (blue: SVD win, orange: SAE win, gray: tie). SVD wins the vast majority of relations across all models and benchmarks.

G.2 Factual Steering Accuracy on Standard Templates

Table 18 reports held-out steering accuracy across all eight model–collection evaluations on the standard Given-style template. Composed cumulative SVD vectors achieve substantial, statistically rigorous gains over learned SAEs across every evaluated model and collection. Overall, SVD improves top-1 accuracy (Acc@1) from 0.606 to 0.742 (+0.136 margin, 95% bootstrap CI $[+0.117, +0.155]$), and top-5 accuracy (Acc@5) from 0.774 to 0.886 (+0.112 margin, 95% bootstrap CI $[+0.094, +0.130]$). SVD wins in all eight evaluation groups, with Acc@1 margins reaching +0.217 in Gemma 3 27B on Synth200.

Figure 15 illustrates the distribution of individual relation outcomes across all 794 evaluated units, showing that SVD decisively outperforms SAE across diverse relation types. Figure 16 summarizes the macro-averaged gains across all model groups.

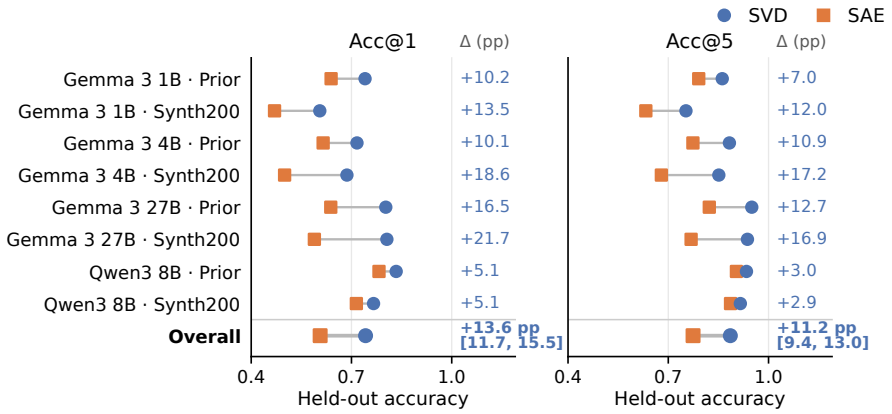


Figure 16: **Original-template SVD–SAE held-out Acc@1/Acc@5 across model groups.** Paired dumbbells show macro averages across all eight model–collection groups, demonstrating consistent +13.6 and +11.2 percentage point gains for SVD with 95% bootstrap intervals.

Table 20: **Relation-level paired baseline outcomes.** On the 68 prior relations in Gemma 3 4B, each entry is SVD wins/ties/baseline wins under each method's separately train-selected configuration. Every triplet sums to 68.

Baseline	Acc@1	Acc@5
Residual neurons	39/12/17	27/33/8
Residual PCA	43/10/15	33/24/11
SVD (one layer)	36/10/22	27/27/14
Module neurons	27/16/25	21/38/9
Weight columns	29/17/22	12/39/17

G.3 Representation Baselines on Gemma-3 4B

Table 19: **Relation intervention baselines on Gemma-3 4B PT.** Held-out first-token accuracy, averaged equally over 68 prior relations on the original given-style evaluation template. Each method uses its own train-selected configuration. The SAE row comes from the matched cross-model comparison. All percentages are rounded to two decimal places.

Representation	Acc@1 (%)	Acc@5 (%)
SVD	**71.65**	**88.29**
Residual SAE	61.50	77.40
Residual neurons	65.81	81.03
Residual PCA	58.03	75.07
SVD (one layer)	62.91	81.93
Module neurons	71.63	83.83
Weight columns	70.37	86.42

Table 19 benchmarks seven representation alternatives on Gemma 3 4B across the 68 prior relations, with relation-level paired win/tie/loss counts in Table 20. Cumulative SVD achieves the highest performance among all evaluated representations, reaching 71.65% Acc@1 and 88.29% Acc@5.

These comparisons demonstrate two key mechanistic principles:

- **Cross-depth module accumulation**: Full cumulative SVD substantially outperforms single-layer SVD restricted to the intervention block (71.65%/88.29% vs. 62.91%/81.93%, +8.74 points Acc@1). This confirms that relational mechanisms draw upon distributed, multi-layer computational components rather than an isolated single-layer circuit.
- **Efficacy of orthogonal singular coordinates**: Cumulative SVD outperforms native weight columns (88.29% vs. 86.42% Acc@5) and module output neurons (88.29% vs. 83.83% Acc@5), winning the majority of non-tied relations against both (29/22 and 27/25 wins

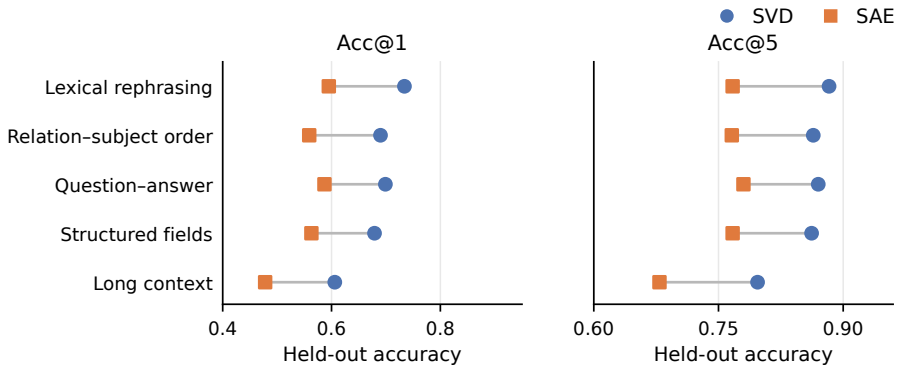


Figure 17: **Frozen-vector transfer across five prompt template families.** SVD steering vectors maintain consistent leads over SAEs across all five unseen structural template families without re-tuning.

on Acc@1; 21/9 wins on Acc@5). Intra-module singular decomposition isolates clean, mutually orthogonal relational directions, eliminating redundant cross-talk present in raw neuron bases and unrotated weight columns.

G.4 Frozen-Vector Prompt-Template Transfer

Template construction and protocol. To test whether steering vectors capture abstract relational concepts rather than template-specific lexical artifacts, we evaluate frozen vectors across ten prespecified prompt templates grouped into five structural families (Table 21): lexical rephrasing, inverted relation–subject order, question–answer phrasing, structured key-value fields, and extended narrative context. For each relation, the source record, target layer L, vector size k, and strength α are held strictly fixed from the Given-style training run.

Table 21: **Ten prespecified relation-replacement templates.** The generator substitutes the relation's original subject-type word for {type} and capitalizes its first letter for {Type}. Evaluation fills {subject} and {relation}; \n denotes a newline. Each template ends immediately before the answer.

Family / ID	Template skeleton
Lexical 1	`For {subject}, the {relation} of this {type} is`
Lexical 2	`Considering {subject}, the {relation} of this {type} is`
Order 1	`The {relation} of this {type} is requested.\n Use this subject: {subject}\n Answer:`
Order 2	`The {relation} for this {type} is requested. Subject: {subject}\n Answer:`
Question 1	`Given {subject}, what is the {relation} of this {type}? Answer:`
Question 2	`What is the {relation} of the following {type}?\n Subject: {subject}\n Answer:`
Fields 1	`Subject: {subject}\n Subject type: {type}\n Relation: {relation}\n Answer:`
Fields 2	`Relation: {relation}\n {Type}: {subject}\n Value:`
Context 1	`Consider the following {type}.\n Subject: {subject}\n The requested relation is {relation}. The answer is`
Context 2	`Use the following information to identify the requested value.\n Subject: {subject}\n Type: {type}\n Relation: {relation}\n The value is`

Transfer results. Table 22 and Figure 17 present transfer results across the 794 relation units. Frozen SVD vectors exhibit exceptional transfer fidelity, outperforming SAEs across all five template families:

$$\text{Acc@1: } 0.682 \text{ (SVD)} \quad \text{vs.} \quad 0.556 \text{ (SAE)} \quad \text{vs.} \quad 0.306 \text{ (Full Residual)}, \tag{20}$$

Table 22: **Frozen-vector transfer across unseen template families.** Macro-averaged held-out accuracy across 794 relation–model units under five distinct syntactic and structural template families. Entries report SVD / SAE / Full-Residual-Vector scores.

Family	Acc@1	Acc@5
Lexical rephrasing	0.734 / 0.595 / 0.270	0.883 / 0.767 / 0.434
Relation–subject order	0.690 / 0.559 / 0.349	0.864 / 0.766 / 0.572
Question–answer	0.699 / 0.587 / 0.365	0.870 / 0.780 / 0.591
Structured fields	0.679 / 0.563 / 0.323	0.862 / 0.767 / 0.575
Long context	0.606 / 0.478 / 0.222	0.797 / 0.679 / 0.415

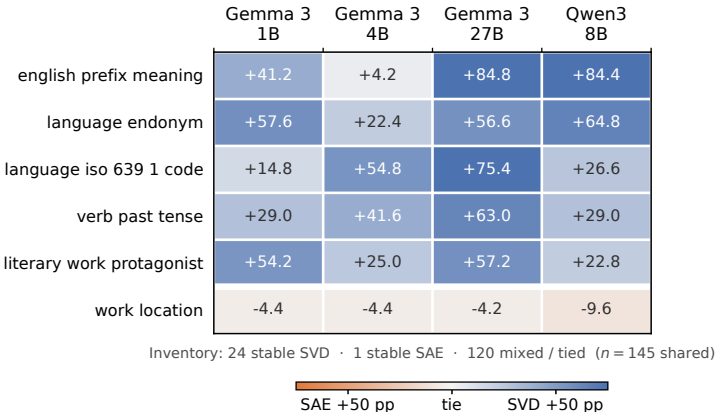


Figure 18: **Cross-model relation specialization.** Held-out $\Delta\text{Acc@1} = \text{SVD} - \text{SAE}$ for relations exhibiting consistent preference across all four evaluated models. Among 145 shared relations, 24 consistently favor SVD in all models, while only one consistently favors SAE.

and Acc@5 of 0.855 (SVD) vs. 0.752 (SAE) vs. 0.517 (Full Residual). SVD wins on 531 relation units (vs. 223 for SAE) on Acc@1, and 435 vs. 184 on Acc@5. Even on challenging long-context prompts, SVD maintains a +12.8 point lead over SAE (0.606 vs. 0.478). This confirms that SVD vectors encode robust, invariant relational mappings rather than surface prompt patterns.

G.5 Relation-Level Domain Specialization

H Extraction Cost and Dictionary Stability

H.1 Extraction Cost Accounting

We evaluate the computational cost of extracting paired attention-output and MLP-output dictionaries at Qwen3-8B-Base layers 18 and 24. While learned SAEs require training two JumpReLU models on billions of tokens captured from backbone forward passes, intrinsic SVD extracts dictionaries directly from frozen weight operators in closed form without training data. Table 23 reports the training-token budgets, counted arithmetic FLOPs for SAE training, and algorithmic reference work for SVD extraction.

Table 23: **Qwen3-8B-Base attention-plus-MLP dictionary extraction.** SAE values report counted arithmetic FLOPs for backbone activation caching and two SAE training runs. SVD values report algorithmic reference work for the QR/SVD factorization path. The final column reports the compute ratio of SAE training to SVD extraction.

Layer	Method	Corpus tokens	Extraction work	SAE/SVD ratio
L18	SAE, 16k, L_0=150	2.0B	1.916×10^{19}	7.28×10^{6}
L18	SAE, 65k, L_0=20	4.0B	5.765×10^{19}	2.19×10^{7}
L18	SVD	0	2.630×10^{12}	1
L24	SAE, 16k, L_0=150	2.0B	2.419×10^{19}	9.20×10^{6}
L24	SAE, 65k, L_0=20	4.0B	6.771×10^{19}	2.57×10^{7}
L24	SVD	0	2.630×10^{12}	1

SAE arithmetic work. We profile floating-point operations via PyTorch `FlopCounterMode` across backbone passes and SAE training steps, counting each multiply–accumulate as two FLOPs. Qwen3-8B-Base has residual dimension $d = 4096$ and intermediate MLP dimension $d_{\text{ff}} = 12288$. For a 4,096-token training batch, dense encoder and decoder matrix multiplications for a width-w SAE require approximately $12Bdw$ FLOPs ($B = 4096$). Because attention and MLP SAEs share the backbone forward pass per batch, the total arithmetic work for a paired training run over N tokens is:

$$F_{\text{SAE,pair}}(\ell, w, N) = \frac{N}{4096}\big(F_{\text{backbone}}(\ell) + 2F_{\text{SAE,site}}(w)\big). \tag{21}$$

Target token budgets are $N = 2 \times 10^9$ for width 16,384 and $N = 4 \times 10^9$ for width 65,536. Table 24 details the measured per-batch arithmetic operations. This arithmetic count establishes a conservative lower bound on the true compute burden by omitting optimizer state updates, elementwise nonlinearities, gradient synchronization across GPUs, and I/O data movement.

Counted component	Per 4,096-token batch	Per token
Backbone through L18	32,641.75 GFLOPs	7,969.18 MFLOPs
Backbone through L24	42,949.67 GFLOPs	10,485.76 MFLOPs
One 16k SAE forward/backward	3,298.535 GFLOPs	805.306 MFLOPs
One 65k SAE forward/backward	13,194.140 GFLOPs	3,221.226 MFLOPs

Table 24: **Counted arithmetic work for SAE dictionary pairs.** Backbone costs capture residual stream activations through decoder layers 0–18 or 0–24.

SVD algorithmic reference work. For each attention query head h, the output write operator is $W_{O,h}W_{V,g(h)}$ (4096×128 and 128×4096). We compute thin QR factorizations of both factors, form the 128×128 product of their triangular factors, compute its SVD via `gesvdj`, and map the left singular vectors back through the first QR basis. Because four query heads share each key–value head in GQA, right-factor QR factorizations are cached. For the 4096×12288 MLP down-projection, we compute a direct float32 reduced SVD using `gesvd`. Singular vectors are canonically oriented such that their largest-magnitude coordinate is positive (Appendix A). Standard numerical linear algebra specifies leading-order counts: $C_{\text{QR},Q+R} \simeq 4mn^2 - \frac{4}{3}n^3$, $C_{\text{QR},R} \simeq 2mn^2 - \frac{2}{3}n^3$, and $C_{\text{SVD}} \simeq \frac{20}{3}n^3$. Attention QR factorizations and back-projections total 18.74 GFLOPs, while MLP reduced SVD requires $C_{\text{MLP}} \simeq 6mn^2 + 20n^3 = 2{,}611.34$ GFLOPs. Together, complete attention-plus-MLP SVD extraction requires only 2,630.08 GFLOPs (2.63×10^{12} FLOPs).

Observed wall-clock extraction time. Table 25 reports CUDA-event measurements for complete SVD extraction on an NVIDIA RTX 4090 GPU (median over 20 repetitions). SVD extracts the complete paired dictionary in approximately 1.65 seconds (peak memory < 1.53 GiB). In stark contrast, production SAE training across four feature-sharded GPUs requires 8.9–44.6 hours (35.5–178.3 GPU-hours; Table 26), establishing an extraction latency speedup exceeding five orders of magnitude.

Table 25: **Single-GPU SVD extraction time.** Median milliseconds on an RTX 4090 GPU (20 repetitions; peak memory 1.53 GiB).

Layer	Attention	MLP	Pair total
L18	99.5	1,521.8	1,642.2
L24	101.5	1,555.0	1,658.1

Table 26: **Four-GPU SAE training time.** Wall-clock elapsed and GPU-hours from production logs across four sharded GPUs.

Training run	Hours	GPU-h
L18, 16k, $L_0 = 150$	8.86	35.46
L18, 16k, $L_0 = 20$	9.92	39.67
L18, 65k, $L_0 = 20$	26.22	104.89
L24, 16k, $L_0 = 150$	10.45	41.80
L24, 65k, $L_0 = 20$	44.57	178.28

H.2 SAE Seed Stability

We evaluate the replicability of learned feature representations by comparing independently trained Qwen3-8B-Base layer-18 JumpReLU SAE dictionaries across three random seeds (42, 123, 456) under identical configurations (width 16,384, target $L_0 = 150$, 2.0B training tokens), differing solely in random initialization and data shuffling order. While SVD feature extraction is strictly deterministic and closed-form with zero seed variance, learned SAEs exhibit substantial instability.

For two decoder dictionaries $D^{(a)}, D^{(b)} \in \mathbb{R}^{m \times d}$ ($m = 16{,}384$) at the same module site, we normalize decoder rows and compute signed cosine similarities $C_{ij} = (d_i^{(a)}/\|d_i^{(a)}\|_2) \cdot (d_j^{(b)}/\|d_j^{(b)}\|_2)$. Because JumpReLU activations are strictly non-negative, signed cosine measures feature alignment without polarity ambiguity. We compute the exact maximum-weight bipartite matching via linear sum assignment:

$$S(D^{(a)}, D^{(b)}) = \frac{1}{m} \max_{\pi \in \mathfrak{S}_m} \sum_{i=1}^{m} C_{i,\pi(i)}. \tag{22}$$

To assess alignment independent of one-to-one constraints, we also compute symmetric nearest-neighbor cosine (the mean of row- and column-wise maximum similarities).

Table 27: **Replicability and similarity distribution of learned SAE dictionaries across random seeds.** Evaluated on Qwen3-8B-Base layer-18 JumpReLU SAEs across three random seeds under exact maximum-weight bipartite matching. Symmetric NN reports unconstrained mean best matches. Rightmost columns report the proportion of matched features failing similarity thresholds.

Site	Seeds	Mean	Median	Symm. NN	5th–95th pct.	< 0.5	< 0.7	< 0.9
Attention	42–123	0.690	0.772	0.721	0.196–0.951	23.1%	40.8%	75.9%
Attention	42–456	0.688	0.772	0.719	0.196–0.949	23.5%	40.7%	76.5%
Attention	123–456	0.688	0.769	0.719	0.194–0.949	23.4%	41.0%	76.8%
MLP	42–123	0.850	0.976	0.867	0.192–0.995	12.4%	17.2%	28.9%
MLP	42–456	0.851	0.977	0.868	0.190–0.995	12.4%	17.3%	29.0%
MLP	123–456	0.852	0.977	0.870	0.194–0.995	12.1%	17.0%	28.5%

As shown in Table 27, learned SAEs exhibit marked instability across random training seeds. Under exact one-to-one alignment, mean cosine similarity across the three seed pairs is only 0.689 for attention and 0.851 for MLP outputs. Strikingly, 40.8% of matched attention features and 17.2% of MLP features have cosine similarities below 0.70, while over 76% of attention features and 28% of MLP features fail to reach 0.90 similarity. This demonstrates that a substantial fraction of learned SAE latents represent non-replicable optimization artifacts rather than invariant model features. In contrast, intrinsic SVD produces mathematically unique, deterministic dictionaries with zero seed variance.