On the Versatility of Sparse Autoencoders for In-Context Learning

Ikhyun Cho[1] **Gaeul Kwon**[2] **Julia Hockenmaier**[1]
[1]University of Illinois at Urbana-Champaign [2]University of Southern California
ihcho2@illinois.edu gaeulkwo@usc.edu juliahmr@illinois.edu

Abstract

Sparse autoencoders (SAEs) are emerging as a key analytical tool in the field of mechanistic interpretability for large language models (LLMs). While SAEs have primarily been used for interpretability, we shift focus and explore an understudied question: "Can SAEs be applied to practical tasks beyond interpretability?" Given that SAEs are trained on billions of tokens for sparse reconstruction, we believe they can serve as effective *extractors*, offering a wide range of useful knowledge that can benefit practical applications.

Building on this motivation, we demonstrate that SAEs can be effectively applied to in-context learning (ICL). In particular, we highlight the utility of the SAE-reconstruction loss by showing that it provides a valuable signal in ICL—exhibiting a strong correlation with LLM performance and offering a powerful unsupervised approach for prompt selection. These findings underscore the versatility of SAEs and reveal their potential for real-world applications beyond interpretability. Our code is available at https://github.com/ihcho2/SAE-GPS.

1 Introduction

Sparse autoencoders (SAEs), an entirely unsupervised approach, are rapidly gaining popularity as a key tool in mechanistic interpretability (MI), which aims to explain and understand how large language models (LLMs) process and generate responses (Cunningham et al., 2023; Sharkey et al., 2025). Trained with a sparsity loss term to reconstruct the input (i.e., LLM embedding) in a fully unsupervised manner, SAEs have been shown to be effective, *to some extent*, in decomposing embeddings into sparse features, where each feature corresponds to human-understandable, mono-semantic concept (Sharkey and Beren, 2022; Bricken et al., 2023). Although SAEs have been predominantly used for MI, there is no strong reason to assume they are limited to this domain[1]. Thus, in this work, we explore the understudied question of whether SAEs can be practical for real-world applications beyond MI. Promisingly, our findings reveal that SAEs are highly informative and can be effectively applied to in-context learning (ICL), a widely used approach of LLMs. Promisingly, our findings reveal that SAEs provide meaningful insights and can be effectively applied to in-context learning (ICL), a widely used paradigm in large language models.

ICL—an emergent ability of LLMs to perform diverse tasks by learning on-the-fly from just a few exemplars within a single prompt—has become a dominant paradigm in natural language processing (Brown et al., 2020). Yet, understanding the underlying mechanisms of ICL remains an open challenge (Von Oswald et al., 2023). In this paper, we analyze ICL behaviors of LLMs through the lens of SAEs, uncovering their interesting patterns and insights.

Our key motivation is that SAEs serve as effective *extractors*, trained on billions of tokens, that decompose LLM embeddings into sparse, explicit features. We believe these well-trained extractors provide new sources of knowledge that can directly benefit multiple aspects of ICL. In particular, we focus on the **SAE-Reconstruction Loss (SAE-RL)**—defined as the L2-norm between the input LLM embedding and the SAE's reconstructed embedding—and demonstrate that it offers valuable information to enhance the ICL framework.

We hypothesize that *"the richer the knowledge contained in an embedding, the more challenging it becomes for SAEs to disentangle it into distinct features—resulting in a higher SAE-RL"*. This, in turn, allows us to effectively measure the overall quality of the embeddings produced by a given prompt,

[1]Although the sparsity loss term of SAEs is designed for sparse dictionary learning, we believe it is relatively flexible, allowing broader applications—which is the primary focus of this paper.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 19531–19538
November 4-9, 2025

making ICL significantly more predictable. These findings highlight the versatility of SAEs, paving the way for exciting new avenues for exploration.

2 Related Work

In-Context Learning in LLMs In-context learning (ICL) is a paradigm in which large language models (LLMs) are provided with a few exemplars within a single prompt and generate responses conditioned on these examples without any task-specific fine-tuning (Brown et al., 2020). By leveraging pre-trained knowledge, ICL enables efficient generalization across various tasks, making it highly adaptable in natural language processing.

However, the precise mechanism behind ICL remain debated and not fully understood (Dong et al., 2022; Wang et al., 2023; Von Oswald et al., 2023). Additionally, ICL faces various practical challenges such as exemplar selection (Liu et al., 2021; Ye et al., 2023), exemplar order issues (Min et al., 2022), and prompt sensitivity (Zhou et al., 2022; Cho et al., 2024, 2025).

In this paper, we examine ICL through the lens of SAEs, aiming to uncover insights that not only deepen our understanding of ICL but also enhance its performance by addressing the challenges mentioned above.

Sparse Autoencoders in LLM-Interpretability SAEs have emerged as a key approach to sparse dictionary learning, which aims to decompose LLM embeddings into sparse monosemantic features (Huben et al., 2024). They have proven surprisingly effective at uncovering ground truth features in controlled, toy experiment settings (Sharkey and Beren, 2022), although many challenges remain when applying them to real LLMs. Nonetheless, due to their simplicity and effectiveness, SAEs have become one of the dominant baselines in MI.

SAEs use a squared error reconstruction loss and a sparsity penalty:

$$L_{SAE} = \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{x}^{(i)} - \hat{\mathbf{x}}^{(i)}\|_2^2 + \lambda \cdot L_{sparsity} \tag{1}$$

and the SAE-reconstruction loss is defined as:

$$\text{SAE-RL} = \|\mathbf{x}^{(i)} - \hat{\mathbf{x}}^{(i)}\|_2^2 \tag{2}$$

where $\mathbf{x}^{(i)}$ represents the original input, and $\hat{\mathbf{x}}^{(i)}$ is the reconstructed output of the SAE:

$$\hat{\mathbf{x}}^{(i)} = W_{dec}\big(\sigma(W_{enc}(\mathbf{x}^{(i)}))\big) \tag{3}$$

For the activation function σ, various techniques are used, such as JumpReLU (Lieberum et al., 2024) and TopK-ReLU (Gao et al., 2024).

Sparse Autoencoders beyond Interpretability A growing body of recent work explores the application of SAEs to practical tasks beyond MI (Cho and Hockenmaier), but their application in ICL remains underexplored. For example, Demircan et al. (2024) show that certain SAE features align with temporal difference loss in Markov decision problems when provided as exemplars in LLM prompts. Kharlapenko et al. (2025) use SAEs to reconstruct task vectors (Todd et al., 2023) for simple tasks like single-token replacement, but this method struggles to scale to more complex tasks (Brumley et al., 2024).

In contrast, we apply SAEs to general, complex ICL tasks, such as classification and multiple-choice QA, without relying on additional frameworks. Our work also addresses key challenges in ICL, including effective prompt selection and making ICL more predictable.

Off-the-Shelf Value of SAEs A key advantage of SAEs is their *off-the-shelf* usability. Since they are trained on broad, general-purpose corpora such as *The Pile* (Gao et al., 2020), SAEs can be readily applied without the need for task-specific fine-tuning. Owing to this advantage, we believe that demonstrating the effectiveness of SAEs in enhancing aspects of general ICL carries substantial merit.

3 SAE-GPS: SAE-Guided Prompt Search

In this paper, we show that SAEs can add values to ICL, making it more predictable in multiple ways. Specifically, we verify that the SAE-reconstruction loss (SAE-RL) exhibits a strong correlation with LLM performance across various tasks. Based on this finding, we introduce **SAE-GPS** (**SAE**-**G**uided **P**rompt **S**earch), which enables selecting effective prompts (e.g., a set of ICL exemplars) on-the-fly in an entirely unsupervised manner.

Intuition The core intuition behind SAE-GPS is that the more knowledge an embedding contains, the more challenging it becomes for SAEs to decompose it effectively, resulting in a higher SAE-RL. Therefore, SAE-RL allows comparing different prompts directly in an unsupervised manner. We verify this core intuition in Section 5.1.

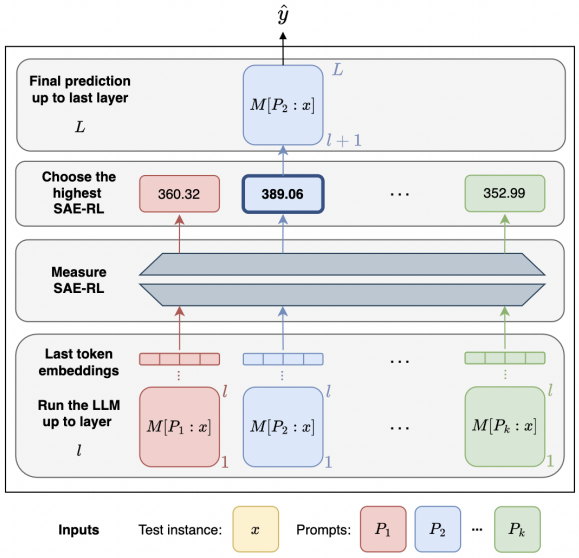


Figure 1: The overall architecture of SAE-GPS. Prompt with the highest SAE-RL is selected.

Methodology Motivated by our intuition, we present SAE-GPS (**SAE**-**G**uided **P**rompt **S**earch), an inference-time method that relies solely on SAE-RL to accurately select effective prompts on-the-fly. The overall architecture of SAE-GPS is illustrated in Figure 1. Given a test instance x and k multiple candidate prompts $P_1, P_2, \ldots, P_k$, the goal is to select the prompt P^* that yields the best performance. We utilize our key intuition that prompts with higher SAE-RL (i.e., using more knowledge) tend to predict more accurately. Accordingly, we run the LLM M with each prompt P_i up to layer l (i.e., $g_i = M_{1:l}[P_i; x]$), then measure the SAE-RL of the last tokens' embeddings (i.e., $||g_{i,last} - SAE(g_{i,last})||_2^2$). The prompt $\hat{P}$ that produces the highest SAE-RL is selected, and further processed through the remaining layers for the final prediction (i.e., $\hat{y} = M_{1:L}[\hat{P}; x]$). We show in Section 5.2 that SAE-GPS consistently outperforms the baselines, and can also be effectively performed at layers below the final layer. This offers a unique computational benefit by requiring processing only up to layer l for the remaining $(k-1)$ prompts, thereby reducing the computational burden.

4 Experimental Settings

Datasets We select two widely used classification datasets in ICL: aspect-based sentiment classification (ABSC) (SemEval-14-Laptops and Restaurants (Pontiki et al., 2014)) and one widely used multiple-choice question-answering dataset: MMLU (Hendrycks et al., 2021). We use the multilingual version of MMLU provided by OpenAI (OpenAI, 2024) for multilingualism analysis. We create the validation set by selecting 300 instances per label from the training set, and the final evaluation is conducted on the test set. Detailed explanations of each task can be found in Appendix A.

Models and Settings One requirement of our work is having access to well-trained SAEs for the LLM. Since training SAEs is computationally expensive, we utilize the open-sourced SAEs released to encourage research by Lieberum et al. (2024) and etc. Consequently, we employ the mid-size model—Gemma2-9B-IT (Team, 2024) and Llama3-8B-IT—along with its corresponding public SAEs. We evaluate our approach on the general few-shot setting using two exemplars per label. We use layer 32's SAE unless specified otherwise.

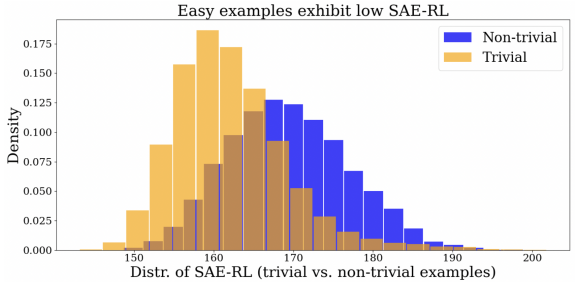
One requirement of our work is having access to well-trained SAEs for the underlying LLM. Since training SAEs is computationally expensive, we leverage publicly released SAEs that were open-sourced to encourage research by Lieberum et al. (2024) and others. Accordingly, we employ mid-sized models—Gemma2-9B-IT (Team, 2024) and LLaMA-3-8B-Instruct (AI)—together with their corresponding public SAEs. We evaluate our approach in a standard few-shot setting using two exemplars per label.

5 Main Results and Analyses

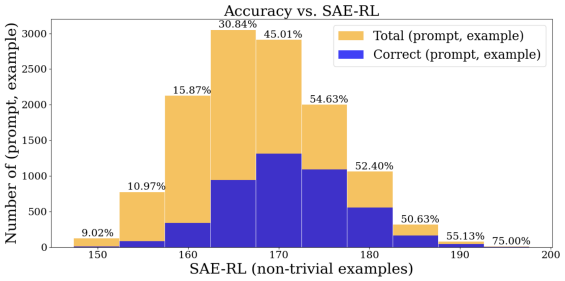
5.1 SAE-RL provides valuable information

As discussed in Section 3, we hypothesize that embeddings containing richer knowledge are likely to yield higher SAE-RL values. To test this, we employed 20 different prompts (i.e., randomly selected sets of ICL exemplars) and compared their average SAE-RL with overall test performance. Figure 2(c) shows a very strong correlation between performance and SAE-RL, confirming our intuition (see also Figure 2(d) in Section 5.4).

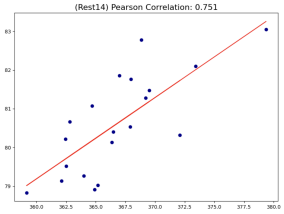
Furthermore, Figures 2(a) and (b) offer a fine-grained, example-level analysis. In Figure 2(a), trivial examples (defined as those for which all 20 prompts yield correct answers) exhibit lower SAE-RL than non-trivial examples, indicating that SAE-RL can serve as a signal to distinguish between easy and challenging examples for LLMs. Figure 2(b) further shows that among non-trivial examples, prompts with higher SAE-RL are more likely to produce correct answers.



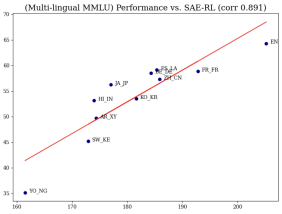
(a) Easy examples exhibit small SAE-RL (Rest14).



(b) Accuracy vs. SAE-RL (Rest14, example-level).



(c) Accuracy vs. SAE-RL (Rest14, Corpus-level)



(d) Accuracy vs. SAE-RL (MMLU-M, Corpus-level)

Figure 2: **(a):** Easy examples show lower SAE-LR. **(b), (c), (d):** Prompts with higher SAE-RL tends to have better performance, aligning with our core intuition.

These results validate our hypothesis and support the use of SAE-RL as the key basis for our proposed method, SAE-GPS.

5.2 SAE-GPS significantly outperforms the baselines *even when using the earlier layers*

Table 1 summarizes the performance of SAE-GPS on Gemma2-9B-IT (Team, 2024), demonstrating that SAE-GPS can select effective prompts with high accuracy. Furthermore, we observe that SAE-GPS can perform effectively using earlier layers rather than just the final layer, offering a computational advantage. Surprisingly, we find that using layers 28 or 32 is more effective than using the final layer (layer 42). These results suggest that the SAEs from earlier layers provide sufficient —and possibly more task-relevant— information, giving SAE-GPS a unique computational advantage over baselines that require processing through all layers.

We also experiment with the inverse version of SAE-GPS (Flipped SAE-GPS), which selects the prompt with the lowest SAE-RL. The results, summarized in the bottom of Table 1, indicate that prompts with lower SAE-RLS generally lead to worse performance.

	Rest14	Lap14
Gemma2-9B-IT	Acc&F1	Acc&F1
Baselines		
ICL Avg.	$80.73_{0.65}$	$75.44_{0.35}$
ICL + Majority Vote	$80.59_{0.91}$	$76.20_{0.43}$
Ours: ICL + SAE-GPS		
Layer 42	$82.41_{0.52}$	$76.95_{0.87}$
Layer 35	$82.68_{0.44}$	$\underline{77.96_{0.46}}$
Layer 32	$\underline{83.55_{0.33}}$	$\mathbf{78.42_{0.45}}$
Layer 28	$\mathbf{83.77_{0.37}}$	$77.98_{0.38}$
Layer 21	$80.67_{1.08}$	$76.11_{0.41}$
Layer 14	$81.45_{0.11}$	$75.57_{0.65}$
ICL + Flipped SAE-GPS		
Layer 42	$78.95_{0.80}$	$73.27_{1.47}$
Layer 35	$78.59_{1.00}$	$72.80_{2.19}$
Layer 32	$78.42_{0.93}$	$72.05_{1.42}$
Layer 28	$78.28_{0.30}$	$72.50_{1.52}$
Layer 21	$81.22_{1.18}$	$74.11_{1.31}$
Layer 14	$79.76_{0.31}$	$74.20_{1.39}$

Table 1: With $k = 15$ random exemplar sets given, our proposed SAE-GPS consistently outperforms majority voting when using layers over 28. Note that SAE-GPS does not use any additional knowledge except SAE reconstruction loss. Experiments are conducted across five random seeds, with standard errors shown as subscripts for the average of accuracy and F1 score.

1. Gemma2-9B-IT	Rest14	Lap14	MMLU (STEM)
ICL Avg.	80.73	75.44	63.94
ICL + Majority Vote	80.59	76.20	64.28
SAE-GPS (Flipped)	78.28	72.50	64.11
SAE-GPS	**83.77**	**77.98**	**65.20**
2. Llama3-8B-Instruct	Rest14	Lap14	MMLU (STEM)
ICL Avg.	76.84	75.59	45.83
ICL + Majority Vote	77.23	75.50	45.98
SAE-GPS (Flipped)	75.62	74.87	44.36
SAE-GPS	**79.10**	**77.21**	**47.15**

Table 2: We provide additional experiments on MMLU (STEM topics) as well as with another model, LLaMA-3-8B-Instruct (AI). These results show that our SAE-GPS method is generally effective across different models and extends beyond classification tasks to multiple-choice question answering.

5.3 General Effectiveness and Compatibility of SAE-GPS

Table 2 presents additional experiments on MMLU (STEM topics) as well as results with another model, LLaMA-3-8B-Instruct (AI) (using the SAE from layer 25). These results demonstrate that our SAE-GPS method is broadly effective across different models and extends beyond classification tasks to multiple-choice question answering.

In Table 3, we further examine the synergy between SAE-GPS and exemplar selection strategies. Specifically, we reproduced the KATE approach (Liu et al., 2021), a standard unsupervised exemplar selection baseline. KATE identifies the top-k most similar exemplars based on cosine similarity between the embeddings of the test query and the candidate exemplars. We implemented this using dense embeddings from the LLM. To generate candidate prompts, we repeatedly sampled six exemplars at random from the top-30 most similar exemplars. Our results show that SAE-GPS achieves even stronger performance when combined with KATE, with the exception of Rest14, where KATE underperforms relative to the random selection baseline. Nevertheless, SAE-GPS still improves performance by 1.05%, indicating that even when the candidate prompt set is of poor quality, SAE-GPS can contribute meaningfully to performance gains.

1. Gemma2-9B-IT	Rest14	Lap14	MMLU (STEM)
KATE	80.07	77.71	64.33
KATE + Majority Vote	80.32	77.88	64.54
KATE + SAE-GPS (Flipped)	79.63	76.87	64.02
KATE + SAE-GPS	**81.12**	**79.96**	**65.65**

Table 3: SAE-GPS creates synergy with exemplar-selection methods (Liu et al., 2021).

5.4 SAE-RL accurately identifies multi-lingual performance

In this study, we investigate whether SAE-RL provides valuable insights beyond its original domain, particularly in multilingualism in LLMs. Using M-MMLU (OpenAI, 2024), which offers parallel data in multiple languages, we examine whether SAE-RL can illuminate the multilingual capabilities of LLMs. As shown in Figure 2 (d), SAE-RL exhibits a remarkably strong correlation with language performance, reinforcing our intuition that languages with stronger capabilities (e.g., English, Chinese) tend to have a better task understanding —leading to a richer feature representation and, consequently, a higher SAE-RL. We believe this finding is valuable for two reasons: (1) it further supports SAE-RL as a generally informative signal across diverse domains, and (2) it enables precise identification of languages where the LLM struggles. This insight can guide LLM developers in optimizing data allocation for specific languages.

6 Conclusion

This paper investigates whether sparse autoencoders (SAEs) can benefit in-context learning (ICL) in large language models (LLMs). We present several novel findings and approaches. First, we show that SAE-reconstruction loss (SAE-RL) is highly informative and can be leveraged to predict effective prompts during inference. We also demonstrate that SAE-RL can distinguish between easy and challenging examples for LLMs. Furthermore, SAE-RL appears to be a valuable knowledge source across various domains—including multilingualism—thereby opening exciting avenues for future research. To the best of our knowledge, our work is among the first to explore the versatility of SAEs in ICL, highlighting their potential and paving the way for further advancements in the field.

7 Limitations

The major limitation of our study is that it requires sparse autoencoders (SAEs), which may not always be available due to the high computational cost of training them. Consequently, our work relies on the publicly available SAEs provided by Lieberum et al. (2024), limiting our experiments to the Gemma2 model (Team, 2024). Exploring our findings and evaluating the effectiveness of our approaches on other model types (once their SAEs become available) would be an intriguing direction for future research.

8 Acknowledgment

This research was supported in part by Other Transaction award HR0011249XXX from the U.S. Defense Advanced Research Projects Agency (DARPA) Friction for Accountability in Conversational Transactions (FACT) program. Additionally, this research used the Delta advanced computing and data resource which is supported by the National Science Foundation (award OAC 2005572) and the State of Illinois. Delta is a joint effort of the

University of Illinois Urbana-Champaign and its National Center for Supercomputing Applications.

References

Meta AI. Llama 3: Open and efficient foundation language models. *arXiv preprint arXiv:2407.12345*.

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. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Madeline Brumley, Joe Kwon, David Krueger, Dmitrii Krasheninnikov, and Usman Anwar. 2024. Comparing bottom-up and top-down steering approaches on in-context learning tasks. *arXiv preprint arXiv:2411.07213*.

Ikhyun Cho and Julia Hockenmaier. Analyzing multilingualism in large language models with sparse autoencoders. In *Second Conference on Language Modeling*.

Ikhyun Cho, Gaeul Kwon, and Julia Hockenmaier. 2024. Tutor-icl: Guiding large language models for improved in-context learning performance. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9496–9506.

Ikhyun Cho, Changyeon Park, and Julia Hockenmaier. 2025. The power of bullet lists: A simple yet effective prompting approach to enhancing spatial reasoning in large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3047–3057.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2023. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*.

Can Demircan, Tankred Saanum, Akshay K Jagadish, Marcel Binz, and Eric Schulz. 2024. Sparse autoencoders reveal temporal difference learning in large language models. *arXiv preprint arXiv:2410.01280*.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

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. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. 2024. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

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*.

Dmitrii Kharlapenko, Stepan Shabalin, Fazl Barez, Neel Nanda, and Arthur Conmy. 2025. Scaling sparse feature circuits for studying in-context learning.

Tom Lieberum, Senthooran 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*.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*.

OpenAI. 2024. Multilingual mmlu dataset. Hugging Face.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Lee Sharkey and Dan Braun Beren. 2022. [interim research report] taking features out of superposition with sparse autoencoders.

Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, Stella Biderman, Adria Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, Eric J. Michaud, Stephen Casper, Max Tegmark, William Saunders, David Bau, Eric Todd, Atticus Geiger, Mor Geva, Jesse Hoogland, Daniel Murfet, and Tom McGrath. 2025. Open problems in mechanistic interpretability.

Gemma Team. 2024. Gemma.

Eric Todd, Millicent L Li, Arnab Sen Sharma, Aaron Mueller, Byron C Wallace, and David Bau. 2023. Function vectors in large language models. *arXiv preprint arXiv:2310.15213*.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. 2023. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pages 35151–35174. PMLR.

Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. 2023. Label words are anchors: An information flow perspective for understanding in-context learning. *arXiv preprint arXiv:2305.14160*.

Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. 2023. Compositional exemplars for in-context learning. In *International Conference on Machine Learning*, pages 39818–39833. PMLR.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2022. Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910*.

A Tasks and Datasets

In our study, we employ an aspect-based sentiment classification (ABSC) task and a multiple-choice question-answering (multiple-choice QA) task. For ABSC, we specifically use SemEval-14-Laptops and Restaurants datasets (Lap14 and Rest14) (Pontiki et al., 2014). For multiple-choice QA, we use the multilingual version of Massive Multitask Language Understanding dataset (M-MMLU) (Hendrycks et al., 2021; OpenAI, 2024) focusing on the STEM domain for multilingualism analysis.

Lap14 and Rest14 are the tasks to evaluate the sentiment (positive, negative, or neutral) of the given laptop and restaurant reviews toward a specified target within the sentence.

M-MMLU is a collection of MMLU's test set translated into 14 languages, where the task is to answer multiple-choice questions across 57 subjects. These subjects are organized into four main domains: Humanities, Social Sciences, STEM, and Others. For this study, we focus exclusively on the STEM domain, which contains 19 of 57 subjects.

Validation Set Settings The validation set is constructed by selecting 300 instances per label from the training set, except when fewer than 300 instances are available. For M-MMLU, we use the entire provided MMLU development and validation sets as our validation set. The final evaluation is conducted on the test set. Detailed statistics are provided in Table 4. All experiments are conducted on a single NVIDIA A100 GPU.

	Dataset		Label Words	
Task	Train	Test	Label	Count
Lap14	2313	638	Positive	341
			Negative	128
			Neutral	169
Rest14	3602	1120	Positive	728
			Negative	196
			Neutral	196
M-MMLU	430	3153	Abstract Algebra	100
			Anatomy	135
			astronomy	152
			College Biology	144
			College Chemistry	100
			College Computer Science	100
			College Mathematics	100
			College Physics	102
			Computer Security	100
			Conceptual Physics	235
			Electrical Engineering	145
			Elementary Mathematics	378
			High School Biology	310
			High School Chemistry	203
			High School Computer Science	100
			High School Mathematics	270
			High School Physics	151
			High School Statistics	216
			Machine Learning	112

Table 4: Detailed information on the sizes of the training and test datasets for each task, as well as the sizes of the test datasets for each label within each task. The M-MMLU dataset is available in one language, with identical structures for the other 13 languages.

B SAE Settings

Gemma Scope (Lieberum et al., 2024) offers a diverse set of SAEs for the Gemma model family, with at least two SAEs per layer trained using different hyperparameters. For our experiments, we use the 131k-SAEs labeled as "canonical" unless stated otherwise. All experiments are conducted

on layer 32, which is known to be a good place to work with (Gao et al., 2024).