Do Sparse Autoencoders Generalize? A Case Study of Answerability

Lovis Heindrich [1] **Philip Torr** [2 3] **Fazl Barez** [2 *] **Veronika Thost** [4 *]

Abstract

Sparse autoencoders (SAEs) have emerged as a promising approach in language model interpretability, offering unsupervised extraction of sparse features. For interpretability methods to succeed, they must identify abstract features across domains, and these features can often manifest differently in each context. We examine this through "answerability"—a model's ability to recognize answerable questions. We extensively evaluate SAE feature generalization across diverse, partly self-constructed answerability datasets for Gemma 2 SAEs. Our analysis reveals that residual stream probes outperform SAE features within domains, but generalization performance differs sharply. SAE features show inconsistent out-of-domain transfer, with performance varying from almost random to outperforming residual stream probes. Overall, this demonstrates the need for robust evaluation methods and quantitative approaches to predict feature generalization in SAE-based interpretability.

Figure 1: SAE features vs linear probes trained on SQUAD answerability data: generalization performance varies drastically with different answerability OOD datasets. While the features generalize well to some datasets (Equation, IDK, Celebrity), generalization to BoolQ is almost random.

1. Introduction

Language models increasingly drive real-world applications, yet their black-box nature remains a fundamental barrier to deployment. This lack of visibility has sparked intense interest in interpretability methods (Olah et al., 2018), with sparse autoencoders (SAEs) emerging as a particularly promising direction (Cunningham et al., 2023; Bricken et al., 2023). The core idea is appealingly simple: train an autoencoder to reconstruct neural activations through a sparse bottleneck, forcing the model to learn disentangled, interpretable features.

Recent work demonstrates that SAEs can effectively capture a wide range of features in language models such as errors in code, sycophancy, and gender bias (Templeton et al., 2024), as well as syntax patterns and sentiment (Lan et al., 2024; Marks et al., 2024a). In parallel, circuit-level analyses have begun to reveal mechanistic underpinnings of neural network behavior (Elhage et al., 2021; Olsson et al., 2022; Marks et al., 2024b). While interpretability methods like SAEs are often motivated by AI safety concerns, recent work suggests that even advanced interpretability approaches may have fundamental limitations for ensuring AI safety (Barez et al., 2025). Despite these advances and critiques, a crucial question remains unaddressed: can these methods fully capture the abstract concepts that would help us understand how language models think and process information in a robust way? We study this question by focusing on answerability—a model's ability to recognize whether it can answer a question. This capability is fundamental to language model behavior and exists across diverse tasks. If SAEs truly capture meaningful abstractions, they should be able to extract features representing this capability. Specifically, we evaluate how well a model's notion of answerability generalizes across different datasets by (1) training answerability probes on an in-domain dataset and (2) evaluating the probes on four out-of-distribution datasets, including two synthetic datasets designed by us. Our experimental results reveal a more nuanced picture than previous work suggests. While residual stream probes outperform SAE features within specific domains, we obtain more mixed results when we look at generalization. **We**

[*]Equal advising [1]Work done during a research visit at University of Oxford. [2]University of Oxford, Oxford, United Kingdom [3]WhiteBox [4]MIT-IBM Watson AI Lab. Correspondence to: Lovis Heindrich <lovis.heindrich@student.uni-tuebingen.de>.

Published at ICML 2025 Workshop on Reliable and Responsible Foundation Models. Copyright 2025 by the author(s).

find good SAE features for individual domains, but varying transfer abilities across datasets. Similarly, residual stream probes exhibit high variance in generalization despite strong in-domain performance. These findings raise important questions about SAE research, particularly for abstract or more complex concepts that manifest differently across contexts.

2. Preliminaries

Sparse autoencoders (SAE) have recently been proposed as an interpretability method that aims to decompose the dense and usually uninterpretable activations $\mathbf{x}$ of an LLM into a sparse and interpretable representation $\mathbf{f}$ (Cunningham et al., 2023; Bricken et al., 2023; Yun et al., 2021). This is achieved by training an autoencoder to reconstruct model activations, such that the learned representation is significantly wider and sparser than the original activations.

A standard formulation of the autoencoder consists of an encoder and decoder layer with a ReLU activation function (Equations (1) and (2), Bricken et al., 2023), where $W_e \in \mathbb{R}^{d_{sae} \times d_{model}}$ with $d_{sae} \gg d_{model}$ is the encoder layer with bias term $\mathbf{b}_e$, and W_d is the decoder layer with bias term $\mathbf{b}_d$.

$$\mathbf{f} = \mathrm{ReLU}(W_e(\mathbf{x} - \mathbf{b}_d) + \mathbf{b}_e) \tag{1}$$

$$\hat{\mathbf{x}} = W_d\mathbf{f} + \mathbf{b}_d \tag{2}$$

While initially sparsity has been achieved through L1 regularization in the loss function (Equation (3), Bricken et al., 2023), recent research on SAEs has focused on designing improved methods to achieve sparsity (Lieberum et al., 2024; Gao et al., 2024).

$$\mathcal{L} = \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 + \lambda \|\mathbf{f}\|_1 \tag{3}$$

3. Related Works

SAE Training Much SAE research is dedicated to improving training efficiency and effectiveness, but the latter is usually measured in terms of reconstruction quality and hence disconnected from downstream scenarios (Rajamanoharan et al., 2024; Lieberum et al., 2024). Only most recent work addresses this issue. Gao et al. (2024) apply downstream probing, yet the classification tasks considered are most simple (e.g sentiment, language identification) and likely do not test any generalization. Similarly, Makelov et al. (2024) use downstream data without considering generalization.

SAE Downstream Evaluation Several studies evaluate SAE features, including downstream settings. Yet, research often focuses on simple syntactic features or does not evaluate how general the discovered features are (Yun et al., 2021; Bricken et al., 2023; Kissane et al., 2024). There are various works focusing on more abstract concepts such as indirect object identification (Cunningham et al., 2023) and subject-verb agreement (Marks et al., 2024b), but those are still directly related to syntax. In contrast, answerability often depends on domain-specific background knowledge (e.g., math or factual knowledge) and hence better suits the study on generalization. Demircan et al. (2024) consider the representation of quality estimates from reinforcement learning, and hence a rather complex concept, but they focus on task-specific (vs generalizable) SAEs.

Closest to our work are the following studies. Bricken et al. (2024) focus on comparing SAE features to linear probes as bioweapon classifiers. Similarly to us, they show that the SAE probes are competitive but more complex and brittle; for instance, already format mismatch between the transformer/SAE/probe training data may degrade performance. When evaluated on multilingual out-of-distribution data (similar to in-domain data but in different languages), they find that SAE features can generalize well in specific settings in the mostly lexical task of bioweapon classification. Kantamneni et al. (2024; 2025) similarly conduct experiments comparing to traditional probing on activations and demonstrate that SAEs seem to work in certain scenarios (e.g., with very small datasets or corrupted data). They also consider out-of-distribution data, yet focus more on multilingual and syntactic variations. Similar to us, they obtain mixed results without clear conclusion which probes are better. Generally SAEs do not outperform regular probing. Our evaluation lifts this work to a more complex task and a greater variety of distributions.

In summary, prior work has largely neglected generalization beyond multilingual/syntactic scenarios. Our study closes this gap in the context of a suitable, complex concept, answerability, which likely manifests differently across contexts.

4. Methodology

We evaluate SAE probes for answerability detection with a specific focus on generalization. To evaluate generalization of individual SAE features, we select features which activations lead to high answerability detection on the training dataset, and evaluate if the same features can also detect answerability on a range of out-of-distribution datasets. We compare SAE feature generalization to a baseline of training a linear probe on the residual stream of the LLM. Figure 2 provides an overview of the methodology.

SAE Probes We use the "Gemma Scope" SAEs pretrained by Lieberum et al. (2024) for the instruction-tuned model Gemma 2 (Team et al., 2024), and specifically the largest available SAEs with a width (number of dimensions) of

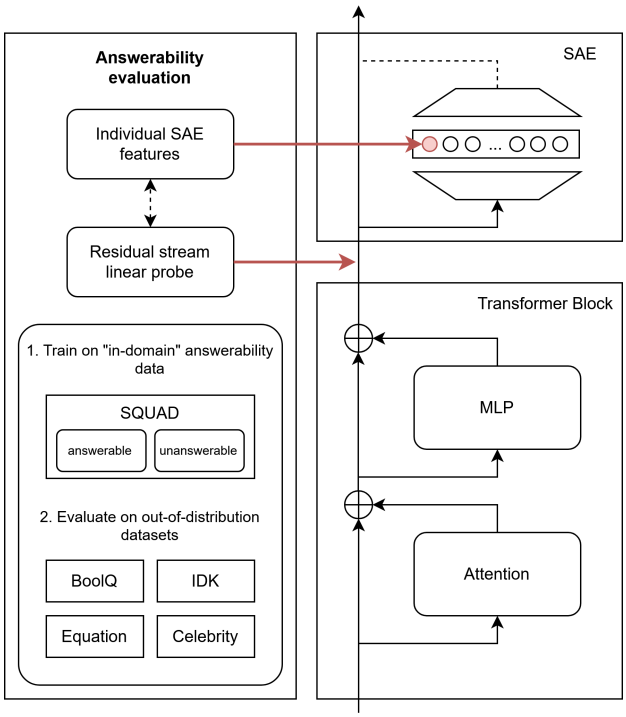


Figure 2: Visualization of our evaluation approach on a simplified transformer and SAE architecture. Activations from individual SAE features and the full residual stream are used to train answerability probes on SQUAD. Generalization performance is evaluated on multiple OOD datasets.

131k. Lieberum et al. (2024) provide SAEs trained on layers 20 and 31. Note that answerability more generally (i.e., beyond specific types of answerability) is a rather high-level concept, which we assume to be represented in intermediate and later layers.

We search for SAE features using 2k samples of SQUAD (balanced, leaving 1.8k for testing). We collect the feature activations on the last token position and then use 5-fold cross validation for finding SAE features that are predictive for answerability, thus obtaining 1-sparse SAE probes (Gurnee et al., 2023). We then train final probes[1] (i.e., scale and bias) for best performing features, which are used for the out-of-distribution evaluation. See Appendix A for details. "Top" features are selected based on training set performance.

Baselines: Linear Probes We train simple linear residual stream probes on the (in-domain) training dataset we also use for finding the SAE features. To ensure robustness, we employ bootstrap analysis across different training splits. Since we also focus on SAE features for the residual stream, this probing represents an upper bound for the SAE probing performance on in-domain data. Observe that these probes achieve 85-90% accuracy on the in-domain SQUAD data, and thus provide a strong benchmark for comparison.

We do not evaluate the SAE or residual stream probes against alternative methods for answerability detection (e.g. Chuang et al., 2024), as our goal was to evaluate SAE generalization capabilities, instead of developing a novel method for answerability detection.

Table 1: Number of total examples per dataset. The number of answerable and unsanwerable examples is balanced within each used dataset.

Dataset	Size
SQUAD (train)	2000
BoolQ (train)	2000
SQUAD (test)	1800
SQUAD (variations)	1800
BoolQ (test)	2000
IDK	484
Equation	2000
Celebrity	600

Datasets We focus on context-based question answering in the English language; we further present a short experiment using Spanish language test data in Appendix B.6. Unless otherwise specified, we treat SQUAD as the in-domain dataset used for training linear probes, and treat all other datasets as out-of-domain (OOD) test datasets. We use established data as well as datasets specifically constructed for out-of-distribution evaluation; for examples see Table 2. Table 1 further shows the number of examples for each dataset used in our evaluation.

- **SQUAD** (Rajpurkar et al., 2018): Dataset consisting of a short context passage and a question relating to the context. We follow the training data split and prompting template provided by Slobodkin et al. (2023).
- **IDK** (Sulem et al., 2021): Dataset with questions in the style of SQUAD, containing both answerable and unanswerable examples. We specifically use the non-competitive and unanswerable subsets of the ACE-whQA dataset.
- **BoolQ_3L** (Sulem et al., 2022): Yes/no questions with answerable and unanswerable subsets.
- **Math Equations**: Synthetic dataset contrasting solvable equations with equations containing unknown variables.
- **Celebrity Recognition**: Context-based queries requiring knowledge about celebrities. For construction, we

[1]We use SAE probes and SAE features synonymously.

Table 2: Answerable and unanswerable examples from the used datasets.

Dataset	Answerable	Unanswerable
SQUAD	**Passage:** The first beer pump known in England is believed to [. . .]. **Question:** When was John Lofting born?	**Passage:** Starting in 2010/2011, Hauptschulen were merged [. . .]. **Question:** In what school year were Hauptschulen last combined with Realschulen and Gesamtschulen?
IDK	**Passage:** Singapore has reported 16 deaths. **Question:** Where are the deaths?	**Passage:** Showed the arrest of the prime suspect. **Question:** Where was the arrest?
BoolQ	**Passage:** On April 20, 2018, ABC officially renewed *Grey's Anatomy* for a network primetime drama record-tying fifteenth season. **Question:** Is season 14 the last of *Grey's Anatomy*?	**Passage:** Discover is the fourth largest credit card brand in the U.S., behind Visa, MasterCard, and American Express, with nearly 44 million cardholders. **Question:** Are pasilla chiles and poblano chiles the same?
Equation	**Given equations:** n = 53 v = 90 **Final equation:** n / v =	**Given equations:** n = 17 u = 38 **Final equation:** n * t =
Celebrity	**Article:** Yesterday, I saw an article about Gerard Butler. They really are a great actor. **Question:** Do you know what their age is?	**Article:** Yesterday, I saw an article about Tania Scott. They really are a great actor. **Question:** Do you know what their age is?

use a public dataset of actors and movies from IMDB[2], and generate a list of the 1000 most popular actors after 1990, as measured by the total number of ratings their movies received. We construct an additional dataset of non-celebrity names by randomly generating first and last name combinations using the most common North American names from Wikipedia[3].

5. Evaluation

In the following, we present of our main experiments; see Appendix B for additional findings.

Linear vs SAE Probes: Generalization, Figure 3 We focus first on layer 31. In domain, the best SAE features reach an accuracy of around 0.8 while the linear probes reach 0.9. Note that this is not surprising, since the probes have more parameters that are actually trained and thus optimized for this data. Nevertheless, it shows some advantage of probes in case in-domain data is available.

We see rather great variation across our out-of-distribution datasets. Our custom Equation data stands out in that several SAE features and also the probe reach high performance. This seems to show that the mathematical context makes answerability easier to detect in general (as top accuracies are very high), and that the internal representation of answerability between the two datasets is compatible (some of the SAE features selected for high answerability detection on SQUAD also detect answerability on the Equation data). Notably, the best SAE features generalize better to Equation data than the residual stream probe. This only applies to the features in layer 31 and trained on SQUAD, the generalization performance is considerably worse on layer 20, see Figure 7 in Appendix B.1.

Similarly, some, but few, top SAE features reach considerable performance out of distribution on IDK (matching the performance of the linear probe) and the Celebrity data (outperforming the linear probe). Yet, the generalization performance on BoolQ is considerably bad. On the other hand, the linear probe performs bad on Celeb. Figure 9 in Appendix B.1 shows the median value over 10 bootstrap samples including quartiles in the error bars; overall it correlates with performance.

These discrepancies in generalization can not be explained by the overall difficulty of the datasets, when selecting top SAE features for each dataset individually (i.e. training the 1-sparse probes on each dataset instead of evaluating the 1-sparse probes trained on SQUAD), we detect high performing features for each dataset (see Figure 11 in the

[2] https://www.kaggle.com/datasets/darinhawley/imdb-films-by-actor-for-10k-actors

[3] https://en.wikipedia.org/wiki/Lists_of_most_common_surnames_in_North_American_countries and https://en.wikipedia.org/wiki/List_of_most_popular_given_names?utm_source=chatgpt.com

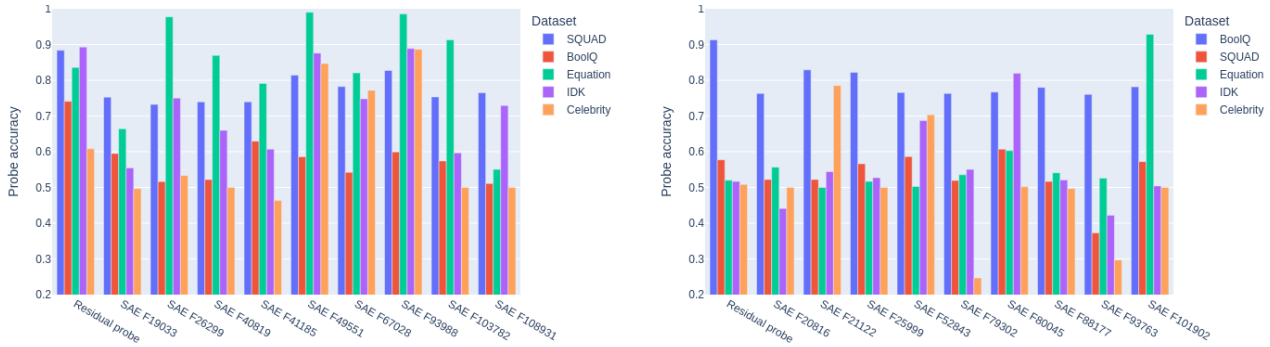


Figure 3: Out-of-distribution comparison between the top SAE features and the linear probe on layer 31. SAE features are trained from pre-ReLU activations. Two in-domain training datasets are evaluated, SQUAD (left) and BoolQ (right), while all other datasets are used as OOD data.

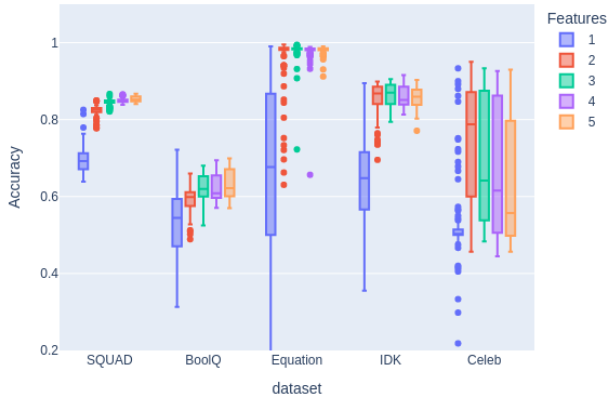


Figure 4: OOD generalization performance for combinations of multiple top SAE features, with different amounts of features (1 to 5). The median value across top feature groups with quartile ranges in the error bars is shown.

Appendix). Instead, the results indicate that the internal representation of answerability varies between the different datasets, such that neither SAE features nor the residual stream probes represent the concept in a fully generalizable way.

Prior work (Sclar et al., 2023) found that LLMs can be highly sensitive to minor variations in their prompts. We evaluated generalization performance on small prompt variations of the SQUAD dataset (see Table 3) as additional out-of-distribution datasets. Figure 10 in Appendix B.2 shows that while both SAE features and linear probes often fail to generalize to our true OOD datasets, they generalize well to this much easier case of out-of-distribution data.

We repeated our analysis for layer 20 of Gemma 2 (Figure 7) and generally observe worse performance. Interestingly, the numbers for Celeb are significantly worse than all others for both the SAE and the linear probes. Since we see one exception (an SAE feature with higher than random performance), we hypothesize that there are special features encoding knowledge about celebrities which do not happen to be among our top answerability features. In fact, a closer investigation reveals that there are good features for BoolQ and our domain-specific Equation and Celeb datasets on layer 20 already (see Figure 11 in Appendix B.3), but they are not the same features as the ones found by training on SQuAD.

Finally, we confirmed our findings by also training our probes on BoolQ (also 2k samples) and evaluating generalization on the other datasets. We mainly see that varying the training data can make generalization considerably worse, even when training on the same task with seemingly similar, but potentially lower-quality data. The unanswerable samples in BoolQ were constructed by combining contexts and questions of similar dataset samples, and therefore capture only one type of unanswerability.

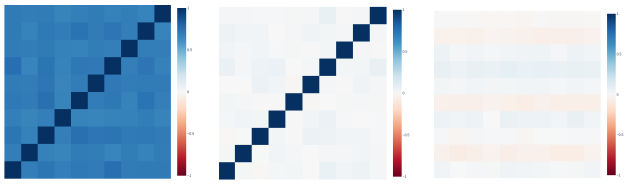
Overall, our experiments demonstrate one main critical issue with OOD data: *the standard procedure for finding good SAE features can easily fail to select features that generalize, even if good features are available*. The fact that good features exist while the linear probes also fail shows some potential of SAEs. Yet finding good, generalizing features represents an open challenge.

Top Features, Figure 3 Interestingly, the top three features on the in-domain SQuAD data happen to also generalize better here. While the in-domain performance on SQUAD of the top SAE features is lower than the residual-stream probe, some of the generalization performance surpasses not only the probe's generalization but also the in-domain performance, indicating that some SAE features represent more general concepts that also detect answerability on the OOD data.

These results do not hold beyond the top-1 feature more generally, see Appendix B.1. For BoolQ, the variability of

the results precludes clear conclusions.

SAE Feature Combinations, Figure 4 Given the partly domain-specific nature of our out-of-distribution datasets, we hypothesized that combinations of features might work better as general probes. However, while increasing the number of SAE features improves the in-domain performance, OOD performance doesn't improve upon the best performing individual feature (top of blue error bars) here; layer 31, pre-activation. Other examples in Appendix B.4 show similar trends, and even some degradation. This underlines our above finding that the OOD setting requires better methods for SAE feature search, potentially in combination with new SAE training methods that encourage general features to be learned.



(a) Residual probe similarities (b) SAE feature similarities (c) SAE residual similarities

Figure 5: Pairwise cosine similarities of (a) ten residual stream probes across different seeds, (b) top SAE features and (c) top SAE features (y-axis) with linear probes across training seeds (x-axis).

Feature Similarity, Figures 5 & 14 We investigated whether how similar individual SAE features with high answerability performance are to each other or to the residual stream probes with cosine similarities. With an average similarity of 0.7, we find that different residual stream probes learn highly similar features, but different high performing SAE features are highly dissimilar from each other and also highly different from the linear probes. Interestingly, the best generalizing SAE feature turns out to have the highest (though overall very low) similarity with the probes. Figure 14 in Appendix B.5 shows that averaging multiple high performing SAE features yields greater similarity with linear probes.

6. Conclusions

We extensively evaluated SAE features for Gemma 2 in the out-of-distribution scenario using a variety of established and custom datasets. On the bright side, we find good SAE features for answerability across these domains. However, we show from various angles that the standard SAE feature search can fail to identify generalizing features, even if they exist. Additionally, fully general features might not be learned by standard SAE training methods, limiting their potential applications for steering and probing. We hypothesize that this is due to both suboptimal training objectives and feature splitting with complex concepts (Bricken et al., 2023; Chanin et al., 2024). This shows the need for better technology for evaluating SAE features before SAEs are robustly applicable in practice. Overall, our results indicate that comprehensive out-of-distribution evaluation is critical for future SAE-based interpretability applications.

Limitations

Our evaluation is bottlenecked by the availability of high-quality SAEs trained on instruction-tuned LLMs. Therefore, we could only evaluate generalization on Gemma 2, where open-source SAEs for are available (Lieberum et al., 2024). While there are available SAEs for other open-source LLMs like Llama 3 (Dubey et al., 2024), these are generally trained on the base models or of uncertain quality. In preliminary tests with Llama 3, we could not find usable answerability features for a probing comparison. Furthermore, the SAE features highly depend on the training hyperparameters and data they were trained on. Nevertheless, we would expect certain features to exist when aiming to use such SAEs in practice. Finally, we used rather simple techniques for SAE probe training, which did not yield best results. But this is the point we intend to make in this work, that the existing technology is insufficient in detecting generalizable SAE features. A key issue for SAE probing could be feature splitting, a phenomenon where SAEs of different sizes learn features in different granularities, often splitting more general features into multiple more specialized features (Marks et al., 2024b; Chanin et al., 2024). If abstract concepts like answerability are split into many separate features, this can cause problems for feature-based practical applications.

References

Barez, F., Fu, T., Prabhu, A., Casper, S., Sanyal, A., Bibi, A., O'Gara, A., Kirk, R., Bucknall, B., Fist, T., Ong, L., Torr, P., Lam, K.-Y., Trager, R., Krueger, D., Mindermann, S., Hernandez-Orallo, J., Geva, M., and Gal, Y. Open problems in machine unlearning for ai safety. 2025. URL https://arxiv.org/abs/2501.04952.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Hatfield-Dodds, Z., Tamkin, A., Nguyen, K., McLean, B., Burke, J. E., Hume, T., Carter, S., Henighan, T., and Olah, C. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Bricken, T., Marcus, J., Mishra-Sharma, S., Tong, M., Perez, E., Sharma, M., Rivoire, K., and Henighan, T. Using dictionary learning features as classifiers. *Transformer Circuits Thread*, 2024. https://transformer-circuits.pub/2024/features-as-classifiers/index.html.

Carrino, C. P., Costa-Jussà, M. R., and Fonollosa, J. A. Automatic spanish translation of the squad dataset for multilingual question answering. *arXiv preprint arXiv:1912.05200*, 2019.

Chanin, D., Wilken-Smith, J., Dulka, T., Bhatnagar, H., and Bloom, J. A is for absorption: Studying feature splitting and absorption in sparse autoencoders. *arXiv preprint arXiv:2409.14507*, 2024.

Chuang, Y.-S., Qiu, L., Hsieh, C.-Y., Krishna, R., Kim, Y., and Glass, J. Lookback lens: Detecting and mitigating contextual hallucinations in large language models using only attention maps. *arXiv preprint arXiv:2407.07071*, 2024.

Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

Demircan, C., Saanum, T., Jagadish, A. K., Binz, M., and Schulz, E. Sparse autoencoders reveal temporal difference learning in large language models. *arXiv preprint arXiv:2410.01280*, 2024.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

Gao, L., la Tour, T. D., Tillman, H., Goh, G., Troll, R., Radford, A., Sutskever, I., Leike, J., and Wu, J. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*, 2024.

Gurnee, W., Nanda, N., Pauly, M., Harvey, K., Troitskii, D., and Bertsimas, D. Finding neurons in a haystack: Case studies with sparse probing. *arXiv preprint arXiv:2305.01610*, 2023.

Kantamneni, S., Engels, J., Rajamanoharan, S., and Nanda, N. Sae probing: What is it good for? absolutely something! https://www.lesswrong.com/posts/NMLq8yoTecAF44KX9/sae-probing-what-is-it-good-for-absolutely-someth 2024.

Kantamneni, S., Engels, J., Rajamanoharan, S., Tegmark, M., and Nanda, N. Are sparse autoencoders useful? a case study in sparse probing, 2025. URL https://arxiv.org/abs/2502.16681.

Kissane, C., Krzyzanowski, R., Bloom, J. I., Conmy, A., and Nanda, N. Interpreting attention layer outputs with sparse autoencoders. *arXiv preprint arXiv:2406.17759*, 2024.

Lan, M., Torr, P., Meek, A., Khakzar, A., Krueger, D., and Barez, F. Sparse autoencoders reveal universal feature spaces across large language models, 2024. Preprint, arXiv:2410.06981.

Lieberum, T., Rajamanoharan, S., Conmy, A., Smith, L., Sonnerat, N., Varma, V., Kramár, J., Dragan, A., Shah, R., and Nanda, N. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. *arXiv preprint arXiv:2408.05147*, 2024.

Makelov, A., Lange, G., and Nanda, N. Towards principled evaluations of sparse autoencoders for interpretability and control. *arXiv preprint arXiv:2405.08366*, 2024.

Marks, L., Paren, A., Krueger, D., and Barez, F. Enhancing neural network interpretability with feature-aligned sparse autoencoders, 2024a. Preprint, arXiv:2411.01220.

Marks, S., Rager, C., Michaud, E. J., Belinkov, Y., Bau, D., and Mueller, A. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. *arXiv preprint arXiv:2403.19647*, 2024b.

Olah, C., Mordvintsev, A., and Schubert, L. The building blocks of interpretability. *Distill*, 2018. URL https://distill.pub/2018/building-blocks/.

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Rajamanoharan, S., Conmy, A., Smith, L., Lieberum, T., Varma, V., Kramár, J., Shah, R., and Nanda, N. Improving dictionary learning with gated sparse autoencoders. *arXiv preprint arXiv:2404.16014*, 2024.

Rajpurkar, P., Jia, R., and Liang, P. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*, 2018.

Sclar, M., Choi, Y., Tsvetkov, Y., and Suhr, A. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. *arXiv preprint arXiv:2310.11324*, 2023.

Slobodkin, A., Goldman, O., Caciularu, A., Dagan, I., and Ravfogel, S. The curious case of hallucinatory (un) answerability: Finding truths in the hidden states of over-confident large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 3607–3625, 2023.

Sulem, E., Hay, J., and Roth, D. Do we know what we don't know? studying unanswerable questions beyond squad 2.0. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 4543–4548, 2021.

Sulem, E., Hay, J., and Roth, D. Yes, no or idk: The challenge of unanswerable yes/no questions. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1075–1085, 2022.

Team, G., Riviere, M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N. L., McDougall, C., MacDiarmid, M., Freeman, C. D., Sumers, T. R., Rees, E., Batson, J., Jermyn, A., Carter, S., Olah, C., and Henighan, T. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Yun, Z., Chen, Y., Olshausen, B. A., and LeCun, Y. Transformer visualization via dictionary learning: contextualized embedding as a linear superposition of transformer factors. *arXiv preprint arXiv:2103.15949*, 2021.

A. Preliminaries

1- sparse SAE probes To evaluate how well SAE features predict a certain abstract feature, we utilize 1-sparse probes (Gurnee et al., 2023). Specifically, we collect activations of a specific SAE feature on a contrastive dataset containing both answerable and not answerable examples, and fit a slope coefficient and intercept to predict the dataset label using linear regression. The Gemma 2 SAEs are trained using a JumpReLU activation function (Lieberum et al., 2024). We can sample SAE activations after the activation function (post-relu) or before (pre-relu). Since there are more learnt features to be found in the latter setting, the main paper figures focus on that. However, we report all results for the post-relu setting in the appendix.

Residual stream probes Our residual stream probes are trained on model activations sampled from the model's residual stream. To avoid overfitting, we train the regression model using 5-fold cross validation and perform a hyperparameter optimization by sweeping over regularization parameters with 26 logarithmically spaced steps between 0.0001 and 1. To measure the variability of residual stream probes, we repeat our analysis 10 times with different randomly sampled training datasets.

N-sparse SAE probes To train SAE probes with more than 1 feature, we follow the general methodology of our 1-sparse probes. As testing all possible SAE feature combinations is computationally infeasible, we iteratively increase the number of features while testing only the most promising candidates for higher features combinations. Specifically, to find combinations of k features, we use the top 50 best performing features of size $k-1$ and test all possible new combinations with the 500 best performing single SAE features. We use a constant regularization parameter of 1 for the probes, regardless of the number of features.

Feature similarities To calculate feature similarities, we use the cosine similarity of the corresponding SAE encoder weight and the slope coefficients of the linear probes trained on the residual stream. SAE features are only compared to other SAE features of the same SAE, and residual stream probes trained at the same location in the model as the SAE. To compare how similar differently sized groups of SAE features are to the residual stream probes, we calculate the mean absolute cosine sim of the top 10 best performing SAE features of a certain group size (1 to 5) with the 10 residual stream probes trained on different training subsets.

B. Additional analysis

B.1. Answerability Detection at Different Layers

We repeat our SAE feature analysis in Layer 20 of the model, as well as providing additional analysis for SAE features activations sampled after the activation function. Figure 7 shows the Layer 20 results using activations sampled before the activation function, while Figures 6 and 8 show analogous results when sampling SAE activations after the activation function. Sampling after the activation reduces the number of relevant features our probe finds, since many features are inactive. However, this does not change the overall results, as we still find features with good general-

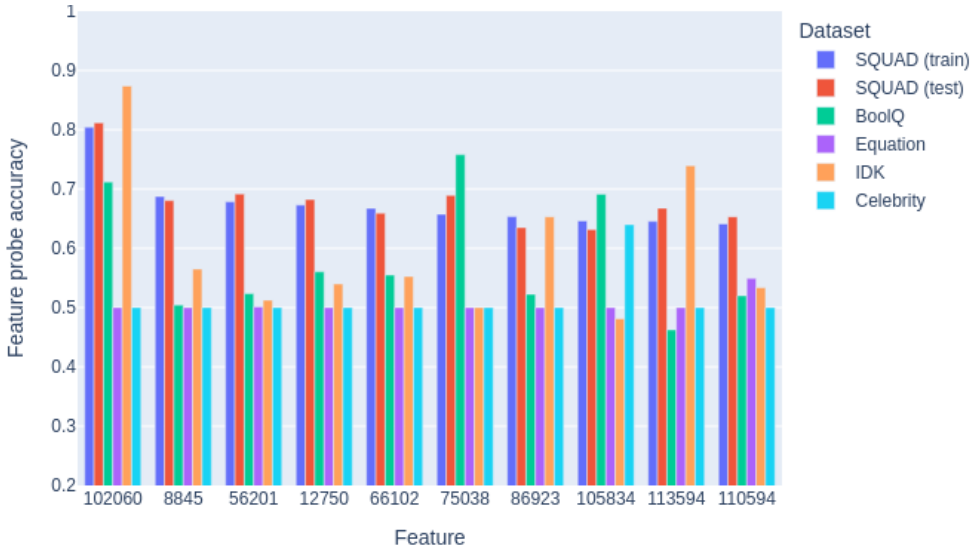


Figure 6: Answerability detection accuracies for top SAE features (Layer 20, post-activation).

ization performance.

Figure 9 shows the probing accuracy for the residual stream linear probe for both Layer 20 and 31. The evaluation is repeated across 10 seeds with different training set splits. While the SAE features, as part of the pre-trained autoencoder model, do not heavily depend on the probing dataset, this is not necessarily true for the residual stream probe. The's probe performance across the out-of-distribution datasets varies strongly, indicating that the generalization performance heavily depends on the minor differences in the training data.

B.2. Prompt variations

We investigated if the SAE features or the residual stream probes are sensitive to small variations in the prompt. To evaluate this question, we created five variations of the prompt template used for the SQuAD training data (see Table 3). The results can be found in Figure 10, and indicate neither the residual stream probe nor the SAE features are sensitive to this kind of variation.

B.3. In-domain SAE feature accuracies

Figure 11 shows the accuracy of 1-sparse SAE feature probes for each dataset individually, demonstrating that each of our contrastive datasets is detectable with a probing accuracy of over 80%.

B.4. SAE Feature Combination Analyses

Figure 12 shows additional probing analysis for the best performing groups of SAE features up to a group size of five. Group performance is generally dominated by the best performing features and does not majorly exceed the performance of the strongest feature.

Figure 13 shows additional analysis for SAE feature combinations in Layer 20, analogous to the results for Layer 31 given in Figure 4.

B.5. Cosine Similarities

We conducted an additional similarity analysis for the top SAE feature groups of different sizes. The results can be found in Figure 14 and show a clear trend of larger groups of features becoming more similar to the linear probes. This provides some weak evidence that by default, linear probes might learn more specialized directions that can be represented as a linear combination of more general SAE features.

B.6. Spanish language generalization

We performed an additional generalization experiment using a Spanish language variant of SQUAD (Carrino et al., 2019). Similar to our other OOD experiments, we used residual and SAE probes trained on the English SQUAD dataset and evaluated their generalization capabilities to the Spanish SQUAD data. Our results did not show large differences in generalization between individual features or the residual stream probe (see Figure 15). Instead, all top probes and the residual probe generalized roughly equally well, with a performance gap of slightly below 10%.

Other experiments We validated our setup by searching for bias-related features as it was done in related works. We also experimented with (inofficial) SAEs for an instruction-tuned Llama model, but could not find SAE features with

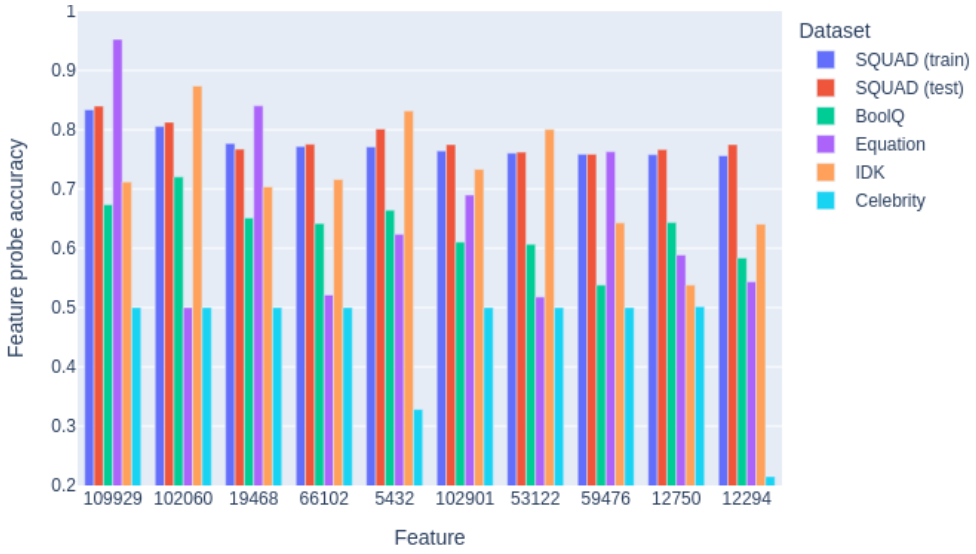


Figure 7: Answerability detection accuracies for top SAE features (Layer 20, pre-activation).

sufficient in-domain probing accuracy. Finally, we also performed analysis on Gemma 2 2B and also the base models, but performance on the answerability task was relatively low in these models (the best SAE features achieved around 70% probing accuracy).

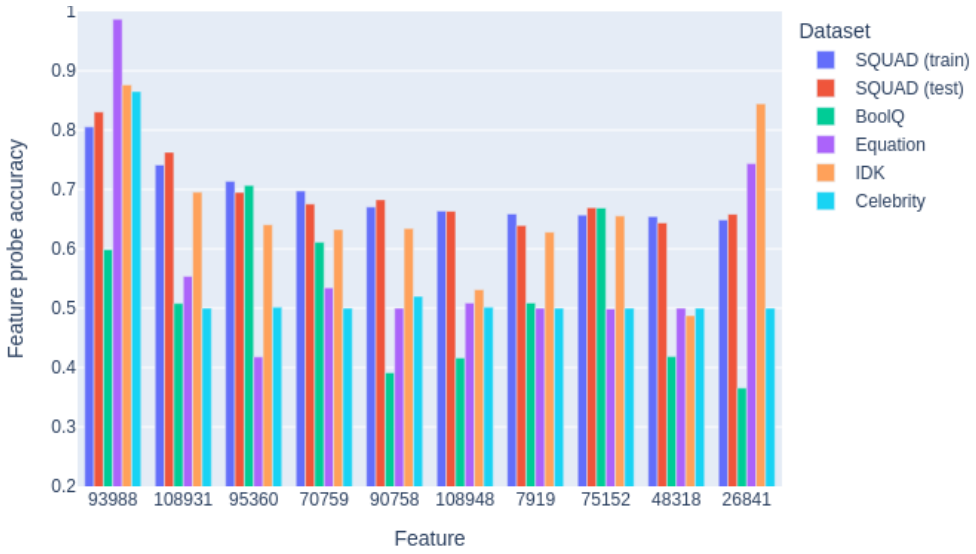


Figure 8: Answerability detection accuracies for top SAE features (Layer 31, post-activation).

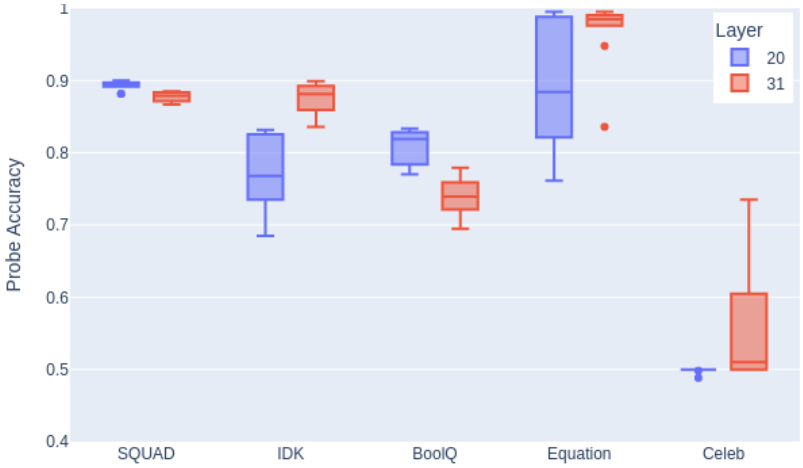


Figure 9: Linear probe trained on Layer 20 and Layer 31 residual stream (SQuAD) and evaluated on IDK, BoolQ, Celebrity, and Equation. The plot shows the median accuracy including the first and third quartile.

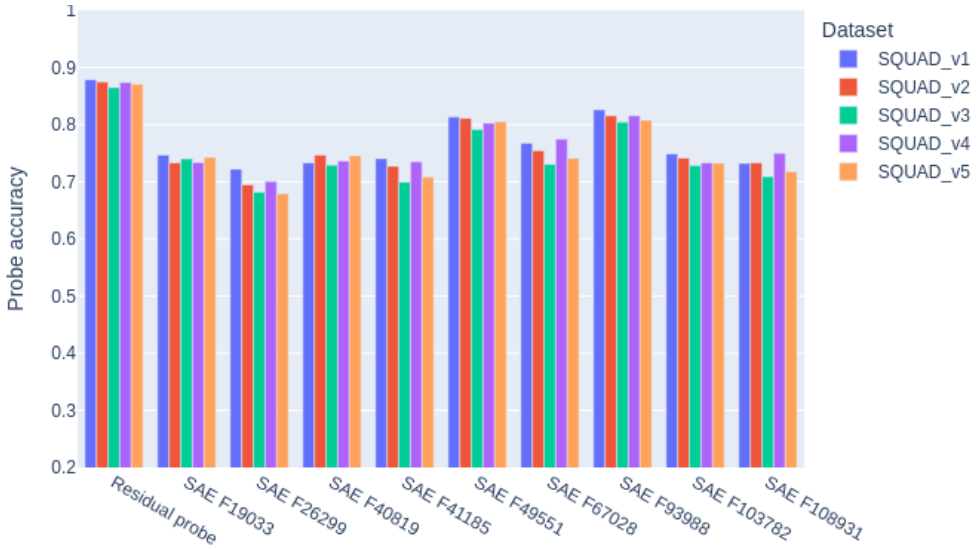


Figure 10: Performance of top SAE features and the residual stream linear probe on variations of prompt used with the SQuAD dataset (layer 31, pre-activation).

Default	Given the following passage and question, answer the question: Passage: {passage} Question: {question}
Variation 1	Please read this passage and respond to the query that follows: Passage: {passage} Question: {question}
Variation 2	Based on the text below, please address the following question: Text: {passage} Question: {question}
Variation 3	Consider the following excerpt and respond to the inquiry: Excerpt: {passage} Inquiry: {question}
Variation 4	Review this content and answer the question below: Content: {passage} Question: {question}
Variation 5	Using the information provided, respond to the following: Information: {passage} Query: {question}

Table 3: SQuAD prompt template variations.

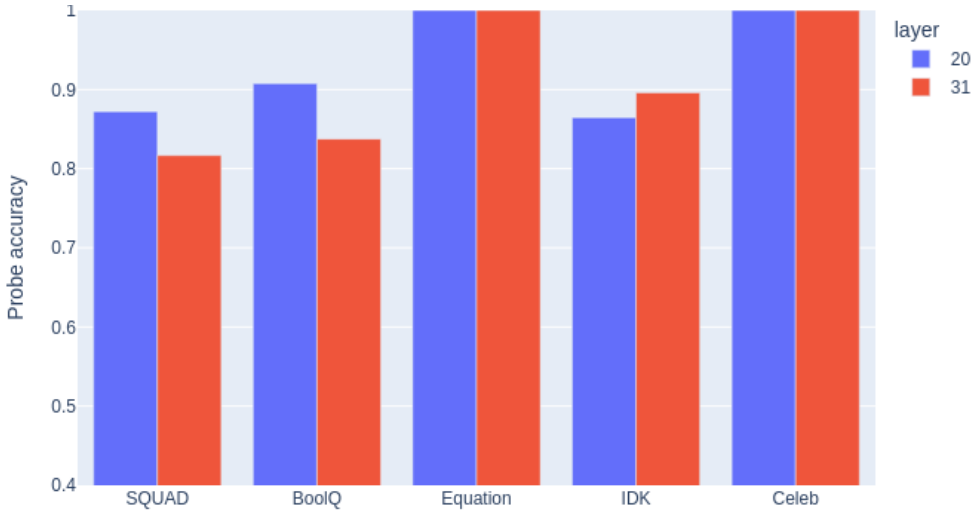


Figure 11: Performance of the top SAE feature's probing accuracy when training and evaluating features on each dataset individually (pre-activation).

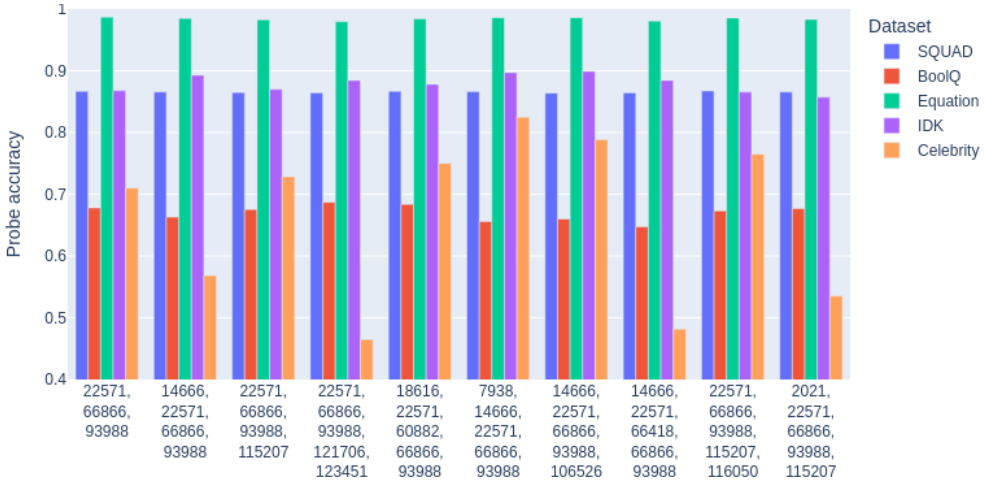


Figure 12: Performance of top feature combinations (layer 31, pre-activation).

Figure 13: Accuracies of SAE probes trained on different numbers of SAE features (Layer 20, pre-activation).

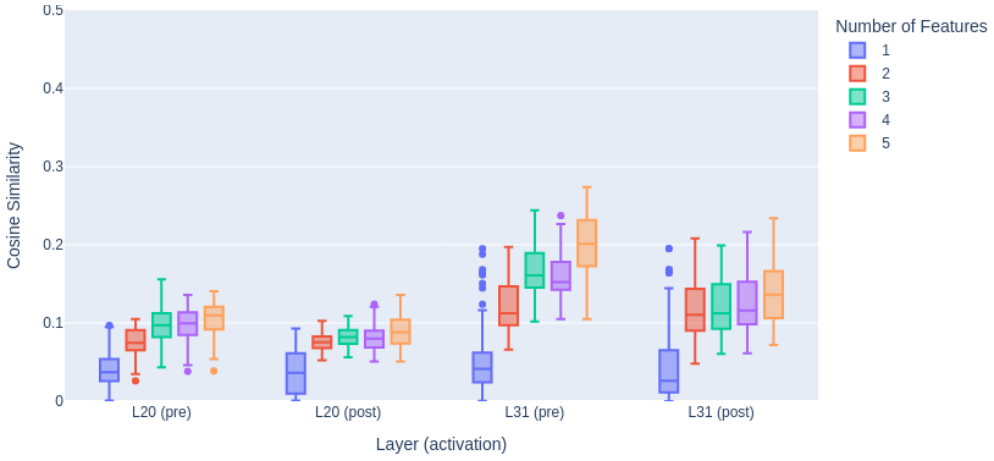


Figure 14: Absolute cosine similarities of top 10 SAE features at different layers, compared with the residual stream probe.

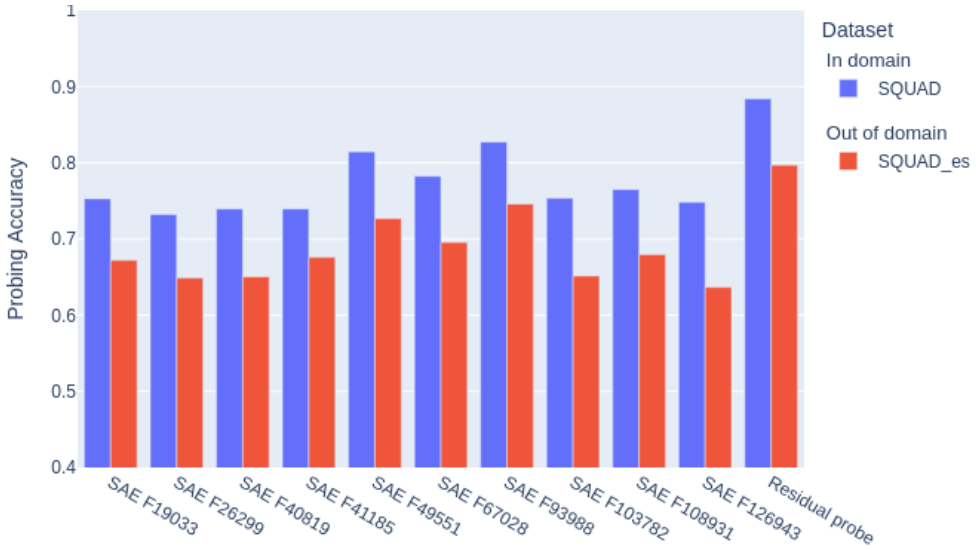


Figure 15: Probing accuracies for top SAE features and the residual stream probe when evaluated on Spanish OOD data.