Evaluation of SAE-based Refusal Features in Base and Chat LLMs: Dataset-specific Characteristics for SAEs

Master Thesis Proposal

Tilman Kerl
Technical University of Vienna
Department of Informatics and Data Science

Supervisors:

Prof. Dr. Peter Knees (Main Supervisor)
Ilya Lasy, MSc (Second Supervisor)

December 18, 2024

1 Problem Statement

A key challenge in safely deploying large language models (LLMs) lies in ensuring that they consistently refuse to respond to unsafe prompts while still engaging constructively with safe ones [2].

Although fine-tuning (FT) on curated datasets has been the predominant solution for achieving refusal behavior in large language models, it often struggles to generalize beyond its training distribution, particularly when exposed to adversarial or multi-turn prompts. Without explicit adversarial training, FT is prone to overfitting, resulting in brittle performance under adversarial conditions and complex multi-turn interactions [27, 30]. Even with adversarial training, models still exhibit notable limitations in maintaining robustness and generalization. Further, achieving reliable refusal behavior often requires reinforcement learning from human feedback (RLHF) to better align models with human preferences and mitigate issues such as hallucination or misalignment [7].

To address these limitations and foster understanding of the workings of these models, recent research has begun exploring methods that intervene on the model's activations at inference time - referred to as feature steering - rather than relying solely on weight updates through fine-tuning [19, 10, 9, 21].

An emerging method for feature steering leverages sparse autoencoders (SAEs) [21, 2] trained on a model's activations. SAEs aim to discover a more interpretable, disentangled representation of the neural computations underlying a model's behaviour. Central to this effort is the superposition hypothesis [11]: large neural networks often pack a vast number

of distinct features into a limited set of neurons by representing them as overlapping vectors in a high-dimensional space. This compression, known as polysemanticity, can cause a single neuron or dimension to encode multiple, conceptually unrelated features [11, 6]. While polysemantic encoding efficiently utilizes available resources, it hinders mechanistic interpretability — making it challenging to isolate the features that mediate specific behaviours, such as refusals.

SAEs provide a potential solution by learning a sparse, over-complete basis for the model's hidden representations. In doing so, they strive to "decompress" this polysemantic tangle into more monosemantic features - i.e., more directly interpretable latent dimensions that correspond to distinct, coherent concepts or behaviours ([6, 2], also refer to [20, 10]).

By training an SAE on the model's activations, we can identify latent features that strongly correlate with refusal behaviour. Once identified, these features can either be manipulated at inference time by "steering" them - clamping or shifting their activation values - to encourage or discourage certain responses without retraining the model [21, 1]. Alternatively, these feature directions can be employed directly in the model by multiplying their values to the respective part of the model [1].

This approach complements existing safety methods, offering the potential to improve generalisation and adapt while leaving the model's underlying parameters untouched.

2 Goals and Expected Outcome

This research aims to evaluate the effectiveness of SAEs by enhancing refusal behaviour in LLMs. By examining the dependency of SAE performance on dataset characteristics, this works seeks to identify key properties such as distributional diversity and the prevalence of refusals that contribute to robust and interpretable feature disentanglement. Additionally, the study will assess the safety and robustness of SAE-informed refusal features against adversarial prompts and unintended over-refusals, ensuring that safety improvements do not compromise core model capabilities. These findings aim to advance the field of mechanistic interpretability, offering insights for developing safer, more transparent AI systems.

3 Research Questions

The following section presents our research questions and explains them.

RQ 1: How does the choice of training dataset (e.g., the original pre-training corpus vs. other refined datasets) affect the sparse autoencoder's ability to isolate and represent refusal features within a language model's latent activations?
The focus is on comparing the impact of the original pre-training data against alternative datasets (e.g., chat-refinement data) in shaping the SAE's capacity to identify refusal-related features from model activations. To properly conduct this evaluation we rely on fully open-source models.

RQ 2: Which characteristics of the underlying training data (e.g., distributional properties, topical diversity, presence of refusals) are most predictive of the strength and clarity of SAE-extracted refusal features?
The training data indirectly shapes the model's latent activations, which are the basis for

SAE-extracted refusal features. Understanding which dataset characteristics—such as distributional properties, topical diversity, or the frequency of refusals—most strongly influence these activations can help identify the conditions under which refusal-related features emerge most clearly and robustly.

RQ 3: How do extracted refusal features of sparse autoencoders trained on activations gathered from a model exposed to refined-datasets compare to those trained on activations gathered from a model exposed to the original pre-training corpus, in terms of robustness, interpretability, downstream performance and controllability of refusal-related features?
The activations gathered from a model reflect its internal representations, which are influenced by the datasets it was exposed to during training. By comparing SAEs trained on activations from models conditioned on refined datasets versus the original pre-training corpus, we can evaluate how the choice of data impacts the robustness, interpretability, and downstream controllability of refusal-related features.

4 Research Methods

Following, we detail our methodological approach. As evaluation is a crucial part in this research please also refer to section 5 for further details.

4.1 Literature Research

A thorough literature review precedes this proposal, encompassing key advancements in SAEs for feature disentanglement, their applications in mechanistic interpretability, and recent findings on steering model behavior via sparse representations. An extensive literature and related work section will be included in the thesis. Please also see section 6.

4.2 Models and Configurations

At the core of our experiences are the Pythia model family [5] which was specifically designed to be analysed and has it's entire code [5, 4, 13] and data [3] open source. Here we look at the different sizes (from 410m - to 2.8b). Firstly, we need to extract which model is the first one to exhibit a refusal behaviour that can properly be evaluated. E.g. pythia 70m often just repeats the prompt or produces gibberish.

Further models we might analyse is the e.g. OLMo model [14] which is also a fully open model .

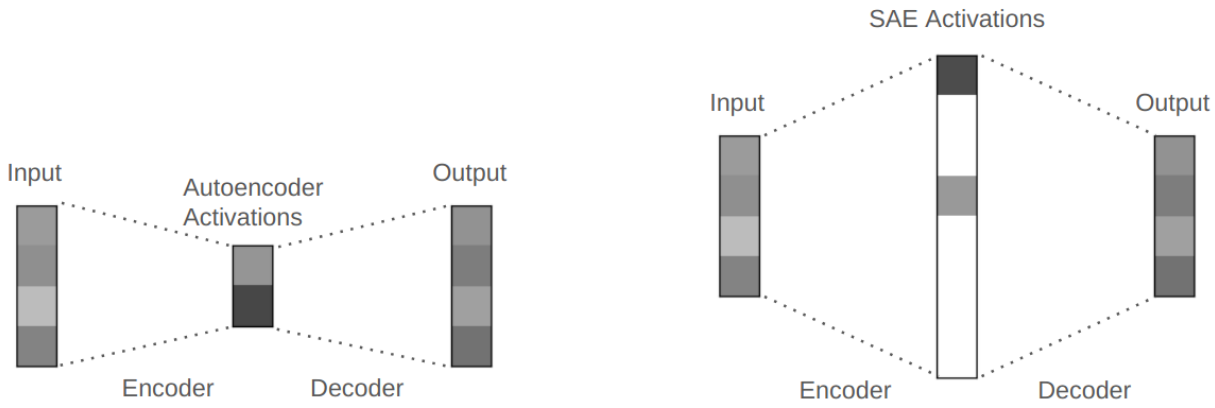
4.3 Training SAEs

SAEs reconstruct neural network activations into a sparse, overcomplete basis, isolating *disentangled features* that address the challenge of polysemantic neurons [11, 20, 22, 2, 19].

SAEs consist of an encoder and decoder, where the encoder maps input activations $x \in \mathbb{R}^d$ to a sparse higher-dimensional representation $h \in \mathbb{R}^n$ (with $n > d$), and the decoder reconstructs x from h. Formally,

$$h = \text{ReLU}(W_{\text{enc}}x + b_{\text{enc}}), \quad x' = W_{\text{dec}}h,$$

where W_{enc}, W_{dec}, and b_{enc} are learned parameters. The training objective minimizes the



(a) Typical Autoencoder Structure

(b) Label distribution of the final dataset

Figure 1: Comparative visualisation of Autoencoders (a) to Sparse Autoencoders (b), taken from Karvonen [20].

reconstruction loss while enforcing sparsity via an ℓ_1-norm penalty:

$$L(x) = \|x - x'\|^2 + \alpha\|h\|_1,$$

where α controls the trade-off between accuracy and sparsity. The resulting sparse features, corresponding to specific linear directions in activation space, disentangle overlapping representations and enable interpretable analysis. These features can be *steered* (amplified or suppressed, for more details refer to subsection 6.2 to analyse and manipulate behaviours (e.g., refusal tendencies in language models). SAEs thus provide an effective tool for isolating monosemantic features and advancing mechanistic interpretability.

Please also refer to Figure 1a and Figure 1b for a comparative visualisation of autoencoders versus sparse-AEs.

4.3.1 Pre-Training SAE

Train SAEs on activations from the base - model (pre-training) dataset.

4.3.2 Post-Training SAE

Train SAEs on activations derived from a post-training dataset.

4.4 Refusal Feature Extraction

Following Ardeiti or Microsoft or Nanda we extract the refusual direction.

5 Evaluation

The evaluation is structured into three distinct phases, following well-established and proven methodologies and metrics (see [9] for reference).

5.1 Phase 1: SAE Quality Metrics

This phase focuses on evaluating the SAEs independently to verify their effectiveness in reconstructing and sparsely encoding model activations.

5.1.1 Reconstruction Error

We measure how accurately each SAE reconstructs the original activation patterns of the language models. Reconstruction quality will be quantified by minimizing the loss between input activations and their SAE-generated reconstructions.

5.1.2 Sparsity

To ensure and evaluate how much the extracted features remain interpretable and disentangled, we calculate the fraction of non-zero activations in the sparse vector representation produced by the SAEs. Sparsity metrics will follow prior work, as described by Arditi et al. [1].

5.2 Phase 2: Refusal

This phase investigates how refusal-related features, extracted via SAEs, influence the models' refusal behavior across safe and unsafe instructions. By enabling or disabling SAE-informed refusal features, we can measure their direct impact on refusal and over-refusal rates, isolating the effect of these features on model behavior. We closely follow the methodology of Kissane et al. [9] and Arditi et al. [1].

5.2.1 Refusal Rate

We evaluate the models' ability to refuse unsafe prompts, measuring refusal frequency using benchmarks like Wild Guard [16] and JailbreakBench [8]. Refusal will be detected by identifying common refusal phrases in model completions.

5.2.2 Over-refusal Rate

To ensure the refusal features do not inadvertently reduce the models' compliance with benign prompts, we measure the over-refusal rate for safe instructions. This evaluation compares model outputs for paired safe and unsafe instructions.

5.3 Phase 3: Robustness and Downstream Performance

In this phase, we assess the broader impact of SAE-informed refusal features on model robustness, reasoning, and general performance across tasks. Comparisons will be conducted across four configurations: the base model, the chat-tuned model, and their respective SAE-enhanced counterparts. We explicitly disable identified refusal features or use neutral feature clamping (e.g., setting activations to zero or a baseline value) to quantify their specific contributions to task performance and refusal robustness. These ablations help evaluate whether refusal features interfere with unrelated capabilities or improve safety alignment.

5.3.1 Massive Multitask Language Understanding (MMLU)

We employ the MMLU benchmark [17] to evaluate the models' overall reasoning and performance across diverse tasks. This analysis focuses on determining how the integration of refusal features impacts capabilities unrelated to refusals.

5.3.2 Jailbreak Robustness

We test model robustness against adversarial prompts designed to bypass refusal mechanisms using benchmarks like JailbreakBench [8]. Multi-turn evaluation setups, similar to those in

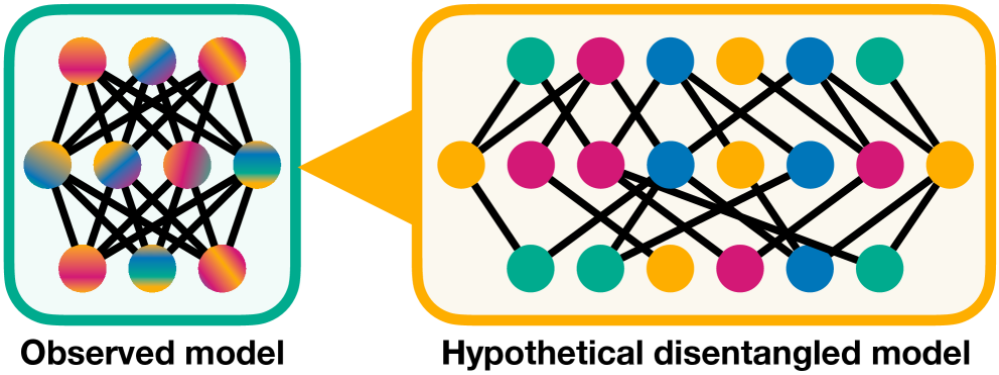


Figure 2: Visual representation of superposition and disentanglement in neural networks, taken from Bereska et al. [2].

[9], will be used to analyze the models' resilience to jailbreak attempts and the effectiveness of SAE-informed refusal features.

5.3.3 Baseline Comparisons

To establish a clear understanding of the effects introduced by the SAEs, we conduct baseline comparisons with standard models (base and chat-tuned) that do not include SAE-informed modifications. This includes ablation studies to isolate and evaluate the specific contributions of the refusal features:

6 State of the Art

This section covers current approaches in the field of mechanistic interpretability related to our research. While not recent, we would like to mention the foundational work of Elhage et al. [12], and Nanda (i.a. [25, 26, 24, 23]). Further, we would explicitly like to mention the recent study from Bereska et al. [2] that reviews the current state and established concepts and frameworks in the field.

6.1 Sparse Autoencoders and Superposition

The concept of sparse autoencoders (SAEs) was introduced by Makhzani et al. in 2013 [22] but their application gained renewed interest following Elhage et al.'s 2022 superposition hypothesis [11] and following work [28, 6].

Theoretical analyses using toy models [11] validate the occurrence of superposition under conditions of feature sparsity and limited neurons, with SAEs successfully recovering disentangled representations in these scenarios. Also scaling this approach to more complex, state-of-the-art models has proven effective, with recent work demonstrating that SAEs can extract interpretable and monosemantic features even in large-scale systems [28].

For a visual representation, please refer to Figure 2. The observed model (left) demonstrates polysemanticity, where individual neurons encode multiple features, leading to overlapping and entangled representations. In contrast, the hypothetical disentangled model (right) illustrates monosemanticity, where each neuron cleanly encodes a single, distinct feature.

Features SAEs

6.2 Steering Large Language Models

Steering generally refers to manipulating the output of LLMs by emplyoying XXXXXX [2]. Various ways to address this have been proposed, and developed [18, 31, 29, 15]. Use cases for steering are manifold - jailbraiking, performance improvement, alignment and safety.

6.3 Refusal of Harmful Content

As already mentioned in section 1, current approaches mostly use some form of adversarial prompt tuning [27, 30] and or in combination with RLHF [7]. Other current research shows xxxx

7 Relevance to the Curriculum

This work tries to explore and uncover the workings of generative decoder transformer models working with natural language. Thus, bringing together various aspects from different courses and core elements of the Data Science and Informatics curriculum. Courses such as `Machine Learning` (188.702), `Natural Language Processing and Information Extraction` (194.093), and `Applied Deep Learning` (194.077) have provided a strong foundation in machine learning techniques, including deep learning architectures central to transformer models. Additionally, `Advanced Information Retrieval` (188.980), which explores state-of-the-art retrieval techniques like BERT-based models, directly supports this research's focus on identifying and refining task-specific circuits within language models.

The course `Research Topics in Natural Language Processing` (194.135) has introduced advanced methodologies specific to NLP, equipping me with essential research skills in e.g. examining model interpretability and ethical behaviour in language models.

Additional expertise in handling complex data systems, as developed and studied in `Advanced Database Systems` (184.780) and `Data-intensive Computing` (194.048), is critical for the large-scale model processing and analysis involved in this work. Furthermore, courses like `Information Visualization` (186.143) and `Visual Data Science` (186.868) provide a solid foundation in visualising complex data, which will be crucial in effectively communicating the results of model circuit analysis and interventions.

8 Overall Review

The conducted literature research establishes a solid foundation for this work by demonstrating the potential of SAEs to disentangle and analyse latent features, particularly for safety-critical behavior such as refusals. Existing studies have highlighted the importance of dataset composition [10] in shaping learned behavior, but the interplay between training data characteristics and SAE performance remains under explored. This proposal will extend these findings by systematically addressing the outlined research questions, examining how dataset choices and properties influence the extraction, robustness, and controllability of refusal features, ensuring a comprehensive understanding and progression of this research area.

References

[1] Andy Arditi, Oscar Obeso, A. S. D. P. N. P. W. G., and Nanda, N. Refusal in language models is mediated by a single direction, 2024.

[2] Bereska, L., and Gavves, E. Mechanistic interpretability for ai safety – a review, 2024.

[3] Biderman, S., Bicheno, K., and Gao, L. Datasheet for the pile. *arXiv preprint arXiv:2201.07311* (2022).

[4] Biderman, S., Prashanth, U. S., Sutawika, L., Schoelkopf, H., Anthony, Q., Purohit, S., and Raff, E. Emergent and predictable memorization in large language models.

[5] Biderman, S., Schoelkopf, H., Anthony, Q. G., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning* (2023), PMLR, pp. 2397–2430.

[6] 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).

[7] Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P., et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217* (2023).

[8] Chao, P., Debenedetti, E., Robey, A., Andriushchenko, M., Croce, F., Sehwag, V., Dobriban, E., Flammarion, N., Pappas, G. J., Tramèr, F., Hassani, H., and Wong, E. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. In *NeurIPS Datasets and Benchmarks Track* (2024).

[9] Connor Kissane, Robert Krzyzanowski, A. C., and Nanda, N. Base llms refuse too. Alignment Forum, 2024.

[10] Connor Kissane, Robert Krzyzanowski, N. N., and Conmy, A. Saes are highly dataset dependent: A case study on the refusal direction. Alignment Forum, 2024.

[11] Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Amodei, D., Wattenberg, M., and Olah, C. Toy models of superposition. *Transformer Circuits Thread* (2022).

[12] 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).

[13] Gao, L., Tow, J., Abbasi, B., Biderman, S., Black, S., DiPofi, A., Foster, C., Golding, L., Hsu, J., Le Noac'h, A., Li, H., McDonell, K., Muennighoff, N., Ociepa, C., Phang, J., Reynolds, L., Schoelkopf, H., Skowron, A., Sutawika, L., Tang, E., Thite, A., Wang, B., Wang, K., and Zou, A. A framework for few-shot language model evaluation, Sept. 2021.

[14] Groeneveld, D., Beltagy, I., Walsh, P., Bhagia, A., Kinney, R., Tafjord, O., Jha, A., Ivison, H., Magnusson, I., Wang, Y., Arora, S., Atkinson, D., Authur, R., Chandu, K. R., Cohan, A., Dumas, J., Elazar, Y., Gu, Y., Hessel, J., Khot, T., Merrill, W., Morrison, J. D., Muennighoff, N., Naik, A., Nam, C., Peters, M. E., Pyatkin, V., Ravichander, A., Schwenk, D., Shah, S., Smith, W., Strubell, E., Subramani, N., Wortsman, M., Dasigi, P., Lambert, N., Richardson, K., Zettlemoyer, L., Dodge, J., Lo, K., Soldaini, L., Smith, N. A., and Hajishirzi, H. Olmo: Accelerating the science of language models. *arXiv preprint* (2024).

[15] Han, C., Xu, J., Li, M., Fung, Y., Sun, C., Jiang, N., Abdelzaher, T., and Ji, H. Word embeddings are steers for language models. 16410–16430.

[16] Han, S., Rao, K., Ettinger, A., Jiang, L., Lin, B. Y., Lambert, N., Choi, Y., and Dziri, N. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms, 2024.

[17] Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300* (2020).

[18] Ilharco, G., Ribeiro, M. T., Wortsman, M., Gururangan, S., Schmidt, L., Hajishirzi, H., and Farhadi, A. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089* (2022).

[19] Joseph Bloom, C. T., and Chanin, D. Saelens. `https://github.com/jbloomAus/SAELens`, 2024.

[20] Karvonen, A. An intuitive explanation of sparse autoencoders for llm interpretability, Jun 2024. Accessed: 2024-06-18.

[21] Kyle O'Brien, David Majercak, X. F. R. E. J. C. H. N. D. C. E. H., and Poursabzi-Sangde, F. Steering language model refusal with sparse autoencoders, 2024.

[22] Makhzani, A., and Frey, B. J. k-sparse autoencoders. *CoRR abs/1312.5663* (2013).

[23] Nanda, N. 200 cop in mi: Analysing training dynamics. LessWrong, 2022.

[24] Nanda, N. 200 cop in mi: Looking for circuits in the wild. Neel Nanda's Blog, 2022.

[25] Nanda, N. A comprehensive mechanistic interpretability explainer & glossary. Neel Nanda's Blog, December 2022.

[26] Nanda, N. A longlist of theories of impact for interpretability. AI Alignment Forum, March 2022.

[27] Raman, M., Maini, P., Kolter, J., Lipton, Z., and Pruthi, D. Model-tuning via prompts makes NLP models adversarially robust. In *Proc. of the Conf. on Empirical Methods in NLP* (Singapore, Dec. 2023), H. Bouamor, J. Pino, and K. Bali, Eds., Association for Computational Linguistics, pp. 9266–9286.

[28] 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).

[29] Todd, E., Li, M. L., Sharma, A. S., Mueller, A., Wallace, B. C., and Bau, D. Function vectors in large language models. *arXiv preprint arXiv:2310.15213* (2023).

[30] Zhong, Z., Chen, T., and Wang, Z. Mat: Mixed-strategy game of adversarial training in fine-tuning, 2023.

[31] Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A., Goel, S., Li, N., Byun, M. J., Wang, Z., Mallen, A., Basart, S., Koyejo, S., Song, D., Fredrikson, M., Kolter, J. Z., and Hendrycks, D. Representation engineering: A top-down approach to AI transparency. *CoRR abs/2310.01405* (2023).