McNair Shah

Pittsburgh, PA mcnairs@andrew.cmu.edu 205-778-8194 McNair Shah

Education

Carnegie Mellon University, School of Computer Science — *Aug 2025 – Present*
Alabama School of Fine Arts (Math/Science Specialty) — *Aug 2021 – May 2025*

Research Positions

Time Series Foundation Model Development — *2025*
Auton Lab, Carnegie Mellon University

- Developing the second version of the *MOMENT* series of time series foundation models
- A large scale model that can perform tasks like understanding, forecasting, and more on general time series data.

Harmfulness Interpretability Research
Algoverse AI — *2025*

- Constructed and analyzed a multi-dimensional harmfulness subspace by probing on harmfulness subconcepts
- Designed dominant-direction steering and subspace ablation methods based on interpretability findings

Human Connectome Upsampling Research
Cognition, Brain, and Autism Laboratory @ Univ. Alabama Birmingham — *2024*

- Used machine learning methods to transform low-resolution connectomes into high-resolution ones
- Utilized graph-based score-based generative modeling to upsample connectome graphs

EEG Decoding Research
ELDEN Lab @ Univ. Alabama — *2023*

- Created multiple machine learning models to analyze subject EEG data on four classification tasks
- Beat baseline ERP decoding methods by a significant margin
- Created a framework to extract information about the temporal structure of brain processes

Publications

Death by a Thousand Directions: Exploring the Geometry of Harmfulness in LLMs through Subconcept Probing — NeurIPS 2025 Mechanistic Interpretability

McNair Shah*, Saleena Angeline, Adhitya Kumar, Naitik Chheda, Kevin Zhu, Vasu Sharma, Sean O'Brien, Will Cai

ArXiv

Projects

Page-By-Page — *2025-*

- An AI-powered application aimed at making recent research more accessible to the general public.
- Created a scrollable experience focused on simplifying and summarizing complex research papers.

CMU Maps — *2025-*

- Member of a team working on CMU Maps, a custom maps service for CMU.
- Includes various features such as area information and custom pathing through the inside of buildings.

Independent Research in Human-AI Interaction — *2022-2024*

- Experimented with using GANs with auxiliary classifiers in sentimental language synthesis.
- Used reinforcement learning in articulatory synthesis; synthesized speech by simulating human vocal tract.

Technologies

Languages: Python, Java, HTML/CSS, Javascript, TypeScript, C#, C++, C, LaTeX

Technologies: React, Next, Flask, Express, Spring Boot, Pytorch, Huggingface, PostgreSQL

Awards

- USA Physics Olympiad Silver Medalist (Top $\sim 1\%$ of skilled physics high schoolers)
- Top 3 High School Division NYU CSAW Cybersecurity Capture the Flag
- Top 16 National Science Bowl, MIT Science Bowl
- AIME qualifier
- Eagle Scout