T. Lieberum, S. Rajamanoharan, A. Conmy, L. Smith, N. Sonnerat, V. Varma, J. Kramár, A. Dragan, R. Shah, and N. Nanda. Gemma Scope: Open Sparse Autoencoders Everywhere All At Once on Gemma 2, Aug. 2024. URL http://arxiv.org/abs/2408.05147. arXiv:2408.05147 [cs].

J. Lin and J. Bloom. Announcing Neuronpedia: Platform for accelerating research into Sparse Autoencoders, Mar. 2024. URL https://www.alignmentforum.org/posts/BaEQoxHhWPrkinmxd/announcing-neuronpedia-platform-for-accelerating-research.

J. Lindsey, A. Templeton, J. Marcus, T. Conerly, and J. Batson. Sparse Crosscoders for Cross-Layer Features and Model Diffing, Oct. 2024. URL https://transformer-circuits.pub/2024/crosscoders/index.html.

L. Liu, X. Liu, J. Gao, W. Chen, and J. Han. Understanding the difficulty of training transformers. In B. Webber, T. Cohn, Y. He, and Y. Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5747–5763, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.463. URL https://aclanthology.org/2020.emnlp-main.463/.

A. Makelov. Sparse Autoencoders Match Supervised Features for Model Steering on the IOI Task. June 2024. URL https://openreview.net/forum?id=JdrVuEQih5.

A. Makhzani and B. Frey. k-Sparse Autoencoders, Mar. 2014. URL http://arxiv.org/abs/1312.5663. arXiv:1312.5663 [cs].

S. C. Marshall and J. H. Kirchner. Understanding polysemanticity in neural networks through coding theory, Jan. 2024. URL http://arxiv.org/abs/2401.17975. arXiv:2401.17975 [cs].

K. Meng, D. Bau, A. Andonian, and Y. Belinkov. Locating and Editing Factual Associations in GPT. *Advances in Neural Information Processing Systems*, 35:17359–17372, Dec. 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

A. Mudide, J. Engels, E. J. Michaud, M. Tegmark, and C. S. de Witt. Efficient Dictionary Learning with Switch Sparse Autoencoders. *arXiv preprint arXiv:2410.08201*, 2024. URL https://arxiv.org/abs/2410.08201.

A. Mueller. Missed Causes and Ambiguous Effects: Counterfactuals Pose Challenges for Interpreting Neural Networks, 2024. URL https://arxiv.org/abs/2407.04690.

A. Ng. Sparse autoencoder, 2011. URL https://graphics.stanford.edu/courses/cs233-21-spring/ReferencedPapers/SAE.pdf.

K. O'Brien, D. Majercak, X. Fernandes, R. Edgar, J. Chen, H. Nori, D. Carignan, E. Horvitz, and F. Poursabzi-Sangde. Steering Language Model Refusal with Sparse Autoencoders, Nov. 2024. URL https://arxiv.org/abs/2411.11296.

B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, June 1996. ISSN 1476-4687. doi: 10.1038/381607a0. URL https://www.nature.com/articles/381607a0. Publisher: Nature Publishing Group.

B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37(23):3311–3325, Dec. 1997. ISSN 0042-6989. doi: 10.1016/S0042-6989(97)00169-7. URL https://www.sciencedirect.com/science/article/pii/S0042698997001697.

K. Park, Y. J. Choe, and V. Veitch. The Linear Representation Hypothesis and the Geometry of Large Language Models, Nov. 2023. URL http://arxiv.org/abs/2311.03658. arXiv:2311.03658 [cs, stat].

G. Paulo and N. Belrose. Sparse Autoencoders Trained on the Same Data Learn Different Features, 2025. URL https://arxiv.org/abs/2501.16615.

G. Paulo, A. Mallen, C. Juang, and N. Belrose. Automatically Interpreting Millions of Features in Large Language Models, Oct. 2024.

J. Pennington, R. Socher, and C. D. Manning. GloVe: Global Vectors for Word Representation. In A. Moschitti, B. Pang, and W. Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, Oct. 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162.

S. Rajamanoharan, A. Conmy, L. Smith, T. Lieberum, V. Varma, J. Kramar, R. Shah, and N. Nanda. Improving Sparse Decomposition of Language Model Activations with Gated Sparse Autoencoders. June 2024a. URL https://openreview.net/forum?id=Ppj5KvzU8Q.

S. Rajamanoharan, T. Lieberum, N. Sonnerat, A. Conmy, V. Varma, J. Kramár, and N. Nanda. Jumping Ahead: Improving Reconstruction Fidelity with JumpReLU Sparse Autoencoders, July 2024b. URL http://arxiv.org/abs/2407.14435. arXiv:2407.14435 [cs].

A. Scherlis, K. Sachan, A. S. Jermyn, J. Benton, and B. Shlegeris. Polysemanticity and Capacity in Neural Networks, July 2023. URL http://arxiv.org/abs/2210.01892. arXiv:2210.01892 [cs].

L. Sharkey, D. Braun, and B. Millidge. Taking features out of superposition with sparse autoencoders, Dec. 2022. URL https://www.alignmentforum.org/posts/z6QQJbtpkEAX3Aojj/interim-research-report-taking-features-out-of-superposition.

A. Templeton, T. Conerly, J. Marcus, J. Lindsey, T. Bricken, B. Chen, A. Pearce, C. Citro, E. Ameisen, A. Jones, H. Cunningham, N. L. Turner, C. McDougall, M. MacDiarmid, A. Tamkin, E. Durmus, T. Hume, F. Mosconi, C. D. Freeman, T. R. Sumers, E. Rees, J. Batson, A. Jermyn, S. Carter, C. Olah, and T. Henighan. Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet, May 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

M. Wattenberg and F. Viégas. Relational Composition in Neural Networks: A Survey and Call to Action. June 2024. URL https://openreview.net/forum?id=zzCEiUIPk9.

M. Weber, D. Fu, Q. Anthony, Y. Oren, S. Adams, A. Alexandrov, X. Lyu, H. Nguyen, X. Yao, V. Adams, B. Athiwaratkun, R. Chalamala, K. Chen, M. Ryabinin, T. Dao, P. Liang, C. Ré, I. Rish, and C. Zhang. RedPajama: an Open Dataset for Training Large Language Models, 2024. URL https://arxiv.org/abs/2411.12372.

Z. Yun, Y. Chen, B. Olshausen, and Y. LeCun. Transformer visualization via dictionary learning: contextualized embedding as a linear superposition of transformer factors. In E. Agirre, M. Apidianaki, and I. Vulić, editors, *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 1–10, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.deelio-1.1. URL https://aclanthology.org/2021.deelio-1.1.

F. Zhang and N. Nanda. Towards Best Practices of Activation Patching in Language Models: Metrics and Methods. In *The Twelfth International Conference on Learning Representations*, Oct. 2023. URL https://openreview.net/forum?id=Hf17y6u9BC.

Z. Zhong and J. Andreas. Algorithmic Capabilities of Random Transformers, 2024. URL https://arxiv.org/abs/2410.04368.

A Broader Impact

This work investigates a method currently used for mechanistic interpretability of LLMs, yielding results that challenge certain assumptions about sparse autoencoders. By demonstrating that SAEs can produce similar aggregate auto-interpretability scores for both random and trained transformers, our findings raise important questions about what these SAE evaluation methods are actually capturing.

By better understanding the metrics of SAE quality, we hope that this work will contribute to a more informed search of better SAE-like methods and thus help to make these models more interpretable and to mitigate the potential harm these models could cause. Since our work is an empirical study of the capabilities of a presently used method, and it shows that the method provides interpretation of both random and trained transformers, we think the risk that this work could lead to negative social impact is minimal.

B Auto-interpretability ROC curves

Figures 6, 8, 12 show the similarity between 'fuzzing' AUROC for the trained and randomized SAEs for the 70M, 160M, and 1B models. Figures 7, 9, 13, show the similarity between 'detection' AUROC for the trained and randomized SAEs for the 70M, 160M, and 1B models.

B.1 Pythia 70m

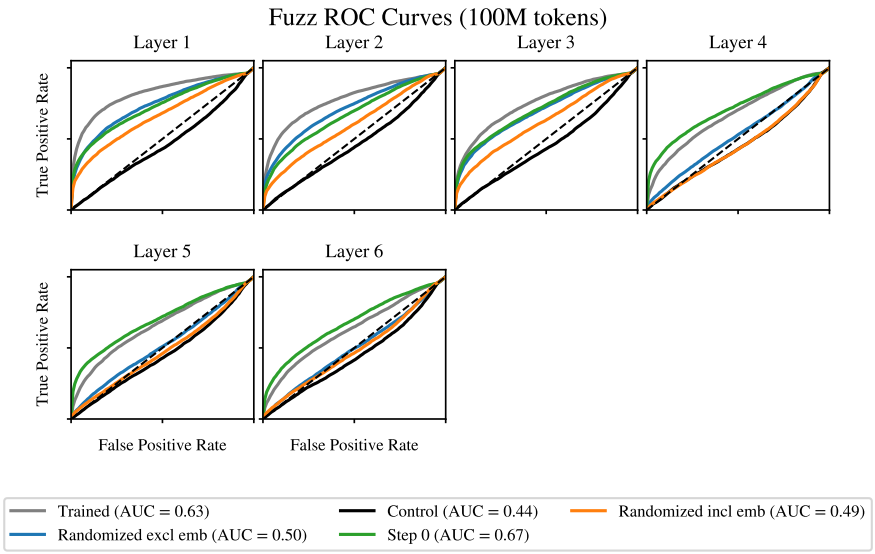

Figure 6: ROC curves for 'fuzzing' auto-interpretability for Pythia-70m over 100 SAE latents. These results demonstrate the similarity in performance between the SAE variants, as well as the overall degradation in performance as the layer index increases.