the importance of benchmarking interpretability techniques against strong, appropriately constructed null models, such as the randomly initialized transformers used here. Without such baselines, it is difficult to confidently attribute discovered features to the process of learning.

REFERENCES

J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim. Sanity Checks for Saliency Maps, 2020. URL https://arxiv.org/abs/1810.03292.

A. J. Bell and T. J. Sejnowski. An Information-Maximization Approach to Blind Separation and Blind Deconvolution. *Neural Computation*, 7(6):1129–1159, Nov. 1995. ISSN 0899-7667. doi: 10.1162/neco.1995.7.6.1129. URL https://ieeexplore.ieee.org/abstract/document/6796129. Conference Name: Neural Computation.

N. Belrose, L. Quirke, A. Garriga-Alonso, neverix, HongchuanZeng, A. Duong, lewington, P. Minervini, T. Vogel, T. Fel, Yang, avgalichin, and T. V. Browne. Eleutherai/sparsify. https://github.com/EleutherAI/sparsify, sep 22 2025. URL https://github.com/EleutherAI/sparsify.

S. Biderman, H. Schoelkopf, Q. G. Anthony, H. Bradley, K. O'Brien, E. Hallahan, M. A. Khan, S. Purohit, U. S. Prashanth, E. Raff, A. Skowron, L. Sutawika, and O. V. D. Wal. Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling. In *Proceedings of the 40th International Conference on Machine Learning*, pages 2397–2430. PMLR, July 2023. URL https://proceedings.mlr.press/v202/biderman23a.html. ISSN: 2640-3498.

S. Bills, N. Cammarata, D. Mossing, H. Tillman, L. Gao, G. Goh, I. Sutskever, J. Leike, J. Wu, and W. Saunders. Language models can explain neurons in language models, May 2023. URL https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html.

D. Braun, J. Taylor, N. Goldowsky-Dill, and L. Sharkey. Identifying Functionally Important Features with End-to-End Sparse Dictionary Learning, May 2024. URL http://arxiv.org/abs/2405.12241. arXiv:2405.12241 [cs].

T. Bricken, A. Templeton, J. Batson, B. Chen, A. Jermyn, T. Conerly, N. Turner, C. Anil, C. Denison, and A. Askell. Towards Monosemanticity: Decomposing Language Models With Dictionary Learning, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features.

Caden, L. Quirke, G. Paulo, A. Mallen, neverix, N. Belrose, johnny, A. Sumer, A. Duong, W. Li, S. Biderman, L. Farnik, I. E. Ashimine, A. Muhamed, D. Ghilardi, F. Xiao, and M. M. Rahman. Eleutherai/delphi. https://github.com/EleutherAI/delphi, sep 22 2025. URL https://github.com/EleutherAI/delphi.

L. Chan. Superposition is not "just"' neuron polysemanticity. Apr. 2024. URL https://www.alignmentforum.org/posts/8EyCQKuWo6swZpagS/superposition-is-not-just-neuron-polysemanticity.

M. Chaudhary and A. Geiger. Evaluating Open-Source Sparse Autoencoders on Disentangling Factual Knowledge in GPT-2 Small, Sept. 2024. URL http://arxiv.org/abs/2409.04478. arXiv:2409.04478 [cs].

D. Choi, V. Huang, K. Meng, D. D. Johnson, J. Steinhardt, and S. Schwettmann. Scaling Automatic Neuron Description, Oct. 2024. URL https://transluce.org/neuron-descriptions.

H. Cunningham, A. Ewart, L. Riggs, R. Huben, and L. Sharkey. Sparse Autoencoders Find Highly Interpretable Features in Language Models, Oct. 2023. URL http://arxiv.org/abs/2309.08600. arXiv:2309.08600 [cs].

T. Dooms and D. Wilhelm. Tokenized SAEs: Disentangling SAE Reconstructions. June 2024. URL https://openreview.net/forum?id=5Eas7HCe38.

N. Elhage, T. Hume, C. Olsson, N. Schiefer, T. Henighan, S. Kravec, Z. Hatfield-Dodds, R. Lasenby, D. Drain, C. Chen, R. Grosse, S. McCandlish, J. Kaplan, D. Amodei, M. Wattenberg, and C. Olah. Toy Models of Superposition, Sept. 2022. URL http://arxiv.org/abs/2209.10652. arXiv:2209.10652 [cs].

J. Engels, I. Liao, E. J. Michaud, W. Gurnee, and M. Tegmark. Not All Language Model Features Are Linear, May 2024. URL http://arxiv.org/abs/2405.14860. arXiv:2405.14860 [cs].

L. Farnik, T. Lawson, C. Houghton, and L. Aitchison. Jacobian Sparse Autoencoders: Sparsify Computations, Not Just Activations. In *Forty-second International Conference on Machine Learning*, June 2025. URL https://openreview.net/forum?id=TPuFRuNano.

A. Foote, N. Nanda, E. Kran, I. Konstas, and F. Barez. N2G: A Scalable Approach for Quantifying Interpretable Neuron Representations in Large Language Models, Apr. 2023. URL http://arxiv.org/abs/2304.12918. arXiv:2304.12918 [cs].

L. Gao, T. D. la Tour, H. Tillman, G. Goh, R. Troll, A. Radford, I. Sutskever, J. Leike, and J. Wu. Scaling and evaluating sparse autoencoders, June 2024. URL http://arxiv.org/abs/2406.04093. arXiv:2406.04093 [cs].

D. Ghilardi, F. Belotti, M. Molinari, T. Ma, and M. Palmonari. Group-SAE: Efficient Training of Sparse Autoencoders for Large Language Models via Layer Groups, Sept. 2025.

W. Gurnee, N. Nanda, M. Pauly, K. Harvey, D. Troitskii, and D. Bertsimas. Finding Neurons in a Haystack: Case Studies with Sparse Probing, June 2023. URL http://arxiv.org/abs/2305.01610. arXiv:2305.01610 [cs].

S. Heimersheim and N. Nanda. How to use and interpret activation patching, Apr. 2024. URL http://arxiv.org/abs/2404.15255. arXiv:2404.15255 [cs].

J. Huang, Z. Wu, C. Potts, M. Geva, and A. Geiger. RAVEL: Evaluating Interpretability Methods on Disentangling Language Model Representations, Aug. 2024. URL http://arxiv.org/abs/2402.17700. arXiv:2402.17700 [cs].

A. Hyvärinen and E. Oja. Independent component analysis: algorithms and applications. *Neural Networks*, 13(4):411–430, June 2000. ISSN 0893-6080. doi: 10.1016/S0893-6080(00)00026-5. URL https://www.sciencedirect.com/science/article/pii/S0893608000000265.

E. T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge, UK, 2003. ISBN 0521592712. The maximum entropy property of the Gaussian distribution is discussed on pages 208 and 365, in chapters 7 and 11.

A. Karvonen, C. Rager, J. Lin, C. Tigges, J. Bloom, D. Chanin, Y.-T. Lau, E. Farrell, A. Conmy, C. McDougall, K. Ayonrinde, M. Wearden, S. Marks, and N. Nanda. SAEBench: A Comprehensive Benchmark for Sparse Autoencoders, Dec. 2024a. URL https://www.neuronpedia.org/sae-bench/info.

A. Karvonen, C. Rager, S. Marks, and N. Nanda. Evaluating Sparse Autoencoders on Targeted Concept Erasure Tasks, Nov. 2024b. URL https://arxiv.org/abs/2411.18895.

A. Karvonen, B. Wright, C. Rager, R. Angell, J. Brinkmann, L. Smith, C. M. Verdun, D. Bau, and S. Marks. Measuring Progress in Dictionary Learning for Language Model Interpretability with Board Game Models. *arXiv preprint arXiv:2408.00113*, 2024c. URL https://arxiv.org/abs/2408.00113.

C. Kissane, R. Krzyzanowski, J. I. Bloom, A. Conmy, and N. Nanda. Interpreting Attention Layer Outputs with Sparse Autoencoders. June 2024. URL https://openreview.net/forum?id=fewUBDwjji.

T. Lawson, L. Farnik, C. Houghton, and L. Aitchison. Residual Stream Analysis with Multi-Layer SAEs. In *The Thirteenth International Conference on Learning Representations*, Oct. 2024. URL https://openreview.net/forum?id=XAjfjizaKs.

V. Lecomte, K. Thaman, R. Schaeffer, N. Bashkansky, T. Chow, and S. Koyejo. What Causes Polysemanticity? An Alternative Origin Story of Mixed Selectivity from Incidental Causes, Feb. 2024. URL http://arxiv.org/abs/2312.03096. arXiv:2312.03096 [cs].

H. Lee, A. Battle, R. Raina, and A. Ng. Efficient sparse coding algorithms. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006. URL https://proceedings.neurips.cc/paper_files/paper/2006/hash/2d71b2ae158c7c5912cc0bbde2bb9d95-Abstract.html.