

DARSHAN GOWDA M

AI Engineer | GenAI & RAG Systems | LLM Applications | Cloud ML
+91 7760572027 | darshangowdam17@gmail.com | linkedin.com/in/darshangowdam1704 |
github.com/Darshan1704

PROFESSIONAL SUMMARY

AI Engineer at Tata Consultancy Services with 2 years of experience building production GenAI systems, RAG pipelines, and agentic AI workflows across regulated domains including biopharmaceutical, aviation, and medical device industries. Shipped 4 end-to-end AI solutions integrating Azure OpenAI (GPT-4o), AWS SageMaker, LangGraph, and vector databases — reducing manual effort and enabling data-driven decisions for enterprise clients. Skilled in LLM prompt engineering, predictive modelling, statistical analysis, and scalable cloud ML deployment with a focus on Responsible AI and governance.

TECHNICAL SKILLS

GenAI & LLMs	Azure OpenAI (GPT-4o), Cohere command-r-plus, RAG Pipelines, Agentic AI, LangGraph, LangChain
Prompt Engineering	Extraction prompts, generation prompts, temperature tuning, JSON-mode, verbatim extraction, few-shot prompting
Vector & Search	Qdrant, FAISS (IndexFlatL2), BM25Plus, TF-IDF, RapidFuzz, Hybrid Search, Cosine Similarity
ML & Statistics	Random Forest, Voting Ensemble, Predictive Modelling, Feature Engineering, Hypothesis Testing, MSE/MAE, Scikit-learn
Cloud Platforms	AWS (S3, SageMaker, boto3) Azure (Azure OpenAI, Azure AI Studio)
Languages & Dev	Python, SQL, Java FastAPI, Streamlit, Node.js, Express
Data & Visualization	Pandas, NumPy, Plotly, Kaleido, Matplotlib, Seaborn, Power BI, Excel
PDF & Document AI	PyMuPDF (fitz), PyPDF2, FPDF, ReportLab, langextract
Tools & Practices	Git, VS Code, PyCharm, Postman Responsible AI, Governance, POC Development

WORK EXPERIENCE

Assistant System Engineer — AI & Data | Tata Consultancy Services

August 2024 – Present (2 years) Hyderabad, India

- Architected a Medical Device Regulatory Submission Assistant using a RAG pipeline (Qdrant, Cohere, FastAPI) to automate FDA 510(k) predicate discovery, eliminating an estimated 80% of manual document search effort
- Developed a Vectorless PageIndex RAG system on Azure OpenAI GPT-4o with BM25Plus and LLM-driven hierarchical tree navigation, improving aviation AD document retrieval precision over baseline keyword search
- Engineered a 5-agent LangGraph Agentic Report Generation System with FAISS and BM25 hybrid retrieval and self-correcting validation loops, achieving 2-3x faster throughput vs. sequential processing
- Released a cloud-native Predictive Analytics Dashboard on AWS SageMaker and S3, enabling HMW aggregate clearance prediction across 5 purification steps without wet-lab experiments — cutting experiment cycle time
- Standardized LLM prompt engineering practices across projects (temperature=0.0 for extraction, temperature=0.2 for generation), reducing hallucination incidents and improving output reliability in regulated environments

KEY PROJECTS (PROFESSIONAL — TCS)

Medical Device Regulatory Submission Assistant — RAG, Qdrant, Cohere, FastAPI, PyMuPDF, TF-IDF, RapidFuzz

- Constructed an end-to-end RAG pipeline for FDA 510(k) predicate discovery across a corpus of cleared medical device PDFs; extracted Intended Use, Technical and Design Characteristics using Cohere command-r-plus at temperature=0.0
- Deployed a two-layer similarity system: dense vector cosine search for Intended Use matching and TF-IDF sentence-level matching for technical characteristics; auto-generated compliant 510(k) Summary Reports via LLM

Vectorless PageIndex RAG — Azure OpenAI GPT-4o, BM25Plus, langextract, PyMuPDF

- Designed a retrieval-augmented pipeline without embedding vectors — using LLM-generated hierarchical document trees with entity enrichment (aircraft models, ATA codes, action keywords) and BM25Plus re-ranking for EASA AD documents
- Applied 6 systematic improvements over baseline RAG: entity-enriched tree nodes, iterative retrieval with sufficiency checks, query decomposition, normalized metadata filtering, BM25 post-filtering, and context budget management

Agentic Report Generation System — LangGraph, FAISS, BM25, Azure OpenAI GPT-4o, FastAPI, Streamlit

- Orchestrated a 5-agent LangGraph pipeline (Planner, Retrieval, Writer, Validation, Assembly) with conditional routing that autonomously self-corrects report sections scoring below 7/10 on tone, purpose, and quality metrics
- Accelerated report generation by 2-3x through concurrent async section processing (asyncio and ThreadPoolExecutor) with hybrid retrieval combining 70% FAISS dense vectors and 30% BM25 sparse keyword search

DSP Predictive Analytics Dashboard — AWS SageMaker, AWS S3, Scikit-learn, Streamlit, Plotly, boto3

- Produced an ML-powered Streamlit dashboard in the biopharmaceutical domain to predict High Molecular Weight (HMW) aggregate clearance across 5 purification steps using SageMaker-hosted models (Random Forest, Voting Ensemble; MSE 0.05-0.13)
- Integrated AWS S3 for centralized experimental data access and real-time SageMaker endpoint inference via boto3; embedded correlation heatmaps, clearance factor analytics, and scientific PDF export via FPDF and Kaleido

EDUCATION

B.E. — Electronics and Communication Engineering CGPA: 7.51 / 10

Don Bosco Institute of Technology (DBIT), Bengaluru | December 2020 – May 2024

CERTIFICATIONS

- AWS Partner: Generative AI Essentials — Amazon Web Services
- Microsoft Certified: Azure Fundamentals (AZ-900) — Microsoft
- Microsoft Certified: Power Platform Fundamentals — Microsoft
- Data Science Certification — Internshala

ACHIEVEMENTS

- Secured 3rd place in Inter-Departmental Mini Project Exhibition — outperforming 25 competing teams
- Ranked 5th in DevHacks 1.0 National Level Hackathon — competing against 150+ participants across India