

Ariz Ahmad

(352) 871-5033 | ariz.ahmad.ml@gmail.com | [linkedin.com/in/ariz-ahmad](https://www.linkedin.com/in/ariz-ahmad)

TECHNICAL SKILLS

Machine Learning & Core AI Frameworks & Libraries: PyTorch, TensorFlow, Keras, Scikit-Learn, Pandas, NumPy, SparkML, MLflow
Domains: Computer Vision (3D CNNs, SAM), NLP, Recommender Systems, Transformers, Deep Learning
Generative AI & LLM Engineering Frameworks & Tools: LangChain, LangGraph, CrewAI, Hugging Face, Model Control Protocol (MCP), OpenAI API, Anthropic Claude API
Techniques & Paradigms: Retrieval-Augmented Generation (RAG), Semantic Search, Multi-Agent Systems, AI Safety (Guardrails), Fine-Tuning (PEFT/LoRA)
Data Pipelines & Platform Engineering: Apache Spark, PySpark, Databricks (Delta Lake), Airflow, Kafka
Databases: SQL, Cosmos DB (NoSQL), Vector Databases (Azure AI Search, Pinecone)
Cloud & DevOps: AWS (SageMaker), GCP (Vertex AI), Azure
Infrastructure & Orchestration: Docker, Kubernetes, Terraform, Microservices, Serverless, FastAPI, Flask
Monitoring & Observability: Prometheus, Grafana, MLflow Registry, Telemetry

EXPERIENCE

Software Engineer

2022 – 2026

Microsoft

Atlanta, GA

- Built end-to-end RAG pipeline on Azure AI Search (hybrid/semantic search) to parse 500+ PDF invoices, serving grounded, zero-hallucination GPT-4o answers via FastAPI with cross-document mathematical aggregation.
- Built XGBoost churn model on 8K tenants using SHAP TreeExplainer for risk attribution; automated personalized GPT-4o retention outreach via Semantic Kernel for accounts with probability ≥ 0.60 , scaling output to Cosmos DB and executing a segment-level Responsible AI fairness audit.
- Designed a 4-agent AutoGen pipeline (GPT-4o) on Azure Functions featuring automatic speaker selection and search tools; achieved a 65% completion rate (avg 11.2 turns) while enforcing Human-in-the-Loop (HITL) compliance gates and logging complete audit trails to Cosmos DB.
- Built FastAPI safety middleware using Azure Content Safety and GPT-4o-as-judge to block harmful inputs and ungrounded RAG outputs; achieved precision=1.00 and 0% False Positive Rate under a fail-secure architecture with structured audit logging.

Research Assistant

2021 – 2022

University of Florida

Gainesville, FL

- Built end-to-end lung nodule detection pipeline on LUNA16 dataset; trained a 3D CNN (AUC=0.88) outperforming a ViT counterpart, integrated SAM for zero-shot slice segmentation, and deployed via FastAPI and ONNX.

Software Developer

2016 – 2019

Samsung Research Institute

Noida, India

- Delivered 10+ production software releases serving 50M+ users at Samsung Research Institute; 99.5% crash-free rate and 40% bug reduction through rigorous testing and performance optimization

PROJECTS

- Built unified LLM platform routing requests across Bedrock, OpenAI, and Ollama; implemented a stateful LangGraph research agent, exposed the platform as an MCP server for Claude Desktop, and tracked unit costs and p95 latency using Prometheus.
- Built end-to-end Databricks pipeline ingesting 100K transactions into Delta Lake; engineered 9 risk features via Spark SQL, trained an XGBoost model (AUC=0.84), and deployed via Model Serving with MLflow reproducibility.

EDUCATION

University of Florida

Gainesville, FL, US

Master of Science in Computer Science, Machine Learning Specialization

2019 – 2021

Birla Institute of Technology

Mesra, JH, India

Bachelor of Engineering in Computer Science

2012 – 2016