# Adam Davies

adamdavies@microsoft.com
LinkedIn | Google Scholar
Homepage: https://ahdavies6.github.io/


**Senior Applied Scientist** at Microsoft IDEAS Research, NYC.

Research areas: *natural language processing, (mechanistic) interpretability, cognitive science, OOD generalization, causal machine learning, synthetic data, multimodal representation learning, computational social science, human-AI collaboration* and *AI for education.*

## EDUCATION

**University of Illinois Urbana-Champaign**, Urbana, IL 08/2021 - 05/2026
*Ph.D. in Computer Science (May 2026)*
Dissertation: *Interpreting, Evaluating, and Controlling Features Learned and Leveraged by Foundation Models* [5].
Advisors: Profs. ChengXiang Zhai and Julia Hockenmaier.
Committee: Julia Hockenmaier (chair), ChengXiang Zhai (director), Han Zhao, and Aaron Mueller.

**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

**Senior Applied Scientist** at Microsoft IDEAS Research 05/2026 - Present

- Studying how LLM-based agents represent and reason over user theory of mind (e.g., goals, intentions, and preferences), including posttraining approaches to improve user understanding and alignment.
- Introduced an agentic architecture for faithful and intervenable cognitive user simulation [4, under review].
- Created a synthetic benchmark evaluating how enterprise agents track project state across heterogeneous data types and user perspectives [9, under review].

**Doctoral Researcher** at INVITE 08/2025 - 05/2026

- Studied how LLMs internally represent the latent structure of symbolic reasoning tasks in order to predict and improve their compositional generalization [5, (chapter 8)].
- Helped identify open problems in contextual approaches to privacy and security in agentic AI systems [1, **Google** (technical report)].

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

- Studied how Transformer models can implement leading models of human cognition to solve in-context learning symbolic reasoning tasks [17, **JAIR** (journal)].

**Doctoral Researcher** at INVITE 05/2024 - 05/2025

- Studied the reliability of causal intervention-based mechanistic interpretability methods for interpreting LLM task behaviors [3, **IJCNLP-AACL'25** (conference)].
- Introduced theoretically-motivated sparse autoencoder architecture and evaluation framework for interpreting LLM embedding representations [12, **COLM'25** (conference)].
- Introduced generalizable LLM steering methods for distribution-shift robustness and bias mitigation [11, **ICML'25** (conference)].
- Studied LLM-based agent simulation for education [14, **AAAI'25** (conference oral)] and social science research [15, **SocialSim@COLM'25** (workshop)].
- Worked with domain-area experts to define and operationalize principles of socially responsible foundation models [8, **HAIC@ICLR'25** (workshop)] for educational LLM applications [13, **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 [2, **IAI@NeurIPS'24** (workshop oral)].
  - introducing a general causal probing framework for LLM interpretation and analysis and new causal probing methods based on adversarial machine learning [7, **IAI@NeurIPS'24** (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 [6, preprint].
- Evaluated the abstract shape recognition abilities of vision-language models by synthesizing benchmarks using conditional generative models [10, **NeurIPS'24** (conference)], and studied how synthetic data from text-to-image models can improve distribution-shift robustness of image classifiers [18, **ICML'24** (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 [19, **JCSS** (journal)] and [16, **PASC'22** (conference oral)].

## PUBLICATIONS


[1] Eugene Bagdasarian, Lillian Tsai, Sahar Abdelnabi, Ananth Balashankar, Borja de Balle Pigem, Sebastian Benthall, Igor Bilogrevic, Eleanor Birrell, Elie Bursztein, Kamalika Chaudhuri, Madiha Zahrah Choksi, Mihai Christodorescu, Rachel Cummings, **Adam Davies**, Dj Dvijotham, Seliem El-Sayed, Ferdinando Fioretto, Matt Franchi, Julien Freudiger, Adria Gascon, Roxana Geambasu, Sahra Ghalebikesabi, Hamed Haddadi, Jamie Hayes, Ashish Hooda, Amir Houmansadr, Somesh Jha, Peter Kairouz, Chloé Kiddon, Tadayoshi Kohno, Abdullatif Köksal, Alison Lentz, Haoran Li, Tianshi Li, Nathan Malkin, Sarah Meiklejohn, Niloofar Mireshghallah, Sewoong Oh, Katherine Ortiz, Ivan Petrov, Francesco Pinto, Daniel Ramage, Mariana Raykova, Franziska Roesner, Edo Roth, Jacqueline Rowe, Khawaja Shams, Reza Shokri, Ilia Shumailov, Yan Shvartzshnaider, Dawn Song, Jose Such, Octavian Suciu, Pierre Tholoniat, Matt Tierney, Trishita Tiwari, Hal Triedman, Ren Yi, Wen Zhang, Xuhui Zhou, Kassem Fawaz, Stefan Mellem, Helen Nissenbaum, Nina Taft, Sarah de Haas, and Marco Gruteser. Open and emergent problems in agentic privacy and security: A contextual angle. Technical report, Google, 2026. URL `https://research.google/pubs/open-and-emergent-problems-in-agentic-privacy-and-security-a-contextual-angle/`.

[2] 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`.

[3] 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/`.

[4] Yikun Chi, **Adam Davies**, Debarati Das, Lin Ai, and Scott Counts. COGENT: A modular cognitive architecture for agentic user simulation. Under review, 2026.

[5] **Adam Davies**. *Interpreting, Evaluating, and Controlling Features Learned and Leveraged by Foundation Models*. PhD thesis, University of Illinois Urbana-Champaign, 2026. URL `https://ahdavies6.github.io/assets/docs/dissertation.pdf`.

[6] **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`.

[7] **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`.

[8] **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.

[9] Jiexin Ding, Lin Ai, **Adam Davies**, and Scott Counts. Taking the PULSE of enterprise workflows: Evaluating persistent project-state understanding. Under review, 2026.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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`.

[18] 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`.

[19] 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

- **Agents:** Claude Code, GitHub Copilot, OpenCode
- **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

- **Steering LLMs with Adaptive Feature Specification** 10/2025
  *(Tutorial, Summit for AI Institutes Leadership)*
- **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

Mentored the following students during Microsoft research internships:

- Yikun Chi (Stanford PhD) 05/2026 - 08/2026
  - COGENT [4] (topic: *cognitive architectures for user simulation*; under review)
- Jiexin Ding (University of Washington PhD) 05/2026 - 08/2026
  - PULSE [9] (topic: *persistent project-state understanding*; under review)

Advised the following undergraduate students:

- Chirag Rastogi (UIUC BS) 07/2023 - 10/2024
  - Publication [3] (topic: *evaluating interpretability methods*; earlier workshop version [2])
- Jize Jiang (UIUC BS → MS) 01/2023 - 05/2023
  - Undergraduate thesis (topic: *formal reasoning with LLMs*)
  - First publication [7] (topic: *language model interpretability*)

Co-advised the following undergraduate students:

- Arshia Hemmat (Oxford internship) 01/2024 - 08/2024
  - First conference publication [10] (topic: *evaluating abstract shape recognition*)
- Jianhao Yuan (Oxford BS → PhD) 10/2022 - 05/2023
  - Undergraduate thesis [18] (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)