Evaluation of Sparse Autoencoder-based Refusal Features in LLMs: Dataset-dependence study

Master's Thesis Proposal

Tilman Kerl, BSc.
TU Wien Informatics
Research Unit Data Science

Supervisors:

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

MARCH 8, 2025

1 Problem Statement

Large language models (LLMs) play an increasingly prominent role in applications requiring careful consideration of safety and ethical standards. A key challenge in safely deploying LLMs lies in ensuring they consistently refuse to respond to unsafe prompts while engaging constructively with safe ones [5]. Refusal, in this context, refers to the model's ability to identify and reject inputs that could lead to harmful, unethical, or otherwise inappropriate outputs. This behavior is critical for mitigating risks in high-stakes settings where the generation of unsafe or misleading content could have severe consequences.

Existing techniques for promoting refusal behavior predominantly rely on fine-tuning approaches, including supervised fine-tuning, reinforcement learning with human feedback (RLHF), and adversarial training [36, 42, 4, 39, 28, 10].

An emerging alternative involves steering model behavior directly at inference time by intervening on its internal activations. Sparse autoencoders (SAEs) represent a promising approach within this paradigm, as they can extract and manipulate latent features associated with specific behaviors, such as refusals [14, 27]. SAEs enable a disentangled representation of neural activations, isolating refusal-related features in a manner that enhances both interpretability and controllability [16]. Unlike fine-tuning, SAE-based steering methods preserve the model's underlying parameters, making them less resource-intensive and more adaptable to changing requirements.

Despite the potential of SAE-based methods, their effectiveness and robustness are heavily influenced by the characteristics of the datasets used during model training and feature extraction [15]. Pre-training datasets shape the internal representations of language models, which, in turn, determine the clarity and strength of the features identified by SAEs. However, it remains unclear how the choice of training data—such as the original pre-training corpus versus instruction datasets — affects the quality of SAE-extracted refusal features Additionally, little is known about the specific dataset properties that most strongly influence the emergence of disentangled and monosemantic refusal-related features [15, 14].

This research aims to address these gaps by systematically evaluating SAE-based refusal feature extraction across varying training datasets. By analysing the relationship between dataset characteristics and the robustness, interpretability, and controllability of refusal-related features, this work seeks to provide insights into the conditions under which these features can be most effectively isolated and utilized.

2 State of the Art

This section covers current approaches in the field of mechanistic interpretability and others related to our research. While not recent, we would like to mention the foundational work of Elhage et al. [17], and Nanda (i.a. [32, 33, 31, 30]). Further, we would explicitly like to mention the recent study from Bereska et al. [5] that reviews the current state and established concepts and frameworks in the field.

2.1 Refusal of Harmful Content

Fine-tuning on curated datasets remains the dominant approach for promoting safe refusal behavior in LLMs. Supervised fine-tuning reinforces desired outputs by training on labelled data, while RLHF aligns model behavior with human preferences through reward modeling [4, 10]. Adversarial training supplements these methods by exposing models to harmful prompts designed to mimic potential adversarial inputs [28]. Despite these advances, fine-tuning methods are limited in their ability to generalize beyond the training distribution. For instance, jailbreaking attacks reveal persistent vulnerabilities in models not explicitly trained on adversarially crafted prompts [29], and even when adversarial training is applied, the robustness of these models in multi-turn interactions remains inadequate [36, 42].

2.2 Sparse Autoencoders and Superposition

Mechanistic interpretability research provides a complementary perspective by seeking to uncover how neural networks process and represent information. A significant challenge in this domain is the phenomenon of polysemanticity [16, 9], wherein individual neurons encode multiple overlapping features. The superposition hypothesis formalizes this observation, positing that neural networks encode features as overlapping vectors in high-dimensional spaces to maximize representational efficiency (see also Figure 1).

While efficient, this encoding hinders interpretability, making it difficult to isolate and control specific behaviours, such as refusal. Sparse autoencoders (SAEs) offer a solution by learning a disentangled representation of neural computations. The specific form of

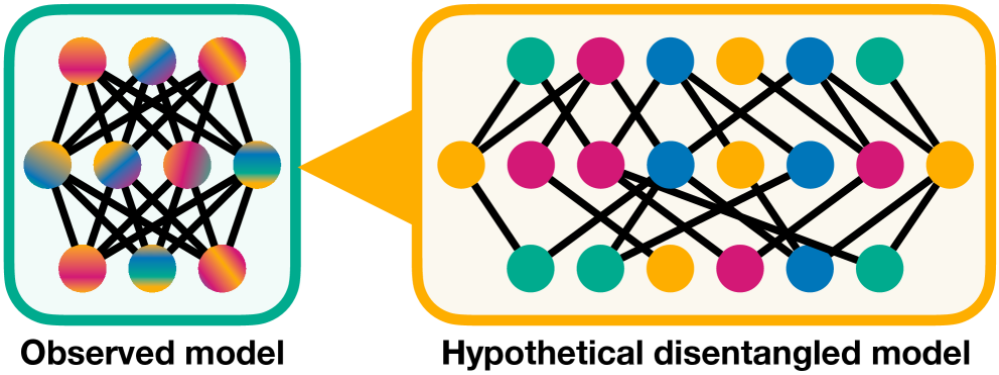


Figure 1: Visual representation of superposition and disentanglement in neural networks, taken from Bereska et al. [5].

SAEs we employ was introduced by Bricken et al. [9], building on a substantial body of prior research [2, 18, 37, 41, 35, 40]. By decomposing polysemantic representations into monosemantic dimensions, SAEs enable individual latent features to correspond to distinct and interpretable concepts [9, 16] (see also Figure 3b). This is achieved by training SAEs on the activations of a model (refer to section 5, and Figure 2), which are derived from the dataset used to train the model itself. As a result, SAEs are sometimes described as being trained on a dataset, even though technically they are trained on the model's activations, which indirectly reflect the dataset.

2.3 Steering Large Language models

Recent advances demonstrate the utility of SAEs for feature steering, a technique that manipulates a model's activations during inference. By training on a model's activations, SAEs can identify latent features associated with specific behaviours, such as refusal, and enable targeted manipulations without retraining the model [27]. These disentangled representations also enhance interpretability, providing a basis for understanding and controlling neural computations [15, 9]. For instance, refusal-related features identified by SAEs can be clamped or shifted during inference, ensuring safer outputs without modifying the underlying parameters [1, 5, 27].

Beyond SAE-based methods, other feature steering approaches have emerged, including task-specific vector manipulation and latent representation adjustment. These techniques exploit the interpretability of high-dimensional representations to guide model outputs toward alignment and safety objectives [24, 43, 38, 21]. Collectively, steering methods demonstrate the flexibility and scalability needed to address challenges that traditional fine-tuning methods face.

2.4 Related Concepts

2.4.1 Interpretability

In the context of SAEs, interpretability implies that individual latent dimensions correspond to distinct features, such as refusal tendencies, and can be understood and controlled explicitly [9].

2.4.2 Robustness

Robustness refers to the ability of extracted features to maintain their effectiveness across varying conditions, including unseen datasets, adversarial inputs, and noisy prompts. In this context, robust SAE features generalise well beyond their training data (as suggested by prior work [13]) and reliably mediate intended behaviours (here: refusals) without degradation.

2.4.3 Dataset Characteristics

Topical diversity. Topical diversity measures the range of subjects, domains, or contexts represented in a dataset. A dataset with high topical diversity includes content spanning various domains (e.g., legal, medical, general knowledge), promoting generalisability in the model's extracted features [3].

Distribution diversity. By distributional diversity we refer to the statistical spread and variability of data within a dataset, including factors like language styles, prompt structures, and frequency distributions of target behaviours (e.g., refusals). Various metrics to compute the diversity in datasets have been proposed and assessed [12].

3 Goals and Expected Outcome

This research aims to evaluate the effectiveness of SAEs for improving 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.

1. **Identification of Dataset Characteristics for Robust SAE Features:** Establish the properties of datasets, such as topical diversity and prevalence of refusal instances, that contribute to robust, interpretable, and effective SAE-extracted features for refusal behaviour.

2. **Evaluation of SAE-based Refusal Features:** Quantify the robustness, interpretability, and generalisability of SAE-extracted refusal features across diverse datasets and adversarial conditions, providing evidence for their practical viability in enhancing model safety.

3. **Impact Assessment of SAE Features on Model Behaviour:** Assess how SAE-informed refusal features affect downstream performance, safety alignment, and adversarial robustness in language models, establishing their utility without degrading unrelated capabilities.

4. **Contribution to Open and Reproducible Research:** Deliver open-source resources, including datasets, metrics, and evaluation frameworks, to support transparency and reproducibility in mechanistic interpretability research.

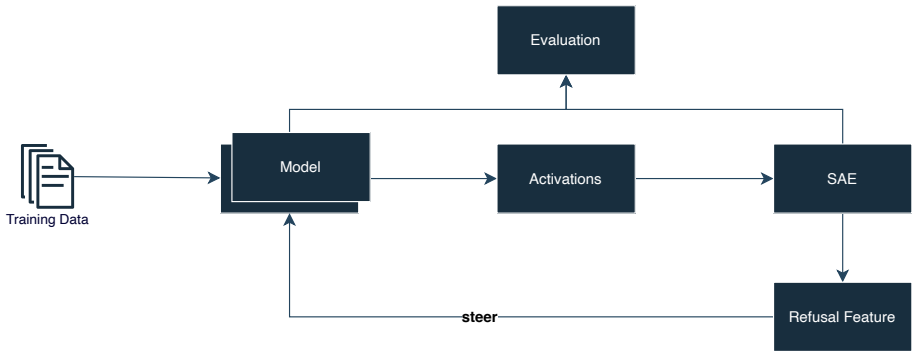


Figure 2: Training and Evaluation flow of models, SAEs and extracted refusal features.

4 Research Questions

The following section presents our research questions and explains them.

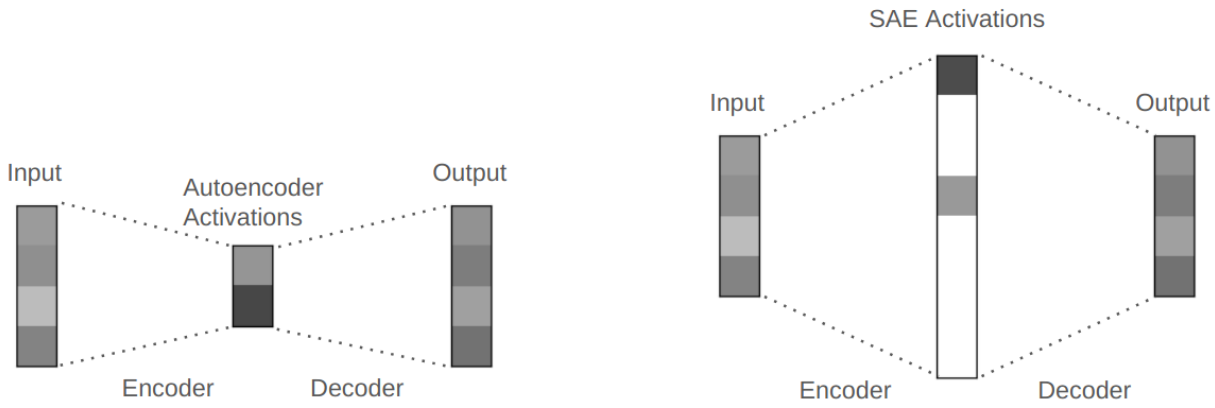
RQ 1: How does the choice of training dataset affect the sparse autoencoder's ability to isolate and represent refusal features within a language model's latent activations?
The focus is on comparing the impact of the original pre-training data against alternative datasets (e.g., instruction data) in 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 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 instruction-datasets compare to those trained on 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 instruction 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.

5 Methods

Following, we detail our methodological approach. Please also refer to Figure 2 for an overview of the connection and flow of the different components. As evaluation is a cru-



(a) Typical Autoencoder Structure

(b) Label distribution of the final dataset

Figure 3: Comparative visualisation of Autoencoders (a) to Sparse Autoencoders (b), taken from Karvonen [26].

cial part in this research please also refer to section 6 for further details.

5.1 Models and Configurations

At the core of our experiences are the Pythia model family [8] which was specifically designed to be analysed and has it's entire code [8, 7, 19] and data [6] 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 [20] which is also a fully open model.

5.2 Training SAEs

SAEs reconstruct neural network activations into a sparse, overcomplete basis, isolating *disentangled features* that address the challenge of polysemantic neurons [16, 26, 5, 25].

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,

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

where W_{enc}, W_{dec}, and b_{enc} are learned parameters. σ denotes the activation function, ReLU is a common choice. The training objective minimizes the 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 2.3 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 3a and Figure 3b for a comparative visualisation of autoencoders versus sparse-AEs.

5.3 Gathering Sparse Autoencoders

As outlined in subsection 5.1, most of the designated models already come with pre-trained, open-source SAEs. Consequently, no additional training is required from our side during the initial stages of the study.

5.4 Dataset Collection

This stage involves two key tasks:

1. bundling, collecting, and organizing the various pre-training datasets associated with the designated models, and

2. identifying and gather the alternative datasets, which serve as the comparison basis.

5.5 Computing Dataset Characteristics

A core aspect of this work relies on extracting relevant characteristics from the datasets. Thus, we utilize established methods, as described in section 2, to compute and extract the dataset properties of interest for our evaluation.

5.6 Refusal Feature Extraction

Following prior work [1, 27, 15, 34], we first analyse SAE feature activations during refusal responses to archetypal unsafe prompts. For instance, prompts such as ”How to build a bomb” serve as a controlled test case to observe which features are activated when the model refuses to generate a harmful response. By examining the activation patterns, we can identify the features most strongly correlated with refusal behavior.

Further, we apply a difference-in-means approach to compare SAE-generated encodings for harmless and harmful prompts. This allows us to determine which features are distinctly associated with refusal by identifying differences in their activation patterns across the two prompt categories. The combination of these methods ensures robust identification of refusal-mediating features.

5.7 Refusal Feature Steering

Once refusal-related features are identified, we can manipulate their activation values to evaluate the model’s ability to produce controlled refusal behavior, as seen in prior work [1, 27, 15, 34]. Specifically, we clamp the activation value of refusal-mediating features to predefined constants. Increasing the activation value of these features will test the model’s capacity to amplify refusal behavior, while decreasing the value will assess the impact on the suppression of refusals.

The clamped sparse vectors are then be passed through the SAE decoder, which reconstructs them into dense vectors. These dense vectors will be reintroduced into the model’s pipeline, enabling us to observe the downstream effects of manipulating specific features.

By systematically adjusting activation values, we can evaluate the extent to which refusal behavior can be reliably and predictably controlled.

6 Evaluation

We assess the performance of the gathered models, SAEs trained on the models activations and models steered by refusal features extracted from the SAEs along the following dimensions. Phase 1 however only applies to the SAEs itself.

The evaluation is structured into three distinct phases, following well-established and proven methodologies and metrics (see [14] for reference).

6.1 Phase 1: SAE Quality Metrics

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

6.1.1 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.

6.1.2 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].

6.2 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. [14] and Arditi et al. [1].

6.2.1 Refusal Rate

We evaluate the models' ability to refuse unsafe prompts, measuring refusal frequency using benchmarks like Wild Guard [22] and JailbreakBench [11]. Refusal will be detected by identifying common refusal phrases in model completions.

6.2.2 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.

6.3 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 respec-

tive 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.

6.3.1 Massive Multitask Language Understanding (MMLU)

We employ the MMLU benchmark [23] 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.

6.3.2 Jailbreak Robustness

We test model robustness against adversarial prompts designed to bypass refusal mechanisms using benchmarks like JailbreakBench [11]. Multi-turn evaluation setups, similar to those in [14], will be used to analyze the models' resilience to jailbreak attempts and the effectiveness of SAE-informed refusal features.

6.3.3 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:

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), `NLP 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 Visualisation` (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 [15] 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] ARORA, S., LI, Y., LIANG, Y., MA, T., AND RISTESKI, A. Linear algebraic structure of word senses, with applications to polysemy. *Transactions of the Association for Computational Linguistics 6* (2018), 483–495.

[3] BACHE, K., NEWMAN, D., AND SMYTH, P. Text-based measures of document diversity. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY, USA, 2013), KDD '13, Association for Computing Machinery, p. 23–31.

[4] BAI, Y., JONES, A., NDOUSSE, K., ASKELL, A., CHEN, A., DASSARMA, N., DRAIN, D., FORT, S., GANGULI, D., HENIGHAN, T., JOSEPH, N., KADAVATH, S., KERNION, J., CONERLY, T., EL-SHOWK, S., ELHAGE, N., HATFIELD-DODDS, Z., HERNANDEZ, D., HUME, T., JOHNSTON, S., KRAVEC, S., LOVITT, L., NANDA, N., OLSSON, C., AMODEI, D., BROWN, T., CLARK, J., MCCANDLISH, S., OLAH, C., MANN, B., AND KAPLAN, J. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022.

[5] BERESKA, L., AND GAVVES, E. Mechanistic interpretability for ai safety – a review, 2024.

[6] BIDERMAN, S., BICHENO, K., AND GAO, L. Datasheet for the pile. *arXiv preprint arXiv:2201.07311* (2022).

[7] BIDERMAN, S., PRASHANTH, U. S., SUTAWIKA, L., SCHOELKOPF, H., ANTHONY, Q., PUROHIT, S., AND RAFF, E. Emergent and predictable memorization in large language models.

[8] 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.

[9] 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).

[10] 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).

[11] 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).

[12] CHEN, H., WAHEED, A., LI, X., WANG, Y., WANG, J., RAJ, B., AND ABDIN, M. I. On the diversity of synthetic data and its impact on training large language models, 2024.

[13] CONNOR KISSAN, ROBERT KRZYZANOWSKI, A. C., AND NANDA, N. Saes (usually) transfer between base and chat models. Alignment Forum, 2024.

[14] CONNOR KISSANE, ROBERT KRZYZANOWSKI, A. C., AND NANDA, N. Base llms refuse too. Alignment Forum, 2024.

[15] CONNOR KISSANE, ROBERT KRZYZANOWSKI, N. N., AND CONMY, A. Saes are highly dataset dependent: A case study on the refusal direction. Alignment Forum, 2024.

[16] 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).

[17] 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).

[18] FARUQUI, M., TSVETKOV, Y., YOGATAMA, D., DYER, C., AND SMITH, N. Sparse overcomplete word vector representations. *arXiv preprint arXiv:1506.02004* (2015).

[19] 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.

[20] 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).

[21] 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.

[22] 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.

[23] 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).

[24] 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).

[25] JOSEPH BLOOM, C. T., AND CHANIN, D. Saelens. `https://github.com/jbloomAus/SAELens`, 2024.

[26] KARVONEN, A. An intuitive explanation of sparse autoencoders for llm interpretability, Jun 2024. Accessed: 18.12.2024.

[27] 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.

[28] MAZEIKA, M., PHAN, L., YIN, X., ZOU, A., WANG, Z., MU, N., SAKHAEE, E., LI, N., BASART, S., LI, B., FORSYTH, D., AND HENDRYCKS, D. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal, 2024.

[29] MOWSHOWITZ, Z. Jailbreaking chatgpt on release day. `https://www.lesswrong.com/posts/RYcoJdvmoBbi5Nax7/jailbreaking-chatgpt-on-release-day`, 2022. Accessed: 18.12.2024.

[30] NANDA, N. 200 cop in mi: Analysing training dynamics. LessWrong, 2022.

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

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

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

[34] NEVERIX, KHARLAPENKO, D., CONMY, A., AND NANDA, N. Sae features for refusal and sycophancy steering vectors, October 2024. Accessed: 18.12.2024.

[35] PANIGRAHI, A., SIMHADRI, H. V., AND BHATTACHARYYA, C. Word2sense: sparse interpretable word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (2019), pp. 5692–5705.

[36] 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.

[37] SUBRAMANIAN, A., PRUTHI, D., JHAMTANI, H., BERG-KIRKPATRICK, T., AND HOVY, E. Spine: Sparse interpretable neural embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2018), vol. 32.

[38] 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).

[39] TOUVRON, H., MARTIN, L., STONE, K., ALBERT, P., ALMAHAIRI, A., BABAEI, Y., BASHLYKOV, N., BATRA, S., BHARGAVA, P., BHOSALE, S., BIKEL, D., BLECHER, L., FERRER, C. C., CHEN, M., CUCURULL, G., ESIOBU, D., FERNANDES, J., FU, J., FU, W., FULLER, B., GAO, C., GOSWAMI, V., GOYAL, N., HARTSHORN, A., HOSSEINI, S., HOU, R., INAN, H., KARDAS, M., KERKEZ, V., KHABSA, M., KLOUMANN, I., KORENEV, A., KOURA, P. S., LACHAUX, M.-A., LAVRIL, T., LEE, J., LISKOVICH, D., LU, Y., MAO, Y., MARTINET, X., MIHAYLOV, T., MISHRA, P., MOLYBOG, I., NIE, Y., POULTON, A., REIZENSTEIN, J., RUNGTA, R., SALADI, K., SCHELTEN, A., SILVA, R., SMITH, E. M., SUBRAMANIAN, R., TAN, X. E., TANG, B., TAYLOR, R., WILLIAMS, A., KUAN, J. X., XU, P., YAN, Z., ZAROV, I., ZHANG, Y., FAN, A., KAMBADUR, M., NARANG, S., RODRIGUEZ, A., STOJNIC, R., EDUNOV, S., AND SCIALOM, T. Llama 2: Open foundation and fine-tuned chat models, 2023.

[40] 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).

[41] ZHANG, J., CHEN, Y., CHEUNG, B., AND OLSHAUSEN, B. A. Word embedding visualization via dictionary learning. *arXiv preprint arXiv:1910.03833* (2019).

[42] ZHONG, Z., CHEN, T., AND WANG, Z. Mat: Mixed-strategy game of adversarial training in fine-tuning, 2023.

[43] 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).