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

Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. Transcoders find interpretable llm feature circuits. *arXiv preprint arXiv:2406.11944*, 2024.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

Joshua Engels, Isaac Liao, Eric J Michaud, Wes Gurnee, and Max Tegmark. Not all language model features are linear. *arXiv preprint arXiv:2405.14860*, 2024.

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

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. *arXiv preprint arXiv:2304.14767*, 2023.

Wes Gurnee and Max Tegmark. Language models represent space and time. *arXiv preprint arXiv:2310.02207*, 2023.

Connor Kissane, Robert Krzyzanowski, Joseph Isaac Bloom, Arthur Conmy, and Neel Nanda. Interpreting attention layer outputs with sparse autoencoders. *arXiv preprint arXiv:2406.17759*, 2024.

Honglak Lee, Chaitanya Ekanadham, and Andrew Ng. Sparse deep belief net model for visual area v2. *Advances in neural information processing systems*, 20, 2007.

Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 991–999, 2015.

Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. *arXiv preprint arXiv:2210.13382*, 2022.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2, 2024. URL `https://arxiv.org/abs/2408.05147`.

Johnny Lin. Neuronpedia: Interactive reference and tooling for analyzing neural networks, 2023. URL `https://www.neuronpedia.org`. Software available from neuronpedia.org.

Aleksandar Makelov, George Lange, and Neel Nanda. Towards principled evaluations of sparse autoencoders for interpretability and control. *arXiv preprint arXiv:2405.08366*, 2024.

Alireza Makhzani and Brendan Frey. k-sparse autoencoders, 2014. URL `https://arxiv.org/abs/1312.5663`.

Samuel Marks, Can Rager, Eric J Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. *arXiv preprint arXiv:2403.19647*, 2024.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372, 2022.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*, 2023.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001, 2020.

Bruno A Olshausen and David J Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, 37(23):3311–3325, 1997.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Tom Lieberum, Vikrant Varma, János Kramár, Rohin Shah, and Neel Nanda. Improving dictionary learning with gated sparse autoencoders. *arXiv preprint arXiv:2404.16014*, 2024a.

Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. Jumping ahead: Improving reconstruction fidelity with jumprelu sparse autoencoders. *arXiv preprint arXiv:2407.14435*, 2024b.

Lewis Smith. The 'strong' feature hypothesis could be wrong. `https://www.alignmentforum.org/posts/tojtPCCRpKLSHBdpn/the-strong-feature-hypothesis-could-be-wrong`, [Accessed 23-09-2024], 2024.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Adly Templeton. *Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet*. Anthropic, 2024.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*, 2022.

Martin Wattenberg and Fernanda B Viégas. Relational composition in neural networks: A survey and call to action. *arXiv preprint arXiv:2407.14662*, 2024.

A Appendix / supplemental material

A.1 Glossary of Terms

Active Latents (L0): For an input x and SAE activation function $f(x)$, the number of non-zero elements in $f(x)$. Typically measured as average L0 across a batch: $L0 = \frac{1}{n}\sum_i \|f(x_i)\|_0$.

Canonical Unit: Hypothetical, fundamental building blocks of a LLMs computation that are unique, complete, and atomic.

Cross-Entropy Degradation: The increase in cross-entropy loss when replacing the model activations with the reconstruction of the SAE.

Decoder Directions: The columns of the decoder matrix W^{dec} that map from latent to input space. Two decoder directions with high cosine similarity suggest related features.

Dictionary Size: The dimensionality of the latent space in an SAE, determining the maximum number of unique features that can be learned.

Feature Splitting: A phenomenon where a broad latent learned by a smaller SAE splits into more fine-grained latents in a larger SAE.

Latent: The encoder-decoder pair corresponding to single element in the SAE's dictionary, i.e. a learned feature of the SAE rather than a feature of the data.

Mechanistic Interpretability: The study of reverse-engineering neural networks into interpretable algorithms, focusing on identifying and understanding computational features and circuits.

Meta-latents: Features learned by a meta-SAE when trained on the decoder directions of another SAE.

Monosemantic: Property of a feature that responds selectively to a single coherent concept. Contrasts with polysemantic features.

Novel Latents: Features in a larger SAE with maximum cosine similarity below threshold θ to any feature in a smaller SAE, indicating capture of previously unrepresented information.

Polysemanticity: The phenomenon where individual neurons or features respond to multiple unrelated concepts.

Reconstruction Latents: Features in a larger SAE with maximum cosine similarity above threshold θ to features in a smaller SAE, representing refined or specialized versions of existing features.

Residual Stream: In the context of transformer architectures, the residual stream refers to the main information flow that bypasses the self-attention and feed-forward layers through residual connections.

SAE Stitching: A technique for analyzing feature relationships across SAEs of different sizes by systematically transferring latents based on decoder similarity.

Sparsity Coefficient (λ): Hyperparameter in the loss function of some SAE architectures $L = \|x - \hat{x}\|^2 + \lambda S(f(x))$ controlling the trade-off between reconstruction accuracy and activation sparsity.

TopK: A sparsification approach that maintains exactly k non-zero activations per input by zeroing all but the k largest values: $\text{TopK}(x)_i = x_i$ if x_i is among k largest elements, 0 otherwise.

A.2 SAE Variants

ReLU SAEs (Bricken et al., 2023) use the L1-norm $S(\boldsymbol{f}) := ||\boldsymbol{f}||_1$ as an approximation to the L0-norm for the sparsity penalty. This provides a gradient for training unlike the L0-norm, but suppresses latent activations harming reconstruction performance (Rajamanoharan et al., 2024a). Furthermore, the L1 penalty can be arbitrarily reduced through reparameterization by scaling the decoder parameters, which is resolved in Bricken et al. (2023) by constraining the decoder directions to the unit norm. Resolving this tension between activation sparsity and value is the motivation behind more recent architecture variants.

TopK SAEs (Gao et al., 2024; Makhzani & Frey, 2014) enforce sparsity by retaining only the top k activations per sample. The encoder is defined as:

$$\boldsymbol{f}(\boldsymbol{x}) := \text{TopK}(\mathbf{W}^{\text{enc}}\mathbf{x} + \mathbf{b}^{\text{enc}}) \tag{7}$$

where TopK zeroes out all but the k largest activations in each sample. This approach eliminates the need for an explicit sparsity penalty but imposes a rigid constraint on the number of active latents