Harshitha Kolukuluru

4134728391 hkolukuluru@gmail.com harshithakolukuluru HarshithaKolukuluru hkolukuluru

Education

University of Massachusetts Amherst **Sep 2024 - May 2026**
Master of Science in Computer Science *CGPA: 3.967*
Teaching Assistant - CS 689: Advanced Machine Learning

Indian Institute of Technology Indore **Jul 2018 - Jun 2022**
Bachelors in Electrical Engineering *CGPA: 3.7*

Technical Skills

Programming: Python, SQL, Go, C/C++, Typescript, Bash

Technologies & Frameworks: AWS, Kafka, Elasticsearch, PostgreSQL, Redis, Flask, REST APIs,Google CP, OAuth2

Tools: Git, Docker, Kubernetes, Helm, ArgoCD, Terraform, Linux, Prometheus, Grafana, PySpark

AI Skills: PyTorch, Hugging Face, Keras, TensorFlow, LlamaIndex, Scikit, LangChain, NumPy, Pandas, MCP

Experience

Adobe **Jan 2026 - Present**
ML Engineer Extern *Amherst, MA*
- Implement **learned orchestration** and **context-pruning** policies for long-horizon agents to cut inference cost
- Formulate **RL-based reward** functions enabling policy-driven agent control over prompt-only orchestration.

BioNLP Lab, UMass Amherst **Feb 2025 - Dec 2025**
ML Engineer - Applied Decision Systems *Amherst, MA*
- Built and owned a **memory-augmented decision system** using retrieval-augmented generation (RAG) to condition agent actions on historical context, driving a **17.4%** lift over baseline policies in controlled evaluations.
- Designed and executed **controlled experiments** to isolate the impact of retrieval-based state augmentation, quantifying performance gains across multiple metrics and validating **causal improvements over baseline decision policies.**
- Built a **knowledge-graph–backed ML pipeline** (Neo4j + RAG) to ground decision policies in structured state.

Rakuten Mobile, Inc. **Jan 2023 - Jul 2024**
Software Engineer *Tokyo, Japan*
- Engineered a **Django**-based **Celery** scheduling system to manage CPaaS SMS workflows, cutting **MTTD by 2.8×**.
- Architected and implemented a **distributed backend** for SMS receipt ingestion and processing with **Python, PostgreSQL, and Redis**, reducing end-to-end latency by **20%** under production traffic.
- Deployed and operated **7+ microservices** in production on **Kubernetes** with **Helm**, achieving **99.9% uptime** and cutting deployment time by **6×** via **ArgoCD** and **Jenkins CI**.
- Established a comprehensive **observability stack** with **Prometheus, Grafana, and ELK**, improving platform performance and reducing **MTTA** by **25%**.
- Authored **Terraform** to provision and manage **reproducible** cloud infrastructure, reducing misconfigurations by **40%**

Univ.AI **Aug 2022 - Dec 2022**
Product Management Intern *Bangalore, India*
- Supported analytics-informed initiatives through data analysis, contributing to a **33%** increase in program engagement.

Projects

Two-Stage Retrieval Pipeline for Fit-Aware Fashion Recommendations | *GitHub*
- Improved **cross-modal retrieval** on a **44K** e-commerce text–image dataset by fine-tuning **CLIP** with **contrastive learning** to align image–text embeddings for scalable candidate generation.
- Achieved **Recall@1 = 0.42** and **Recall@20 = 0.95** by implementing a **two-stage retrieval pipeline** with embedding search and a **fit-aware neural re-ranker** using bounding boxes and dominant colors.

Learned Sparse Retrieval with Vector Quantization (LSR-VQ) | *GitHub*
- Developed a **hybrid sparse-dense retrieval framework** on the **MS MARCO** passage ranking benchmark (**2M passages**), combining **BM25** with transformer embeddings discretized via **Vector Quantization (VQ)** to enable symbolic and trainable sparse retrieval.
- Outperformed **BM25** on MS MARCO with **MRR@10 = 0.49**, achieving high **sparse retrieval efficiency** (**38.7 avg. query terms**) while preserving semantic relevance under large-scale evaluation.

Real-Time E-Commerce Analytics and Recommendation System | *GitHub*
- Built a **real-time recommendation pipeline** for low-latency personalization using **Kafka** for streaming and **Flink** for processing, improving **recommendation accuracy by 21%**.
- Implemented **collaborative filtering** with matrix factorization, optimizing training with multiple optimizers (**Adam achieving MSE = 1.77**); reduced online query latency by **27%** using **Redis**-backed caching.