significant performance degradation, both for inference and training?

Acknowledgements

We are incredibly grateful to Joseph Bloom, Johnny Lin and Curt Tigges for their help creating an interactive demo of Gemma Scope on Neuronpedia (Lin and Bloom, 2023), creating tooling for researchers like feature dashboards, and help making educational materials. We are grateful to Alex Tomala for engineering support and Tobi Ijitoye for organizational support during this project. Additionally, we would like to thank Meg Risdal, Kathleen Kenealy, Joe Fernandez, Kat Black and Tris Warkentin for support with integration with Gemma, and Omar Sanseviero, Joshua Lochner and Lucain Pouget for help with integration into HuggingFace. We also thank beta testers Javier Ferrando, Oscar Balcells Obeso and others for additional feedback. We are grateful for help and contributions from Phoebe Kirk, Andrew Forbes, Arielle Bier, Aliya Ahmad, Yotam Doron, Ludovic Peran, Anand Rao, Samuel Albanie, Dave Orr, Matt Miller, Alex Turner, Shruti Sheth, Jeremy Sie and Glenn Cameron.

Author contributions

Tom Lieberum (TL) led the writing of the report, and implementation and running of evaluations. TL also led optimization of SAE training code and fast distributed data loading with significant contributions from Vikrant Varma (VV) and Lewis Smith (LS). Senthooran Rajamanoharan (SR) developed the JumpReLU architecture, led SAE training and significantly contributed to writing and editing the report. SAEs were trained using a codebase that was designed and implemented by TL and VV with significant contributions from Arthur Conmy (AC), which in turn relies on an interpretability codebase written in large part by János Kramár (JK). JK also wrote Mishax, a python library that was used to seamlessly adapt our codebase to the newest Gemma models, which was open-sourced with contribution from Nicolas Sonnerat (NS). AC led the early access and open sourcing of code and weights with significant contribution from LS, in addition to training and evaluating the transcoders and IT SAEs with significant contribution from SR. LS wrote the Gemma Scope tutorial. Neel Nanda (NN) wrote the list of open problems in Section 5 and led coordination with the various stakeholders required to make the launch possible. Anca Dragan (AD), Rohin Shah (RS) and NN provided leadership and advice throughout the project and edited the report.

References

E. Anders, C. Neo, J. Hoelscher-Obermaier, and J. N. Howard. Sparse autoencoders find composed features in small toy models. LessWrong, 2024.

Y. Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, Mar. 2022. doi: 10.1162/coli_a_00422. URL https://aclanthology.org/2022.cl-1.7.

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. https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html, 2023.

T. Bricken, A. Templeton, J. Batson, B. Chen, A. Jermyn, T. Conerly, N. Turner, C. Anil, C. Denison, A. Askell, R. Lasenby, Y. Wu, S. Kravec, N. Schiefer, T. Maxwell, N. Joseph, Z. Hatfield-Dodds, A. Tamkin, K. Nguyen, B. McLean, J. E. Burke, T. Hume, S. Carter, T. Henighan, and C. Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features/index.html.

T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess,

J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners, 2020. URL https://arxiv.org/abs/2005.14165.

B. Bussmann, P. Leask, J. Bloom, C. Tigges, and N. Nanda. Stitching saes of different sizes, July 2024. URL https://www.alignmentforum.org/posts/baJyjpktzmcmRfosq/stitching-saes-of-different-sizes.

T. Conerly, A. Templeton, T. Bricken, J. Marcus, and T. Henighan. Update on how we train SAEs. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/april-update/index.html#training-saes.

A. Conmy and N. Nanda. Activation steering with SAEs. *Alignment Forum*, 2024. URL https://www.alignmentforum.org/posts/C5KAZQib3bzzpeyrg/progress-update-1. Progress Update #1 from the GDM Mech Interp Team.

A. Conmy, A. N. Mavor-Parker, A. Lynch, S. Heimersheim, and A. Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability, 2023. URL https://arxiv.org/abs/2304.14997.

H. Cunningham and T. Conerly. Circuits Updates - June 2024: Comparing TopK and Gated SAEs to Standard SAEs. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/june-update/index.html#topk-gated-comparison.

H. Cunningham, A. Ewart, L. Riggs, R. Huben, and L. Sharkey. Sparse autoencoders find highly interpretable features in language models, 2023.

J. Dunefsky, P. Chlenski, and N. Nanda. Transcoders find interpretable llm feature circuits, 2024. URL https://arxiv.org/abs/2406.11944.

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. *Transformer Circuits Thread*, 2022. https://transformer-circuits.pub/2022/toy_model/index.html.

J. Engels, I. Liao, E. J. Michaud, W. Gurnee, and M. Tegmark. Not all language model features are linear, 2024. URL https://arxiv.org/abs/2405.14860.

L. Gao, S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. Nabeshima, S. Presser, and C. Leahy. The pile: An 800gb dataset of diverse text for language modeling, 2020. URL https://arxiv.org/abs/2101.00027.

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, 2024. URL https://arxiv.org/abs/2406.04093.

Gemini Team. Gemini: A family of highly capable multimodal models, 2024. URL https://arxiv.org/abs/2312.11805.

Gemma Team. Gemma: Open models based on gemini research and technology, 2024a. URL https://arxiv.org/abs/2403.08295.

Gemma Team. Gemma 2: Improving open language models at a practical size, 2024b. URL https://storage.googleapis.com/deepmind-media/gemma/gemma-2-report.pdf.

W. Gurnee, N. Nanda, M. Pauly, K. Harvey, D. Troitskii, and D. Bertsimas. Finding neurons in a haystack: Case studies with sparse probing. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=JYs1R9IMJr.

W. Gurnee, T. Horsley, Z. C. Guo, T. R. Kheirkhah, Q. Sun, W. Hathaway, N. Nanda, and D. Bertsimas. Universal neurons in gpt2 language models, 2024. URL https://arxiv.org/abs/2401.12181.

M. Hanna, O. Liu, and A. Variengien. How does gpt-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model, 2023. URL https://arxiv.org/abs/2305.00586.

K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, 2015. URL https://arxiv.org/abs/1502.01852.

R. Hendel, M. Geva, and A. Globerson. In-context learning creates task vectors. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://openreview.net/forum?id=QYvFUlF19n.

J. Huang, A. Geiger, K. D'Oosterlinck, Z. Wu, and C. Potts. Rigorously assessing natural language explanations of neurons, 2023. URL https://arxiv.org/abs/2309.10312.

E. Hubinger. A transparency and interpretability tech tree. *Alignment Forum*, 2022.

S. Jain, E. S. Lubana, K. Oksuz, T. Joy, P. H. S. Torr, A. Sanyal, and P. K. Dokania. What makes and breaks safety fine-tuning? a mechanistic study, 2024. URL https://arxiv.org/abs/2407.10264.

A. Jermyn, C. Olah, and T. Henighan. Attention head superposition, May 2023. URL https://transformer-circuits.pub/2023/may-update/index.html#attention-superposition.

A. Karvonen, B. Wright, C. Rager, R. Angell, J. Brinkmann, L. R. Smith, C. M. Verdun, D. Bau, and S. Marks. Measuring progress in dictionary learning for language model interpretability with board game models. In *ICML 2024 Workshop on Mechanistic Interpretability*, 2024. URL https://openreview.net/forum?id=qzsDKwGJyB.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization, 2017. URL https://arxiv.org/abs/1412.6980.

C. Kissane, R. Krzyzanowski, J. I. Bloom, A. Conmy, and N. Nanda. Interpreting attention layer outputs with sparse autoencoders, 2024a. URL https://arxiv.org/abs/2406.17759.

C. Kissane, R. Krzyzanowski, A. Conmy, and N. Nanda. Saes (usually) transfer between base and chat models, 2024b. URL https://www.alignmentforum.org/posts/fmwk6qxrpW8d4jvbd/saes-usually-transfer.

P. Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand, Sept. 13-15 2005. URL https://aclanthology.org/2005.mtsummit-papers.11.

K. Li, O. Patel, F. Viégas, H. Pfister, and M. Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=aLLuYpn83y.

J. Lin and J. Bloom. Analyzing neural networks with dictionary learning, 2023. URL https://www.neuronpedia.org. Software available from neuronpedia.org.

A. Makelov, G. Lange, and N. Nanda. Towards principled evaluations of sparse autoencoders for interpretability and control. In *ICLR 2024 Workshop on Secure and Trustworthy Large Language Models*, 2024. URL https://openreview.net/forum?id=MHIX9H8aYF.

S. Marks, C. Rager, E. J. Michaud, Y. Belinkov, D. Bau, and A. Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models, 2024.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space, 2013. URL https://arxiv.org/abs/1301.3781.

N. Nanda. A longlist of theories of impact for interpretability. *Alignment Forum*, 2022.