David Wu

✉ david_wu@berkeley.edu • 🌐 davidxwu.github.io • davidxwu

Education

University of California, Berkeley **Berkeley, CA**
Ph.D. in EECS, coadvised by Prasad Raghavendra and Anant Sahai. GPA: 4.0/4.0 *May 2026 (expected)*
Supported by an NSF GRFP and an OpenAI Superalignment Grant

Massachusetts Institute of Technology **Cambridge, MA**
B.S. in Computer Science and Mathematics, GPA: 5.0/5.0 *May 2022*

Research Interests

LLM reasoning, RL for coding and math, weak-to-strong generalization, synthetic data, knowledge distillation, Markov chains and sampling, algorithms for machine learning

Skills

Python (PyTorch, numpy, pandas), Go, Java, LaTeX, git.

Work Experience

Windsurf **Mountain View, CA**
ML Research Intern *Summer 2025*
Created and designed verifiable coding evals and agentic judges for the Windsurf coding agent in golang. Post-trained models and built new golang and Python data and eval pipelines for new coding capabilities. Maintained eval and training infrastructure during transition period for Cognition acquisition.

Hudson River Trading **New York, NY**
Algorithm Developer Intern *Summer 2021*
Developed low level high frequency signals in C++ and Python for trading cryptocurrency perpetuals. Implemented completely automated trading bot with two teammates in C++ for live trading. Modeled market impact of cryptocurrency liquidations with rigorous statistical techniques. Developed a novel algorithm to predict liquidation events based on open interest and price data.

Selected Publications

Papers listed by contribution

1. David X. Wu, Shreyas Kapur, Anant Sahai, and Stuart Russell. **Synthetic Error Injection Fails to Elicit Self-Correction In Language Models**. *arXiv preprint arXiv:2512.02389. In submission.*, 2025
   tl;dr: Synthetic data approaches using SFT fail to induce robust error correction capabilities for reasoning tasks, even with golden solutions and perfect credit assignment.

2. David X. Wu and Anant Sahai. **Provable weak-to-strong generalization via benign overfitting**. *International Conference on Learning Representations (ICLR)*, 2025 (Previously appeared at NeurIPS 2024 M3L workshop).
   tl;dr: Extreme weak-to-strong generalization can occur because ground-truth concepts are more easily learnable than incorrect concepts with strong model representations.

3. David X. Wu and Anant Sahai. Precise asymptotic generalization for multiclass classification with overparameterized linear models. In *Neural Information Processing Systems (NeurIPS)*, 2023 **(Spotlight)**

4. David X. Wu, Chulhee Yun, and Suvrit Sra. On the training instability of shuffling SGD with batch normalization. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pages 37787–37845. PMLR, 23–29 Jul 2023
   tl;dr: Empirical+theoretical analysis of divergent training dynamics arising from the interaction of optimization and architecture choices.

Papers listed alphabetically

1. Amit Rajaraman and David X. Wu. Markov Chains Approximate Message Passing. *arXiv preprint arXiv:2512.02384. In submission*, 2025
   tl;dr: The performance of Glauber dynamics for certain low-rank matrix recovery tasks reduce to a simple scalar recursion which also dictates the performance of message passing algorithms.

2. Brice Huang, Sidhanth Mohanty, Amit Rajaraman, and David X. Wu. **Weak Poincaré Inequalities, Simulated Annealing, and Sampling from Spherical Spin Glasses**. *ACM Symposium on Theory of Computing (STOC)*, 2025
   tl;dr: New framework for understanding warm starts and annealing for sampling, with applications to diffusion.

3. Kuikui Liu, Sidhanth Mohanty, Prasad Raghavendra, Amit Rajaraman, and David X. Wu. Locally Stationary Distributions: A Framework for Analyzing Slow-Mixing Markov Chains. *IEEE Symposium on Foundations of Computer Science (FOCS)*, 2024
   tl;dr: New framework for analyzing the statistical performance of sampling algorithms before convergence.

4. Kuikui Liu, Sidhanth Mohanty, Amit Rajaraman, and David X. Wu. Fast Mixing in Sparse Random Ising Models. *IEEE Symposium on Foundations of Computer Science (FOCS)*, 2024

5. Sidhanth Mohanty, Prasad Raghavendra, and David X. Wu. Robust recovery for stochastic block models, simplified and generalized. *ACM Symposium on Theory of Computing (STOC)*, 2024

Awards

○ OpenAI Superalignment Grant ($150, 000 award)	April 2024
○ NSF GRFP fellowship ($159, 000 award)	July 2022
○ Robert M. Fano MIT UROP Award	July 2021
○ Regneron Science Talent Search, 5th place finalist ($90, 000 award)	March 2018