ayanarshad2002@gmail.com
+91 7377869686

Md Ayan Arshad

BS in Data Science & Its Application
Indian Institute of Technology, Madras

LinkedIn
GitHub

Summary

GenAI Engineer with ~2 years of production experience focused on building reliable backend systems for RAG and agentic workflows.
Architected multi-tenant RAG Chatbot at Softeon (500+ enterprise tenants) with hybrid retrieval, tenant isolation, and observability. Built FastAPI based RAG system with Claude MCP integrations for agentic tool use.
Passionate about context engineering, agent workflows and shipping high quality production AI systems. I write in-depth technical blogs sharing real production tradeoffs and lessons learned
Read my blogs here: dev.to
Available for US EST / EU hours.

Professional Experience

Softeon | GenAI *May 2025 – Present | Chennai, India*

Multi-Tenant RAG Chatbot serving 500+ Enterprise tenants

- Architected and led delivery of a production multi-tenant WMS chatbot, **reported directly to AVP of Data Science.**
- Designed HLD and LLD for multi-tenant architecture using AWS (Cognito, S3, DynamoDB, Lambda, EC2); defined data flows, access boundaries, and deployment strategy.
- Solved tenant isolation at the vector DB layer (Pinecone namespaces), not code-level filters, eliminating cross-tenant data-leak risk across 500+ enterprise tenants
- Built hybrid retrieval with reranking and grounding validation; pushed **faithfulness from 0.67 → 0.91** and unsupported claim rate below 4%.
- **Led a 2-engineer team:** task breakdown, sprint planning, code reviews, and on-time delivery.

AI-Driven Ticket Resolution System

- Led the **end-to-end design and delivery** of an AI-powered ticket resolution system for Softeon's WMS support division, owning the project from **architecture design (HLD/LLD) to production rollout**.
- Built a **statistically evaluated semantic retrieval system** over historical tickets and documentation using PostgreSQL and ChromaDB, significantly reducing manual lookup time and unresolved ticket backlog.
- Performed **EDA on historical support data** to analyze ticket distributions, resolution patterns, and retrieval effectiveness; optimized relevance thresholds and retrieval depth.
- Designed evaluation workflows to assess **retrieval precision, response usefulness, and failure cases**, enabling continuous system improvement.
- Implemented monitoring and logging to support reliability and future extensibility.

GenAI Engineer · (SBL) Second Brain Labs *Sep 2024 - Feb 2025 | Remote*

- Built a LinkedIn outreach chatbot integrated with the LinkedIn API, personalized messages, multi-turn conversation handling, lead qualification across multiple client accounts simultaneously.
- System ran on GPT-4 with real campaign traffic.

Key Projects

kapa .ai (YC S23) Inspired Multi-Tenant RAG System - FastAPI + MCP

- Built a **production-ready RAG backend** using FastAPI that powers streaming, context-rich answers for documentation, directly relevant to agentic context engineering and LLM workflows.
- Designed and optimized the full retrieval pipeline through systematic experimentation (12+ combinations); selected **Heading-Aware chunking + OpenAI embeddings + Cohere reranker + Qdrant**, achieving **RAGAS scores (0.91 Faithfulness, 0.95 Context Recall, 0.89 Context Precision)** and significantly improved retrieval quality.
- Implemented **Model Context Protocol (MCP)** server for seamless integration with Claude/Cursor, enabling agentic tool use.
- Developed a clean, **scalable 3-layer architecture (Orchestrator → Strategy → Implementation)** with dependency inversion, making it easy to swap LLMs, vector stores, and caching layers.

- Added enterprise-grade production features: **multi-tenancy** (isolated Qdrant collections per tenant), API key authentication, Redis caching, PostgreSQL conversation memory, async ingestion pipeline, rate limiting, Prometheus metrics, and LangSmith tracing.
- Containerized the full stack with Docker Compose (PostgreSQL, Qdrant, Redis) and maintained high reliability with comprehensive unit + integration tests.

Credit Card Fraud Detection - Full MLOps | *LearnYard*

- Conducted statistical EDA to analyze class imbalance, feature distributions, and leakage risks.
- Built and evaluated models using ROC-AUC and precision-recall trade-offs with reproducible experiment tracking (MLflow).
- Implemented an automated ML pipeline with CI/CD and retraining workflows using DVC and Kubernetes.

Vehicle Insurance Purchase Prediction | *Production ML*

- Built an ML pipeline to predict insurance purchase likelihood using tabular vehicle and customer data.
- Focused on **probability calibration and threshold tuning** to align predictions with business risk.
- Deployed the model using FastAPI with CI/CD and cloud infrastructure.

System Threat Forecaster (Malware Prediction) | *IIT Madras*

- Built an end-to-end ML pipeline to predict system infection using telemetry data.
- Conducted EDA, feature engineering, and ensemble modeling using **XGBoost and LightGBM**.
- **Scored 93/100** in the final project evaluation.

Technical Skills

- **GenAI & RAG**: LangChain, LangGraph, LangSmith, Custom Agent Frameworks, RAGAS, OpenAI API, Claude, AWS Bedrock, Hybrid Retrieval, Grounding Validation, Cohere
- **LLMOps & Cloud:** AWS (EC2, Lambda, S3, DynamoDB, Cognito, ECR, CloudWatch), Docker, Kubernetes, MLflow, DVC, ZenML, CI/CD
- **ML & Modeling:** ML Algorithms, XGBoost, LightGBM, scikit-learn, Statistical Modeling, Feature Engineering, EDA, Bias-Variance Analysis
- **Systems Design:** HLD/LLD Architecture, Multi-Tenant Systems, API Design, Microservice Interactions, Observability
- **Programming:** Python, SQL, FastAPI, Bash
- **Databases:** VectorDB, PostgreSQL, DynamoDB, Qdrant, ChromaDB, MongoDB

Scholastic Achievements

- **Topped** Programming in Python & System Command course at IIT Madras
- **Scored 85+** in Machine Learning Practice Coding Exam
- **Scored 93/100** in ML Project at IIT Madras.

Writing & Open Work

- **Technical blog on DEV Community** where I deep dives on RAG, AgenticAI, multi-tenancy, retrieval failures, and system design for GenAI/AgenticAI.
- Publishing production **GenAI engineering content on LinkedIn** (2,300+ followers), the decisions and tradeoffs tutorials never cover.
- **1 free RAG architecture review per month**, 48-hour turnaround, finding where systems break before users do.