Anirudh Joshi

Pittsburgh, PA | (412) 692-1884 | ajoshi3@andrew.cmu.edu | LinkedIn | GitHub

SUMMARY - Machine Learning Engineer with hands-on experience developing and deploying LLM-based systems, retrieval pipelines, and multimodal AI applications. Skilled in optimizing training pipelines, fine-tuning models for performance and efficiency, and delivering scalable ML solutions. Passionate about bridging research and production to build intelligent, high-impact systems.

EDUCATION

Carnegie Mellon University — **Pittsburgh, PA**
Master of Science Artificial Intelligence Engineering | GPA: 3.73/4 — August 2024 – December 2025
Coursework: Large Language Models, Machine Learning, Deep Learning, Multimodal ML, Introduction to Computer Systems
Teaching Assistant: OOPS & DSA (Summer 2025), System & Toolchains for AI (Fall 2025)

PES University — **Bangalore, India**
Bachelor of Technology Computer Science Engineering | GPA: 8.51/10 — August 2019 - May 2023
Awards: Distinction Award (5 semesters), Specialization in Machine Intelligence & Data Science

WORK EXPERIENCE

Pervaziv AI — **Pittsburgh, PA**
Machine Learning Engineer Intern — May 2025 – August 2025

- Boosted code-retrieval accuracy by 140% (MRR@10) by fine-tuning nomic-embed-text-v1.5 and a custom re-ranker with contrastive losses (MNRL, Triplet)
- Reduced insecure code generations by 25% across SecurityEval and LLMSecEval benchmarks using DPO + LoRA fine-tuning of Gemma-2-12B via Unsloth GGUFs
- Scaled LLM training with DeepSpeed and PyTorch FSDP, enabling 16× larger batches and cutting training time by 50%
- Deployed embedding, reranking, and code generation models with vLLM and FastAPI for low-latency inference in production

Hewlett Packard Enterprise — **Bangalore, India**
Software Developer (Cloud Ops) — August 2023 - May 2024

- Built a chatbot using Mistral 7B and OpenSearch to query Aruba Central cloud logs, reducing developer debug time by 40%
- Integrated ML-Ops pipelines mapping IoT device health metrics to Kafka topics, enabling real-time anomaly detection and smart alerts for 1 million devices

Software Development Intern — January 2023 - July 2023

- Designed a real-time network monitoring application using Streamlit and Kafka; employed Random Forest for anomaly detection

PESU Venture Lab — **Bangalore, India**
Machine Learning Intern — July 2022 - December 2022

- Built a journal recommender system with BERT multi-label classification achieving 90% precision on 50K+ academic articles
- Developed a high-performance Glassdoor scraper and used NER + topic modeling to extract in-demand skills from 1,000+ job descriptions

PATENTS AND PUBLICATIONS

- **Attention-Based Evolutionary Approach for Image Classification** – Achieved SOTA accuracy on CIFAR-10 with 50% fewer generations than baseline NAS IEEE Link
- **System and Method for Clustering and Categorizing Large Datasets** – Patented system combining text embeddings and dimensionality reduction for efficient document retrieval IEEE Link
- **Automated Workflow for Deepfake Detection** – Built a bi-directional LSTM API with 98% accuracy while reducing parameters by 100×

PROJECTS

- **Multimodal Document QA on Research Papers –** Improved QA accuracy by 15% by fine-tuning Qwen-VL-7B and CLIP for better image-text retrieval
- **Reward Model-Guided Slide Generation –** Trained SmolVLM-500M as a reward model, boosting AutoPresent performance by 28%
- **Impact of Code on Pre-Training –** Demonstrated a 5–7% improvement on Eleuther AI benchmarks via continuous pre-training of GPT-medium on code data
- **Auto ML Framework –** Deployed a full-stack AutoML framework on AWS leveraging Typescript, Figma, React JS to accomplish data engineering, feature selection and hyperparameter optimization

SKILLS

Programming: Python, Java, C++, JavaScript
ML Frameworks: Pandas, NumPy, PyTorch, MCP SDK, Crew AI, LangChain , TensorFlow, Scikit-learn, JAX, HuggingFace
LLM & Deployment: ArgoCD, Sagemaker, GraphQL, FastAPI, vLLM, FastAPI, Docker, Kubernetes, Ray
Data & Infra: Kafka, PySpark, MongoDB, Redis, PostgreSQL, Pinecone, Chroma, Qdrant, AWS, GCP