

DARSHAN GOWDA M

AI Engineer | Forward Deployed Engineer | GenAI & RAG Systems | Client-Facing AI Delivery

+91 7760572027 | darshangowdam17@gmail.com | linkedin.com/in/darshangowdam1704 | github.com/Darshan1704

PROFESSIONAL SUMMARY

Client-focused AI Engineer with 2 years of experience translating complex enterprise problems into production-grade AI solutions at Tata Consultancy Services. Skilled at working directly with regulated-industry clients (biopharmaceutical, aviation, medical devices) to design, prototype, and deploy GenAI systems — from initial discovery to full production rollout. Brings deep technical expertise in RAG pipelines, agentic workflows, and cloud ML, combined with strong communication skills to bridge business requirements and engineering execution. Thrives in forward-deployed settings where fast iteration, ownership, and cross-functional collaboration are critical.

WHAT I BRING TO A FORWARD DEPLOYED ROLE

Fast Prototyping POC → production in regulated domains; shipped 4 end-to-end GenAI systems	Client Proximity Directly embedded with enterprise clients across FDA, aviation, and biopharma domains
Full-Stack AI LLMs + RAG + vector DBs + cloud APIs + FastAPI/Streamlit front-ends in one engineer	☐ Responsible AI Hallucination controls, governance practices, and validated outputs for regulated industries

IMPACT HIGHLIGHTS

- ~80% manual effort reduction on FDA 510(k) regulatory document discovery — directly presented to client stakeholders
- 2–3× faster report generation through self-correcting multi-agent orchestration deployed in a live client environment
- Top 10 — TCS Internal Hackathon (AI/LLM & RAG track) from nationwide submissions, demonstrating innovation leadership
- Eliminated full wet-lab experiment cycles in biopharma client engagement using cloud-hosted predictive ML models

TECHNICAL SKILLS

GenAI & LLMs	Azure OpenAI (GPT-4o), Cohere command-r-plus, RAG Pipelines, Agentic AI, LangGraph, LangChain
Prompt Engineering	Few-shot, extraction, generation, JSON-mode, temperature tuning, verbatim extraction, hallucination reduction
Vector & Search	Qdrant, FAISS (IndexFlatL2), BM25Plus, TF-IDF, RapidFuzz, Hybrid Search, Cosine Similarity
ML & Data Science	Random Forest, Voting Ensemble, Feature Engineering, Hypothesis Testing, Scikit-learn, MSE/MAE

Cloud Platforms	AWS (S3, SageMaker, boto3) Azure (Azure OpenAI, Azure AI Studio)
APIs & Services	FastAPI, Streamlit, Node.js, Express Postman, REST API design
Languages	Python (primary), SQL, Java
Client-Facing Tools	Power BI, Excel, Plotly, Matplotlib, Seaborn Scientific PDF reporting (FPDF, Kaleido)
Dev & Governance	Git, VS Code, PyCharm Responsible AI, Governance, Regulated-domain AI delivery

WORK EXPERIENCE

Assistant System Engineer — AI & Data | Tata Consultancy Services

August 2024 – Present (2 years) • Hyderabad, India

- Served as on-site AI engineer for enterprise clients in biopharmaceutical, aviation, and medical device industries — owning full project lifecycle from requirements gathering to production deployment
- Architected and delivered a Medical Device Regulatory Submission Assistant (RAG + Qdrant + Cohere) for a medical device client — eliminating ~80% manual FDA 510(k) predicate search effort
- Deployed a Vectorless PageIndex RAG for an aviation client using Azure OpenAI GPT-4o + BM25Plus, replacing inefficient keyword search with precision LLM-tree document navigation
- Built and deployed a 5-agent LangGraph Agentic Report Generation System with hybrid retrieval and self-correction; client-facing throughput improved by 2–3×
- Delivered an AWS SageMaker-powered Predictive Analytics Dashboard for a biopharma client, eliminating costly wet-lab experiment cycles and enabling data-driven purification decisions
- Led prompt engineering standardisation across projects, enforcing temperature governance (0.0 extraction / 0.2 generation) to meet regulated-industry quality benchmarks

CLIENT PROJECT HIGHLIGHTS

Medical Device Regulatory Submission Assistant

RAG | Qdrant | Cohere command-r-plus | FastAPI | PyMuPDF | TF-IDF | RapidFuzz

- Delivered directly for a medical device enterprise client; automated FDA 510(k) predicate discovery with multi-layer RAG — dense cosine search + TF-IDF sentence-level matching
- Auto-generated compliant 510(k) Summary Reports via LLM; reduced regulatory submission preparation effort by ~80%

Vectorless PageIndex RAG — Aviation

Azure OpenAI GPT-4o | BM25Plus | langextract | PyMuPDF

- Client-embedded solution for aviation AD document retrieval; replaced keyword search with LLM-generated hierarchical tree navigation enriched with aircraft models, ATA codes, and action keywords
- Applied 6 retrieval improvements: entity-enriched trees, iterative sufficiency checks, query decomposition, normalised metadata filtering, BM25 post-filtering, and context budget management

Agentic Report Generation System

LangGraph | FAISS | BM25 | Azure OpenAI GPT-4o | FastAPI | Streamlit

- 5-agent LangGraph pipeline (Planner → Retrieval → Writer → Validation → Assembly) deployed for a client's operational reporting workflow; self-corrects content below 7/10 quality threshold
- Async concurrent generation (asyncio + ThreadPoolExecutor) with 70/30 FAISS/BM25 hybrid retrieval achieved 2–3× throughput gain in live client environment

DSP Predictive Analytics Dashboard — Biopharmaceutical

AWS SageMaker | S3 | Scikit-learn | Streamlit | Plotly | boto3

- Deployed for a biopharma client; ML models (Random Forest, Voting Ensemble; MSE 0.05–0.13) predict HMW clearance across 5 purification steps, removing the need for wet-lab trial experiments
- Delivered full-stack: SageMaker inference via boto3, S3 data pipeline, interactive Streamlit UI with scientific PDF export

INTERNSHIPS

AI-ML Intern | Samsung Innovation Campus — DBIT

Sep 2023 – Oct 2023 • Bengaluru, India

- Developed Credit Card Fraud Detection app with Streamlit — 20% improvement in detection accuracy

ML with Python Intern (Research-Based) | Compssoft Technologies

Apr 2023 – May 2023 • Bengaluru, India

- Validated LTS model performance (V0.1.1.1.23); contributed to 15% increase in prediction accuracy

EDUCATION

Bachelor of Engineering — Electronics & Communication Engineering | CGPA: 7.51 / 10

Don Bosco Institute of Technology (DBIT), Bengaluru | Dec 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
- Python Data Visualisation for Beginners (CPD Accredited)
- Data Science Certification — Internshala
- Project Management Fundamentals

ACHIEVEMENTS

- Ranked in Top 10 at TCS Internal Hackathon — AI/LLM & RAG track; selected from hundreds of internal TCS submissions nationwide
- Secured 3rd place in Inter-Departmental Mini Project Exhibition — outperformed 25 competing teams
- Ranked 5th in DevHacks 1.0 National Level Hackathon — competed against 150+ participants across India

I declare that all information stated above is true to the best of my knowledge and belief.

Darshan Gowda M