

AMAN TRIVEDI

Gen AI Engineer

+91-76655-44777 | amantrivedi2003@gmail.com | [LinkedIn](#) | [Portfolio](#)

PROFESSIONAL SUMMARY

Gen AI Engineer at AYE Finance and **M.Sc. student at BITS Pilani** (Digital). Built and shipped **production multi-agent RAG systems** (LangGraph, AWS Bedrock, Qdrant) with layered safety defenses, a **multilingual OCR/vision support assistant**, a self-hosted NLP-to-SQL layer, and a fine-tuned open-source LLM — reducing inference cost via tiered routing and semantic caching. 2nd Runner-up, AWS Agentic AI Hackathon.

SKILLS

LLM & Agentic AI: LangGraph, LangChain, hand-rolled (framework-free) agent/job orchestration, RAG, MCP, AWS Bedrock (Guardrails), Groq, NVIDIA NIM, Ollama, prompt engineering & injection defense

Eval, Optimization & Fine-tuning: RAGAS (faithfulness, context precision/recall), custom cost/latency/cache-hit metrics pipelines, LoRA/QLoRA/PEFT, quantization, semantic caching, tiered model routing

Retrieval & Multimodal: Qdrant, Supabase pgvector, hybrid search & reranking, Tesseract OCR, vision-LLM/screenshot queries

ML & Backend: XGBoost, PyTorch, Transformers, SHAP, Python, FastAPI, PostgreSQL, Redis, Docker, Kafka, Airflow

Data & Cloud: SQL, PySpark, Power BI, AWS (Lambda, EKS/ECS, S3, Step Functions, SES, SQS, SNS)

EXPERIENCE

Gen AI Engineer | AYE Finance

Oct 2025 – Present

- **Architected and shipped two production multi-agent RAG systems** (LangGraph, AWS Bedrock, Qdrant): a **Credit Helper Agent** over credit/policy documents with layered prompt-injection/output-safety defenses (Bedrock Guardrails + LLM-based validation), delivered via Electron + FastAPI + WhatsApp (Gupshup); and a **NotebookLM-style assistant** for employee documents with citation-grounded retrieval and data-retention controls.
- Built and shipped a **multilingual, OCR/vision-driven support assistant** (hand-rolled agent loop) trained on ticket data — accepts screenshot queries, answers in-language, and shares an OCR-to-LLM ingestion pipeline (Tesseract) reused across RAG and fine-tuning, cutting manual ticket triage by 70% during the pilot.
- Built and shipped a self-hosted NLP-to-SQL layer for the Lending Analytics Dashboard; added an LLM-generated insight layer to the same platform (5,000+ employees).
- Fine-tuned an open-source LLM (QLoRA) for a lending-domain task, and improved production efficiency via tiered model routing and semantic caching (60% cache hit rate).
- Own the **Vault and Repeat services** (~40% of company business, ~5 lakh hits/month) plus **5 other business-critical APIs**; lead 6-7 engineers through sprint planning and code review.

IT Intern – Data Warehousing | AYE Finance

Dec 2024 – Sep 2025

- Led end-to-end design of a **production data warehouse**, schema design, and ETL pipelines handling 10M+ records using **SQL, PySpark, and Python** across LOS, LMS, and CMS systems.
- Mapped complete data lineage and designed a unified, scalable schema; presented architecture to CTO and delivered RBI compliance app 3× faster than planned with a 10-member team.
- Awarded '**Cha Gaye – February**' for outstanding delivery impact.

Technical Intern | RK Agrotech

Jan 2024 – Jun 2024

- Built an **inventory and lead management dashboard** (HTML/CSS/JS + FastAPI) backed by PostgreSQL, with automated data ingestion from Google Forms/Sheets.

EDUCATION

BITS Pilani

M.Sc. in Data Science & AI (Pursuing)

2026 – 2028

MIT-WPU, Pune

B.Tech. in Computer Science & Engineering

2021 – 2025

PROJECTS

Multi-Tenant Agentic Support Platform | LangGraph, FastAPI, AWS, pgvector, Razorpay, Langfuse | [Live demo](#)

- Architected an async multi-tenant support platform: SES email and web chat feed SQS workers into a LangGraph supervisor (router, RAG, vision, action, escalation, guardrail agents) with Postgres-checkpointed state, human-approval interrupts, per-tenant pgvector namespaces, and hybrid retrieval with reranking.
- Secured tool execution with idempotent Razorpay refunds, webhook signature checks, policy limits, PII masking, and prompt-injection defences; gated CI with auto-generated per-tenant evals (Ragas) and traced cost and latency in Langfuse.

Agentic Document Intelligence Platform | FastAPI, Groq, NVIDIA NIM, Supabase/pgvector, S3

- Built and deployed a two-service platform (hand-rolled orchestration, no agent framework): RAG chat, YouTube→Notes, and Text→PPT generation over a normalized document pipeline (DOCX/PPTX/XLSX), backed by Supabase/pgvector.
- Routed inference across Groq + self-hosted NVIDIA NIM behind one abstraction layer, added semantic caching (50% cache hit rate), and evaluated with RAGAS across 300+ automated test queries.

ACHIEVEMENTS & RECOGNITION

- 2nd Runner-up, AWS Agentic AI Hackathon
- 6+ internal certifications within 10 months at AYE Finance for delivery excellence and impact
- HackerRank SQL 5★