

Sudarsh Kunnavakkam

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

EDUCATION

California Institute of Technology

Physics / Computer Science

Pasadena, CA

In progress

University High School

High School Diploma

Irvine, CA

Sep 2020 — Jun 2024

- Selected Coursework: Mathematical Physics, Linear Algebra, Differential Equations, Multivariable Calculus, Theoretical Computer Science
- Graduated **Summa Cum Laude**

WORK EXPERIENCE

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

Research Assistant (Contract)

Sep 2023 — Present

Model Evaluation and Threat Research (METR)

Berkeley, CA

- Designed evaluations for estimating agentic performance of language models
- Worked on evaluations for Chain-of-Thought Faithfulness of Large Language Models
- Technologies: Python, SQL, Large Language Models

Undergraduate Research Intern

Nov 2024 — Present

ShapiroLab at Caltech

Pasadena, CA

- Developing novel ultrasound reporters for extracellular signal sensing (e.g. cancer markers)
- Wrote GPU-accelerated simulators for Ca²⁺ dynamics across RyR cells with Connexon channels
- Built CV algorithms for ultrasound signal processing; researching RyR gating modulation
- Pipelines to process 10s of TBs of data per day

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

PUBLICATIONS

Workshops

- 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.

Conference Publications

- 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
- 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
- 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 (WIP?) [2025](#)

LLM Agent Collusion Arena [2025](#)

- Helped implement 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 Systems

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

Non-trivial Fellow 2024

Physics Brawl, top 10 US High School Teams 2024, 2023