Sudarsh Kunnavakkam

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

Work Experience

Research Assistant (Contract) Sep 2023 — Present
Model Evaluation and Threat Research (METR) *Berkeley, CA*

- Lead engineer for internal project to estimate the agentic time horizon of LLMs at much lower cost
- Co-lead engineer of a state of the art evaluation for Chain-of-Thought Faithfulness of Large Langauge Models
- Helped lead teams of contractors red-team LLMs and curate datasets such as DAFT Math of difficult, free-response questions

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

- Building better BCIs by engineering towards 10ms response time ultrasound reporters
- Built a high throughput ultrasound screening platform to scale to 1000s of variants per day
- 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
- Benchmarked emergent collusion between LLMs under various pressures
- Work 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
- Developed new transmission matrix method replacing repeated ellipsometry
- Created transfer-matrix reverse solver to easily get refractive index information under nonlinear conditions

Education

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

University High School Irvine, CA
High School Diploma *Sep 2020 — Jun 2024*

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

- Co-lead engineer on METR research report on chain-of-thought (CoT) faithfulness (Aug 2025), extending Anthropic's seminal evaluation to three frontier models and publishing findings for the wider safety community

LLM Agent Collusion Arena 2025

- Implemented a continuous double auction system for agents
- Implemented oversight, monitors, and other experimental conditions to test influence on collusion
- Added logging and metrics with WandB
- Accepted to ICML 2025 Workshop on Multi-agent Systemsa

EM Simulator 2025

- Reverse mode differentiable FDFD simulators in Jax for inverse design
- Forward and backward diffusion models trained with DDPM and Physics-inspired reward functions to approximate steady state solutions
- Implemented fast FDTD for transient events + implemented Fourier Neural Operators for speedup

Circuit Simulator 2025

- Reverse-mode autodiff for RLC network optimization
- Gradient-based optimization for component selection
- Works in time domain, as well as just to do component selection
- Implemented custom `spsolver` that is differentiable in JaX

Adversarial Attack Using Soft Tokens 2024

- Soft-token embedding technique for adversarial text generation
- Orthogonal Procrustes Alignment for token mapping
- Demonstrated attack generalization across models (PyTorch)

Scanning Tunneling Microscope 2024

- Built working STM for $1,000 using open-source design
- Achieved atomic-resolution imaging capabilities (Circuit Design, Signal Processing, Mechanical Engineering)

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