Harsh Mahesh Tikone

Generative AI Engineer Intern

Phone: (716) 416-1170 Email: harshtikonehs28@gmail.com Linkedin GitHub Buffalo, NY

Summary

AI Engineer with expertise in GenAI applications, RAG pipelines, and AI agents using LLMs and vector databases. Proven experience building production-grade machine learning systems with 95%+ accuracy and workflow optimization.

Education

University at Buffalo **2025-08 – 2026-12**
M.S. Engineering Science (Artificial Intelligence) *Buffalo, NY* ***GPA: 3.83/4.0***

University of Mumbai **2021-08 – 2025-07**
B.Tech. Artificial Intelligence & Machine Learning *Mumbai, India* ***GPA: 3.82/4.0***

Skills

Technical: Python, LLMs, Transformers, GenAI, RAG, AI agents, LangChain, LangGraph, Vector databases, FAISS, Chroma, Embeddings, Prompting, Text generation, Model hosting, vLLM, PyTorch, TensorFlow, Machine Learning, Deep Learning, SQL, FastAPI, REST API, Azure ML, Inference, Evaluation frameworks, Benchmarks
Soft: Ownership, Curiosity, Problem-solving, Debugging, Optimization, Collaboration, Stakeholder communication, Analytical thinking, Systems thinking
Global: Git, GitHub, Jupyter, Docker, Microsoft Office, Google Workspace, Documentation, Reporting, Presentations

Work Experience

TCS **2025-01 – 2025-03**
Project Intern (AI/ML) *Mumbai, India*

- Designed Azure ML workflow with 95% OCR accuracy using Document Intelligence and Face API for automated document processing
- Optimized Blob Storage lifecycle management reducing storage costs by 30% through intelligent data tiering policies
- Built predictive scoring system with 3 GPS-derived signals achieving 60% reduction in manual review time
- Delivered 3 production artifacts including Power BI dashboards and model evaluation reports to stakeholders

Projects

RAG Document Q&A Assistant | *Embeddings, Vector DB (FAISS/Chroma), Python, NLP* **2024**

- Implemented RAG pipeline processing 300-500 token chunks with vector similarity search for document retrieval
- Architected embedding-based retrieval system providing 2-3 citations per response for accuracy verification
- Enhanced answer quality from 60% to 78% correct response rate across 50 evaluation questions
- Deployed FAISS vector database with efficient indexing reducing query latency by 35% for document search

GenAI Content Generator API | *FastAPI, Prompting, JWT, Python, REST API* **2024**

- Developed FastAPI-based GenAI application with 6 endpoints handling 1,000+ test calls for content generation
- Engineered 20+ optimized prompt templates for diverse content generation use cases and quality improvement
- Integrated JWT authentication system ensuring secure API access and user session management
- Implemented comprehensive error handling and response validation achieving 98%+ API reliability