Evaluating Sparse Autoencoders on Concept Removal Tasks

Anonymous Author(s)
Affiliation
Address
email

Abstract

Sparse Autoencoders (SAEs) are an interpretability technique aimed at decomposing neural network activations into interpretable units. However, a major bottleneck for SAE development has been the lack of high-quality performance metrics, with prior work largely relying on unsupervised proxies. In this work, we introduce a family of evaluations based on SHIFT, a downstream task from Marks et al. that measures an SAE's ability to disentangle and reduce spurious correlations. To create an automated evaluation, we extend SHIFT by replacing human judgment with LLMs. Additionally, we introduce the Targeted Probe Perturbation (TPP) metric that quantifies an SAE's ability to disentangle similar concepts, effectively scaling SHIFT to a wider range of datasets. We apply both SHIFT and TPP to multiple open-source models, demonstrating that these metrics effectively differentiate between various SAE training hyperparameters and architectures.

1 Introduction

Sparse Autoencoders have emerged as a popular interpretability tool for neural networks. As the number of open-sourced SAEs increases, researchers must choose the SAE suitable to their experiment. Recently, a broad range of SAE architectures and training approaches have been proposed, including TopK, Gated, BatchTopK, p-Annealing, and JumpReLU SAEs [1, 2, 3, 4, 5]. However, the field lacks trusted metrics to properly evaluate the performance of trained SAEs. Prior work has primarily relied on unsupervised proxy metrics [6, 7, 8], which can sometimes be uncorrelated with desirable characteristics like interpretability [9, 10].

The challenge in evaluating SAEs stems from the absence of a clear ground truth objective in natural language. While SAEs are intended to faithfully decompose model activations into sparse, interpretable units, our limited understanding of model internals makes it difficult to establish a definitive benchmark. To address this, we propose directly measuring SAEs through their applicability to downstream tasks.

Sparse Human Interpretable Feature Trimming (SHIFT) debiases a classifier by removing spurious correlates in models. The method allows evaluators to remove model internal properties without explicitly blacklisting them in advance. We introduce a family of evaluations based on SHIFT that measures an SAE's ability to disentangle concepts and remove spurious correlations.

However, the reliance of SHIFT on datasets containing potential spurious correlations makes it challenging to scale to a wider variety of concepts. Motivated by this, we additionally develop Targeted Probe Perturbation (TPP). For both SHIFT and TPP, we select SAE latents using probe attribution scores, and optionally filter the latents for interpretability using an LLM judge. We evaluate our evaluation metrics against a range of sanity checks, looking at the improvement throughout training and unimodality in sparsity.

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

Our main contributions are as follows:

- Adapting SHIFT as an SAE evaluation and scaling SHIFT to a range of SAEs, datasets and LLMs
- Proposing the Targeted Probe Perturbation metric for SAE quality
- Evaluating open-sourced SAEs across multiple LLMs, datasets, and SAE training checkpoints

2 Background

Sparse autoencoders (SAEs) aim to identify an overcomplete basis of sparse, interpretable features from model internal representations [11, 12, 13]. The quality of SAEs is determined by their faithfulness to the model's internal computations and their ability to disentangle human-interpretable concepts. This disentanglement can be further broken down into correlational and causal aspects, namely the detection of interpretable concepts) and the ability to modify model behavior in targeted ways.

Typical unsupervised proxy metrics include: (1) the cross-entropy loss recovered, which measures how well the original model's loss can be reconstructed using the SAE's predictions, and (2) the L0-norm of feature activations, which quantifies the sparsity of activated features for a given input [14]. Recent work has explored evaluating SAEs on board game models [3], features in manually identified circuits [15] and detecting pre-defined natural language concepts [1].

Concept removal aims to identify and eliminate specific concepts or biases from neural representations while preserving overall performance on downstream tasks. Various approaches have been proposed, ranging from linear methods like Hard Debias [16] and LEACE [17] to more complex non-linear techniques such as Iterative Gradient-Based Projection [18].

3 Method

In this paper, we focus on the causal isolation of concepts as our primary metric for SAE quality. Our approach quantifies how effectively an SAE can manipulate individual concepts within the model's representations. To illustrate our methods, consider the task of selecting the best SAE from a hyperparameter sweep. Given a list of concepts in natural language, we follow three steps to systematically evaluate each SAE in the sweep:

1. Train binary linear probes on model activations to detect pre-defined concepts.
2. Identify a set of SAE latents corresponding to each concept.
3. Measure intended effects on probe performance when ablating SAE latents.

To validate our metrics, we conduct a series of sanity checks. These checks verify that our metric aligns with fundamental properties of SAEs, such as their characteristic sparsity and their expected performance improvements over the course of training. The subsections below describe each step in the evaluation process in detail. Appendix A.2 contains further details about our SAE training process.

3.1 SAE Latent Selection

Measuring the disentanglement of a given natural language concept with an SAE, requires identifying the subset of corresponding SAE latents. We determine relevant SAE latents by ranking their causal effect on a concept-related probe and optionally filter for interpretability as judged by an LLM.

Probing. We train binary linear probes detecting concept c from model activations as described in Appendix A.1. The attribution score of latent with index l on concept c is given by

$$I(\mathbf{l}, c) = (\mathbf{l}_{\text{pos}} - \mathbf{l}_{\text{neg}})(\mathbf{d}_l \cdot \mathbf{P}), \tag{1}$$

where $\mathbf{l}$ denotes the batch of activations of latent l, $\mathbf{d}_l$ denotes the SAE decoder vector corresponding to l, $\mathbf{P}$ is the weight matrix of the binary probe, $\mathbf{l}_{\text{pos}}$ is the mean activation of the latent for inputs

of the targeted concept, and $\mathbf{l}_{\text{neg}}$ is the mean activation for inputs unrelated to the concept. The difference of mean activations ($\mathbf{l}_{\text{pos}} - \mathbf{l}_{\text{neg}}$) helps filter out high-frequency features that activate with equal frequency on positive and negative class inputs, focusing on features that discriminate between the classes.

We select the top N features with the highest attribution scores. [1] Our current approach involves increasing N until we observe statistically significant differences in the performance of various SAEs. We find that the required N increased with model size. However, the optimal selection of N is an open problem, as there isn't a clear optimal number of features that a good SAE should have for a given concept.

Automated intepretability judgement. We optionally employ Claude-3.5 as an LLM judge to decide whether an SAE latent is interpretable. The LLM judge receives a description similar to EleutherAI [20]: Our prompt contains the top 5 contexts that activated the latent from a set of 10k random web text samples [21], as well as the top 5 promoted tokens indicated by direct logit attribution [22], and 4 few-shot examples demonstrate scoring the relation to each concept from 0 to 4. The model is instructed to perform in-context reasoning and outputs a score for each category in json format.

Appendix A.7 shows a full example prompt. We observe that the LLM judge is prone to false negatives, rating concepts with 0 although they are related. We decide to keep features that show any relation with score 1 or higher. We refined the system prompts by manually labelling 60 examples, and looking at prompts where ratings differ. After tweaking the prompts the inter-rater agreement shows a Cohen's κ value of 0.44 for gender and 0.58 for profession, passing our minimum requirements.[2].

3.2 SHIFT

In the SHIFT method [8], a human evaluator debiases a classifier by ablating SAE latents. We automate SHIFT and operationalize the method as an evaluation for SAE quality: Better SAEs translate into a better ability to precisely remove spurious correlates and debias the classifier.

The SHIFT method requires a dataset that maps text to two binary labels. Here, we use the Bias in Bios dataset [23] which maps professional biographies to occupation and gender, and the Amazon reviews dataset [24] which covers reviews from a broad range of product categories with accompanying rating scores. First, we filter both datasets for two binary labels. For example, we select two professions (professor, nurse) and the gender label (male, female) from the Bias in Bios dataset. This dataset is partitioned into a balanced set (containing all combinations of professor/nurse, male/female) and a biased set (only containing the labels male+professor and female+nurse). We then train a linear classifier C_{b} on the biased dataset which we will attempt to debias.

We consider two feature selection methods. The first feature selection method replicates the original SHIFT setup by Marks et al. In it, we select the top n SAE latents L according to their *absolute* probe attribution score to the biased probe. Note, that we use the absolute importance score to select SAE latents corresponding to either label of the binary attribute. We then filter the latents using an LLM judge, retaining only the features that the LLM scores as relating to our desired concept. Second, we use the "feature skyline" method from Sparse Feature Circuits: We select set L containing the top n SAE latents according to their absolute probe attribution score with a probe trained specifically to predict the spurious signal.

Next, we attempt to remove the spurious signal by defining a modified classifier $C_{\text{m}} = C_{\text{b}} \setminus L$ where all selected features are zero-ablated. The accuracy with which the modified classifier C_{m} predicts the desired class when evaluated on the balanced dataset indicates SAE quality. A higher accuracy suggests that the SAE was more effective in isolating and removing the spurious correlation of e.g. gender, allowing the classifier to focus on the intended task of e.g. profession classification. We consider a normalized evaluation score

$$S_{\text{SHIFT}} = \frac{A_{\text{abl}} - A_{\text{base}}}{A_{\text{sky}} - A_{\text{base}}} \tag{2}$$

[1] The computation of attribution scores I can be efficiently calculated using precomputed model activations. However, this is restricted to the SAE layer coinciding with the probe layer. If probe layer and SAE layer differ, the attribution I can be approximated with attribution patching [19], which requires requires additional forward and backward passes on the order of `num_classes * num_batches`.

[2] For reference, $\kappa = 0$ means random chance, $\kappa = 1$ means perfect correlation and $\kappa > 0.4$ denotes reasonable correlation.

where A_{abl} is the probe accuracy after ablation, A_{base} is the baseline accuracy (spurious probe before ablation), and A_{sky} is the skyline accuracy (probe trained directly on the desired concept). This score represents the proportion of improvement achieved through ablation relative to the maximum possible improvement, allowing fair comparison across different classes and models.

	Clean		Mod.	
Model	**Gen.**	**Prof.**	**Gen.**	**Prof.**
Pythia-70M				
Prof. / Nurse	0.77	0.72	0.52	0.91
Arch. / Journ.	0.86	0.63	0.53	0.88
Gemma-2-2B				
Prof. / Nurse	0.64	0.86	0.83	0.65
Arch. / Journ.	0.69	0.80	0.94	0.55

Table 1: Original and modified accuracies of biased probes evaluated on balanced gender and profession datasets. Modified accuracies represent the best accuracies obtained using the SHIFT method on any SAE.

3.3 Targeted Probe Perturbation

SHIFT requires datasets with multiple independent labels, and (artificial) filtering for spurious correlates. We generalize SHIFT to all multiclass NLP datasets by introducing the targeted probe perturbation (TPP) metric. On a high level, we aim to find sets of SAE latents that disentangle the dataset classes. Inspired by SHIFT, we train probes on the model activations and measure the effect of ablating sets of features on the probe accuracy. Ablating a disentangled set of latents should only have an isolated causal effect on one class probe, while leaving other class probes unaffected.

We consider a dataset with M classes. For each class with index $i = 1, \ldots, M$ we select the set L_i of the most relevant SAE latents as described in section 3.1. Note, that we select the top *signed* importance scores, as we are only interested in features that actively contribute to the targeted class.

For each concept c_i, we partition the dataset into samples of the targeted concept and a random mix of all other labels. We define the model with probe corresponding to class c_j with $j = 1, \ldots, M$ as a linear classifier C_j. Futher, $C_{i,j}$ denotes a classifier for c_j where latents L_i are ablated. Then, we iteratively evaluate the accuracy $A_{i,j}$ of all linear classifiers $C_{i,j}$ on the dataset partitioned for the corresponding class c_j. The targeted prope perturbation score

$$S_{\text{TPP}} = \underset{(i=j)}{\text{mean}}(A_{i,j}) - \underset{(i\neq j)}{\text{mean}}(A_{i,j}) \tag{3}$$

represents the effectivenes of causally isolating a single probe. Ablating a disentangled set of features should only show a significant accuracy decrease if $i == j$, namely if the latents selected for class i are ablated in the classifier of the same class i, and remain constant if $i \neq j$.

4 Results

Trusted SAE evaluations should reflect the core characteristics of SAEs. In line with previous work, we focus on SAE characteristics and use them as sanity checks for our evaluations.

- Unimodality w.r.t. sparsity as suggested by experimental observations [3]
- Improvement throughout training

We test for these characteristics by evaluating a sweep of SAEs (Appendix A.2) across a range of sparsities and training checkpoints.

4.1 SHIFT

First, we evaluate both Pythia-70M and Gemma-2-2B SAEs on the SHIFT task. SAEs effectively remove an unwanted signal and improve the accuracy of a biased classifier for both models. The evaluation clearly differentiates SAEs of different sparsities and architectures in all experiments.

Figure 1 shows the SHIFT evaluation of Gemma-2-2B SAEs trained on layer 19. The metric reflects the relative accuracy increase S_{SHIFT} of classifying the desired attribute after ablating spurious SAE latents as given in Equation 2. Here, we select the top 50 spurious latents from the original SHIFT method (Section 3.2) and filtered for interpretablity with an LLM judge (Section 3.1). Figure 1 (left) demonstrates a clear separation between architectures, with TopK and JumpReLU outperforming the Standard architecture. Both TopK and JumpReLU are unimodal with respect to sparsity and peak in the L0 range [20, 100]. Figure 1 (right) verifies that SAEs improve on the SHIFT metric throughout training on average. The first 10% of training (corresponding to 20M tokens) correspond to 85 % of SHIFT score on average.

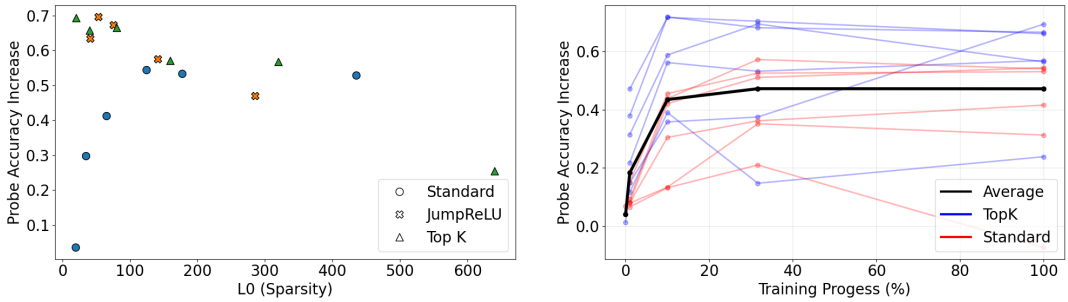


Figure 1: The left scatterplot of loss recovered vs L0, with color corresponding to coverage score, and each point representing a single SAE. We differentiate between SAE training methods with shapes (left) and colors (right).

Our findings replicate across models, datasets, and feature selection methods, as further discussed in Appendix A.5. Interestingly, most SAE features identified by the feature skyline are judged interpretable by the LLM. This trend especially holds for the peak region of the L0 range [20, 100]. The feature skyline method without automated interpretability is very fast and a good proxy for the original SHIFT method.

In the original experiment by Marks et al. [8], SHIFT relied on Standard SAEs at many locations in the model and attribution patching. We find that with improved architectures such as TopK, good performance can be obtained with a single SAE. Moreover, we find that SHIFT scores before and after applying the LLM interpretability filter correlate highly with $r = 0.81$. Appendix A.1 shows a direct comparison for Pythia-70M.

4.2 TPP

Figure 2 shows the TPP score (Equation 3) of Gemma-2-2B SAEs trained on layer 19. Again we ablate up to 50 SAE latents. TPP clearly differentiates TopK and JumpReLU SAEs from the Standard SAEs in the left subplot. This separation is also visible in the TPP score evolution throughout training (right): While TopK and JumpReLU SAEs achieve 80 % of their final TPP score after 10% of training, Standard SAEs achieve 80 % of only after 31% of training. On average, the TPP score increases throughout training. Moreover, we find that the TPP scores before and after applying the LLM interpretability filter correlate highly with $r = 0.88$. Appendix A.4 contains results of TPP on Pythia-70M in Figure 5.

5 Discussion and Limitations

Our experimental results demonstrate that SHIFT and TPP effectively differentiate Standard SAEs from TopK and JumpReLU SAEs. While both metrics identify at least one SAE that successfully solves the task at hand the optimal sparsity for each metric differs significantly. Further investigation

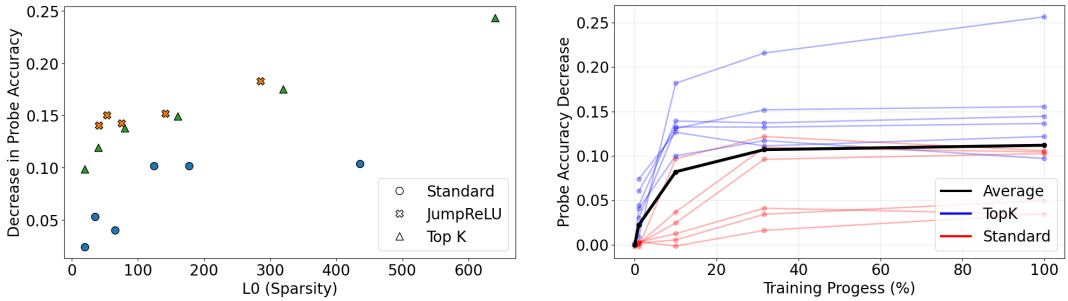


Figure 2: Targeted Probe Perturbation (TPP) scores over sparsity for SAEs of Standard, JumpReLU, and TopK architectures (left). TPP scores as a function of training progress, measured for checkpoints at 0%, 1%, 10%, 31%, and 100% of SAE training over 6 TopK and 6 Standard SAEs (right). Each datapoint (left) and line (right) corresponds to a single SAE, architectures are differentiated by color.

on correlations of the TPP metric with L0 is required. Further, this finding underlines the importance of evaluating SAE evaluations across a range of sparsities.

We generally observe that our LLM judge deems most of the SAE latents that were preselected by probe attribution interpretable.

Our LLM judge has a lower bar for interpretability than the common implementation by Juang et al. [20]. While we simply score a latent's relevance to certain concepts, automated interpretability usually aims to find a concise description in natural language which is predictive of latent activations.

SHIFT and TPP without the LLM judge are simpler, faster, and cheaper, and can be computed in seconds, enabling frequent evaluations such as during SAE training. However, we can get increased confidence in the metric using an LLM judge to ensure that identified latents are interpretable. Thus, there is a tradeoff in adding the LLM judge. We analyze the correlation of the metrics with and without the LLM judge in Appendix A.6.

Our evaluation metrics rely on subjective hypotheses about what SAEs should learn based on human-understandable concepts, which may not accurately reflect the model's true internal representations. This dependency on human-generated concepts can overlook important features and constrains the evaluation's scope.

6 Conclusion

SHIFT and our Targeted Probe Perturbation (TPP) method offer several advantages for evaluating Sparse Autoencoders (SAEs). They can be easily applied to a wide range of datasets, show improvement throughout training, and are computationally efficient. Both metrics can be computed in seconds when using precomputed activations, and we encourage researchers to use our codebase to evaluate their SAEs and monitor their SAE training runs.

However, SHIFT and TPP have some limitations, including complexity (especially when using an LLM judge) and undetermined hyperparameters. Thus, they should only be treated as additional evidence as part of a broader SAE evaluation suite. The lack of ground truth features and significant variability across model sizes makes hyperparameter and feature selection challenging. Therefore, it is important to perform sanity checks on SAE evaluations, like monitoring metrics during training and evaluating SAEs across a range of sparsities.

Our metrics focus on the causal isolation of human-interpretable concepts. However, high-quality SAE latents should also exhibit characteristics such as sparsity, disentanglement of natural language concepts, human interpretability, and correlation with natural language concepts. Developing evaluations that cover all these aspects of SAE quality remains an open challenge.

References

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

[2] Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Tom Lieberum, Vikrant Varma, János Kramár, Rohin Shah, and Neel Nanda. Improving dictionary learning with gated sparse autoencoders, 2024.

[3] Adam Karvonen, Benjamin Wright, Can Rager, Rico Angell, Jannik Brinkmann, Logan Smith, Claudio Mayrink Verdun, David Bau, and Samuel Marks. Measuring progress in dictionary learning for language model interpretability with board game models, 2024.

[4] Bart Bussmann, Patrick Leask, and Neel Nanda. Batchtopk: A simple improvement for topk-saes. *AI Alignment Forum*, Jul 2024.

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

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

[7] Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2, 2024.

[8] Samuel Marks, Can Rager, Eric J. Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models, 2024.

[9] Adam Jermyn, Adly Templeton, Joshua Batson, and Trenton Bricken. Tanh penalty in dictionary learning. `https://transformer-circuits.pub/2024/feb-update/index.html#dict-learning-tanh`, Feb 2024. In: Circuits Updates - February 2024.

[10] Dan Braun, Jordan Taylor, Nicholas Goldowsky-Dill, and Lee Sharkey. Identifying functionally important features with end-to-end sparse dictionary learning, 2024.

[11] Lee Sharkey, Dan Braun, and beren. [interim research report] taking features out of superposition with sparse autoencoders, 12 2022. AI Alignment Forum.

[12] Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

[13] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022. https://transformer-circuits.pub/2022/toy$_m odel/index.html$.

[14] Javier Ferrando, Gabriele Sarti, Arianna Bisazza, and Marta R. Costa-jussà. A primer on the inner workings of transformer-based language models, 2024.

[15] Aleksandar Makelov, Georg Lange, and Neel Nanda. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. In *The Twelfth International Conference on Learning Representations*, 2024.

[16] Tianlu Wang, Xi Victoria Lin, Nazneen Fatema Rajani, Bryan McCann, Vicente Ordonez, and Caiming Xiong. Double-hard debias: Tailoring word embeddings for gender bias mitigation. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5443–5453, Online, July 2020. Association for Computational Linguistics.

[17] Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[18] Shadi Iskander, Kira Radinsky, and Yonatan Belinkov. Shielded representations: Protecting sensitive attributes through iterative gradient-based projection. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5961–5977, Toronto, Canada, July 2023. Association for Computational Linguistics.

[19] Neel Nanda. Attribution patching: Activation patching at industrial scale, 2023.

[20] Caden Juang, Gonçalo Paulo, Jacob Drori, and Nora Belrose. Open source automated interpretability for sparse autoencoder features, jul 2024. Building and evaluating an open-source pipeline for auto-interpretability.

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

[22] nostalgebraist. Interpreting gpt: the logit lens, 2020.

[23] Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19. ACM, January 2019.

[24] Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. Bridging language and items for retrieval and recommendation. *arXiv preprint arXiv:2403.03952*, 2024.

[25] Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling, 2023.

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

A Appendix

A.1 Probe Training

When training probes on biased datasets (e.g., male_professor, female_nurse), the probe can potentially leverage two distinct signals: profession and gender. The probe must trade off between these signals. Therefore, evaluating the biased probe on balanced gender and profession datasets will result in a mean accuracy of at most 75%.

For effective spurious correlation removal and differentiation between SAEs, the probe must capture both signals to some degree. If the probe's accuracy is heavily skewed (e.g., 50% on class 1 and 95% on class 2), ablating SAE latents corresponding to class 2 cannot significantly improve class 1 accuracy, limiting the effectiveness of the evaluation.

We find that the direction of the spurious correlation varies based on the underlying LLM selected. The strength of this spurious correlation is influenced by probe training hyperparameters. For instance, using a large probe batch size on Gemma-2-2B resulted in a weak spurious correlation (< 55% accuracy) for the gender category. Decreasing the batch size led to the probe learning a more balanced mix of both signals, enabling significant improvements in probe accuracy on a desired concept and better differentiation between SAEs. In addition, we find that a high learning rate introduces variance in the direction and strength of the spurious correlation.

In our experiments with Pythia-70M, a smaller model, we found that the probe mostly relied on the gender signal. This scenario provides an intuitive demonstration of removing unwanted gender bias, where we ablate gender features to improve profession classification. However, with larger models like Gemma-2-2B, we observed that our probes primarily picked up on profession signals. In such cases, to significantly change probe accuracy and differentiate between SAEs, we need to ablate the profession signal to improve gender classification.

To decrease the dependence of the SHIFT on our selection of individual dataset classes, we average scores over multiple class pairs. We consider the pairs ("professor', 'nurse") and ("architect', 'journalist") in the Bias and Bios dataset, and the pairs ("Books", "CDs and Vinyl"), ("Software", "Electronics"), ("Pet Supplies", "Office Products"), ("Industrial and Scientific", "Toys and Games"). These classes were selected based on the objective that a classifier C_{b} develops an accuracy of at least 0.6 for both categories in the pair (gender / profession or sentiment / product category).

Table 2: Training parameters of our linear probes.

Parameter	Value
LLM Context Length	128
Input Datapoints	4000
Epochs	5
Optimizer	`Adam`
Adam betas	(0.9, 0.999)
Batch size	16
Learning rate	1e-3

A.2 Sparse Autoencoder Training

As a testbed for our evaluation pipeline, we train and open-source a suite of SAEs on the models pythia-70m-deduped [25] and Gemma-2-2B [26]. We train Vanilla and TopK architectures on 200M tokens from The Pile [21] with expansion factors 8x (Pythia, Gemma) and 32x (Pythia). [3] Additionally, we evaluate JumpReLU SAEs from the Gemma-Scope suite [7].

Table 3: Training parameters of our sparse autoencoders.

Parameter	**Value**
LLM Context Length	128
Number of tokens	200M
Optimizer	`Adam`
Adam betas	(0.9, 0.999)
Linear warmup steps	1,000
Batch size	4,096
Learning rate	3e-4
Expansion factor	{8, 32}

[3] The expansion factor denotes the ratio of SAE latents to input dimension.

A.3 Gemma Results without Auto-interp

For TPP without an LLM judge, we find that a randomly initialized Standard SAE performs relatively well on our metric in Figure 3. We believe this is because the randomly initialized SAE has dense activations with an L0 of 9000, significantly larger than Gemma-2-2B's d_model of 2304. Once training begins and the Standard SAE becomes more sparse, the results with and without the LLM judge begin to correlate.

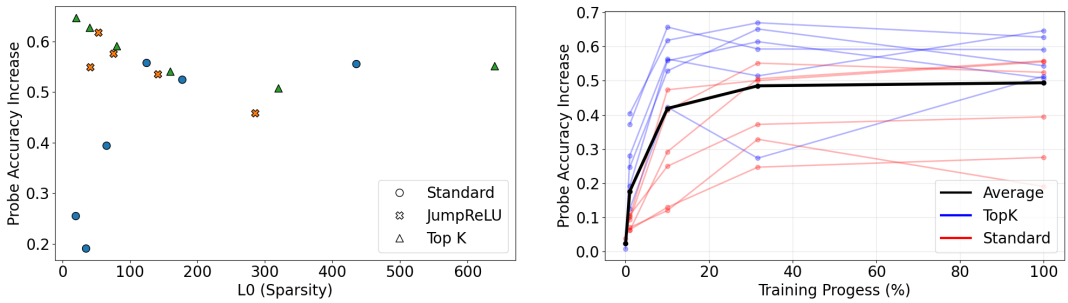


Figure 3: SCR scores without auto-interp as a function of training progress, measured for checkpoints at 0%, 1%, 10%, 31%, and 100% of SAE training over 6 TopK and 6 Standard SAEs (right). Each datapoint (left) and line (right) corresponds to a single SAE, architectures are differentiated by color.

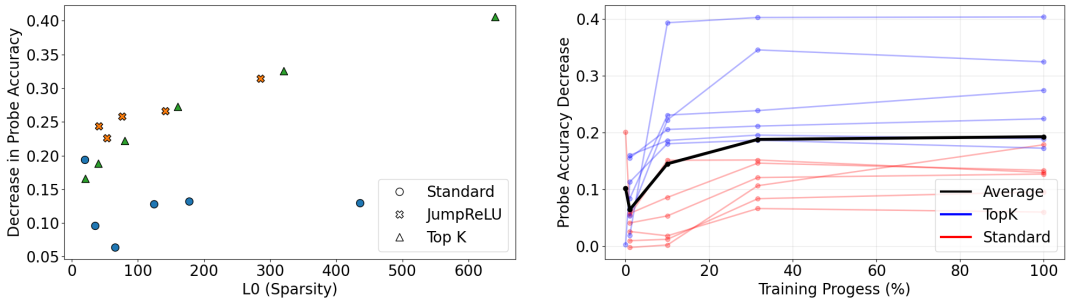


Figure 4: Targeted Probe Perturbation (TPP) scores without auto-interp over sparsity for SAEs of Standard, JumpReLU, and TopK architectures (left). TPP scores as a function of training progress, measured for checkpoints at 0%, 1%, 10%, 31%, and 100% of SAE training over 6 TopK and 6 Standard SAEs (right). Each datapoint (left) and line (right) corresponds to a single SAE, architectures are differentiated by color.

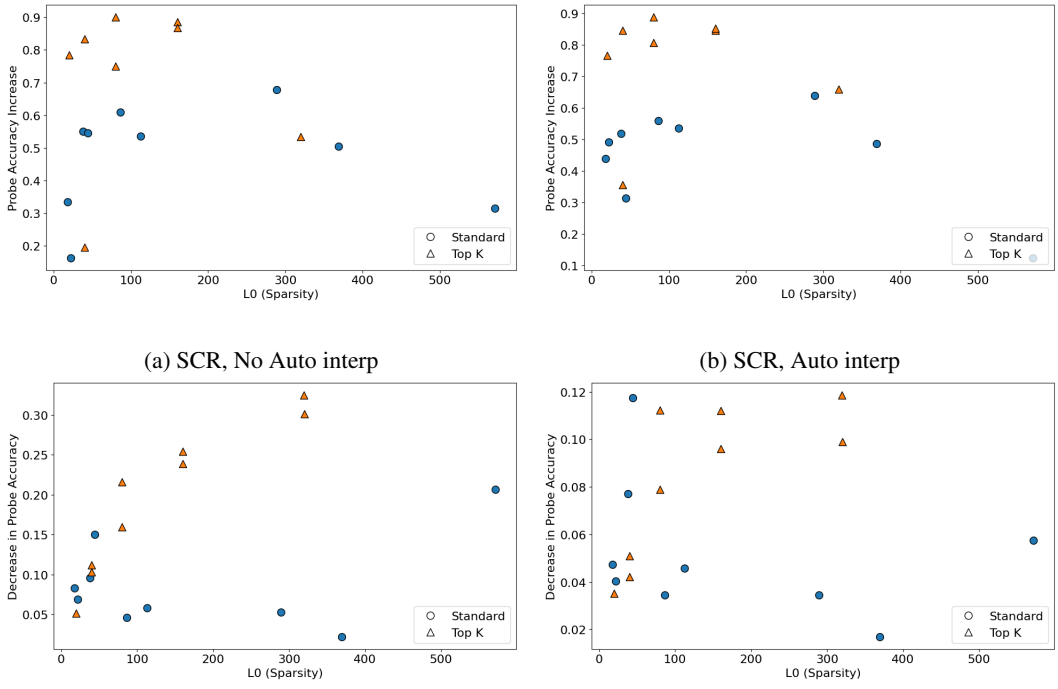
A.4 Pythia-70M Results

Complementary to our results in Section 4 we provide an evaluation sweep for SHIFT and TPP on the Pythia-70M model in Figure 5. Additionally we compare SAE architectures in the number of SAEs required to achieve a high SHIFT score in Table 5.

Table 4: SAE intervention locations and their Ground Truth Accuracy.

SAE intervention locations	**Ground Truth Acc**
None	59%
Standard SAE, Embedding and Layers 0-4	88%
Standard SAE, Embedding only	86%
Standard SAE, Layers 0-4	84%
Standard SAE, Layers 3-4	81%
Standard SAE, Layers 3-4, Resid only	79%
TopK SAE, Layer 4, Resid only	**90%**

Table 5: TopK SAEs significantly improve performance of the SHIFT method in the setting used in Marks et al. [8]. A biased probe is trained on the class pair of ("professor", "nurse"). By ablating gender-related features, we improve the probe's accuracy at profession classification. In Marks et al., 16 Standard SAEs were used on MLP output, attention output, and resid_post in layers 0-4, in addition to the embedding output. We exceed their performance using only a single TopK SAE.



(a) SCR, No Auto interp

(b) SCR, Auto interp

(c) TPP, No Auto interp

(d) TPP, Auto interp

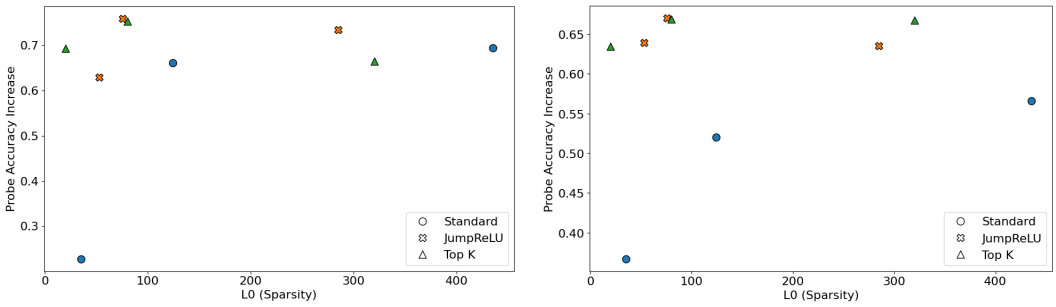
Figure 5: Results for Pythia-70M. The left column contains a scatterplot of loss recovered vs L0, with color corresponding to coverage score, and each point representing different hyperparameters. We differentiate between SAE training methods with shapes.

A.5 Amazon SCR Results

In the Bias in Bios dataset, we evaluate the concepts of profession / gender. In the Amazon reviews dataset, the corresponding concepts are product category / review sentiment.

We did not evaluate SCR or TPP on Pythia-70M with the Amazon dataset, as we found that probes trained on Pythia-70M obtained poor scores for identifying Product Category.

We evaluate the following product category pairs: ("Books","CDs_and_Vinyl"), ("Software", "Electronics"), ("Pet_Supplies", "Office_Products").



(a) SCR, No Auto interp

(b) SCR, Auto interp

Figure 6: Results for Amazon SCR, Gemma-2-2B. The left column contains a scatterplot of loss recovered vs L0, with color corresponding to coverage score, and each point representing different hyperparameters. We differentiate between SAE training methods with shapes.

A.6 Correlation of SHIFT scores with interpretability.

We measure the correlation of SHIFT (Figure 7 and TPP (Figure 8) scores with and without an interpretability filter as judged by an LLM. Incorporating an LLM judge provides an additional layer of validation by ensuring the identified latents are interpretable. The SHIFT and TPP metrics, when used without an LLM judge, offer simplicity, speed, and cost-effectiveness. These streamlined versions can be calculated rapidly, often within seconds, allowing for frequent assessments, such as during SAE training. This introduces a trade-off: while the LLM judge enhances confidence in the metric, it also increases complexity and computational requirements.

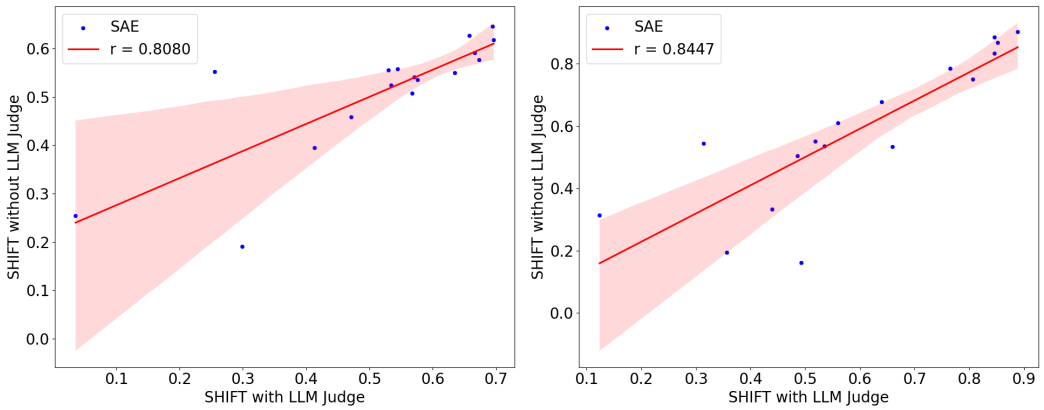


Figure 7: Correlation of SHIFT scores on Gemma-2-2B (left) and Pythia-70M before and after filtering SAE latents with the LLM judge described in Section 3.1. The red area denotes the 95% confidence interval.

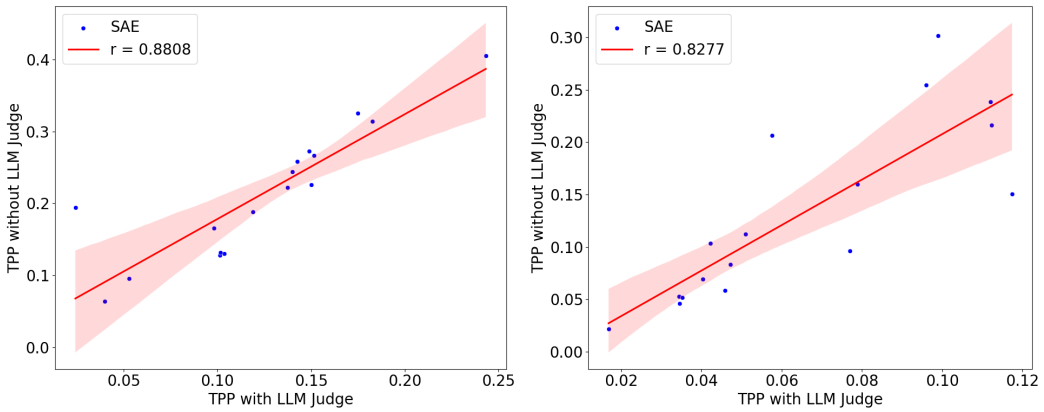


Figure 8: Correlation of TPP scores on Gemma-2-2B (left) and Pythia-70M before and after filtering SAE latents with the LLM judge described in Section 3.1. The red area denotes the 95% confidence interval.

A.7 LLM judge prompt

Our LLM judge prompt contains a system prompt, few-shot examples and a real example for labelling.

System prompt

```
You are a meticulous AI researcher conducting an important investigation into a certain neuron in a language model. Your task is to analyze the neuron and score how strong its behavior is related to a concept in {concepts} on an integer scale from 0 to 4 in json format.

Task description:
You will be given a list of text examples on which the neuron activates. The specific tokens which cause the neuron to activate will appear between delimiters like <<this>>. The activation value of the token is given after each token in parentheses like <<this>>(3).
You will also be shown a list called promoted tokens. The logits promoted by the neuron shed light on how the neuron's activation influences the model's predictions or outputs. It is possible that this list is more informative than the list of text examples.
For each concept, try to judge whether the neurons behavior is related to the concept.
If part of the text examples or predicted tokens are incorrectly formatted, please ignore them.
If you are not able to find any coherent description of the neurons behavior, decide that the neuron is not related to any concept.

Scoring rubric:
Score 4: The majority of examples, activation scores, and promoted tokens are clearly related to the concept.
Score 3: About half of the examples and promoted tokens are directly related to the concept.
Score 2: Only some of the examples are directly related to the concept, and some more are distantly related.
Score 1: NONE of the examples is directly related to the concept, but single tokens can be distantly related to the general domain of the concept.
Score 0: NONE of the text examples can be distantly related in any way to the broader field of the concept.

Structure your response as follows:
Step 1. Give a single sentence summary for the full text examples.
Step 2. Give a separate single sentence summary for the promoted tokens.
Step 3. Discuss your decision in 1-3 sentences.
After finishing all steps above, provide a single json block at the end of your response. The json block should contain your scores on an integer scale from {min_scale} to {max_scale} for each concept as shown in the examples.
```

Few-shot examples

```
Promoted tokens:  broadcasts, broadcasting, Broadcasting, television, broadcast, Television, announ,Television,TV, TV
Example prompts:

Example 1: Radio Nova (Ireland)

Radio Nova was a pirate radio station <<broadcasting>>(2) from Dublin, Ireland. Owned and operated by the UK pirate radio veteran Chris Cary, the station's first broadcasts were during the summer of 1981 on 88.5 MHz FM and 819 kHz AM.

Early history
Prior to Nova's arrival, Irish radio consisted of the government broad<<caster>>(2) <<RT>>(2)   and a number of local AM pirate <<stations>>(3). Radio Nova was the first <<station>>(2) in Ireland to use a
```

```
    high powered signal on FM. By 1982 Radio Nova was pulling in over 40%
     of the available audience around Dublin. In September 1982 Radio

Example 2: the network, and was the interim president of The Weather
    Channel for four months in 2013.

Scott is a 25-year veteran of NBC News. Before founding Peacock, she was
    executive producer and general manager of <<NBC>>(2) News Productions
     and NBC Media Productions. She was a member of the executive team
    for "Dateline" and "Now, with Tom Brokaw and Katie Couric."

Scott joined <<NBC>>(3) News in 1990 as news director for WTVJ-<<TV>>(2),
     <<NBC>>(3)'<<s>>(3) Owned and Operated station in Miami. Her honors
    include a number of national news Emmy awards in addition to a George
     Foster Pe

Chain of thought: Step 1: All activations are on words related to
    television and broadcasting.
Step 2: The top promoted logits are related to television and
    broadcasting.
Step 3: These themes are clearly related to filmmakers. I will rate
    filmmakers as a 4, and all other classes as 0.

{"gender": 0, "professor": 0, "nurse": 0, "accountant": 0, "architect":
    0, "attorney": 0, "dentist": 0, "filmmaker": 4}
```

Real example

```
"Okay, now here's the real task.
As a reminder, we only want to use these classes:
    Beauty_and_Personal_Care, Books, Automotive, Musical_Instruments,
    Software
Promoted tokens: connector, cable, connectors, connections, cables
Example prompts:

Example 1: Hager

The Hager Group is a leading supplier of solutions and services for
    electrotechnical installation in residential and commercial buildings
     as <<well>>(44) as for industrial applications.

As a leading supplier of systems, solutions and services for <<electrical
    >>(47) installations, the Hager Group provides an extensive offer,
    ranging from <<power>>(42) distribution and <<cable>>(103) management
     <<to>>(42) smart building automation and safety and security items
        for an equally extensive field of application suitable for
    residential, commercial and industrial properties.

Besides the Hager brand which stands for a wide range of systems,
    solutions and electrotechnical components in buildings, the Hager
    Group is also home to the brands Daitem and Diagral offering security
     items

Example 2: the plaintiff and his family rather than the electrical
    company. As such our analysis of the facts focuses upon the
    information given to Spink when Floyd called to request service.
    Floyd merely indicated there was a "lack of power to his motor" and a
     "power shortage" was perhaps the cause. Floyd did not convey any
    information to Spink that the electrical problem could have
    originated between the meter box and the nonworking motor.
    Furthermore, since Floyd had not detected the <<hidden>>(34) short in
     <<the>>(51) <<extension>>(88) <<cord>>(46), a <<cord>>(39) he had
    exclusive control over, he could not have given sufficient
    information to raise this possibility to Spink. The two employees
    arrived at
```

```

Chain of thought:
```