[1] Marc Canby*, **Adam Davies***, Chirag Rastogi, and Julia Hockenmaier. Measuring the reliability of causal probing methods: Tradeoffs, limitations, and the plight of nullifying interventions. In *NeurIPS 2024 Workshop on Interpretable AI*, 2024. URL https://openreview.net/forum?id=tmpMQLxVHh.

[2] Marc E. Canby*, **Adam Davies***, Chirag Rastogi, and Julia Hockenmaier. How reliable are causal probing interventions? In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 857–878, Mumbai, India, December 2025. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics. ISBN 979-8-89176-298-5. URL https://aclanthology.org/2025.ijcnlp-long.47/.

[3] **Adam Davies** and Ashkan Khakzar. The cognitive revolution in interpretability: From explaining behavior to interpreting representations and algorithms. *arXiv preprint arXiv:2408.05859*, 2024. URL https://arxiv.org/abs/2408.05859.

[4] **Adam Davies**, Jize Jiang, and ChengXiang Zhai. Competence-based analysis of language models. In *NeurIPS 2024 Workshop on Interpretable AI*, 2024. URL https://openreview.net/forum?id=x6ZM5Is2Po.

[5] **Adam Davies**, Elisa Nguyen, Michael Simeone, Erik Johnston, and Martin Gubri. Social science is necessary for operationalizing socially responsible foundation models. In *ICLR 2025 Workshop on Human-AI Coevolution*, 2025. URL https://openreview.net/forum?id=zbB2vjAq7X.

[6] Arshia Hemmat, **Adam Davies**, Tom A. Lamb, Jianhao Yuan, Philip Torr, Ashkan Khakzar, and Francesco Pinto. Hidden in plain sight: Evaluating abstract shape recognition in vision-language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 88527–88556. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/a13ff984831deea39e6132bafdfdd6d5-Paper-Datasets_and_Benchmarks_Track.pdf.

[7] Tom A. Lamb, **Adam Davies**, Alasdair Paren, Philip Torr, and Francesco Pinto. Focus on this, not that! steering LLMs with adaptive feature specification. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=rbI5mOUA8Z.

[8] Sewoong Lee, **Adam Davies**, Marc E. Canby, and Julia Hockenmaier. Evaluating and designing sparse autoencoders by approximating quasi-orthogonality. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=XhdNFeMclS.

[9] Amogh Mannekote, **Adam Davies**, Juan D Pinto, Shan Zhang, Daniel Olds, Noah L Schroeder, Blair Lehman, Diego Zapata-Rivera, and ChengXiang Zhai. Large language

models for whole-learner support: opportunities and challenges. *Frontiers in Artificial Intelligence*, 7:1460364, 2024. URL https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2024.1460364/full.

[10] Amogh Mannekote, **Adam Davies**, Jina Kang, and Kristy Elizabeth Boyer. Can LLMs reliably simulate human learner actions? A simulation authoring framework for open-ended learning environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025. URL https://eaai-conf.github.io/year/eaai-25.html.

[11] Amogh Mannekote, **Adam Davies**, Guohao Li, Kristy Elizabeth Boyer, ChengXiang Zhai, Bonnie J Dorr, and Francesco Pinto. Do role-playing agents practice what they preach? belief-behavior alignment in LLM-based simulations of human trust. In *First Workshop on Social Simulation with LLMs*, 2025. URL https://openreview.net/forum?id=1BDRPz3hcK.

[12] Sandeep Puthanveetil Satheesan, **Adam Davies**, Alan B Craig, Yu Zhang, and ChengXiang Zhai. Toward a big data analysis system for historical newspaper collections research. In *Proceedings of the Platform for Advanced Scientific Computing Conference*, pages 1–11, 2022. URL https://doi.org/10.1145/3539781.3539795.

[13] Paul Smolensky, Roland Fernandez, Zhenghao Herbert Zhou, Mattia Opper, **Adam Davies**, and Jianfeng Gao. Mechanisms of symbol processing for in-context learning in transformer networks. *Journal of Artificial Intelligence Research*, 84(23), 2025. URL https://jair.org/index.php/jair/article/view/17469.

[14] Jianhao Yuan*, Francesco Pinto*, **Adam Davies***, and Philip Torr. Not just pretty pictures: Toward interventional data augmentation using text-to-image generators. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 57924–57952. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/yuan24e.html.

[15] Yu Zhang, **Adam Davies**, and ChengXiang Zhai. Understanding the social construction of juvenile delinquency: insights from semantic analysis of big-data historical newspaper collections. *Journal of Computational Social Science*, pages 1–43, 2024. URL https://link.springer.com/article/10.1007/s42001-024-00254-x.