

Neelasha Sudarshan

neelasha.sudarshan@gmail.com | [LinkedIn](#) [Github](#)

PROFESSIONAL SUMMARY

Computer Science and Engineering Student with an interest in AI research in collusion and interpretability. Seeking research roles in technical AI safety.

EDUCATION

Santa Clara University, Santa Clara, CA

Bachelor of Science in Computer Science and Engineering (GPA: 3.84)

- Relevant Coursework: Linear Algebra; Abstract Data Types and Data Structures; Embedded Systems

WORK EXPERIENCE

AI/ML Research Fellow

06/2026-Current

Santa Clara University, Department of Computer Science and Engineering, *Santa Clara, CA*

- Granted Undergraduate Research Award to replicate & extend research on collusion and AI control. [[Link to write-up](#), [Link to code](#)]
- Proficient in using open-weight models, supervised fine-tuning, InspectAI eval tooling, and LLM CoT reviews

Private Tutor

04/2023-06/2023

Freelance

- Tutored for 180+ hours in Calculus, Algebra, Spanish, and English.

PROJECTS

White-Box Steering Safeguards for AI [[Link to write-up](#), [Link to code](#)]

- Tested white-box safeguard strategies to prevent prompt-level attacks to AI agents
- Relevant skills: Python, experiment design

Open-Source Data Privacy Evaluation Framework [[Link](#)]

- Contributed to SynEval (synthetic data eval framework), implemented PR-AUC/cross-validation; corrected for imbalanced datasets
- Proficient in: binary classification, MIA (Membership Inference Attack), classifiers

CERTIFICATIONS

BlueDot Impact Technical AI Safety

08/2026

- Completed 30-hour program covering AI safety techniques, key gaps, and various threat models. Created a one-page action plan for contributing to the field.

SKILLS

Languages: C, Python

Operating Systems: MacOS, Linux

Applications: Docker