

UDAYKUMAR MUNCIP

AI/ML Engineer | Generative AI | RAG & Agentic Systems
+1 (480) 221-4727 | umuncip@asu.edu | Tempe, AZ | LinkedIn

Professional Summary

AI/ML Engineer experienced in Python, Generative AI, LLMs, RAG pipelines, and agentic AI application development. Skilled in building AI-powered solutions, REST APIs, multi-agent workflows, and cloud-based applications. Knowledgeable in API security, authentication, access controls, and data protection for reliable, scalable AI systems.

Technical Skills

Languages & Core: Python, SQL, Java, REST APIs, Object-Oriented Programming

Generative AI & LLMs: LangChain, LangGraph, RAG Pipelines, Prompt Engineering, OpenAI API, Hugging Face Transformers, FAISS, ChromaDB, Pinecone, LLM Evaluation

AI/ML Frameworks: PyTorch, TensorFlow, Scikit-learn, XGBoost, Model Training & Evaluation, Anomaly Detection

Cybersecurity & API Security: Secure API Design, Authentication & Authorization, OAuth 2.0, JWT, RBAC, Input Validation, Data Protection, OWASP Top 10

Cloud & DevOps: AWS, Azure, IAM, Docker, Kubernetes, Terraform, CI/CD Pipelines, Git, Secrets Management

Agentic AI & Automation: AI Agents, Multi-Agent Orchestration, Tool Calling, OpenAI SDK, AWS Bedrock, Workflow Automation

Data Engineering & Databases: ETL Pipelines, PostgreSQL, MongoDB, Pandas, NumPy, Data Validation

Professional Experience

AI/ML Engineer Intern | *qBotica, Arizona, USA*

Jul 2026 – Present

- Engineered Python-based Generative AI applications using LangChain, LangGraph, and OpenAI APIs to integrate LLMs with enterprise applications and workflows.
- Designed RAG pipelines using document chunking, embeddings, and FAISS/Pinecone vector search for context-aware enterprise responses.
- Built multi-agent orchestration and multi-model AI pipelines to coordinate specialized agents and execute complex tasks.
- Developed REST APIs and backend services using Python and FastAPI to integrate AI models with enterprise applications.
- Contributed to AI governance and Context Harness architecture, exploring runtime guardrails, observability, evaluation, and audit capabilities.
- Collaborated with engineering and business teams to test, troubleshoot, and optimize AI-powered enterprise solutions.

AI Research Assistant | *Arizona State University, Tempe, AZ*

Nov 2024 – Mar 2026

- Developed and evaluated AI/ML prototypes using Python and machine learning frameworks for research initiatives.
- Implemented RAG pipelines using LangChain, LlamaIndex, FAISS, and Pinecone for semantic retrieval and LLM-powered question answering.
- Built REST APIs using FastAPI to integrate AI models with backend services and research applications.
- Evaluated models using accuracy, precision, recall, and F1-score and documented research findings.
- Supported cloud infrastructure, deployment workflows, and automation with a focus on scalability and reproducibility.
- Collaborated with faculty and research teams to troubleshoot implementations and communicate results.

Software Engineer | *JPMorganChase, India*

Jun 2023 – Jul 2024

- Developed Java Spring Boot microservices and distributed backend systems supporting enterprise supply chain operations.
- Architected Kafka-based event-driven pipelines, increasing throughput by 3× under peak load.
- Built Go microservices and gRPC services for high-performance internal platforms and distributed communication.
- Containerized backend services using Docker, Helm, and Kubernetes for reliable deployments and zero-downtime releases.
- Contributed to cloud-native architecture, performance optimization, service reliability, and operational efficiency.

Key Projects

RAG Pipeline – Enterprise Knowledge Assistant

Python, LangChain, FAISS, Pinecone, OpenAI API

- Built a RAG pipeline using document chunking, embeddings, hybrid dense/sparse retrieval, and reranking to ground LLM responses and reduce hallucinations.
- Exposed source-backed answers through REST APIs and designed retrieval workflows for relevant, traceable responses.

AI Medical Assistant

Python, LangChain, RAG, OpenAI API

- Developed a RAG assistant that retrieves clinical/reference documents for grounded, context-aware responses.
- Designed prompts, scope checks, safety guardrails, and human-review escalation for high-risk or out-of-scope questions.

Multi-Model Orchestration & Agent Framework

Python, LangChain, LangGraph, OpenAI API

- Designed task routing across multiple LLMs and specialized agents, combining outputs into coherent responses.
- Implemented adaptive chaining, tool calling, and task decomposition for complex multi-step workflows.

Voice Agent – Conversational Voice Assistant

Python, LangChain, STT/TTS APIs

- Built a voice agent integrating speech-to-text, LLM reasoning, and text-to-speech with backend tools and data sources.
- Implemented intent detection, multi-turn context, task execution, information retrieval, scheduling, and Q&A.

Auto Blog Writing Agent

Python, LangChain, OpenAI API

- Developed an autonomous content pipeline for topic research, outline generation, drafting, and iterative self-review.
- Used chained prompts to refine structure, identify content gaps, and produce publication-ready drafts.

Education

M.S. in Data Analytics (AI/ML Track), Arizona State University (ASU), Tempe, AZ | GPA: 3.92/4.0

2024 – 2026

B.Tech in Computer Science, Sri Venkateswara College of Engineering, Tirupati, India | CGPA: 9.3/10

2020 – 2024