Alok Shah

+1 412.419.6400 | alokshah@upenn.edu | LinkedIn | Github

Education

University of Pennsylvania — Philadelphia, PA
BSE in Computer Science, BA Mathematics, MSE in Electrical Engineering — *Aug. 2022 – May 2026*

North Allegheny Senior High — Wexford, PA
High School Diploma; GPA: 4.6/4 — *Aug. 2018 – June 2022*

Experience

Stanford Linear Accelerator — April 2024 – August 2024
MLR Intern — *Palo Alto, CA*

- Devising energy-efficient finetuning algorithms using system aware optimization techniques
- Developing generative algorithms for high energy physics to simulate particle accelerator data

MLR @ Penn — October 2023 – Present
President — *Philadelphia, PA*

- Built Penn's first and only ML research club available to both undergraduate and graduate students
- Hosted school-wide speaker events featuring partners from a16z and scientists from FAIR, GRASP, and OpenAI
- Overseeing research projects in watermarking, theoretical bounds on transformers, and state-space modeling
- Leading project in Vision-Language Models for OCR and HTML/Markdown code reproduction

Flagler Health — May 2023 – September 2023
MLE Intern — *San Francisco, CA*

- Designed Embedding application to pipeline and query data to identify patients for novel MSK procedures
- Engineered data using NER models with AWS to anonymize patient notes in compliance with HIPAA
- Fine-Tuned Dolly LLM with LangChain and Hugging Face to diagnose MSK patients and recommend treatment
- Productionized application using MLFlow, ChromaDB, and Databricks to fully deploy tool across 200+ clinics

Discus.ai — June 2023 – Present
Developer — *Philadelphia, PA*

- Created platform to streamline LLM development through auto finetuning and synthetic data generation
- Built & Deployed python package to generate tabular data by modifying CTGAN using LLMs for row comparison
- Implementing Retriever Augmented Dual Instruction Tuning (RADIT) to enhance embedding retrievals for RAG
- Implementing vision-based transformer models (NOUGAT) for OCR/VDU tasks to increase data quality

Research

Screenshot2Code: Hierarchical Semantic Mapping | *University of Pennsylvania - GRASP*

- Building composite model to reproduce HTML code given website screenshots under supervision of Dr. Jianbo Shi
- Retrained SAM using Adapters, Patch Embeddings, and Sobel Operations to segment webpages
- Implementing divide-and-conquer post-processing heuristic to combine segments and corresponding html code
- Improving OOD results for current SOTA model, SightSeer, without retraining the base model

Investigating Language Model Dynamics using Meta-Tokens | *NeurIPS ATTRIB Workshop, 2024*

- Developed novel attention mechanism to enhance LLM performance and interpretability using filler tokens
- Pretrained modified GPT-2 architecture, demonstrating improved empirical performance on MMLU by 1%
- Analyzed attention score and residual stream distributions, revealing that meta-tokens accelerate logit convergence
- Visualized model internals using the logit lens, providing insights into attention dynamics maintain global context

Weak-to-Strong In-Context Optimization of Language Model Reasoning | *NeurIPS 2024 ATTRIB Workshop, 2024*

- Developed a weak-to-strong in-context optimization method to enhance reasoning in LLMs
- Used weak learners to generate reasoning chains, improving strong model performance on MMLU
- Incorporated filler tokens to gauge model reasoning efficiency and optimize response generation
- Achieved a 0.6% boost in performance for Llama-3.1-405B on reasoning benchmarks without addtional fine-tuning

Optimizing Connect Four Solvers with Machine Learning | *Carnegie Mellon University SCS*

- 1 of 64 from over 800 selected for 5-week scholarship to research at CMU through Pennsylvania Governor's School
- Published in CMU's journal comparing the accuracy and runtime of different ML algorithms [link]
- Coded multithreaded implementations of Minimax and Monte Carlo Tree Search (MCTS) algorithms
- Trained Deep RL bot using Q-learning and parallelized inference code across cores for optimized performance

Technical Skills

Languages: Python, Java, C, Rust, R, HTML, Javascript, CSS, Solidity, OCaml

Coursework: Discrete Structures, Data Structures and Algorithms, Ethical Algorithm Design*, Machine Learning (Head TA), Deep Learning*, Computational Linguistics, Computer Vision, Computer Architecture, Operating Systems*, Control Theory*, State Estimation*, Reinforcement Learning*, Multivariable Calculus, Differential Equations, Real Analysis, Abstract Algebra, Advanced Linear Algebra*, Topology*, Probability Theory*, Combinatorics, Convex Optimization*, Stochastic Processes*, Axiom and Differential Geometry*

**denotes graduate/doctoral level courses*