Harshitha Kolukuluru

☎ 4134728391 ✉ hkolukuluru@gmail.com linkedin.com/harshithakolukuluru HarshithaKolukuluru

Education

University of Massachusetts Amherst **Sep 2024 - May 2026**
Master of Science in Computer Science *CGPA: 3.967*

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

Technical Skills

Programming: Python, SQL, C++, Golang, Bash, HTML, PHP, CSS, JavaScript, Hadoop, MATLAB, VectorDB

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

Tools: Git, Docker, Kubernetes, Helm, Argo, CI/CD, Terraform, Linux, Prometheus, Grafana, Jira, Confluence, PySpark

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

CS Concepts: Data Structures, Algorithms, OOPs concepts, Distributed Systems, Machine Learning, System Design, Operating Systems, Databases, Agile, Microservices, MLOps, Deep Learning, Natural Language Processing, Okta, OAuth

Experience

BioNLP Lab, University of Massachusetts Amherst **Feb 2025 - Dec 2025**
Graduate Student Researcher *Amherst, MA*

- Reduced hallucinated medical facts by **33%** and boosted clinical response coherence by designing **agentic patient, doctor, and nurse models** using **GraphRAG**-based architectures and **reinforcement learning**.
- Increased multi-visit agent performance by **17%** through **memory-augmented agents** built with **RAG, test-time adaptation, and longitudinal summaries**, improving interaction consistency and cross-patient generalization.
- Prototyped a **scalable hospital-wide shared memory system** enabling agent-to-agent learning for hundreds of **concurrent** patient-agent interactions via **similarity-based retrieval**, subgraph clustering, and lesson distillation.

Rakuten Mobile, Inc. **Jan 2023 - Jul 2024**
Software Engineer *Tokyo, Japan*

- Accelerated **Mean Time to Detect (MTTD)** by **2.8**× by engineering a **Django based Celery** scheduling system for CPaaS SMS workflows and enforcing **concurrency controls** to prevent task collisions.
- Streamlined end-to-end SMS receipt processing by **20%** by architecting a **distributed backend** using **Python, PostgreSQL, and Redis** and refining **concurrency management**.
- Enabled **scalable, highly available real-time observability** by integrating a **ReactJS** monitoring dashboard with backend services and orchestrating deployments via **Nginx** and **ArgoCD**.
- Sustained **99.9% uptime** and compressed deployment cycles by **6**× as measured by SLAs and CI/CD telemetry, by operating **6+ microservices** on **Kubernetes** with Helm and optimizing pipelines using Argo and GitLab.
- Strengthened platform reliability with a **25% performance gain** and a **40% drop in configuration errors** by instituting end-to-end **observability** with **Prometheus and Grafana** and codifying infrastructure using **Terraform**.

Univ.AI **Aug 2022 - Dec 2022**
Product Management Intern *Bangalore, India*

- Drove **program engagement by 30%**, by leading cross-functional teams to deliver technical solutions and driving **data-driven product strategies** through analytics-informed outreach and stakeholder alignment.

Projects

Two-Stage Fit-Aware Fashion Retrieval System | *GitHub*

- Improved **cross-modal retrieval accuracy** by fine-tuning **CLIP** with **contrastive learning** to align image–text embeddings for scalable candidate retrieval.
- Achieved **Recall@1 = 0.42** and **Recall@20 = 0.95** by implementing a **two-stage pipeline** with embedding retrieval and a **fit-aware neural re-ranker** using bounding boxes and dominant colors.
- Enhanced **ranking robustness and personalization** by training with **pairwise ranking loss** and user-aware embeddings to model fit and style preferences.

Real-Time E-Commerce Analytics and Recommendation System | *GitHub*

- Built a **real-time recommendation pipeline** to support low-latency personalization, using **Kafka** for streaming, **Flink** for processing, and **PySpark** for analytics, improving recommendation accuracy by **25%** and click-through rates by **15%**
- Implemented collaborative filtering with matrix factorization and optimized training with multiple optimizers, achieving **1.77 MSE (Adam)**, while leveraging **HDFS** and **Redis** to reduce query latency by **30%**.

Learned Sparse Retrieval with Vector Quantization (LSR-VQ) | *GitHub*

- Developed a **hybrid sparse-dense retrieval framework** combining BM25 with transformer embeddings discretized via **Vector Quantization (VQ)**, enabling symbolic and trainable sparse retrieval pipelines.
- Attained **MRR@10 of 0.49** (symbolic VQ), outperforming BM25 on the **MS MARCO dev set**.