Shivam Yadav

AI/ML Engineer | Bengaluru, India

+91 9080467733 | shivamyadav22feb@gmail.com | linkedin.com/in/shivam-yadavv | github.com/glichy-gem | shivam-yadav.pages.dev

Summary

AI/ML Engineer with 2+ years building and deploying production Generative AI systems on Azure and AWS, including multimodal RAG, real-time voice agents, and a multi-cloud AI agent orchestration platform. Experienced in LLM guardrails, retrieval architecture, RBAC, and production observability.

Technical Skills

Languages & Backend: Python, SQL, FastAPI, REST APIs

Databases & Search: PostgreSQL, pgvector, MongoDB, Elasticsearch

Generative AI & LLMs: RAG (Semantic, Multimodal, Hybrid), Semantic Search, Hybrid Search, Vector Search, Reranking, Prompt Engineering, Chain-of-Thought, Caching, LLM Guardrails, Hallucination Mitigation, LLM Evaluation

AI Agents: Agent Orchestration, Multi-Agent Workflows, Model Context Protocol (MCP), Tool Calling, Function Calling, Human-in-the-Loop Workflows, Real-Time Voice Agents

Models & Providers: Azure OpenAI, OpenAI Realtime, Amazon Nova, LLaMA, Mistral

Frameworks & Tools: LangChain, LangGraph, LiveKit, Ollama, OCR (Azure Form Recognizer, Pix2Text), Speech AI

Cloud & AI Platforms: Azure AI Foundry, Azure AI Search, Azure Logic Apps, AKS, Entra ID SSO, AWS Bedrock, Bedrock AgentCore, Bedrock Guardrails, OpenSearch, CloudWatch

Data Engineering: Azure Data Factory, Azure Synapse Analytics, Databricks, Azure Purview, ETL Pipeline Design, Data Ingestion, Data Governance

DevOps & Security: Docker, Kubernetes, CI/CD (Continuous Integration and Deployment), OAuth 2.0, Single Sign-On (SSO), Role-Based Access Control (RBAC), Application Monitoring and Logging, Performance Optimization

Experience

Associate Software Engineer (AI/ML), Unisys, Bengaluru **July 2024 - Present**

- Own end-to-end delivery of enterprise Generative AI systems on Azure and AWS, from solution design and retrieval architecture through deployment. Mentored 3 interns, led assigned workstreams, and collaborated with cross-functional stakeholders in an Agile delivery model.
- Built a legal document intelligence platform that extracts 4-level nested rule hierarchies from complex statutory text with human-in-the-loop validation, publishing a reviewed document to the searchable index in 7 to 8 minutes, with clause-level citations traceable to the source section.
- Built a multimodal academic assistant answering student queries across five disciplines over 650 ingested textbooks, with optical character recognition (OCR) to LaTeX conversion for handwritten mathematics.

Student Technical Intern (AI/ML), Unisys, Bengaluru **Jan 2024 - July 2024**

- Engineered scalable ingestion and ETL pipelines with Azure Synapse, Azure Data Factory, and Databricks, moving data through 3 medallion layers under Purview governance for downstream AI workloads.

Key Projects

Multi-Cloud AI Agent Orchestration Platform *Enterprise Control Plane*

Built a control plane for AI agents deployed on AWS Bedrock AgentCore and Azure AI Foundry. Teams discover, register, govern, and chat with agents, chain them into multi-agent workflows, and monitor health and cost live. Bedrock Guardrails enforce every interaction, with a 12-role RBAC model isolating client organizations.

Stack: Python, FastAPI, AWS Bedrock AgentCore, Bedrock Guardrails, CloudWatch, Azure AI Foundry, MCP, PostgreSQL, AKS

Voice-First AI Agent for IT Service Management *Agentic Voice AI*

Developed a conversational assistant that lets employees resolve and file IT tickets entirely by voice. Speech drives autonomous tool selection: the agent searches a 3,000-article RAG knowledge base for self-service resolution first and files a ticket only if needed, with live cards rendering in sync. Enterprise systems are exposed as agent tools via MCP workflows authored in Azure Logic Apps and registered in Azure AI Foundry. Cut latency on the highest-traffic path by 0.4 to 0.8 seconds using run telemetry.

Stack: LiveKit, OpenAI Realtime, Azure AI Foundry, Azure Logic Apps, MCP, RAG, pgvector, ServiceNow, Next.js, React

Multimodal Investigative AI Platform *Cross-Modal RAG*

Engineered an investigative chatbot that answers one natural-language query with corroborating evidence across image, video, audio, and text. Runs parallel top-10 retrieval per modality with cross-modal reranking to top-5, with speech pipelines for transcription and translation in batch and real-time modes.

Stack: Amazon Nova, OpenSearch, Transcribe, Translate, FastAPI, Python, Multimodal Embeddings

Education & Certifications

B.E. in Artificial Intelligence and Machine Learning, Bangalore Institute of Technology, Bengaluru | CGPA: 8.70 **2020 - 2024**

Certifications: Microsoft Certified: Azure AI Engineer Associate (2025) | **Anthropic Claude Certified Architect** (2026)