Adam Davies

adavies4@illinois.edu
+1 (801) 357-9217
https://ahdavies6.github.io/

PhD candidate at UIUC (University of Illinois Urbana-Champaign), advised by Profs. ChengXiang Zhai and Julia Hockenmaier.

Research areas: *natural language processing, (mechanistic) interpretability, distribution-shift robustness, causal machine learning, synthetic data,* and *multimodal representation learning.*

EDUCATION

University of Illinois Urbana-Champaign, Urbana, IL 08/2021 - Present
Ph.D. in Computer Science (anticipated graduation May 2026)

University of Utah, Salt Lake City, Utah 08/2016 - 05/2021
B.S. in Computer Science (May 2021, cum laude)
B.S. in Cognitive Science (May 2021, cum laude)

RESEARCH EXPERIENCE

Doctoral Researcher at INVITE 08/2025 - Present

- Studying how findings and methods from mechanistic interpretability can be leveraged to predict and improve OOD generalization of LLMs.

PhD Research Intern at Microsoft Research 05/2025 - 08/2025

- Studied how LLMs internally represent the latent structure of symbolic reasoning tasks in order to predict and improve their compositional generalization. (Paper in review.)

Doctoral Researcher at INVITE 05/2024 - 05/2025

- Introduced theoretically-motivated sparse autoencoder architecture and evaluation framework for interpreting LLM embedding representations [7, **COLM25** (conference)].
- Introduced generalizable LLM steering methods for distribution-shift robustness and bias mitigation [6, **ICML25** (conference)].
- Studied LLM-based agent simulation for AI + education [9, **AAAI25** (conference oral)] and social science research [10, **SocialSim@COLM25** (workshop)].
- Worked with domain-area experts to define and operationalize principles of socially responsible foundation models [4, **HAIC@ICLR25** (workshop)] for educational LLM applications [8, **Frontiers AI** (journal)].

Doctoral Researcher at UIUC 08/2022 - 05/2024

- Studied foundational mechanistic interpretability, including...
  - defining and evaluating the reliability of leading causal probing methods [1, **IAI@NeurIPS24** (workshop oral)].
  - introducing a general causal probing framework for LLM interpretation and analysis and new causal probing methods based on adversarial machine learning [3, **IAI@NeurIPS24** (workshop)].
  - surveying the history of interpretability and its parallels with cognitive science, up through current categories of interpretability methods and associated goals, key assumptions, and shared challenges [2, preprint].
- Evaluated the abstract shape recognition abilities of vision-language models by synthesizing benchmarks using conditional generative models [5, **NeurIPS24** (conference)], and studied how synthetic data from text-to-image models can improve distribution-shift robustness of image classifiers [12, **ICML24** (conference)].

Doctoral Researcher at NCSA 08/2021 - 08/2022

- Researched intersection of NLP, data mining, and computational social science for studying social construction using "big data" historical newspaper collections [13, **JCSS** (journal)] and [11, **PASC** (conference oral)].

PUBLICATIONS

[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] **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`.

[3] **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`.

[4] **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`.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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`.

TECHNICAL SKILLS

- **Deep Learning in Python:** PyTorch, TensorFlow, Keras, 🤗 Transformers
- **Data Science & Machine Learning in Python:** NumPy, SciPy, scikit-learn, Pandas, 🤗 Datasets
- **Classic NLP in Python:** spaCy, NLTK, CoreNLP, WordNet, gensim
- **Scientific Visualization in Python:** Matplotlib, Seaborn, Plotly, Jupyter
- **Collaboration and Publishing:** Git, LaTeX, Overleaf, and Markdown.

TALKS

- **Measuring the Reliability of Causal Probing Methods** 12/2024
  (Oral, NeurIPS24 Workshop on Interpretable AI)
- **Cognitive Interpretability in the Era of LLMs** 10/2024
  (Guest Lecture, UIUC Seminar in Psychology)
- **Causal Probing for Language Model Interpretability and Analysis** 09/2023
  (Tutorial, University of Oxford)
- **Computational Social Science with Historical Text Analysis** 06/2022
  (Oral, Platform for Advanced Scientific Computing Conference)

TEACHING AND MENTORSHIP

Research Supervision and Mentoring

Advised the following undergraduate students:

- Chirag Rastogi (UIUC BS) 07/2023 - 10/2024
  - Publication [1] (topic: *evaluating interpretability methods*)
- Jize Jiang (UIUC BS → MS) 01/2023 - 05/2023

    - Undergraduate thesis (topic: *formal reasoning with LLMs*)
    - First publication [3] (topic: *language model interpretability*)

Co-advised the following undergraduate students:

- Arshia Hemmat (Oxford internship) 01/2024 - 08/2024
    - First conference publication [5] (topic: *evaluating abstract shape recognition*)
- Jianhao Yuan (Oxford BS → PhD) 10/2022 - 05/2023
    - Undergraduate thesis [12] (topic: *synthetic data for distribution-shift robustness*)

Teaching Assistant at UIUC 08/2023 - 05/2024

- *Applied Machine Learning* (Spring 2024)
- *Natural Language Processing* (Fall 2023)