

Harsh Soni

harsh.soni@berkeley.edu · soharshh.com · linkedin.com/in/Harsh-Soni46

EDUCATION

The Pennsylvania State University

Aug. 2023 – May 2027

Bachelor of Science in Computer Science & Economics

University of California - Berkeley

Jan. 2026 – May 2026

Concurrent Enrollment

RESEARCH EXPERIENCE

Research Fellow, Cambridge AI Safety Hub (MARS V)

Jun. 2026 – Present

Advised by Shivam Raval and Luiza Corpaci

- **Deterministically Measuring Autoformalization Faithfulness:** Developed a methodology for evaluating whether LLMs preserve meaning during autoformalization, using natural-language descriptions generated from Equational Theories Project statements and deterministic grading without human or LLM judges.
- Evaluated nine models using matched narrative and literal descriptions, examining how intermediate abstraction, inference-time reasoning, and structural complexity affect translation accuracy and reasoning-token costs. [**Under review, NeurIPS 2026 Workshop**]

Independent Researcher

Mar. 2026 – Jul. 2026

UC Berkeley

- **ToolFailBench – Diagnosing Tool-Use Failures in LLM Agents:** Developed a diagnostic benchmark and standardized evaluation harness to distinguish how tool-using LLM agents fail, separating behavior into Tool-Skip, Result-Ignore, Output-Fabrication, and Unnecessary-Tool-Use.
- Evaluated 19 models across 1,000 tasks in five professional domains, showing that similar aggregate performance can conceal substantially different failure profiles. [**ICML 2026 Workshops: FAGEN, AIWILD**]

Independent Researcher

Apr. 2026 – Present

Advised by Ilija Lichkovski, Prime Intellect

- **Safety Effects of Prompt Optimization:** Studying whether capability-focused prompt optimization methods (MIPROv2, GEPA) unintentionally degrade LLM safety alignment while improving task performance.
- Running controlled safety evaluations before and after optimization to characterize capability–safety trade-offs, and testing safety-relevant example injection to mitigate alignment degradation.

Research Contributor

Jan. 2026 – Feb. 2026

Collaborated with Idhant Gulati, UC Berkeley

- **Emergent Misalignment in Vision-Language Models:** Contributed to experiments studying how narrow-domain fine-tuning induces broader safety misalignment in VLMs.
- Contributed to dataset curation and ran LoRA fine-tuning experiments across multiple ranks, alongside activation steering experiments, examining misalignment across text and multimodal evaluations. [**ICLR 2026 Workshop**]

Undergraduate Learning Assistant

Aug. 2025 – Nov. 2025

Penn State University; CMPSC 465 (Data Structures and Algorithms)

- Evaluated and graded programming assignments and exams in data structures and algorithms, ensuring accurate and consistent assessment.

PROJECTS

ScarTissue (Dev Tools) | Python, FastAPI, Next.js, ChromaDB, Claude, MCP

May 2026

- Built an agentic pull request review system that flags potential reintroductions of previously fixed bugs using a Claude tool-use loop over bug-fix commits mined from Git history.
- Shipped a Next.js web app with a Modal backend and published an MCP server to PyPI (scartissue-mcp), supporting inline GitHub review comments and retrieval of bug patterns across repositories.

LOFTE (SportsTech) | React Native, Expo, Supabase, Railway, Clerk, FastAPI

Jan. 2026

- Shipped a workout tracking iOS app to the App Store with Nyx, an AI coach that uses training history, WHOOP recovery data, and Apple Health metrics to personalize recommendations.
- Built the frontend, backend, AI integration, and design system end-to-end, translating Figma designs into React Native interfaces.

- Sentiment-Driven Market Volatility Index | *PySpark, NLP, Spark MLlib, APIs*

Jul. 2025
- Built a PySpark pipeline to ingest and process financial text data using distributed NLP and sentiment scoring.
 - Constructed a sentiment-driven volatility index and analyzed its relationship with VIX using correlation analysis and Spark MLlib models.

TECHNICAL SKILLS

Languages: Python, SQL, C/C++, Java, JavaScript/TypeScript
ML & Data: PyTorch, Hugging Face, scikit-learn, vLLM, PySpark, Weights & Biases
Development: FastAPI, React, React Native, Next.js, Node.js, Git, Modal

INVOLVEMENT

- Berkeley AI Safety Initiative (BASIS)**

Jan. 2026 – May 2026
- Member*
 - Participated in research discussions on LLM alignment and evaluation.
- Penn State Derivatives Association**

Sep. 2024 – Dec. 2024
- Member*
 - Explored options and futures pricing, Greeks-based risk management, and market volatility through practical exercises and discussions of macroeconomic developments.
- Lawn Tennis**

Jan. 2012 – Present
- Competitive Player*
 - Achieved a career-best national ranking in the top 50, with district, state, and national tournament wins and two years of international competition in singles and doubles.