Atharv Naphade

anaphade@andrew.cmu.edu 4087265795 atharv-naphade-211a69219 dude123studios

Education

Carnegie Mellon University *B.S. Computer Science, Artificial Intelligence* *May 2028*
GPA: **4.00**

Experience

(Incoming) Software Engineer Intern – Roblox *Summer 2026*
- Team TBD

Jane Street FTTP *Spring 2026*
- Highly Selective 1-week Trading and Technology Program. 1/60 Invitees out of thousands of applicants

Research Fellow – SPAR *Spring 2026*
sparai.org
- Working on Jailbreaks for AI Safety Stream

Research Scientist Intern – CMU MLD *Fall 2025*
- Scaling up RL Post-training of Vision Language Models. Collaboration with NVIDIA Research.

Research Engineer – Refactor(YCombinator S24) *Summer 2025*
ycombinator.com/companies/blast
- Improved Robustness of Lowe's AI at scale by deploying novel RLVR environments.
- Implemented 11+ full stack Infra features in SQL, Redis, Next.JS for scalable evaluation of LLMs including multi-turn evals, error tracking & mitigation, and efficient guard rails. **First hire**.

Machine Learning Engineer – Iowa State University *2024*
- Built video-based deep learning models to **detect and report risky driving** behaviors in real-time. Used Pytorch & Deepstream. My proposed algorithm was **deployed on 260+** highway cameras. Worked under the guidance of Professor Anuj Sharma.

AI Team Lead – ScottyLabs (CMU's largest CS Club) *Fall 2025*
- Building agents for college students to keep track of assignments and classes.

Educator – Social Media
- **17k followers, 1M+ Views** explaining AI Research for everyone to understand. **@agi_atharv**

Awards

- **Putnam 2025 Top 500** Award (Top 270 among all students in North America)
- USAMTS Medalist, 2x BAMO Award winner
- **Stanford University Mathematics Camp** Student Researcher, Focus on Gradient Fields.
- Stanford Math Tournament **1st place/2200** Individual
- 5x **AIME** Qualifier, Top 250 **USAMO** Indeex
- **USACO** Gold (Silver Perfect Score)
- Math Kangaroo **National Champion (1st in USA)**

Projects/Research

Me, Myself, and π: Evaluating and Explaining LLM Introspection *ICLR 2026 Workshop Accepted*
- First Author, Proposed INTROSPECT, a benchmark for LLM-introspection, then showed that models introspect better about themselves rather than of other models, as well as mechanistic evidence of how introspection is learned.

Overcoming Reasoning Mode Collapse in LLM Post training *ICLR 2026 DATA-FM Submission*
- First author. Creates Mode-Collapse metric, then creates Path-Augmented-Scaling (PAS), a data augmentation algorithm which directly reduces Mode-Collapse, and leads to SoTA Pass@16 Performance on AIME24 and Math500.

Rationale Synthesizers or Heuristic Followers? Analyzing LLMs in RAG-based Question-Answering. *ACR ARR 2026*

- Studied dynamics of LLM decision making under conflicting data. Found that LLMs follow simple heuristics rather than making rationale synthesis. (https://arxiv.org/abs/2601.06189)

Confidence estimation from LLM internal representations ICLR Reliable Agents Submission

- Created methods of fine-tuning models to improve probe performance on measuring confidence. Far more efficient than verbal outputting.

On the Emergence of Reasoning – MILA Institute *Summer 2025*

- Proposed Novel Framework for **understanding the Importance** of Each Subthought in Chain of Thought Reasoning with Conditional Probabilistic Framework with Supriyo Chakraborty.

COVID-19 Undercount Estimation - Jnana Prabodhini Foundation *2021 - 2023*

- **Nature Scientific Journal Publication**. Estimated COVID-19 related mortality using a mixture of novel deep learning techniques. "Conventional and frugal methods of estimating COVID-19-related excess deaths and undercount factors", nature.com/articles/s41598-024-57634-6.
- Presented to global scientific advisors at the **G20 Global Health Summit.**

OpenAlign *Fall 2025*

github.com/OpenAlign/AlignLab

- Python package to evaluate and improve LLMs on fairness, trustworthiness, and robustness by implementing 30+ research papers. Grew a community to **450+ developers & users**

Technical Skills

Software Development: C++, Python, Java, JavaScript (Node.js, React, Next.js), SQL, MongoDB, Docker, Git, AWS, Linux, Tailwind CSS

AI Research: Large Language Models (LLMs), Vision-Language Models (VLMs), Reinforcement Learning (DPO, PPO), Synthetic Data, Post-training, Test-time Optimization, Red-teaming, Diffusion Models, Agents

Machine Learning: PyTorch, TensorFlow, Scikit-Learn, NumPy, Pandas, LangChain, Hugging Face, OpenAI API, OpenCV,

References

Provided Upon Request