Sudarsh Kunnavakkam

+1 (949) 254-8232 | Pasadena, CA | kvsudarsh786@gmail.com | github.com/skunnavakkam | sudarsh.com

Work Experience

Research Intern — Sep 2023 — July 2025
Model Evaluation and Threat Research (METR) — *Berkeley, CA*

- Worked on projects to evaluate the agentic time horizon of LLMs
- Co-lead engineer of a state of the art evaluation for Chain-of-Thought Faithfulness of Large Langauge Models
- Led team of contractors to red-team LLMs and write custom datasets

Undergraduate Research Intern — Nov 2024 — Present
ShapiroLab at Caltech — *Pasadena, CA*

- Building better BCIs by engineering towards 10ms response time ultrasound reporters
- Designed custom proteins with RFDiffusion, Alphafold, and ESM3 for 10x faster kinetics

Research Fellow — Feb 2025 — May 2025
Supervised Program for Alignment Research — *Remote*

- Implemented a complex, *continuous double auction* agent arena as a model environment for LLM collusion, accepted to *ICML 2025*

High School Research Intern — Dec 2022 — Jun 2024
Lee Nano-Optics Lab at UC Irvine — *Irvine, CA*

- Scaled 2D ITO fabrication from mm^2 to multi-cm^2 sizes and developed new transfer-matrix methods for ellipsometry and refractive index characterization. Published at a US Government Workshop.

Skills

Machine Learning (PyTorch, Jax, Transformers, Diffusion Models, Reinforcement Learning on LLMs, GRPO, PPO, Interpretability), Python, Rust, C++, Javascript, Full-stack Development, PCB Fabrication, Data Analysis, Signal Processing, Rust, 3D Modeling, Shop Experience, General Wet Lab, Electron Microscopy, AFM, Scanning Probe Microscopy, Triton, vLLM

Education

California Institute of Technology — Pasadena, CA
B.S. in Physics & Computer Science — *In progress*

Selected Publications

1. A. Deng*, S. Von Arx*, B. Snodin, S. Kunnavakkam, T. Lanham, "CoT May Be Highly Informative Despite "Unfaithfulness"" by *METR*
2. K. Agarwal, V. Teo, J. Vaquez, S. Kunnavakkam, V. Srikanth, A. Liu, "Evaluating LLM Agent Collusion in Double Auctions" at *ICML 2025 Workshop on Multi-Agent Systems in the Era of Foundation Models* , Vancouver, Canada, July 2025.
3. C. J. Effarah*, T. Chen*, S. Kunnavakkam*, C. M. Gonzalez, H. W. Lee, "Liquid Metal Printed 2D ITO for Nanophotonic Applications," in *California-US Government Workshop on 2D Materials*, Irvine, California, USA, Sep 2023

Projects

METR: Faithfulness and Monitorability Eval — 2025

- A thorough evaluation building on Anthropic's seminal work on chain-of-thought (CoT) faithfulness, with thorough redteaming throughout.

LLM Agent Collusion Arena — 2025

- A continuous double auction system for agents, oversight, monitors, and other experimental conditions to test influence on collusion, accepted to *ICML 2025*

EM Simulator — 2025

- Reverse mode differentiable FDFD simulators in Jax for inverse design, with fast FDFD and FDTD through diffusion & neural operators. Did tons of optimization and speculative speedups.

Scanning Tunneling Microscope — 2024

- Built working STM for $1,000 using open-source design

Awards

ARENA 6.0 Attendee	2025
Non-trivial Fellow	2024
Physics Brawl, top 10 US High School Teams	2024, 2023
USACO Silver	2023
AIME Qualifier	2023