Atharv Naphade

[LinkedIn] — anaphade@andrew.cmu.edu — 4087265795

Education: Carnegie Mellon University (B.S. Computer Science, Artificial Intelligence) August 2025 - May 2028

Experience

Founder, OpenAlign Fall 2025

- Python package to evaluate+improve LLMs on fairness, trustworthiness, and robustness by implementing 30+ research papers. For example, **Qwen3** on HarmBench can be improved to 0.51 → 0.8
- Created a community of **450+ Users** and I manage **70+ Developers**.

AI Engineer, Blast(YCombinator S24) Palo Alto - Summer 2025

- Improved Lowe's AI: Milow by Implementing SoTA papers using **Reinforcement Learning** to improve LLM red-teaming
- Implemented through **backend and frontend** multiturn-tests using a novel algorithm to adapt for variation in production model output while benchmarking follow-up questions. Used Redis, Next.JS, React, and SQL.

Founder and Software Engineer, IvySpark AI 01/25-Present

- Conceived and developed an AI enabled college counseling experience app to democratize the college application process. The app empowers students to edit essays, match scholarships and opportunities that match their profile, match them with potential mentors, manage their entire college application process in one place. Built with **SQL, NodeJS, React, Tailwind, FirebaseAuth** and Gemini 2.5. Emulates real advice by using RAG on past college results. Used by **2.3k users**.

AI Researcher – Iowa State University 2024-2025

- Built video vision deep learning models to **detect and report risky driving** behaviors in real-time, improving urban traffic safety. My proposed algorithm was **deployed on 300+** Highway Cameras Nationally.

Educator, Social Media

- **17k followers, 1M+ Views** making videos explaining AI Research for everyone to understand.

Research

Vision Language Model Research, CMU Safe AI Lab Fall 2025

- Improving RL and SFT over training of VLM reasoning and action models for physical AI.

Robust AI Image Watermarking, UC Berkeley Summer 2025

- First author of paper introducing a more **robust hidden watermarking** framework using a **Variational Autoencoder** with **adversarial attacks**, on top of the diffusion model. Outperforms Meta's Stable Signature.

Interoperability of Reasoning in LLMS, Quebec AI Institute Summer 2025

- Proposed Novel Framework for **understanding the Importance** of Each Subthought in CoT Reasoning with Conditional Probabilistic Framework.

Nature Scientific Journal — AI Researcher 2021 - 2023

- Worked with CMU, UC Berkeley researchers to estimate COVID-19 related mortality using past mortality data and a mixture of novel periodic Deep Learning techniques
- Presented at the **G20 Global Health Summit** and recognized by the **National Leadership in Health**

Awards

- **Stanford University Mathematics Camp** Student Researcher
- Stanford Math Tournament **1st place/2200**
- 5x **AIME** Qualifier (2x 239 **USAMO** Index),
- The 2025 IBM Thomas J Watson Memorial Scholarship
- **USACO** Gold (Silver Perfect Score)

Expertise

LLM Research| Langchain | AWS | C++ | Python | PyTorch | Java | CSS/JS | NodeJS/React.JS/SQL | Tailwind | Prompt Engineering | Context Engineering | Open Source Tooling | Linux | Next.JS | Excell | Data Science