

# Addressing the Concept Faithfulness Gap through Probe-Guided Sparse Autoencoders

Lorenzo Malandri<sup>1,3</sup>, Fabio Mercorio<sup>1,3</sup>, Antonio Serino<sup>2</sup>

<sup>1</sup>Dept of Statistics and Quantitative Methods, University of Milano-Bicocca, Italy,

<sup>2</sup>Dept of Economics, Management and Statistics, University of Milano-Bicocca, Italy,

<sup>3</sup>CRISP Research Centre [crispresearch.eu](https://crispresearch.eu), University of Milano-Bicocca, Italy

## Abstract

Sparse Autoencoders (SAEs) decompose language model activations into sparse, interpretable features, but these features often struggle to predict task-relevant concepts that remain clearly encoded in the original dense activations. We call this mismatch a *concept faithfulness gap*: a faithful SAE should preserve concept-level representations, such as toxicity or sentiment, rather than producing features that look interpretable but lose predictive relevance. We introduce ProbeSAE, a probing-guided fine-tuning framework that addresses this gap by preserving task-relevant concept structure in a more monosemantic sparse decomposition, enabling more effective causal control over model behaviour without sacrificing interpretability. We evaluate ProbeSAE on two SAE architectures (TopK, JumpReLU), two domains (toxicity, sentiment), and multilingual out-of-distribution settings spanning six languages. Trained exclusively on English data, ProbeSAE: (i) produces more predictive, monosemantic, and interpretable sparse features than pre-trained, naïvely fine-tuned, and concept-conditioned SAEs across all six evaluation languages; (ii) learns language-independent feature directions, as confirmed by cross-lingual causal ablation; (iii) passes SAEbench evaluation, confirming these gains do not arise from task/domain overfitting. Together, these results show that the concept faithfulness gap in SAEs can be addressed through probe-guided training, improving monosemanticity and causal effectiveness without sacrificing interpretability, making probe-guided SAEs a more reliable tool for domain-specific interpretability analysis.

## 1 Introduction

Sparse Autoencoders (SAEs) decompose dense language models activations into an overcomplete

<sup>†</sup>Authors appear in alphabetic order; all authors equally contributed to this work.

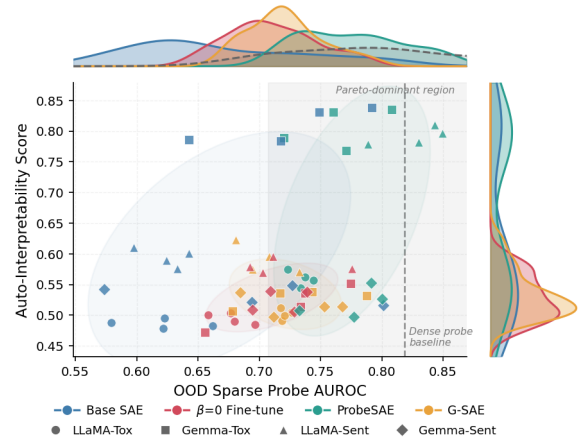


Figure 1: **OOD Sparse Probe AUROC vs. Auto-Interpretability Embedding Score.** Points denote (model, domain, layer) cells. ProbeSAE (green) achieves the best trade-off in 11/16 cells, closely tracking the dense baseline (grey band/line) in the marginal KDE, while other conditions lag.

set of sparsely active features, yielding representations in which each input is explained by a small number of directions (Huben et al., 2024; Bricken et al., 2023; Gao et al., 2025). Standard SAE objectives optimize reconstruction fidelity under sparsity constraints, with the goal of exposing interpretable and approximately monosemantic features. Kantamneni et al. (2025) question whether SAE features are useful in a general sense; here we operationalize utility along a concrete axis: the ability of SAE features to support task-specific downstream analysis. Their key finding, that linear probes trained on SAE features systematically underperform probes trained on dense activations, reveals what we term the *concept faithfulness gap*: if sparse features cannot support classification as well as the dense activations they decompose, the SAE lacks functional fidelity to the task-relevant structure encoded by the model, and its features cannot serve as reliable units of interpretability analysis. The faithfulness gap thus signals a misalignment between the general dictionary

learning objective (Olshausen and Field, 1997) and task-specific interpretability use. This matters in practice because SAE features are increasingly used as the *units of analysis* of interpretability pipelines, from sparse probing to circuit discovery and activation steering: if those units do not retain the task-relevant information carried by the activations they decompose, conclusions drawn from them risk describing the dictionary rather than the model. Throughout, we operationalise a concept as a binary classification task over model activations (toxicity, sentiment), and measure faithfulness as the AUROC of a linear probe trained on SAE features relative to the same probe trained on the dense activations. A natural attempt to close this gap is to fine-tune the SAE directly on task data using reconstruction loss alone. We show, however, that this approach trades one problem for another: while it improves task-information retention, it severely degrades feature interpretability, the property that motivated SAEs in the first place. A principled solution must therefore close the faithfulness gap *without sacrificing* interpretability. In this work, we introduce ProbeSAE, a *probe-guided training* approach that incorporates probing objectives directly into sparse feature learning. Rather than using probes solely for post-hoc evaluation, they act as a training signal that biases the SAE encoder toward task-relevant directions while preserving the sparse autoencoding framework. As Figure 1 previews, ProbeSAE improves OOD AUROC over the Base SAE in 15 of the 16 (model, domain, layer) configurations and leads on both axes jointly in 11 of them; moreover, its AUROC distribution substantially overlaps with the dense probe baseline, while Base SAE,  $\beta = 0$  and G-SAE remain clearly shifted left, indicating that probe-guided training recovers task-relevant information that standard SAE objectives fail to preserve. Our **contribution** is threefold: (i) we formalize the *concept faithfulness gap* between reconstruction-optimized SAE features and features that are causally grounded in model internal state; (ii) we propose ProbeSAE, a probing-guided fine-tuning framework that closes or substantially narrows the faithfulness gap *without sacrificing* feature interpretability; (iii) we validate ProbeSAE across two domains (toxicity and sentiment) on eight SAEs from two architectures (Top-k and JumpReLU), evaluating it along three axes: *probe utility*, *interpretability* and *monosemanticity*, and

*causal grounding* of the learned features. Code and fine-tuned checkpoints are publicly available at <https://github.com/serino28/Probe-Guided-Sparse-Autoencoders> and <https://huggingface.co/serino28/probesae-checkpoints>, respectively.

## 2 Preliminaries and Background

**Sparse Autoencoders.** Sparse Autoencoders decompose activations into an overcomplete basis of recurring directions (Olshausen and Field, 1997; Bricken et al., 2023; Huben et al., 2024). Sparsity is enforced either by a learnable per-feature threshold (Rajamanoharan et al., 2024) or by keeping the  $k$  largest activations (Gao et al., 2025). SAEs also support auto-interpretability pipelines that explain features in natural language and score those explanations against activations (Bills et al., 2023; Paulo et al., 2025; Mercorio et al., 2025), monosemanticity scores (Härle et al., 2025), and activation steering through decoder directions (Templeton et al., 2024; Marks et al., 2025). SAE Bench (Karvonen et al., 2025) provides a unified evaluation framework across these axes. Despite these advances, Kantamneni et al. (2025) demonstrate that SAE features often fail to support downstream classification as well as the raw activations they decompose, motivating our work.

**Concept Detection via Linear Probing.** Linear probes diagnose whether task-relevant information is linearly accessible in a representation (Alain and Bengio, 2016; Braun et al., 2024). Recent work finds behavioural concepts on near-one-dimensional directions: safe-unsafe separation emerges from a single direction across layers (Ermellino et al., 2026), and sycophancy is linearly separable in a few middle-layer attention heads (Genadi et al., 2026). Such directions are not only readable but actionable: injecting role vectors extracted from activations measurably shifts model behaviour (Poterì et al., 2025), which is the property our causal evaluation tests for SAE features. If task-relevant information is concentrated in low-dimensional linear subspaces, a probing objective can orient SAE features toward those subspaces without sacrificing the sparse overcomplete structure.

**Supervised and Task-Guided SAEs.** Several concurrent works inject supervision into SAE training or feature selection, but along axes orthogonal to ours. Yang et al. (2025) *train from scratch*,

binding ontology concepts to dedicated latent slots; Kulkarni et al. (2025) add a concept bottleneck post-hoc, but only for *vision-language* models; He et al. (2025) keep the SAE *frozen*, training steering vectors in a task-relevant subspace of its latent. Härle et al. (2025) propose G-SAE, which conditions latent dimensions on concept labels during training, and introduce the FMS monosemanticity score that we adopt as one of our evaluation axes.

ProbeSAE is, to our knowledge, the first to combine *post-hoc fine-tuning* of community-released SAEs, *joint probe-SAE optimisation* through the full encoder-decoder rather than a frozen latent, and explicit targeting of the *concept faithfulness gap* against a dense-probe upper bound. G-SAE is the only one of these that also operates post-hoc on pretrained text-LLM SAEs, so we include it as a full condition alongside the Base SAE and the matched  $\beta = 0$  baseline, on identical data, layers and protocol (Section 4).

**Formalizing the Concept Faithfulness Gap.** Consider a pretrained language model with intermediate activations  $\mathbf{x} \in \mathbb{R}^d$  at layer  $\ell$ . A Sparse Autoencoder learns to compress and decompress these activations through a sparse bottleneck:

$$\mathbf{z} = E(\mathbf{x}) \in \mathbb{R}^k, \quad \hat{\mathbf{x}} = D(\mathbf{z}) \approx \mathbf{x} \quad (1)$$

where  $k \gg d$  is the dictionary size and  $\|\mathbf{z}\|_0 \ll k$  by construction. Standard SAE training minimizes reconstruction error while encouraging sparsity, typically via L1 regularization or structural mechanisms like JumpReLU activations. The concept faithfulness gap manifests when we compare classification performance between the original dense activations and the sparse SAE features. For a binary classification task with labels  $y \in \{0, 1\}$ , we train two logistic regression probes: a *dense probe*  $p_d : \mathbf{x} \mapsto y$  operating on original activations, and a *sparse probe*  $p_s : \mathbf{z} \mapsto y$  operating on SAE features. We consistently observe:

$$\text{AUROC}(p_d) > \text{AUROC}(p_s) \quad (2)$$

This inequality has profound implications for interpretability. The classification target  $y$  corresponds to a domain concept (e.g. toxicity, sentiment) that the model demonstrably encodes in its dense activations, as evidenced by the dense probe’s accuracy. We therefore use *concept faithfulness* to denote the SAE’s ability to preserve this concept-level structure in its sparse decomposition, distinct from *explanation-faithfulness* in the broader XAI

sense (Jacovi and Goldberg, 2020), which concerns whether a post-hoc explanation reflects a model’s actual reasoning. If the sparse representation cannot support the same level of classification performance as the dense representation, then the SAE has failed to preserve this concept-level structure. Features that “look interpretable” but cannot support accurate downstream prediction may offer a misleading picture of how the model actually processes information.

### 3 Building ProbeSAE

Our approach relies on the hypothesis that the faithfulness gap arises because standard SAE training treats all dimensions of the input equally. The reconstruction objective  $\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2$  minimizes average error across all features, potentially sacrificing dimensions that are critical for specific tasks in favor of dimensions that contribute more to overall variance. Similarly, sparsity pressure may suppress features that are task-relevant but low-magnitude. Our proposed solution is to augment the SAE training objective with a classification term that explicitly rewards preservation of task-relevant information. By training a linear probe jointly with the SAE and backpropagating the classification loss through the encoder and decoder, we create gradient pressure that steers the SAE toward features that support accurate downstream prediction. The classification task serves as a *proxy for faithfulness*: if the sparse representation supports accurate classification, it must have preserved the relevant structure from the original activations.

**Task-Guided Loss Function.** Given a pretrained SAE with encoder  $E$  and decoder  $D$ , and a labeled dataset  $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$  where  $\mathbf{x}_i$  are dense activations extracted from the target layer and  $y_i \in \{0, 1\}$  are binary labels, we fine-tune the SAE while simultaneously training a sparse linear probe  $\mathbf{w} \in \mathbb{R}^k$  with  $k = d_{\text{sae}}$ . The total loss combines three components:

$$\mathcal{L} = \underbrace{\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2}_{\text{Reconstruction}} + \underbrace{\lambda_1 \|\mathbf{z}\|_1}_{\text{Sparsity}} + \underbrace{\beta_c \cdot \text{BCE}(\sigma(\mathbf{w}^\top \mathbf{z} + b), y)}_{\text{Task-Guided Steering}} \quad (3)$$

The Task-Guided Steering term uses the binary cross-entropy (BCE) between the probe’s predicted probability  $\sigma(\mathbf{w}^\top \mathbf{z})$  and the true label  $y$ . The reconstruction term ensures that the SAE continues to faithfully represent the original activations. The sparsity term maintains interpretable, low-dimensional representations. The task-guided term,

our addition, provides gradient signal that backpropagates through the probe into the SAE, rewarding feature configurations for accurate classification. The hyperparameter  $\beta_c$  controls the strength of task guidance. When  $\beta_c = 0$ , training reduces to standard SAE fine-tuning. As  $\beta_c$  increases, the SAE receives stronger pressure to preserve classification-relevant features, potentially at the cost of general reconstruction fidelity. Our experiments explore this trade-off systematically.

**Training Protocol.** We initialize from a pretrained SAE checkpoint obtained from the SAE Lens repository, preserving the feature structure learned during unsupervised pretraining. The sparse probe is initialized with random weights. For each batch of labeled activations, we compute the forward pass through the SAE to obtain sparse features  $\mathbf{z}$  and reconstructions  $\hat{\mathbf{x}}$ , then compute all three loss terms and backpropagate through the entire system. Both SAE parameters (encoder, decoder, and thresholds for JumpReLU) and probe weights are updated jointly.

## 4 Experimental Setup

**Datasets and Splits.** We evaluate ProbeSAE across two domains: toxicity and sentiment. For toxicity we use ToxiGen (Hartvigsen et al., 2022), a large-scale machine-generated dataset of toxic and benign statements targeting 13 minority groups. For sentiment we use Sentiment140 (Go et al., 2009), downsampled to match the size of ToxiGen for cross-domain comparability while retaining sample sizes sufficient for statistically significant evaluation. Both datasets are English-only and serve as the in-distribution (ID) source for SAE fine-tuning and hyperparameter selection. We partition each dataset into three non-overlapping splits: (i) a *tuning set* of 100k balanced samples used exclusively to fine-tune the SAE; (ii) a *probe training set* of 8k balanced samples used to train the linear evaluation probe, held out from SAE fine-tuning so that probe-based comparisons are calibrated across base,  $\beta = 0$ , and ProbeSAE conditions; (iii) an *ID test set* used solely for hyperparameter selection via grid search (Appendix B).

**Models and Pretrained SAEs.** We conduct our experiments using SAEs trained on the residual streams of two base models representing different scales: Gemma-2-2B (Team et al., 2024) (26 layers) and Llama-3.1-8B (Grattafiori et al., 2024) (32 layers). All pretrained SAEs are sourced from the SAE

Lens repository (Bloom et al., 2024). Specifically, we utilize Gemma Scope (Lieberum et al., 2024) JumpReLU SAEs with a dictionary size of 16k features for Gemma-2-2B<sup>1</sup>, and OpenMOSS TOP-K SAEs with a dictionary size of 32k for Llama-3.1-8B<sup>2</sup>.

**Layer Selection Protocol.** To target the layers that encode the most task-relevant information for each domain, we evaluate the residual stream activations by training domain-specific logistic regression probes using their respective probe training sets. We prioritize the top four layers that maximize the AUROC on the test set (see Table 4, Appendix A). The underlying premise is that high probe accuracy indicates a dense concentration of task concepts, making faithful feature decomposition in these specific layers particularly valuable. Following this protocol for toxicity, we isolate layers 13, 16, 24, and 25 for Gemma-2-2B, and layers 16, 17, 18, and 19 for LLaMA-3.1-8B. For sentiment, the optimal layers for Gemma-2-2B shift to 22, 23, 24, and 25; LLaMA retains the same four layers across both domains.

**Hyperparameter Search.** We conduct a grid search over  $\beta_c \in \{1, 5, 10, 15, 20, 50\}$ , samples  $N \in \{60k, 80k, 100k\}$ , and epochs  $E \in \{3, 5\}$ , yielding 36 configurations per layer. The protocol is summarised in Appendix B; per-configuration results for all layers are released as CSV files in the code repository. For each layer, we select the configuration that best balances classification performance (AUROC), reconstruction quality ( $R^2$ ), and sparsity ( $L_0$ ), ensuring that fine-tuning does not substantially alter the SAE’s representational regime.

### 4.1 Evaluation Framework

Evaluating SAE quality is challenging due to the lack of ground truth (Gantla, 2025). We assess ProbeSAE along three axes, probe utility, interpretability and monosemanticity, and causal control, each framed as an empirical question below (Q1, Q2, Q3), and for each comparing against three baselines: the original Base SAEs, SAEs fine-tuned with  $\beta = 0$ , and G-SAE (Härle et al., 2025). The  $\beta = 0$  condition isolates the contribution of the classification loss from that of fine-tuning per se. We compare against G-SAE using its post-hoc fine-tuning

<sup>1</sup><https://huggingface.co/google/gemma-scope-2b-pt-res>

<sup>2</sup>[https://huggingface.co/OpenMOSS-Team/Llama3\\_1-8B-Base-LXR-8x](https://huggingface.co/OpenMOSS-Team/Llama3_1-8B-Base-LXR-8x)

setup across identical data, layers, and pretrained SAEs, following published hyperparameters.

**Data.** To verify that ProbeSAE captures robust, language-agnostic concepts rather than dataset-specific artifacts, we evaluate it on unseen OOD datasets. For toxicity, we use Measuring Hate Speech (Sachdeva et al., 2022) and HateCheck (Röttger et al., 2021) (across English, Spanish, Italian, and French). For sentiment, we utilize the Amazon Reviews dataset (Keung et al., 2020) across five distinct languages: English, German, French, Spanish, and Japanese. Generalization to these zero-shot multilingual settings strongly indicates that ProbeSAE aligns features with underlying semantic structures rather than superficial lexical patterns.

### Q1: Does ProbeSAE improve the probe utility?

We assess the quality of the learned features quantitatively through their probe utility: several probes are trained on the SAE sparse features  $\mathbf{z}$  and evaluated via AUROC on each OOD dataset. Higher AUROC directly reflects higher feature quality, the more the SAE features encode task-relevant information, the more useful they are for downstream prediction. Alongside this, we report reconstruction fidelity ( $R^2$ ) and sparsity ( $L_0$ ) to verify that these representational properties are preserved under fine-tuning. To rule out that observed gains are merely the result of domain overfitting rather than genuinely improved features, we apply the official SAEbench pipeline (Karvonen et al., 2025) on the Pile-10k, a general-purpose corpus disjoint from our task domains, reporting several metrics: a ProbeSAE that has overfit its training domain would exhibit degraded performance here, whereas stable metrics confirm that it retains general representational ability. SAEbench enters our evaluation strictly as a *do-no-harm* control rather than as a headline metric on which methods are ranked: it asks whether task-guided fine-tuning damages the general-purpose SAE, and we report the cases where it partially does (Appendix D).

**Q2: Do ProbeSAE features preserve monosemanticity and interpretability?** A core desideratum of SAEs is that each learned direction correspond to a single, human-interpretable concept rather than a superposition of unrelated ones (Bricken et al., 2023). Probing supervision could in principle improve probe utility while degrading this property; we test whether it does so along two axes. **Auto-interpretability:** we evaluate the top-50 features per SAE, selected by mean activation magnitude,

a number motivated by the Top- $k$  activation function of LLaMA-3.1-8B SAEs ( $k = 50$ ), applied uniformly to Gemma-2-2B for comparability. Following Bills et al. (2023), natural-language explanations are generated with GPT-OSS-120B (Agarwal et al., 2025) from the highest-activating samples of the in-distribution test set. These explanations are then scored independently using the EleutherAI auto-interpretability framework (Paulo et al., 2025) on each OOD dataset, measuring *embedding score*<sup>3</sup> and *recall score*<sup>4</sup>. Scoring on unseen languages samples whether feature interpretations remain coherent beyond the training distribution. **Monosemanticity:** we compute the Fidelity Measurement Score (FMS) (Härle et al., 2025) on SAE latent activations pooled across OOD datasets. FMS iteratively identifies the most concept-predictive features and measures their classification fidelity on held-out data; higher values indicate that individual directions respond selectively to a single coherent concept.

**Q3: Do ProbeSAE features enable causal control of model internal state?** Each SAE learns a set of latent feature directions, encoded as rows of the decoder matrix  $\mathbf{W}_{\text{dec}}$ . The functional value of these features depends not only on whether they correlate with a target concept, but on whether  $\mathbf{W}_{\text{dec}}$  faithfully encodes the directions along which the model actually processes it: only then can ablating or injecting those directions predictably alter model internal state. Features that prove causally effective in this sense provide stronger evidence of functional alignment with task-relevant model behaviour than probing alone, which establishes correlation but does not verify that intervention on those features predictably alters outputs. Following the activation-engineering literature (Templeton et al., 2024; Marks et al., 2025; Poterì et al., 2025), we conduct two complementary interventions directly in residual stream space. Features are ranked per OOD dataset by  $\Delta_i = \bar{z}_i^{y=1} - \bar{z}_i^{y=0}$  and the top- $k$  are selected. *Activation Ablation* removes their reconstruction contribution from correctly-classified target samples ( $\tilde{\mathbf{x}} = \mathbf{x} - \sum_{i=1}^k z_i \mathbf{W}_{\text{dec}}[i, :]$ ); the modified activation is re-classified by the dense linear probe and we report Flip Rate  $\text{TP} \rightarrow \text{TN}$ . *Activation Steering* performs the dual operation on non-target samples, pushing the top- $k$  activations

<sup>3</sup>whether *multilingual-e5-large* can separate activating from non-activating examples given the explanation

<sup>4</sup>whether GPT-OSS-120B can predict which novel samples activate the feature

to their target-class mean ( $\tilde{\mathbf{x}} = \mathbf{x} + \sum_{i=1}^k (\bar{z}_i^{y=1} - z_i) \mathbf{W}_{\text{dec}}[i, :]$ ); we report Flip Rate  $\text{TN} \rightarrow \text{TP}$ . Both are evaluated at  $k=10$  and  $k=50$ , i.e. using the top 10 and the top 50 target-aligned features; higher flip rates indicate stronger functional alignment between feature directions and task-relevant model behaviour.

## 5 Evaluation and Results

In this section we answer the three evaluation questions introduced in Section 4.1 with the corresponding results A1, A2, and A3 below.

### A1: ProbeSAE Consistently Improves probe utility Addressing the Concept Faithfulness Gap.

Table 1 reports mean OOD AUROC,  $R^2$ , and  $L_0$  across all evaluated layers and both domains; the *Dense* row marks the concept faithfulness upper bound, i.e. the AUROC of a logistic probe trained directly on raw residual-stream activations. Across both model families and domains, ProbeSAE achieves the highest OOD AUROC of any sparse condition in 14 of the 16 layer configurations, and in all four model $\times$ domain cells on average, substantially closing the gap to this upper bound. Both exceptions are on Gemma-sentiment: at L23 the Base SAE is ahead by 0.001, and at L24 G-SAE leads by 0.035. On LLaMA, ProbeSAE meets or surpasses the dense probe in seven of the eight configurations, by a wide margin in places (L17 on sentiment: 0.849 vs. 0.781) and falling just short only at L16 on sentiment (0.788 vs. 0.794), consistent with the interpretation that probe-guided disentanglement removes superposition-induced noise: whereas reconstruction-only training leaves active features mixing relevant and irrelevant components, probe-guided disentanglement concentrates the sparse representation on directions that encode the target concept, so that ReLU zeroing selectively eliminates superposition noise and yields higher linear separability than the raw representation. For Gemma, gap-closing is more moderate but ProbeSAE remains the condition closest to the dense upper bound in every configuration. Crucially,  $\beta = 0$  never approaches the dense probe in any setting, establishing that the probing classification loss, not fine-tuning or sparsity alone, is the primary driver of gap-closing. Concept conditioning alone is likewise not sufficient: G-SAE improves clearly over the Base SAE but stays behind ProbeSAE in every model $\times$ domain cell, most widely on LLaMA-sentiment (0.828 vs. 0.704). Its reserved latent stays gated out of the pretrained

sparse code, so the concept does not resurface post-activation, whereas ProbeSAE’s head operates on the full code;  $R^2$  decreases moderately for ProbeSAE relative to  $\beta=0$ , reflecting the expected tension between task alignment and reconstruction;  $L_0$  remains broadly comparable across conditions. SAEBench evaluation on the Pile-10k (Appendix D) confirms that these gains are not accompanied by degradation of general representations, ruling out domain overfitting as their source. We note that OOD AUROC reflects the same probe class used during training. The gains are further corroborated by the complementary evidence from FMS, auto-interpretability scores, and cross-lingual causal interventions, metrics not directly optimized by the probe-guided loss. Per-layer and per-dataset stratification is in Appendix C, where ProbeSAE’s advantage holds on every individual OOD language except Amazon FR on Gemma-sentiment, and includes Japanese, a language sharing no script or morphological structure with the English training corpus. For Gemma, both  $\beta=0$  and G-SAE occasionally fall below the base SAE on non-English datasets, suggesting that fine-tuning without probing guidance can degrade cross-lingual transfer; ProbeSAE avoids this degradation uniformly, indicating that probing supervision steers the SAE toward language-agnostic, semantically abstract feature directions.

### A2: ProbeSAE Improves Feature Monosemanticity and Preserves Interpretability.

Figure 2 shows the distribution of auto-interpretability scores (embedding and recall) for the top-50 features of each condition, pooled across layers and domains. ProbeSAE maintains embedding and recall scores comparable to or above those of the base SAE across all evaluated configurations, indicating that probing supervision does not degrade the human-interpretability of learned features. Both are automatic proxies for interpretability rather than interpretability itself; we validate them against human judgement below (Table 2), where the recall score correlates strongly with human accuracy while the embedding score correlates positively but more weakly, so we read the embedding axis as the more tentative of the two. The  $\beta=0$  condition, by contrast, exhibits a systematic decrease in embedding scores in several configurations, suggesting that standard fine-tuning without probing guidance can yield features with weaker natural-language characterizations. Per-domain and per-dataset stratifi-

| Layer                          | Base           |               |             | $\beta=0$      |               |              | ProbeSAE       |               |              | G-SAE          |               |              | Dense          |
|--------------------------------|----------------|---------------|-------------|----------------|---------------|--------------|----------------|---------------|--------------|----------------|---------------|--------------|----------------|
|                                | OOD $\uparrow$ | $R^2\uparrow$ | $L_0$       | OOD $\uparrow$ | $R^2\uparrow$ | $L_0$        | OOD $\uparrow$ | $R^2\uparrow$ | $L_0$        | OOD $\uparrow$ | $R^2\uparrow$ | $L_0$        | OOD $\uparrow$ |
| <b>Toxicity, LLaMA-3.1-8B</b>  |                |               |             |                |               |              |                |               |              |                |               |              |                |
| L16                            | 0.579 (0.049)  | 0.513 (0.016) | 50.0 (0.0)  | 0.680 (0.052)  | 0.923 (0.013) | 50.0 (0.0)   | 0.744 (0.030)  | 0.842 (0.031) | 50.0 (0.0)   | 0.726 (0.044)  | 0.930 (0.012) | 50.0 (0.0)   | 0.708 (0.021)  |
| L17                            | 0.622 (0.024)  | 0.526 (0.019) | 50.0 (0.0)  | 0.658 (0.050)  | 0.923 (0.013) | 50.0 (0.0)   | 0.734 (0.045)  | 0.848 (0.031) | 50.0 (0.0)   | 0.719 (0.045)  | 0.928 (0.012) | 50.0 (0.0)   | 0.714 (0.028)  |
| L18                            | 0.623 (0.041)  | 0.539 (0.019) | 50.0 (0.0)  | 0.697 (0.056)  | 0.926 (0.011) | 50.0 (0.0)   | 0.737 (0.030)  | 0.852 (0.030) | 50.0 (0.0)   | 0.720 (0.056)  | 0.930 (0.010) | 50.0 (0.0)   | 0.707 (0.042)  |
| L19                            | 0.662 (0.038)  | 0.557 (0.015) | 50.0 (0.0)  | 0.677 (0.061)  | 0.932 (0.010) | 50.0 (0.0)   | 0.723 (0.057)  | 0.866 (0.027) | 50.0 (0.0)   | 0.718 (0.052)  | 0.936 (0.009) | 50.0 (0.0)   | 0.723 (0.053)  |
| <b>Toxicity, Gemma-2-2B</b>    |                |               |             |                |               |              |                |               |              |                |               |              |                |
| L13                            | 0.792 (0.057)  | 0.866 (0.011) | 85.7 (9.8)  | 0.775 (0.059)  | 0.886 (0.009) | 70.0 (9.9)   | 0.808 (0.062)  | 0.886 (0.009) | 79.6 (10.0)  | 0.787 (0.064)  | 0.889 (0.009) | 80.8 (12.1)  | 0.868 (0.050)  |
| L16                            | 0.749 (0.035)  | 0.889 (0.019) | 75.7 (12.8) | 0.737 (0.040)  | 0.903 (0.009) | 57.7 (11.2)  | 0.760 (0.049)  | 0.903 (0.009) | 61.4 (11.8)  | 0.743 (0.043)  | 0.906 (0.009) | 64.9 (13.7)  | 0.855 (0.044)  |
| L24                            | 0.643 (0.095)  | 0.901 (0.024) | 58.9 (10.5) | 0.656 (0.095)  | 0.884 (0.034) | 48.7 (10.6)  | 0.720 (0.094)  | 0.880 (0.029) | 69.1 (11.1)  | 0.678 (0.082)  | 0.887 (0.034) | 63.1 (15.9)  | 0.809 (0.034)  |
| L25                            | 0.718 (0.070)  | 0.899 (0.026) | 89.2 (9.5)  | 0.734 (0.059)  | 0.879 (0.040) | 75.9 (13.9)  | 0.771 (0.057)  | 0.877 (0.038) | 81.0 (12.9)  | 0.717 (0.068)  | 0.881 (0.042) | 97.4 (22.1)  | 0.809 (0.032)  |
| <b>Sentiment, LLaMA-3.1-8B</b> |                |               |             |                |               |              |                |               |              |                |               |              |                |
| L16                            | 0.633 (0.041)  | 0.483 (0.009) | 50.0 (0.0)  | 0.775 (0.060)  | 0.892 (0.027) | 50.0 (0.0)   | 0.788 (0.043)  | 0.824 (0.053) | 50.0 (0.0)   | 0.695 (0.096)  | 0.898 (0.026) | 50.0 (0.0)   | 0.794 (0.046)  |
| L17                            | 0.598 (0.046)  | 0.488 (0.011) | 50.0 (0.0)  | 0.703 (0.054)  | 0.898 (0.026) | 50.0 (0.0)   | 0.849 (0.056)  | 0.839 (0.047) | 50.0 (0.0)   | 0.733 (0.090)  | 0.905 (0.027) | 50.0 (0.0)   | 0.781 (0.044)  |
| L18                            | 0.624 (0.049)  | 0.486 (0.010) | 50.0 (0.0)  | 0.693 (0.070)  | 0.900 (0.030) | 50.0 (0.0)   | 0.830 (0.049)  | 0.845 (0.050) | 50.0 (0.0)   | 0.708 (0.086)  | 0.905 (0.029) | 50.0 (0.0)   | 0.793 (0.049)  |
| L19                            | 0.642 (0.062)  | 0.514 (0.007) | 50.0 (0.0)  | 0.711 (0.048)  | 0.907 (0.030) | 50.0 (0.0)   | 0.843 (0.046)  | 0.859 (0.042) | 50.0 (0.0)   | 0.681 (0.085)  | 0.911 (0.029) | 50.0 (0.0)   | 0.773 (0.050)  |
| <b>Sentiment, Gemma-2-2B</b>   |                |               |             |                |               |              |                |               |              |                |               |              |                |
| L22                            | 0.727 (0.042)  | 0.886 (0.010) | 70.5 (2.2)  | 0.739 (0.044)  | 0.829 (0.021) | 63.7 (3.0)   | 0.791 (0.117)  | 0.832 (0.019) | 76.8 (4.3)   | 0.685 (0.074)  | 0.830 (0.023) | 77.6 (3.8)   | 0.941 (0.006)  |
| L23                            | 0.801 (0.046)  | 0.874 (0.013) | 73.7 (6.0)  | 0.709 (0.044)  | 0.806 (0.023) | 67.4 (4.9)   | 0.800 (0.066)  | 0.813 (0.024) | 75.6 (7.2)   | 0.753 (0.054)  | 0.808 (0.026) | 85.2 (8.9)   | 0.941 (0.007)  |
| L24                            | 0.573 (0.039)  | 0.870 (0.009) | 68.9 (7.6)  | 0.728 (0.040)  | 0.786 (0.029) | 65.5 (5.1)   | 0.732 (0.124)  | 0.786 (0.030) | 71.8 (6.7)   | 0.768 (0.060)  | 0.789 (0.036) | 83.1 (9.8)   | 0.941 (0.009)  |
| L25                            | 0.694 (0.064)  | 0.878 (0.017) | 103.1 (9.9) | 0.694 (0.119)  | 0.797 (0.009) | 109.4 (10.5) | 0.777 (0.085)  | 0.796 (0.009) | 114.3 (10.9) | 0.712 (0.088)  | 0.800 (0.013) | 135.3 (18.0) | 0.937 (0.012)  |

Table 1: **OOD Results.** Mean OOD AUROC, reconstruction fidelity ( $R^2$ ) and sparsity ( $L_0$ ) per layer, over HateCheck EN/ES/IT/FR + MeasuringHate for toxicity and Amazon Reviews EN/DE/ES/FR/JA for sentiment. *Dense* is the dense-probe AUROC, i.e. the concept faithfulness ceiling. Colour is scaled within each (domain, metric) block; darker is better, and sparser for  $L_0$ . Standard deviation over the OOD sets in parentheses; per-dataset breakdowns in Appendix C.

cation is reported in Appendix E, where the cross-lingual consistency of this pattern is particularly striking: for Gemma,  $\beta=0$  embedding scores collapse uniformly across all OOD languages while ProbeSAE maintains Base-level interpretability on every language including those unseen during training; for LLaMA on sentiment, ProbeSAE shows markedly higher embedding scores across all languages including Japanese, whereas  $\beta=0$  remains at or below Base level throughout. The fact that probing supervision preserves feature coherence across typologically diverse languages suggests that it reinforces semantically grounded directions in the model’s representation space. Figure 3 reports FMS scores across all conditions, layers, and domain-model combinations. ProbeSAE achieves the highest FMS of any condition in 15 of the 16 layer configurations and in all four model $\times$ domain cells, with the clearest improvements on LLaMA-3.1-8B on both tasks; for Gemma-2-2B, sentiment also shows clear gains, while toxicity exhibits only marginal variation and the Base SAE is marginally ahead at L16. Crucially,  $\beta=0$  shows virtually no FMS gain over the base SAE in any configuration, establishing that the classification loss, and not fine-tuning per se, is the driver of improved monosemanticity. Notably, FMS gains co-vary quantitatively with OOD AUROC gains: the Spearman correlation between  $\Delta$ FMS and  $\Delta$ AUROC (both computed as ProbeSAE minus Base SAE across all 16 model $\times$ domain $\times$ layer configurations) is  $\rho = 0.79$ , 95 % CI [0.48, 0.92] ( $N = 16$ ,  $p < 0.001$ , two-tailed). This relationship is informative specifically

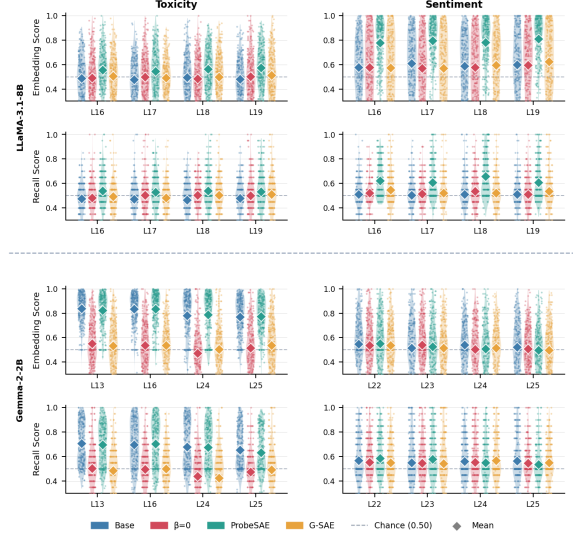


Figure 2: **Auto-interpretability scores** across conditions, pooled over layers, domains, and OOD sets. Boxes: IQR; lines: median; diamonds: mean.

for ProbeSAE, since  $\beta=0$  produces near-zero FMS gains in most configurations, leaving no meaningful variation to correlate. The configurations with the largest FMS improvement are precisely those where ProbeSAE most consistently surpasses the dense probe, while the smallest FMS gains correspond to the widest residual faithfulness gap, suggesting that monosemanticity and cross-lingual transferability share a common origin in classification-guided training. G-SAE behaves like  $\beta=0$  here: it obtains the lower auto-interpretability score in all eight cell $\times$ score comparisons under an identical explainer, and its FMS falls below ProbeSAE’s in all four cells and below even the Base SAE in two of

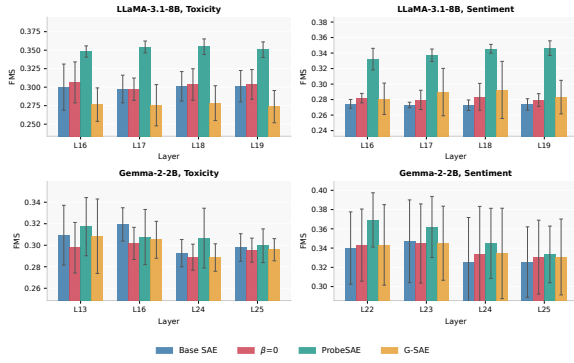


Figure 3: **FMS** per condition, layer, and domain-model combination ( $n=10$  iterations).

them, despite FMS being its own metric. Together, these results show that ProbeSAE simultaneously preserves the natural-language coherence of features and sharpens their selectivity.

| Condition | Bal. acc.    | 95% CI         | Maj. vote    | TPR / TNR   |
|-----------|--------------|----------------|--------------|-------------|
| Base SAE  | 0.500        | [0.460, 0.540] | 0.500        | 0.47 / 0.54 |
| $\beta=0$ | 0.515        | [0.472, 0.555] | 0.525        | 0.46 / 0.58 |
| G-SAE     | 0.523        | [0.463, 0.583] | 0.500        | 0.53 / 0.52 |
| ProbeSAE  | <b>0.593</b> | [0.550, 0.637] | <b>0.662</b> | 0.43 / 0.76 |

Table 2: **Human validation of the automated interpretability scores** (LLaMA-3.1-8B, toxicity; 5 annotators, 32 features, 320 judgements each, Fleiss’  $\kappa = 0.525$ ). Balanced accuracy is averaged over annotators  $\times$  features, with a bootstrap 95% CI; *maj. vote* is the per-snippet consensus.

Because both scores are themselves produced by a language model, we validate them with a blind study on LLaMA-toxicity: five annotators judged 32 features, drawn at random per layer and condition rather than by score, blind to condition and ground truth (320 judgements each, Fleiss’  $\kappa = 0.525$ ). Table 2 shows ProbeSAE alone separating from the Base SAE, widening under majority vote, while the other three sit at chance; annotators reject non-activating snippets far more reliably for ProbeSAE (0.755 against  $\leq 0.575$ ), as expected if its explanations are easier to falsify. Human accuracy correlates strongly with the automated recall score ( $\rho = 0.79$ ,  $p < 10^{-4}$ ) but only weakly with the embedding score, so we read conclusions resting on the latter with more caution. The study covers one model-domain cell in one language.

**A3: ProbeSAE Features Enable Stronger Causal Control.** Table 3 reports ablation and steering flip rates at  $k = 10$  and  $k = 50$  across all conditions, layers, and domains. For toxicity on LLaMA-3.1-8B, ProbeSAE features exhibit substantially higher ablation flip rates than the base SAE, demonstrating that they encode the decoder directions the model

relies on to produce its predictions. Steering flip rates are also markedly elevated, confirming that the same directions are sufficient to induce target-class internal state and suggesting stronger functional alignment than probing alone can establish. The  $\beta = 0$  baseline achieves comparable ablation gains but consistently lower steering rates, indicating that probing supervision specifically strengthens the functional alignment between decoder directions and the model’s actual processing pathways, rather than merely improving feature relevance. At  $k = 10$ , however,  $\beta = 0$  achieves higher ablation flip rates than ProbeSAE on LLaMA toxicity; the ranking inverts at  $k = 50$ . This crossover is consistent with a monosemantic decomposition: probe-guided training distributes concept-relevant information across a larger set of individually selective features, requiring higher  $k$  to accumulate equivalent ablation signal, while  $\beta = 0$  concentrates task information in fewer but potentially polysemantic directions, an interpretation corroborated by ProbeSAE’s consistently higher FMS scores. For Gemma-2-2B on toxicity, steering rates are already high across all conditions, consistent with the architecture’s susceptibility to concept-level interventions at these layers, but ProbeSAE further improves both ablation and steering, providing more complete bilateral causal control. In the sentiment domain, high ablation rates across all conditions for both models indicate that all four conditions capture the residual-stream directions the model uses for this task, while uniformly low steering rates reflect the resistance of strongly negative-valenced representations to additive positive interventions, an asymmetry attributable to the bipolar structure of the task rather than a limitation of feature quality. G-SAE leads ProbeSAE on LLaMA-toxicity steering, the one setting where a competing condition is clearly ahead, but ProbeSAE leads on ablation in both domains and is at least as good in the large majority of cells. Together, these results indicate that ProbeSAE decoder directions are more functionally aligned with task-relevant model behaviour, enabling stronger and more complete causal control than the base SAE,  $\beta = 0$ , or G-SAE. Each flip rate averages the five OOD datasets of its domain, listed in the caption, and the standard deviation in parentheses is taken across them, so it measures how far a condition depends on the evaluation language. Those spreads are of the same order for every condition, so the gaps between conditions reflect the means rather than one dominant

| SAE                        | Layer 1        |                |                |                | Layer 2        |                |                |                | Layer 3        |                |                |                | Layer 4        |                |                |                |
|----------------------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
|                            | Abl.           |                | Str.           |                | Abl.           |                | Str.           |                | Abl.           |                | Str.           |                | Abl.           |                | Str.           |                |
|                            | $k=10$         | $k=50$         | $k=10$         | $k=50$         | $k=10$         | $k=50$         | $k=10$         | $k=50$         | $k=10$         | $k=50$         | $k=10$         | $k=50$         | $k=10$         | $k=50$         | $k=10$         | $k=50$         |
| <b>Toxicity</b>            |                |                |                |                |                |                |                |                |                |                |                |                |                |                |                |                |
| <i>LLaMA-3.1-8B</i>        |                |                |                |                |                |                |                |                |                |                |                |                |                |                |                |                |
| Layers: L16, L17, L18, L19 |                |                |                |                |                |                |                |                |                |                |                |                |                |                |                |                |
| Base                       | 0.02<br>(0.01) | 0.01<br>(0.01) | 0.19<br>(0.09) | 0.42<br>(0.10) | 0.09<br>(0.07) | 0.24<br>(0.15) | 0.16<br>(0.14) | 0.34<br>(0.12) | 0.11<br>(0.08) | 0.05<br>(0.04) | 0.21<br>(0.09) | 0.33<br>(0.13) | 0.12<br>(0.08) | 0.42<br>(0.18) | 0.24<br>(0.11) | 0.44<br>(0.10) |
| $\beta=0$                  | 0.79<br>(0.08) | 0.80<br>(0.08) | 0.41<br>(0.13) | 0.62<br>(0.20) | 0.77<br>(0.09) | 0.80<br>(0.08) | 0.35<br>(0.11) | 0.57<br>(0.08) | 0.79<br>(0.08) | 0.79<br>(0.08) | 0.46<br>(0.10) | 0.72<br>(0.12) | 0.78<br>(0.08) | 0.79<br>(0.08) | 0.34<br>(0.12) | 0.62<br>(0.15) |
| ProbeSAE                   | 0.16<br>(0.10) | 0.89<br>(0.06) | 0.26<br>(0.15) | 0.58<br>(0.28) | 0.09<br>(0.07) | 0.91<br>(0.04) | 0.31<br>(0.20) | 0.73<br>(0.25) | 0.09<br>(0.07) | 0.85<br>(0.05) | 0.34<br>(0.24) | 0.73<br>(0.14) | 0.22<br>(0.15) | 0.84<br>(0.05) | 0.33<br>(0.18) | 0.79<br>(0.17) |
| G-SAE                      | 0.80<br>(0.08) | 0.82<br>(0.08) | 0.75<br>(0.09) | 0.81<br>(0.06) | 0.80<br>(0.08) | 0.80<br>(0.08) | 0.74<br>(0.07) | 0.82<br>(0.03) | 0.70<br>(0.12) | 0.79<br>(0.08) | 0.80<br>(0.14) | 0.86<br>(0.07) | 0.34<br>(0.18) | 0.79<br>(0.08) | 0.76<br>(0.12) | 0.82<br>(0.11) |
| <i>Gemma-2-2B</i>          |                |                |                |                |                |                |                |                |                |                |                |                |                |                |                |                |
| Layers: L13, L16, L24, L25 |                |                |                |                |                |                |                |                |                |                |                |                |                |                |                |                |
| Base                       | 0.04<br>(0.03) | 0.01<br>(0.01) | 0.70<br>(0.11) | 0.86<br>(0.06) | 0.01<br>(0.01) | 0.03<br>(0.04) | 0.67<br>(0.09) | 0.88<br>(0.07) | 0.07<br>(0.08) | 0.33<br>(0.29) | 0.69<br>(0.21) | 0.90<br>(0.17) | 0.05<br>(0.05) | 0.29<br>(0.22) | 0.53<br>(0.32) | 0.74<br>(0.38) |
| $\beta=0$                  | 0.01<br>(0.01) | 0.04<br>(0.02) | 0.76<br>(0.11) | 0.88<br>(0.06) | 0.01<br>(0.01) | 0.11<br>(0.10) | 0.66<br>(0.11) | 0.91<br>(0.05) | 0.06<br>(0.06) | 0.23<br>(0.24) | 0.76<br>(0.10) | 0.93<br>(0.05) | 0.09<br>(0.08) | 0.22<br>(0.20) | 0.62<br>(0.36) | 0.77<br>(0.39) |
| ProbeSAE                   | 0.18<br>(0.04) | 0.38<br>(0.07) | 0.86<br>(0.07) | 0.95<br>(0.03) | 0.02<br>(0.01) | 0.21<br>(0.16) | 0.74<br>(0.11) | 0.92<br>(0.04) | 0.20<br>(0.21) | 0.40<br>(0.29) | 0.92<br>(0.09) | 0.98<br>(0.04) | 0.18<br>(0.14) | 0.44<br>(0.27) | 0.71<br>(0.37) | 0.79<br>(0.39) |
| G-SAE                      | 0.00<br>(0.00) | 0.02<br>(0.01) | 0.81<br>(0.09) | 0.91<br>(0.05) | 0.01<br>(0.01) | 0.06<br>(0.06) | 0.66<br>(0.09) | 0.90<br>(0.06) | 0.06<br>(0.07) | 0.21<br>(0.23) | 0.77<br>(0.11) | 0.93<br>(0.06) | 0.07<br>(0.07) | 0.21<br>(0.19) | 0.58<br>(0.36) | 0.77<br>(0.39) |
| <b>Sentiment</b>           |                |                |                |                |                |                |                |                |                |                |                |                |                |                |                |                |
| <i>LLaMA-3.1-8B</i>        |                |                |                |                |                |                |                |                |                |                |                |                |                |                |                |                |
| Layers: L16, L17, L18, L19 |                |                |                |                |                |                |                |                |                |                |                |                |                |                |                |                |
| Base                       | 0.05<br>(0.02) | 0.71<br>(0.12) | 0.01<br>(0.01) | 0.02<br>(0.01) | 0.05<br>(0.02) | 0.75<br>(0.03) | 0.01<br>(0.01) | 0.03<br>(0.01) | 0.15<br>(0.05) | 0.46<br>(0.12) | 0.01<br>(0.00) | 0.02<br>(0.01) | 0.20<br>(0.04) | 0.82<br>(0.11) | 0.01<br>(0.01) | 0.02<br>(0.01) |
| $\beta=0$                  | 0.68<br>(0.09) | 0.93<br>(0.04) | 0.03<br>(0.01) | 0.06<br>(0.03) | 0.72<br>(0.12) | 0.93<br>(0.06) | 0.03<br>(0.02) | 0.04<br>(0.02) | 0.86<br>(0.13) | 0.87<br>(0.13) | 0.03<br>(0.01) | 0.05<br>(0.03) | 0.76<br>(0.08) | 0.88<br>(0.11) | 0.03<br>(0.01) | 0.03<br>(0.01) |
| ProbeSAE                   | 0.30<br>(0.09) | 0.93<br>(0.05) | 0.04<br>(0.03) | 0.07<br>(0.02) | 0.27<br>(0.05) | 0.93<br>(0.06) | 0.03<br>(0.02) | 0.08<br>(0.01) | 0.74<br>(0.08) | 0.87<br>(0.13) | 0.02<br>(0.01) | 0.08<br>(0.02) | 0.55<br>(0.04) | 0.88<br>(0.12) | 0.02<br>(0.01) | 0.05<br>(0.01) |
| G-SAE                      | 0.02<br>(0.01) | 0.22<br>(0.06) | 0.05<br>(0.04) | 0.05<br>(0.05) | 0.78<br>(0.06) | 0.91<br>(0.05) | 0.05<br>(0.04) | 0.05<br>(0.04) | 0.66<br>(0.07) | 0.75<br>(0.09) | 0.06<br>(0.06) | 0.07<br>(0.05) | 0.03<br>(0.01) | 0.49<br>(0.11) | 0.04<br>(0.03) | 0.04<br>(0.03) |
| <i>Gemma-2-2B</i>          |                |                |                |                |                |                |                |                |                |                |                |                |                |                |                |                |
| Layers: L22, L23, L24, L25 |                |                |                |                |                |                |                |                |                |                |                |                |                |                |                |                |
| Base                       | 0.01<br>(0.00) | 0.04<br>(0.03) | 0.36<br>(0.19) | 0.40<br>(0.21) | 0.01<br>(0.01) | 0.06<br>(0.05) | 0.13<br>(0.09) | 0.18<br>(0.12) | 0.02<br>(0.01) | 0.14<br>(0.06) | 0.11<br>(0.05) | 0.16<br>(0.08) | 0.06<br>(0.04) | 0.10<br>(0.08) | 0.18<br>(0.19) | 0.25<br>(0.20) |
| $\beta=0$                  | 0.01<br>(0.00) | 0.07<br>(0.05) | 0.29<br>(0.15) | 0.34<br>(0.18) | 0.01<br>(0.00) | 0.16<br>(0.16) | 0.12<br>(0.08) | 0.17<br>(0.14) | 0.01<br>(0.00) | 0.22<br>(0.06) | 0.12<br>(0.05) | 0.18<br>(0.11) | 0.04<br>(0.06) | 0.42<br>(0.16) | 0.20<br>(0.20) | 0.29<br>(0.24) |
| ProbeSAE                   | 0.02<br>(0.01) | 0.05<br>(0.03) | 0.30<br>(0.14) | 0.40<br>(0.19) | 0.10<br>(0.12) | 0.24<br>(0.22) | 0.13<br>(0.11) | 0.20<br>(0.16) | 0.01<br>(0.01) | 0.11<br>(0.05) | 0.08<br>(0.03) | 0.13<br>(0.07) | 0.13<br>(0.08) | 0.45<br>(0.15) | 0.18<br>(0.17) | 0.26<br>(0.24) |
| G-SAE                      | 0.01<br>(0.00) | 0.07<br>(0.05) | 0.29<br>(0.16) | 0.36<br>(0.22) | 0.00<br>(0.00) | 0.16<br>(0.16) | 0.13<br>(0.10) | 0.17<br>(0.12) | 0.01<br>(0.00) | 0.20<br>(0.07) | 0.13<br>(0.06) | 0.17<br>(0.09) | 0.05<br>(0.07) | 0.16<br>(0.13) | 0.18<br>(0.16) | 0.26<br>(0.19) |

Table 3: **Causal Results** across both evaluation domains. Each cell shows the mean Flip Rate across the 5 OOD datasets of its domain (HateCheck EN/ES/IT/FR + MeasuringHate for Toxicity; Amazon Reviews EN/DE/ES/FR/JA for Sentiment), with the **standard deviation** in parentheses on the line below. **Ablation (Abl.)**: removing top- $k$  target-aligned features from correctly-classified target samples (Flip Rate TP→TN). **Steering (Str.)**: additively injecting them into correctly-classified non-target samples (Flip Rate TN→TP). Interventions are applied at two strengths,  $k=10$  and  $k=50$ . Columns are layer indices in selection order; the actual layer numbers per (model, domain) are reported in the sub-header of each block. Colour intensity reflects flip-rate magnitude (darker is higher).

language: the advantage is obtained on multilingual data despite training exclusively on English. This cross-lingual consistency suggests that the decoder directions learned by ProbeSAE are more aligned with language-agnostic, semantically abstract model behaviour rather than English-specific surface patterns.

## 6 Conclusion

We formalize the *concept faithfulness gap* as the difference between the performance of a linear probe trained on a model’s dense activations and that of the same probe trained on an SAE’s sparse representation of them. When the sparse representation underperforms, the decomposition has dropped information the dense activations provably contain. As interpretability investigations operate directly on individual sparse features, information lost during decomposition directly undermines interpretation faithfulness. To address this, we introduce ProbeSAE, which folds a Task-Guided Steering term into the sparse autoencoding loss.

Against the original Base SAEs, unguided fine-tuning ( $\beta=0$ ) and concept-conditioned G-SAE, ProbeSAE addresses the faithfulness gap consistently, often matching or surpassing the dense ceiling; it preserves or improves auto-interpretability, attains the highest monosemanticity, and strengthens causal control under both interventions. A classification term makes the probe-utility gain expected; the gains on the other two axes are not, and neither fine-tuned baseline obtains them. Although trained strictly on English, these gains transfer zero-shot to every evaluation language, including scripts as distant as Japanese, whereas unguided fine-tuning degrades cross-lingual interpretability. This suggests that probe guidance reinforces abstract, language-agnostic representations rather than superficial shortcuts to the training label. Together, these results establish probe guidance as a principled mechanism for addressing the concept faithfulness gap in domain-specific interpretability analyses.

## Limitations

ProbeSAE requires labelled data for a specific target concept, supports binary classification only, and produces one fine-tuned SAE per (concept, layer) pair, so it is a domain-specific adaptation rather than a general-purpose SAE improvement. Our evaluation covers two model families in the 2-8B range; whether the gains hold at larger scales, or for other architectures, is untested. Probe utility is measured only with linear probes, the same class used during training, so we cannot say whether the gap closes for non-linear ones.

## Acknowledgments

All experiments were conducted on the Leonardo supercomputer with the support of CINECA-Italian Super Computing Resource Allocation, class C project IsCc9\_MI-PLE (HP10CIQUBQ).

## References

- Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, and 1 others. 2025. [gpt-oss-120b & gpt-oss-20b model card](#). *arXiv preprint arXiv:2508.10925*.
- Guillaume Alain and Yoshua Bengio. 2016. [Understanding intermediate layers using linear classifier probes](#). *arXiv preprint arXiv:1610.01644*.
- Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. 2023. Language models can explain neurons in language models. *OpenAI Blog*.
- Joseph Bloom, Curt Tigges, Anthony Duong, and David Chanin. 2024. [Saelens](#). <https://github.com/decoderresearch/SAELens>.
- Dan Braun, Jordan Taylor, Nicholas Goldowsky-Dill, and Lee Sharkey. 2024. Identifying functionally important features with end-to-end sparse dictionary learning. *Advances in Neural Information Processing Systems*, 37:107286–107325.
- Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermy, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, and 1 others. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Anthropic Research*.
- Andrea Ermellino, Lorenzo Malandri, Fabio Mercorio, and Antonio Serino. 2026. [Safe-unsafe concept separation emerges from a single direction in language models activation space](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3019–3034.
- Sandeep Reddy Gantla. 2025. Exploring mechanistic interpretability in large language models: Challenges, approaches, and insights. In *2025 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI)*, pages 1–8. IEEE.
- Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. 2025. Scaling and evaluating sparse autoencoders. In *The Thirteenth International Conference on Learning Representations*.
- Rifo Ahmad Genadi, Munachiso S Nwadike, Nurdaulet Mukhituly, Tatsuya Hiraoka, Hilal AlQuabeh, and Kentaro Inui. 2026. [Sycophancy hides linearly in the attention heads](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6896–6912.
- Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. [The llama 3 herd of models](#). *arXiv preprint arXiv:2407.21783*.
- Ruben Härle, Felix Friedrich, Manuel Brack, Stephan Wäldchen, Björn Deiseroth, Patrick Schramowski, and Kristian Kersting. 2025. Measuring and guiding monosemanticity. *Advances in Neural Information Processing Systems*, 38:167468–167508.
- Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. [Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326.
- Zirui He, Mingyu Jin, Bo Shen, Ali Payani, Yongfeng Zhang, and Mengnan Du. 2025. [Sae-ssv: Supervised steering in sparse representation spaces for reliable control of language models](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2207–2236.
- Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. 2024. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*.
- Alon Jacovi and Yoav Goldberg. 2020. [Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness?](#) In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4198–4205.
- Subhash Kantamneni, Joshua Engels, Senthoooran Rajamanoharan, Max Tegmark, and Neel Nanda. 2025. Are sparse autoencoders useful? a case study in sparse

- probing. In *Forty-second International Conference on Machine Learning*.
- Adam Karvonen, Can Rager, Johnny Lin, Curt Tigges, Joseph Isaac Bloom, David Chanin, Yeu-Tong Lau, Eoin Farrell, Callum Stuart McDougall, Kola Ayonrinde, and 1 others. 2025. Saebench: A comprehensive benchmark for sparse autoencoders in language model interpretability. In *International Conference on Machine Learning*, pages 29223–29264. PMLR.
- Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. 2020. [The multilingual amazon reviews corpus](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 4563–4568.
- Akshay Kulkarni, Tsui-Wei Weng, Vivek Narayanaswamy, Shusen Liu, Wesam A Sakla, and Kowshik Thopalli. 2025. [Interpretable and steerable concept bottleneck sparse autoencoders](#). *arXiv preprint arXiv:2512.10805*.
- Tom Lieberum, Senthoran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. 2024. [Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2](#). *arXiv preprint arXiv:2408.05147*.
- Samuel Marks, Can Rager, Eric Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. 2025. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *The Thirteenth International Conference on Learning Representations*.
- Fabio Mercorio, Filippo Pallucchini, Daniele Poterì, Antonio Serino, and Andrea Seveso. 2025. [Sfal: Semantic-functional alignment scores for distributional evaluation of auto-interpretability in sparse autoencoders](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 576–583.
- Bruno A Olshausen and David J Field. 1997. [Sparse coding with an overcomplete basis set: A strategy employed by v1?](#) *Vision research*, 37(23):3311–3325.
- Gonçalo Santos Paulo, Alex Troy Mallen, Caden Juang, and Nora Belrose. 2025. Automatically interpreting millions of features in large language models. In *International Conference on Machine Learning*, pages 48393–48421. PMLR.
- Daniele Poterì, Andrea Seveso, and Fabio Mercorio. 2025. [Can role vectors affect LLM behaviour?](#) In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 17735–17747, Suzhou, China. Association for Computational Linguistics.
- Senthoran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. 2024. [Jumping ahead: Improving reconstruction fidelity with jumprelu sparse autoencoders](#). *arXiv preprint arXiv:2407.14435*.
- Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Talat, Helen Margetts, and Janet Pierrehumbert. 2021. [Hatecheck: Functional tests for hate speech detection models](#). In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, pages 41–58.
- Pratik Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia Von Vacano, and Chris Kennedy. 2022. [The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism](#). In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP@ LREC2022*, pages 83–94.
- Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. [Gemma 2: Improving open language models at a practical size](#). *arXiv preprint arXiv:2408.00118*.
- Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, and 3 others. 2024. [Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet](#). *Transformer Circuits Thread*.
- Minglai Yang, Xinyu Guo, Zhengliang Shi, Jinhe Bi, Steven Bethard, Mihai Surdeanu, and Liangming Pan. 2025. [Alignsae: Concept-aligned sparse autoencoders](#). *arXiv preprint arXiv:2512.02004*.

## Appendix

### A Layer Selection

Tables 4 and 5 report the AUROC scores of logistic regression probes trained on the residual stream activations of each analyzed layer, for toxicity and sentiment respectively. The analysis is restricted to the middle-to-late layers of each model (L16-L31 for LLaMA-3.1-8B, L13-L25 for Gemma-2-2B), as earlier layers predominantly encode syntactic and positional structure while task-relevant semantic abstractions tend to become linearly separable in deeper layers. The four layers with the highest probe AUROC for each (model, domain) combination are selected for SAE fine-tuning and highlighted in bold.

For toxicity, LLaMA-3.1-8B shows a contiguous plateau of high performance from L16 to L19 with a gradual decline thereafter, indicating that toxicity information is densely concentrated in the upper-middle block. Gemma-2-2B exhibits a more distributed pattern, with high-performing layers at both the middle (L13, L16) and the late portion (L24, L25) of the network, suggesting multi-stage refinement of toxicity-relevant features. For sentiment, LLaMA-3.1-8B again peaks at the same L16-L19 window, confirming a stable semantic processing region across both domains. Gemma-2-2B on sentiment shifts entirely to the final layers (L22-L25), where probe accuracy is highest, reflecting a different internal organization for valence-related features relative to toxicity.

| Llama 3.1 8B |              | Gemma 2 2B |              |
|--------------|--------------|------------|--------------|
| Layer        | AUROC        | Layer      | AUROC        |
| 16           | <b>0.902</b> | 13         | <b>0.811</b> |
| 17           | <b>0.901</b> | 14         | 0.796        |
| 18           | <b>0.901</b> | 15         | 0.790        |
| 19           | <b>0.896</b> | 16         | <b>0.809</b> |
| 20           | 0.894        | 17         | 0.781        |
| 21           | 0.891        | 18         | 0.783        |
| 22           | 0.881        | 19         | 0.785        |
| 23           | 0.880        | 20         | 0.789        |
| 24           | 0.877        | 21         | 0.781        |
| 25           | 0.875        | 22         | 0.790        |
| 26           | 0.880        | 23         | 0.785        |
| 27           | 0.878        | 24         | <b>0.802</b> |
| 28           | 0.875        | 25         | <b>0.811</b> |
| 29           | 0.867        |            |              |
| 30           | 0.870        |            |              |
| 31           | 0.861        |            |              |

Table 4: **Probe Performance by Layer: Toxicity.** AUROC of linear probes trained to classify toxicity on the residual stream. The four selected layers per model are in bold.

| Llama 3.1 8B |              | Gemma 2 2B |              |
|--------------|--------------|------------|--------------|
| Layer        | AUROC        | Layer      | AUROC        |
| <b>16</b>    | <b>0.838</b> | 12         | 0.813        |
| <b>17</b>    | <b>0.836</b> | 13         | 0.822        |
| <b>18</b>    | <b>0.837</b> | 14         | 0.821        |
| <b>19</b>    | <b>0.837</b> | 15         | 0.827        |
| 20           | 0.833        | 16         | 0.834        |
| 21           | 0.831        | 17         | 0.833        |
| 22           | 0.829        | 18         | 0.829        |
| 23           | 0.827        | 19         | 0.830        |
| 24           | 0.827        | 20         | 0.832        |
| 25           | 0.828        | 21         | 0.840        |
| 26           | 0.830        | <b>22</b>  | <b>0.844</b> |
| 27           | 0.829        | <b>23</b>  | <b>0.842</b> |
| 28           | 0.829        | <b>24</b>  | <b>0.850</b> |
| 29           | 0.832        | <b>25</b>  | <b>0.851</b> |
| 30           | 0.829        |            |              |
| 31           | 0.830        |            |              |

Table 5: **Probe Performance by Layer: Sentiment.** AUROC of linear probes trained to classify sentiment on the residual stream. The four selected layers per model are in bold.

### B Hyperparameter Search

We grid-search  $\beta_c \in \{1, 5, 10, 15, 20, 50\} \times N \in \{60k, 80k, 100k\} \times E \in \{3, 5\}$ , 36 configurations per selected layer, for both models across both domains (576 runs in total). For each layer we select the configuration that best balances AUROC, sparsity and reconstruction ability. AUROC,  $L_0$  and  $R^2$  for every configuration, together with the base-SAE reference metrics of each layer and a flag marking the selected configuration, are released as CSV files in the code repository under [https://github.com/serino28/Probe-Guided-Sparse-Autoencoders/tree/main/grid\\_search](https://github.com/serino28/Probe-Guided-Sparse-Autoencoders/tree/main/grid_search).

### C Full OOD Results

Tables 6-9 report the per-dataset OOD results (AUROC,  $R^2$ ,  $L_0$ ) for all four model-domain combinations, stratified by individual OOD language and dataset, together with a *Dense* row representing the concept faithfulness upper bound (logistic probe on raw residual-stream activations; no  $R^2$  or  $L_0$  reported). These tables complement the layer-averaged summaries in the main text by showing the full distribution of performance across languages and the per-dataset relationship between ProbeSAE and the dense probe ceiling. Aggregated over the four tables, ProbeSAE attains the highest AUROC of any SAE condition on 61 of the 80 language-layer pairs, G-SAE on 13, the Base SAE on 5 and  $\beta=0$  on 1. For LLaMA-3.1-8B

on toxicity (Table 6), ProbeSAE improves on the Base SAE at every language-layer pair and on  $\beta = 0$  at 19 of 20, and its layer-averaged AUROC meets or exceeds the dense probe at all four layers, with the largest absolute gains concentrated in the Spanish and French HateCheck datasets; G-SAE is the closest competitor in this cell, leading on 7 of the 20 pairs, concentrated on HateCheck-EN and on L19. For LLaMA-3.1-8B on sentiment (Table 7), ProbeSAE exceeds the dense probe in three of four layers across all five Amazon Reviews languages, including Japanese, the language most distant from the English training distribution, and is the best SAE condition on 19 of 20 pairs; this is also the cell where its margin over G-SAE is widest (layer-averaged 0.83 vs. 0.70). For Gemma-2-2B on toxicity (Table 8), the dense probe ceiling is substantially higher ( $\approx 0.81$ -0.94); ProbeSAE closes a meaningful fraction of the gap on most layer-language pairs, while both  $\beta = 0$  and G-SAE fall below the base SAE on 10 of the 16 non-English pairs. For Gemma-2-2B on sentiment (Table 9), the dense ceiling is very high ( $\approx 0.93$ -0.96) and no sparse condition approaches it; ProbeSAE is the best SAE condition on 13 of 20 pairs, and this is the one cell where G-SAE is competitive with it, leading on Amazon FR at L22 and L24 and on German and Spanish at L24. Amazon FR is the systematic weak spot of ProbeSAE here, which is outscored on it at three of the four layers, by G-SAE at L22 and L24 and by the Base SAE at L23.

| L   | Cond        | HC-en          |               |       | HC-es          |               |       | HC-it          |               |       | HC-fr          |               |       | MH             |               |       |
|-----|-------------|----------------|---------------|-------|----------------|---------------|-------|----------------|---------------|-------|----------------|---------------|-------|----------------|---------------|-------|
|     |             | AUC $\uparrow$ | $R^2\uparrow$ | $L_0$ | AUC $\uparrow$ | $R^2\uparrow$ | $L_0$ | AUC $\uparrow$ | $R^2\uparrow$ | $L_0$ | AUC $\uparrow$ | $R^2\uparrow$ | $L_0$ | AUC $\uparrow$ | $R^2\uparrow$ | $L_0$ |
| L16 | Base        | 0.571          | 0.519         | 50.0  | 0.562          | 0.516         | 50.0  | 0.544          | 0.530         | 50.0  | 0.553          | 0.514         | 50.0  | 0.666          | 0.488         | 50.0  |
|     | $\beta = 0$ | 0.711          | 0.912         | 50.0  | 0.650          | 0.927         | 50.0  | 0.634          | 0.908         | 50.0  | 0.646          | 0.925         | 50.0  | 0.757          | 0.941         | 50.0  |
|     | ProbeSAE    | 0.728          | 0.820         | 50.0  | 0.762          | 0.851         | 50.0  | 0.717          | 0.806         | 50.0  | 0.724          | 0.848         | 50.0  | 0.789          | 0.885         | 50.0  |
|     | G-SAE       | 0.787          | 0.917         | 50.0  | 0.707          | 0.936         | 50.0  | 0.687          | 0.917         | 50.0  | 0.694          | 0.935         | 50.0  | 0.757          | 0.944         | 50.0  |
|     | Dense       | 0.689          |               |       | 0.743          |               |       | 0.693          |               |       | 0.708          |               |       | 0.704          |               |       |
| L17 | Base        | 0.621          | 0.537         | 50.0  | 0.630          | 0.528         | 50.0  | 0.586          | 0.545         | 50.0  | 0.620          | 0.528         | 50.0  | 0.653          | 0.494         | 50.0  |
|     | $\beta = 0$ | 0.659          | 0.913         | 50.0  | 0.628          | 0.928         | 50.0  | 0.640          | 0.908         | 50.0  | 0.621          | 0.927         | 50.0  | 0.743          | 0.941         | 50.0  |
|     | ProbeSAE    | 0.766          | 0.820         | 50.0  | 0.734          | 0.858         | 50.0  | 0.683          | 0.814         | 50.0  | 0.695          | 0.856         | 50.0  | 0.790          | 0.889         | 50.0  |
|     | G-SAE       | 0.767          | 0.920         | 50.0  | 0.693          | 0.933         | 50.0  | 0.672          | 0.914         | 50.0  | 0.695          | 0.932         | 50.0  | 0.767          | 0.943         | 50.0  |
|     | Dense       | 0.754          |               |       | 0.732          |               |       | 0.689          |               |       | 0.696          |               |       | 0.698          |               |       |
| L18 | Base        | 0.594          | 0.547         | 50.0  | 0.585          | 0.538         | 50.0  | 0.618          | 0.561         | 50.0  | 0.627          | 0.538         | 50.0  | 0.689          | 0.509         | 50.0  |
|     | $\beta = 0$ | 0.752          | 0.916         | 50.0  | 0.657          | 0.931         | 50.0  | 0.655          | 0.913         | 50.0  | 0.656          | 0.930         | 50.0  | 0.763          | 0.941         | 50.0  |
|     | ProbeSAE    | 0.757          | 0.822         | 50.0  | 0.742          | 0.863         | 50.0  | 0.706          | 0.821         | 50.0  | 0.708          | 0.861         | 50.0  | 0.774          | 0.892         | 50.0  |
|     | G-SAE       | 0.783          | 0.922         | 50.0  | 0.679          | 0.935         | 50.0  | 0.672          | 0.918         | 50.0  | 0.688          | 0.934         | 50.0  | 0.781          | 0.943         | 50.0  |
|     | Dense       | 0.767          |               |       | 0.722          |               |       | 0.675          |               |       | 0.660          |               |       | 0.713          |               |       |
| L19 | Base        | 0.667          | 0.562         | 50.0  | 0.643          | 0.558         | 50.0  | 0.648          | 0.575         | 50.0  | 0.627          | 0.555         | 50.0  | 0.725          | 0.533         | 50.0  |
|     | $\beta = 0$ | 0.688          | 0.926         | 50.0  | 0.639          | 0.936         | 50.0  | 0.624          | 0.919         | 50.0  | 0.655          | 0.935         | 50.0  | 0.778          | 0.946         | 50.0  |
|     | ProbeSAE    | 0.791          | 0.848         | 50.0  | 0.708          | 0.875         | 50.0  | 0.660          | 0.833         | 50.0  | 0.682          | 0.872         | 50.0  | 0.775          | 0.904         | 50.0  |
|     | G-SAE       | 0.769          | 0.931         | 50.0  | 0.681          | 0.940         | 50.0  | 0.670          | 0.923         | 50.0  | 0.689          | 0.938         | 50.0  | 0.779          | 0.948         | 50.0  |
|     | Dense       | 0.803          |               |       | 0.730          |               |       | 0.670          |               |       | 0.679          |               |       | 0.732          |               |       |

Table 6: **Full OOD Results: LLaMA-3.1-8B, Toxicity.** AUROC, reconstruction fidelity ( $R^2$ ) and sparsity ( $L_0$ ) for each OOD dataset. The *Dense* row shows the faithfulness upper bound (logistic probe on raw residual-stream activations; no  $R^2$  or  $L_0$ ). Colour intensity reflects relative performance within each (metric, dataset) column (darker = better for AUROC and  $R^2$ ; darker = lower for  $L_0$ ).

| L   | Cond        | Amz-en         |               |       | Amz-de         |               |       | Amz-es         |               |       | Amz-fr         |               |       | Amz-ja         |               |       |
|-----|-------------|----------------|---------------|-------|----------------|---------------|-------|----------------|---------------|-------|----------------|---------------|-------|----------------|---------------|-------|
|     |             | AUC $\uparrow$ | $R^2\uparrow$ | $L_0$ | AUC $\uparrow$ | $R^2\uparrow$ | $L_0$ | AUC $\uparrow$ | $R^2\uparrow$ | $L_0$ | AUC $\uparrow$ | $R^2\uparrow$ | $L_0$ | AUC $\uparrow$ | $R^2\uparrow$ | $L_0$ |
| L16 | Base        | 0.666          | 0.478         | 50.0  | 0.656          | 0.486         | 50.0  | 0.648          | 0.477         | 50.0  | 0.632          | 0.475         | 50.0  | 0.563          | 0.497         | 50.0  |
|     | $\beta = 0$ | 0.813          | 0.899         | 50.0  | 0.782          | 0.899         | 50.0  | 0.806          | 0.908         | 50.0  | 0.807          | 0.907         | 50.0  | 0.669          | 0.844         | 50.0  |
|     | ProbeSAE    | 0.822          | 0.847         | 50.0  | 0.787          | 0.836         | 50.0  | 0.821          | 0.854         | 50.0  | 0.796          | 0.852         | 50.0  | 0.716          | 0.729         | 50.0  |
|     | G-SAE       | 0.711          | 0.904         | 50.0  | 0.721          | 0.904         | 50.0  | 0.778          | 0.916         | 50.0  | 0.735          | 0.914         | 50.0  | 0.530          | 0.852         | 50.0  |
|     | Dense       | 0.840          |               |       | 0.785          |               |       | 0.816          |               |       | 0.810          |               |       | 0.719          |               |       |
| L17 | Base        | 0.599          | 0.480         | 50.0  | 0.619          | 0.494         | 50.0  | 0.615          | 0.481         | 50.0  | 0.637          | 0.482         | 50.0  | 0.519          | 0.505         | 50.0  |
|     | $\beta = 0$ | 0.753          | 0.905         | 50.0  | 0.709          | 0.905         | 50.0  | 0.731          | 0.916         | 50.0  | 0.709          | 0.914         | 50.0  | 0.612          | 0.852         | 50.0  |
|     | ProbeSAE    | 0.896          | 0.857         | 50.0  | 0.845          | 0.851         | 50.0  | 0.885          | 0.866         | 50.0  | 0.866          | 0.865         | 50.0  | 0.754          | 0.755         | 50.0  |
|     | G-SAE       | 0.777          | 0.912         | 50.0  | 0.726          | 0.910         | 50.0  | 0.819          | 0.923         | 50.0  | 0.757          | 0.921         | 50.0  | 0.584          | 0.858         | 50.0  |
|     | Dense       | 0.773          |               |       | 0.776          |               |       | 0.832          |               |       | 0.807          |               |       | 0.715          |               |       |
| L18 | Base        | 0.636          | 0.477         | 50.0  | 0.642          | 0.489         | 50.0  | 0.666          | 0.480         | 50.0  | 0.638          | 0.481         | 50.0  | 0.538          | 0.502         | 50.0  |
|     | $\beta = 0$ | 0.773          | 0.911         | 50.0  | 0.687          | 0.907         | 50.0  | 0.737          | 0.919         | 50.0  | 0.678          | 0.916         | 50.0  | 0.588          | 0.846         | 50.0  |
|     | ProbeSAE    | 0.858          | 0.865         | 50.0  | 0.818          | 0.857         | 50.0  | 0.877          | 0.872         | 50.0  | 0.848          | 0.872         | 50.0  | 0.750          | 0.756         | 50.0  |
|     | G-SAE       | 0.762          | 0.917         | 50.0  | 0.664          | 0.911         | 50.0  | 0.802          | 0.923         | 50.0  | 0.729          | 0.921         | 50.0  | 0.582          | 0.853         | 50.0  |
|     | Dense       | 0.823          |               |       | 0.798          |               |       | 0.824          |               |       | 0.813          |               |       | 0.707          |               |       |
| L19 | Base        | 0.704          | 0.508         | 50.0  | 0.673          | 0.517         | 50.0  | 0.620          | 0.509         | 50.0  | 0.669          | 0.510         | 50.0  | 0.546          | 0.524         | 50.0  |
|     | $\beta = 0$ | 0.753          | 0.919         | 50.0  | 0.699          | 0.914         | 50.0  | 0.760          | 0.925         | 50.0  | 0.705          | 0.924         | 50.0  | 0.640          | 0.853         | 50.0  |
|     | ProbeSAE    | 0.857          | 0.874         | 50.0  | 0.841          | 0.870         | 50.0  | 0.883          | 0.883         | 50.0  | 0.869          | 0.883         | 50.0  | 0.766          | 0.784         | 50.0  |
|     | G-SAE       | 0.709          | 0.923         | 50.0  | 0.632          | 0.917         | 50.0  | 0.792          | 0.929         | 50.0  | 0.704          | 0.928         | 50.0  | 0.568          | 0.860         | 50.0  |
|     | Dense       | 0.844          |               |       | 0.759          |               |       | 0.781          |               |       | 0.773          |               |       | 0.705          |               |       |

Table 7: **Full OOD Results: LLaMA-3.1-8B, Sentiment.** AUROC, reconstruction fidelity ( $R^2$ ) and sparsity ( $L_0$ ) for each OOD dataset. The *Dense* row shows the faithfulness upper bound (logistic probe on raw residual-stream activations; no  $R^2$  or  $L_0$ ). Colour intensity reflects relative performance within each (metric, dataset) column (darker = better for AUROC and  $R^2$ ; darker = lower for  $L_0$ ).

| L   | Cond        | HC-en          |               |       | HC-es          |               |       | HC-it          |               |       | HC-fr          |               |       | MH             |               |       |
|-----|-------------|----------------|---------------|-------|----------------|---------------|-------|----------------|---------------|-------|----------------|---------------|-------|----------------|---------------|-------|
|     |             | AUC $\uparrow$ | $R^2\uparrow$ | $L_0$ | AUC $\uparrow$ | $R^2\uparrow$ | $L_0$ | AUC $\uparrow$ | $R^2\uparrow$ | $L_0$ | AUC $\uparrow$ | $R^2\uparrow$ | $L_0$ | AUC $\uparrow$ | $R^2\uparrow$ | $L_0$ |
| L13 | Base        | 0.868          | 0.847         | 72.4  | 0.809          | 0.874         | 84.9  | 0.779          | 0.872         | 85.4  | 0.793          | 0.873         | 85.7  | 0.709          | 0.864         | 100.1 |
|     | $\beta = 0$ | 0.860          | 0.878         | 58.6  | 0.787          | 0.895         | 67.5  | 0.757          | 0.890         | 68.9  | 0.770          | 0.891         | 69.0  | 0.698          | 0.873         | 86.0  |
|     | ProbeSAE    | 0.905          | 0.879         | 68.5  | 0.814          | 0.895         | 76.9  | 0.791          | 0.890         | 78.1  | 0.798          | 0.892         | 78.4  | 0.732          | 0.875         | 95.9  |
|     | G-SAE       | 0.886          | 0.882         | 68.2  | 0.798          | 0.899         | 77.1  | 0.762          | 0.894         | 78.9  | 0.783          | 0.895         | 79.0  | 0.709          | 0.877         | 101.0 |
|     | Dense       | 0.941          |               |       | 0.884          |               |       | 0.852          |               |       | 0.858          |               |       | 0.805          |               |       |
| L16 | Base        | 0.777          | 0.857         | 55.9  | 0.790          | 0.902         | 79.2  | 0.733          | 0.902         | 75.4  | 0.741          | 0.901         | 76.6  | 0.702          | 0.884         | 91.4  |
|     | $\beta = 0$ | 0.785          | 0.894         | 42.0  | 0.759          | 0.914         | 58.2  | 0.749          | 0.903         | 56.7  | 0.709          | 0.909         | 57.7  | 0.684          | 0.894         | 73.6  |
|     | ProbeSAE    | 0.831          | 0.893         | 45.3  | 0.777          | 0.914         | 61.6  | 0.764          | 0.904         | 60.2  | 0.725          | 0.909         | 61.2  | 0.704          | 0.894         | 78.7  |
|     | G-SAE       | 0.798          | 0.897         | 46.5  | 0.760          | 0.917         | 64.6  | 0.754          | 0.906         | 63.5  | 0.720          | 0.912         | 64.7  | 0.686          | 0.897         | 85.3  |
|     | Dense       | 0.927          |               |       | 0.861          |               |       | 0.838          |               |       | 0.842          |               |       | 0.809          |               |       |
| L24 | Base        | 0.733          | 0.907         | 42.7  | 0.702          | 0.913         | 63.9  | 0.632          | 0.912         | 58.1  | 0.487          | 0.914         | 58.5  | 0.659          | 0.860         | 71.2  |
|     | $\beta = 0$ | 0.798          | 0.939         | 30.5  | 0.667          | 0.889         | 54.4  | 0.605          | 0.849         | 48.7  | 0.542          | 0.869         | 53.3  | 0.666          | 0.875         | 56.8  |
|     | ProbeSAE    | 0.852          | 0.926         | 49.8  | 0.742          | 0.886         | 74.0  | 0.712          | 0.847         | 72.3  | 0.587          | 0.873         | 71.2  | 0.708          | 0.868         | 78.0  |
|     | G-SAE       | 0.813          | 0.942         | 34.7  | 0.674          | 0.893         | 70.2  | 0.647          | 0.853         | 69.9  | 0.591          | 0.871         | 70.8  | 0.667          | 0.878         | 70.0  |
|     | Dense       | 0.861          |               |       | 0.812          |               |       | 0.780          |               |       | 0.779          |               |       | 0.814          |               |       |
| L25 | Base        | 0.824          | 0.921         | 86.3  | 0.716          | 0.908         | 84.0  | 0.675          | 0.902         | 85.0  | 0.639          | 0.909         | 84.5  | 0.734          | 0.854         | 106.1 |
|     | $\beta = 0$ | 0.836          | 0.947         | 52.5  | 0.720          | 0.877         | 75.0  | 0.709          | 0.848         | 86.6  | 0.687          | 0.856         | 85.4  | 0.717          | 0.867         | 79.8  |
|     | ProbeSAE    | 0.865          | 0.943         | 58.9  | 0.781          | 0.877         | 81.8  | 0.731          | 0.848         | 92.1  | 0.752          | 0.855         | 87.4  | 0.726          | 0.863         | 84.7  |
|     | G-SAE       | 0.831          | 0.951         | 59.4  | 0.705          | 0.882         | 101.9 | 0.670          | 0.840         | 115.6 | 0.664          | 0.859         | 109.5 | 0.714          | 0.873         | 100.4 |
|     | Dense       | 0.855          |               |       | 0.817          |               |       | 0.775          |               |       | 0.783          |               |       | 0.813          |               |       |

Table 8: **Full OOD Results: Gemma-2-2B, Toxicity.** AUROC, reconstruction fidelity ( $R^2$ ) and sparsity ( $L_0$ ) for each OOD dataset. The *Dense* row shows the faithfulness upper bound (logistic probe on raw residual-stream activations; no  $R^2$  or  $L_0$ ). Colour intensity reflects relative performance within each (metric, dataset) column (darker = better for AUROC and  $R^2$ ; darker = lower for  $L_0$ ).

| L   | Cond      | Amz-en |         |       | Amz-de |         |       | Amz-es |         |       | Amz-fr |         |       | Amz-ja |         |       |
|-----|-----------|--------|---------|-------|--------|---------|-------|--------|---------|-------|--------|---------|-------|--------|---------|-------|
|     |           | AUC↑   | $R^2$ ↑ | $L_0$ | AUC↑   | $R^2$ ↑ | $L_0$ | AUC↑   | $R^2$ ↑ | $L_0$ | AUC↑   | $R^2$ ↑ | $L_0$ | AUC↑   | $R^2$ ↑ | $L_0$ |
| L22 | Base      | 0.784  | 0.903   | 69.6  | 0.742  | 0.886   | 68.9  | 0.724  | 0.878   | 73.0  | 0.714  | 0.881   | 72.8  | 0.670  | 0.883   | 68.3  |
|     | $\beta=0$ | 0.817  | 0.860   | 62.2  | 0.711  | 0.830   | 62.6  | 0.719  | 0.830   | 67.3  | 0.716  | 0.825   | 66.3  | 0.731  | 0.801   | 60.2  |
|     | ProbeSAE  | 0.845  | 0.863   | 81.1  | 0.888  | 0.824   | 77.8  | 0.869  | 0.832   | 78.7  | 0.603  | 0.827   | 76.9  | 0.753  | 0.814   | 69.7  |
|     | G-SAE     | 0.647  | 0.863   | 77.3  | 0.769  | 0.831   | 76.8  | 0.700  | 0.832   | 80.7  | 0.728  | 0.829   | 81.4  | 0.579  | 0.797   | 71.9  |
|     | Dense     | 0.951  |         |       | 0.940  |         |       | 0.942  |         |       | 0.934  |         |       | 0.938  |         |       |
| L23 | Base      | 0.851  | 0.895   | 71.4  | 0.784  | 0.872   | 74.0  | 0.837  | 0.866   | 79.1  | 0.800  | 0.865   | 79.2  | 0.734  | 0.871   | 65.0  |
|     | $\beta=0$ | 0.705  | 0.845   | 63.9  | 0.750  | 0.801   | 69.9  | 0.752  | 0.807   | 70.4  | 0.691  | 0.795   | 72.2  | 0.647  | 0.783   | 60.6  |
|     | ProbeSAE  | 0.836  | 0.854   | 73.5  | 0.856  | 0.812   | 75.2  | 0.849  | 0.808   | 83.8  | 0.711  | 0.804   | 80.5  | 0.748  | 0.790   | 65.1  |
|     | G-SAE     | 0.747  | 0.848   | 81.4  | 0.790  | 0.802   | 89.3  | 0.810  | 0.810   | 91.0  | 0.750  | 0.803   | 92.8  | 0.669  | 0.776   | 71.4  |
|     | Dense     | 0.951  |         |       | 0.937  |         |       | 0.943  |         |       | 0.934  |         |       | 0.940  |         |       |
| L24 | Base      | 0.567  | 0.877   | 65.4  | 0.571  | 0.864   | 70.3  | 0.566  | 0.865   | 77.9  | 0.527  | 0.861   | 73.1  | 0.635  | 0.881   | 58.1  |
|     | $\beta=0$ | 0.783  | 0.812   | 61.9  | 0.753  | 0.788   | 66.1  | 0.687  | 0.798   | 72.5  | 0.719  | 0.795   | 67.5  | 0.698  | 0.736   | 59.2  |
|     | ProbeSAE  | 0.858  | 0.810   | 68.9  | 0.777  | 0.791   | 74.8  | 0.751  | 0.798   | 79.5  | 0.525  | 0.798   | 73.8  | 0.751  | 0.733   | 61.9  |
|     | G-SAE     | 0.789  | 0.816   | 78.1  | 0.802  | 0.796   | 87.8  | 0.835  | 0.803   | 92.9  | 0.722  | 0.801   | 88.3  | 0.689  | 0.726   | 68.4  |
|     | Dense     | 0.953  |         |       | 0.940  |         |       | 0.943  |         |       | 0.929  |         |       | 0.939  |         |       |
| L25 | Base      | 0.755  | 0.864   | 95.6  | 0.735  | 0.868   | 110.4 | 0.730  | 0.876   | 110.3 | 0.628  | 0.873   | 109.8 | 0.622  | 0.907   | 89.3  |
|     | $\beta=0$ | 0.727  | 0.794   | 93.4  | 0.794  | 0.792   | 121.1 | 0.813  | 0.800   | 112.0 | 0.561  | 0.812   | 114.6 | 0.577  | 0.789   | 106.0 |
|     | ProbeSAE  | 0.861  | 0.787   | 99.5  | 0.794  | 0.794   | 121.6 | 0.854  | 0.799   | 124.8 | 0.697  | 0.810   | 119.4 | 0.681  | 0.792   | 106.3 |
|     | G-SAE     | 0.760  | 0.802   | 120.2 | 0.806  | 0.790   | 152.7 | 0.752  | 0.806   | 146.4 | 0.643  | 0.818   | 145.3 | 0.597  | 0.787   | 112.1 |
|     | Dense     | 0.955  |         |       | 0.938  |         |       | 0.941  |         |       | 0.923  |         |       | 0.930  |         |       |

Table 9: **Full OOD Results: Gemma-2-2B, Sentiment.** AUROC, reconstruction fidelity ( $R^2$ ) and sparsity ( $L_0$ ) for each OOD dataset. The *Dense* row shows the faithfulness upper bound (logistic probe on raw residual-stream activations; no  $R^2$  or  $L_0$ ). Colour intensity reflects relative performance within each (metric, dataset) column (darker = better for AUROC and  $R^2$ ; darker = lower for  $L_0$ ).

## D SAEBench Evaluation

Tables 10-13 report SAEBench metrics evaluated on The Pile (10k tokens) for all four model-domain combinations. The Pile is entirely disjoint from the task domains (toxicity, sentiment) and from all training and evaluation data, making these results a direct measure of whether fine-tuning alters the SAE’s general representational quality. Reported metrics are: KL-divergence score and cross-entropy delta (both  $\uparrow$ , measuring how faithfully the SAE-patched model reproduces the original model’s output distribution), cosine similarity between original and reconstructed activations ( $\uparrow$ ), mean squared error ( $\downarrow$ ),  $L_0$ , and  $L_1$ .

For Gemma-2-2B (Tables 10 and 11), all four conditions maintain closely similar KL-Div and CE scores across all layers and both domains, with only minor differences between Base,  $\beta=0$ , ProbeSAE and G-SAE. Cosine similarity and MSE likewise remain stable, confirming that fine-tuning does not meaningfully alter the SAE’s general reconstruction behavior on out-of-domain text. The one visible divergence is sparsity: on Gemma, G-SAE operates at a markedly higher  $L_0$  than the other three conditions (layer-averaged  $\approx 127$  vs.  $\approx 100$  on toxicity,  $\approx 122$  vs.  $\approx 96$ -108 on sentiment), so its reconstruction scores are obtained with a denser code. For LLaMA-3.1-8B (Tables 12 and 13), the base SAE starts from a lower reconstruction baseline on our evaluation data, attributable to evaluation context mismatch (last-token task-specific vs. all-token web-text, as discussed in the Limitations), and all three fine-tuned conditions substantially improve MSE over it by adapting to this context, G-SAE the most (layer-averaged  $7.29 \rightarrow 1.70$  on toxicity, against 2.22 for  $\beta=0$  and 2.59 for ProbeSAE). On LLaMA, however, ProbeSAE shows meaningfully lower KL-Div, CE and cosine similarity than every other condition,  $\beta=0$  and G-SAE included (e.g., KL-Div drops from  $\sim 0.79$  to  $\sim 0.51$  at L17 toxicity, a  $\sim 35\%$  relative reduction, while G-SAE remains the highest of the four at a layer-averaged 0.83), reflecting a real trade-off between task alignment and preservation of the model’s general predictive distribution. This cost is expected given the stronger classification pressure applied, and does not indicate domain overfitting, the drop is proportional to the task-alignment gain documented in the main results, and the values are stable across Pile-10k rather than concentrated on task-domain tokens. On Gemma, the same metrics remain broadly stable across all conditions. Practitioners applying ProbeSAE to LLaMA should be aware of this trade-off and calibrate  $\beta_c$  accordingly.

| L   | Cond      | KL-Div $\uparrow$ | CE $\uparrow$ | Cosim $\uparrow$ | MSE $\downarrow$ | $L_0$ | $L_1$  |
|-----|-----------|-------------------|---------------|------------------|------------------|-------|--------|
| L22 | Base      | 0.958             | 0.958         | 0.907            | 35.138           | 83.1  | 1865.6 |
|     | $\beta=0$ | 0.934             | 0.936         | 0.861            | 38.279           | 85.7  | 1913.6 |
|     | ProbeSAE  | 0.939             | 0.941         | 0.857            | 36.030           | 93.6  | 2032.7 |
|     | G-SAE     | 0.941             | 0.941         | 0.868            | 37.692           | 96.9  | 2083.8 |
| L23 | Base      | 0.967             | 0.967         | 0.899            | 25.442           | 87.6  | 2073.6 |
|     | $\beta=0$ | 0.942             | 0.940         | 0.853            | 30.481           | 90.3  | 2138.0 |
|     | ProbeSAE  | 0.948             | 0.949         | 0.848            | 30.106           | 95.7  | 2223.9 |
|     | G-SAE     | 0.953             | 0.951         | 0.861            | 29.079           | 104.4 | 2365.3 |
| L24 | Base      | 0.965             | 0.967         | 0.903            | 26.463           | 82.8  | 2340.4 |
|     | $\beta=0$ | 0.942             | 0.939         | 0.859            | 33.049           | 84.8  | 2401.6 |
|     | ProbeSAE  | 0.945             | 0.944         | 0.857            | 32.657           | 90.5  | 2510.0 |
|     | G-SAE     | 0.957             | 0.956         | 0.869            | 31.087           | 102.8 | 2731.9 |
| L25 | Base      | 0.965             | 0.965         | 0.924            | 17.874           | 130.3 | 3693.5 |
|     | $\beta=0$ | 0.939             | 0.937         | 0.879            | 27.987           | 142.9 | 3894.2 |
|     | ProbeSAE  | 0.942             | 0.941         | 0.880            | 28.219           | 151.5 | 4065.6 |
|     | G-SAE     | 0.957             | 0.956         | 0.891            | 26.355           | 182.0 | 4622.3 |

Table 10: **SAEBench Results: Gemma-2-2B, Sentiment.** Evaluated on The Pile (10k tokens). KL-Div and CE Loss scores measure how well the SAE preserves the model’s predictive distribution (higher = better). Colour intensity reflects relative performance within each column.

| L   | Cond      | KL-Div $\uparrow$ | CE $\uparrow$ | Cosim $\uparrow$ | MSE $\downarrow$ | $L_0$ | $L_1$  |
|-----|-----------|-------------------|---------------|------------------|------------------|-------|--------|
| L13 | Base      | 0.953             | 0.956         | 0.918            | 11.448           | 96.6  | 823.1  |
|     | $\beta=0$ | 0.943             | 0.940         | 0.898            | 10.949           | 95.8  | 804.3  |
|     | ProbeSAE  | 0.946             | 0.943         | 0.896            | 10.491           | 102.1 | 844.5  |
|     | G-SAE     | 0.945             | 0.948         | 0.896            | 11.205           | 115.3 | 934.1  |
| L16 | Base      | 0.956             | 0.959         | 0.922            | 20.521           | 90.6  | 1089.5 |
|     | $\beta=0$ | 0.944             | 0.942         | 0.896            | 20.218           | 88.4  | 1059.8 |
|     | ProbeSAE  | 0.948             | 0.944         | 0.893            | 19.413           | 91.8  | 1094.3 |
|     | G-SAE     | 0.946             | 0.948         | 0.897            | 20.562           | 105.5 | 1207.4 |
| L24 | Base      | 0.965             | 0.967         | 0.903            | 26.463           | 82.8  | 2340.4 |
|     | $\beta=0$ | 0.950             | 0.948         | 0.873            | 30.379           | 83.2  | 2357.5 |
|     | ProbeSAE  | 0.950             | 0.951         | 0.870            | 29.171           | 98.9  | 2649.4 |
|     | G-SAE     | 0.961             | 0.961         | 0.880            | 28.952           | 110.9 | 2893.2 |
| L25 | Base      | 0.965             | 0.965         | 0.924            | 17.874           | 130.3 | 3693.5 |
|     | $\beta=0$ | 0.946             | 0.942         | 0.893            | 24.098           | 127.6 | 3584.7 |
|     | ProbeSAE  | 0.947             | 0.947         | 0.893            | 24.433           | 131.0 | 3677.7 |
|     | G-SAE     | 0.961             | 0.961         | 0.902            | 23.287           | 176.9 | 4558.3 |

Table 11: **SAEBench Results: Gemma-2-2B, Toxicity.** Evaluated on The Pile (10k tokens). KL-Div and CE Loss scores measure how well the SAE preserves the model’s predictive distribution (higher = better). Colour intensity reflects relative performance within each column.

| L   | Cond        | KL-Div $\uparrow$ | CE $\uparrow$ | Cossim $\uparrow$ | MSE $\downarrow$ | $L_0$ | $L_1$  |
|-----|-------------|-------------------|---------------|-------------------|------------------|-------|--------|
| L16 | Base        | 0.587             | 0.595         | 0.794             | 8.624            | 50.0  | 50.821 |
|     | $\beta = 0$ | 0.647             | 0.649         | 0.819             | 7.224            | 50.0  | 48.642 |
|     | ProbeSAE    | 0.464             | 0.474         | 0.669             | 7.182            | 50.0  | 69.169 |
|     | G-SAE       | 0.663             | 0.673         | 0.808             | 6.909            | 50.0  | 50.631 |
| L17 | Base        | 0.653             | 0.665         | 0.809             | 6.886            | 50.0  | 53.726 |
|     | $\beta = 0$ | 0.663             | 0.667         | 0.820             | 5.947            | 50.0  | 51.520 |
|     | ProbeSAE    | 0.480             | 0.488         | 0.723             | 5.682            | 50.0  | 65.033 |
|     | G-SAE       | 0.686             | 0.697         | 0.804             | 5.482            | 50.0  | 53.788 |
| L18 | Base        | 0.653             | 0.665         | 0.823             | 6.285            | 50.0  | 58.824 |
|     | $\beta = 0$ | 0.677             | 0.681         | 0.822             | 5.506            | 50.0  | 55.400 |
|     | ProbeSAE    | 0.493             | 0.511         | 0.743             | 5.274            | 50.0  | 68.067 |
|     | G-SAE       | 0.688             | 0.701         | 0.812             | 5.066            | 50.0  | 58.368 |
| L19 | Base        | 0.693             | 0.708         | 0.831             | 7.366            | 50.0  | 61.155 |
|     | $\beta = 0$ | 0.696             | 0.702         | 0.825             | 6.338            | 50.0  | 57.714 |
|     | ProbeSAE    | 0.553             | 0.567         | 0.752             | 5.981            | 50.0  | 68.635 |
|     | G-SAE       | 0.722             | 0.736         | 0.814             | 5.797            | 50.0  | 60.772 |

Table 12: **SAEBench Results: LLaMA-3.1-8B, Sentiment.** Evaluated on The Pile (10k tokens). KL-Div and CE Loss scores measure how well the SAE preserves the model’s predictive distribution (higher = better). Colour intensity reflects relative performance within each column.

| L   | Cond        | KL-Div $\uparrow$ | CE $\uparrow$ | Cossim $\uparrow$ | MSE $\downarrow$ | $L_0$ | $L_1$  |
|-----|-------------|-------------------|---------------|-------------------|------------------|-------|--------|
| L16 | Base        | 0.587             | 0.595         | 0.794             | 8.624            | 50.0  | 50.821 |
|     | $\beta = 0$ | 0.677             | 0.677         | 0.829             | 2.377            | 50.0  | 48.555 |
|     | ProbeSAE    | 0.451             | 0.455         | 0.671             | 2.549            | 50.0  | 86.707 |
|     | G-SAE       | 0.833             | 0.842         | 0.819             | 1.627            | 50.0  | 52.859 |
| L17 | Base        | 0.653             | 0.665         | 0.809             | 6.886            | 50.0  | 53.726 |
|     | $\beta = 0$ | 0.793             | 0.802         | 0.831             | 2.072            | 50.0  | 51.418 |
|     | ProbeSAE    | 0.513             | 0.521         | 0.677             | 2.363            | 50.0  | 88.129 |
|     | G-SAE       | 0.849             | 0.859         | 0.818             | 1.504            | 50.0  | 56.314 |
| L18 | Base        | 0.653             | 0.665         | 0.823             | 6.285            | 50.0  | 58.824 |
|     | $\beta = 0$ | 0.794             | 0.806         | 0.830             | 1.936            | 50.0  | 54.691 |
|     | ProbeSAE    | 0.431             | 0.448         | 0.686             | 2.307            | 50.0  | 92.947 |
|     | G-SAE       | 0.852             | 0.866         | 0.823             | 1.597            | 50.0  | 60.079 |
| L19 | Base        | 0.693             | 0.708         | 0.831             | 7.366            | 50.0  | 61.155 |
|     | $\beta = 0$ | 0.766             | 0.777         | 0.832             | 2.500            | 50.0  | 57.301 |
|     | ProbeSAE    | 0.530             | 0.540         | 0.694             | 3.123            | 50.0  | 93.334 |
|     | G-SAE       | 0.795             | 0.813         | 0.825             | 2.070            | 50.0  | 63.135 |

Table 13: **SAEBench Results: LLaMA-3.1-8B, Toxicity.** Evaluated on The Pile (10k tokens). KL-Div and CE Loss scores measure how well the SAE preserves the model’s predictive distribution (higher = better). Colour intensity reflects relative performance within each column.

## E Auto-Interpretability Results

Figures 4-7 report per-dataset auto-interpretability scores (Embedding Score and Recall Score) for all four model-domain combinations, stratified by individual OOD language. Each strip shows per-feature scores for the top-50 features of each SAE condition; diamonds indicate the condition mean. The dashed line at 0.50 marks chance level. For LLaMA-3.1-8B on toxicity (Figure 4), ProbeSAE achieves higher embedding scores than Base,  $\beta = 0$  and G-SAE on the four HateCheck subsets, including the non-English ones, confirming that probing supervision produces features with more coherent natural-language descriptions regardless of language; MeasuringHate is the exception, where all four conditions sit at chance and the ordering carries no signal. G-SAE stays at chance ( $\approx 0.50$ ) throughout this cell, so the gain is specific to probe guidance rather than to label supervision as such. For LLaMA-3.1-8B on sentiment (Figure 5), the improvement of ProbeSAE is especially pronounced: embedding scores are markedly higher than every other condition on all five Amazon Reviews languages and at every layer (20/20 layer-dataset combinations, for both the embedding and the recall score), while  $\beta = 0$  and G-SAE remain at or near Base level on every dataset. For Gemma-2-2B on toxicity (Figure 6), the most striking pattern is that both baselines trained on the task labels collapse:  $\beta = 0$  and G-SAE embedding scores fall to near chance across all languages ( $\approx 0.47$ - $0.57$ ), while ProbeSAE maintains embedding quality comparable to the base SAE ( $\approx 0.77$ - $0.84$ ), with Base retaining a small edge on two of the five datasets. For Gemma-2-2B on sentiment (Figure 7), differences between conditions are smaller in magnitude and no condition dominates; ProbeSAE, Base and G-SAE are broadly comparable, with Base marginally ahead on the embedding score, while  $\beta = 0$  shows modest decreases on several datasets.

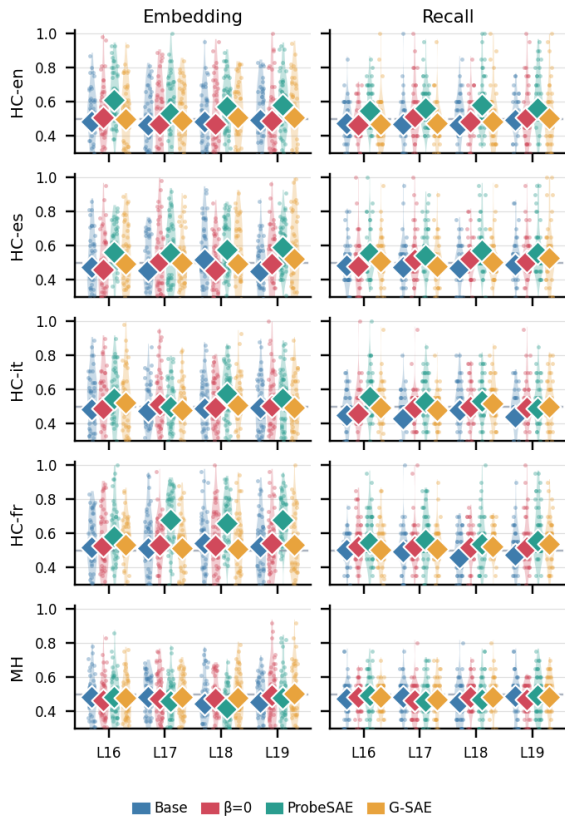


Figure 4: **Auto-interpretability, LLaMA-3.1-8B, toxicity.** Embedding and Recall score per OOD dataset (rows) and layer, for the four SAE conditions; diamonds mark the condition mean, the dashed line chance.

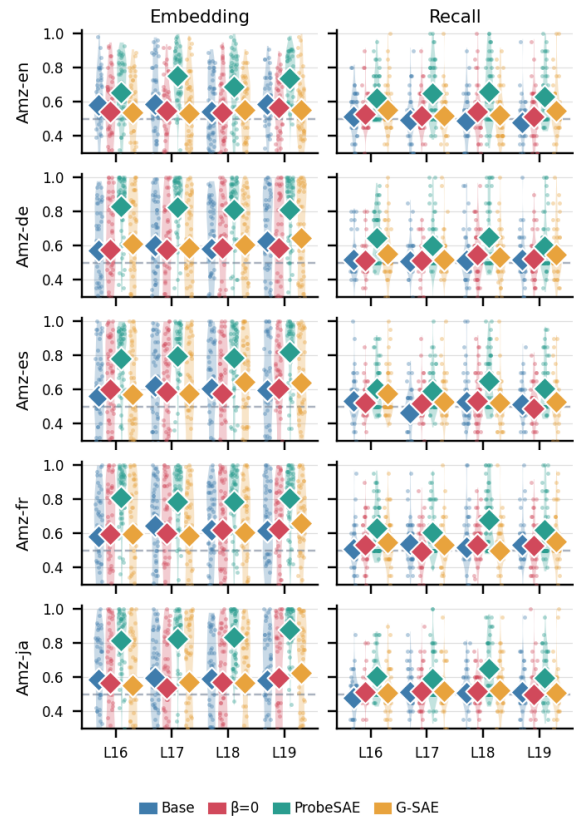


Figure 5: **Auto-interpretability, LLaMA-3.1-8B, sentiment.** Layout as in Figure 4.

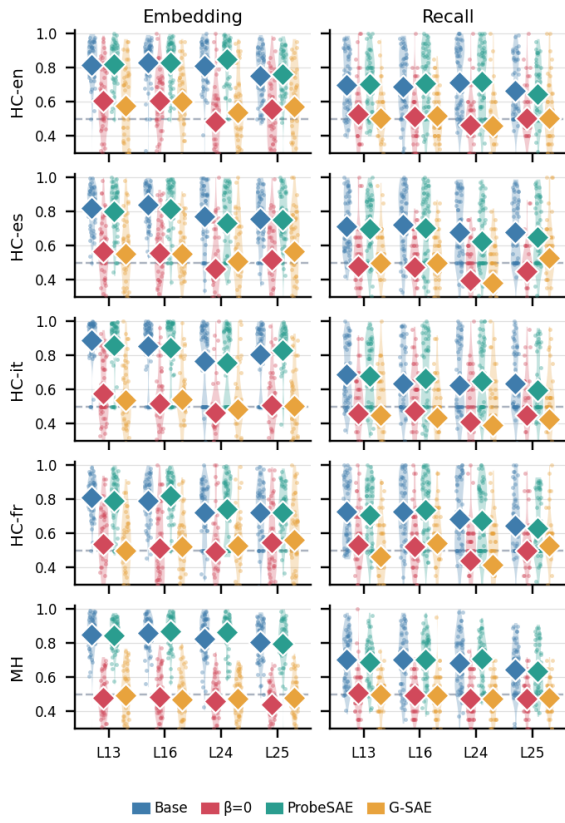


Figure 6: **Auto-interpretability, Gemma-2-2B, toxicity.** Layout as in Figure 4.

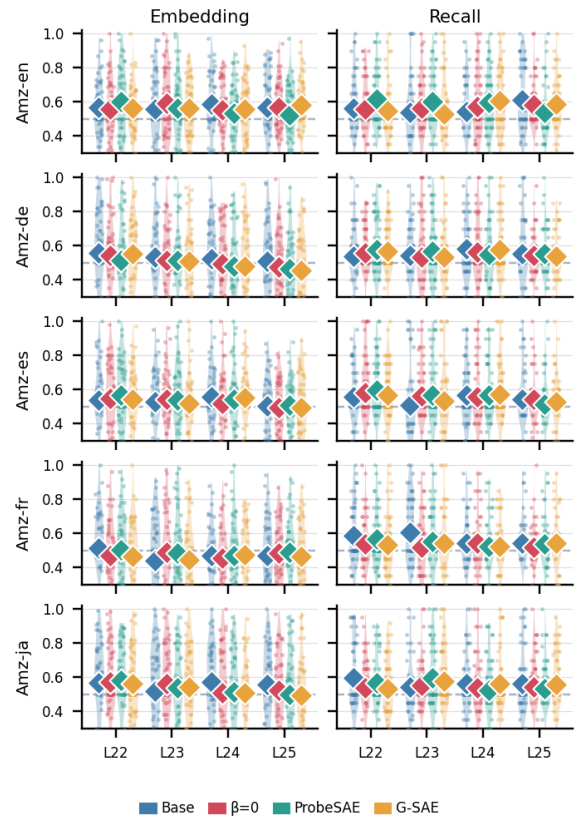


Figure 7: **Auto-interpretability, Gemma-2-2B, sentiment.** Layout as in Figure 4.