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*

- Designed and assisted evaluations for estimating agentic time-horizons of language models
- Co-lead engineer of a state of the art evaluation for Chain-of-Thought Faithfulness of Large Langauge Models
- Worked with contractors to 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*

- Developed ultrasound reporter cells for biochemical signal sensing
- Wrote high throughput computer vision screens for optical and ultrasound imaging to scale to 1000s of cells / day
- Designed custom Protein-Protein linkers with ESM3 and Alphafold
- Imaged cells using xAM mode ultrasound
- **Skills**: CAD, Signal Processing, Wet Lab

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

- Conducted research on the safety of multi-agent systems, focusing on LLM-based agents' cooperation and collusion and developed a benchmarking environment to analyze agents' actions during negotiation.
- Implemented a complex, *continuous double auction* agent arena as a model environment for LLM collusion

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

- Scaled 2D ITO fabrication from mm² to multi-cm² sizes
- Developed new refractive index characterization method replacing repeated ellipsometry
- Created transfer-matrix reverse solver to enhance ellipsometric data interpretation

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*

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. D. Dang, Q. Dang, A. Anopchenko, C. M. Gonzalez, S. Love, C. Effarah, S. Kunnavakkam, W. Wang, J. Calixto, and H. W. Lee, Epsilon-Near-Zero Photonics in Planar and Optical Fiber Platforms,'' presented at the *53rd Winter Colloquium on the Physics of Quantum Electronics (PQE 2024)*, Snowbird, Utah, USA, Jan 2024
4. 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
5. A. Anopchenko, C. M. Gonzalez, D. Dang, Q. Dang, S. Love, L. Zhang, S. Gurung, K. Nguyen, T. Chen, J. Calixto, S. Kunnavakkam, A. Palmer, and H. W. Lee, "Epsilon-Near-Zero Optics in Planar and Optical Fiber platforms," in *SPIE Optics + Photonic Conference 2023*, San Diego, California, USA, Aug 2023.

Projects

METR: Faithfulness and Monitorability Eval 2025

- Co-authored 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
- One of two lead engineers on the project, responsible for building out the evaluation framework
- Ran 100+ hours of red-team prompting with Gray Swan, uncovering worst-case CoT obfuscation tactics and hardening monitoring methods.
- Coordinated building "DAFT Math" free-response dataset to replace MCQs, removing answer-guessing and tightening evaluation rigor.

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