Anupam Verma

AI Engineer

Mumbai, Maharashtra, India
9445864375
anupam215769@gmail.com
LeetCode
Github
Linkedin
Portfolio Website
Tableau Public

SKILLS

Programming
- Python, C/C++

Generative AI
- RAG, LangChain, LangGraph
- Fine-tuning, Unsloth, QLoRA
- Hugging Face, TRL, SFT
- MCP, FastMCP

Machine Learning
- TensorFlow, PyTorch, Scikit-learn

API Design & Protocols
- FastAPI, REST APIs
- Pydantic, JSON Schema
- OpenAPI/Swagger
- Webhooks, HTTP/HTTPS

Databases
- MySQL, PostgreSQL, MongoDB
- Neo4j, ArangoDB
- ChromaDB, PGVector, Qdrant

Cloud & Deployments
- AWS, Azure
- Docker, CI/CD
- Git, GitHub

Data Analysis & Visualization
- Tableau, Power BI

EDUCATION

Post Graduate Diploma in Data Science, Symbiosis Centre for Distance Learning
Jul 2023 – Jun 2025
Passed with 83.2%

B.Tech in CSE, Vel Tech University
Jul 2019 – Jun 2023
Passed with 9.08 CGPA

EXPERIENCE

Financial Software and Systems (FSS), AI Engineer
Sep 2025 – Present
- Architected multiple agentic **RAG systems** tailored to distinct enterprise roles—**Kilo Code** for developers, **CEO Copilot**, and **HR Work Buddy** using **LangGraph**-based multi-agent orchestration, hybrid retrieval, and secure LLM with **FastAPI** endpoints.
- Developed **MCP-compatible** retrieval tools for **Kilo Code**, enabling the IDE agent to perform codebase RAG by querying **PGVector (semantic search)** and **ArangoDB (functions and dependencies)** to surface relevant code.
- Optimized **long-context** LLM workflows for **CEO Copilot** and **HR Work Buddy** using **recursive summarization** and **dynamic token pruning**, enabling **low-latency <2s** reasoning over **80k+ token** inputs.
- Engineered a hybrid search tool combining **PostgreSQL metadata filtering** with **PGVector semantic retrieval**, improving query precision by **95%** for **time-sensitive** data and enabling **multi-hop reasoning** over emails, attachments and internal documents.
- **Fine-tuned Qwen2.5-7B** using **Unsloth** and **QLoRA (4-bit PEFT)** via **Supervised Fine-Tuning (TRL SFTTrainer)** for chargeback banking intent recognition, training on domain-specific data.
- Improved decision reliability using **counterfactual data, hard negatives, cost-sensitive evaluation, and hierarchical confidence gates**, reducing high-risk misclassifications by **~40%** with audit-ready abstention.
- **Served** and **optimized Qwen** models via **vLLM**, integrating **LiteLLM** for monitoring, fallback handling, and guardrails, achieving **24× throughput** and **~65% lower latency**.
- Implemented **per-user logging** and added **Azure AD authentication**, enhancing security, observability.

Comcast, Engineer 1 - Software Development & Engineering
Jan 2023 – Aug 2025

Generative AI
- Engineered an **Agentic RAG system** using **LangChain** and **LangGraph** to query millions of lines of code, reducing manual reviews by **40%** through bug detection, test generation, and streamlined documentation of new RDK contributions with **Azure OpenAI GPT-4.1 and o3** models.
- Developed an interactive chatbot with **Streamlit** and **FastAPI** as the **RESTful API**, enabling **on-the-fly RAG** by generating vector embeddings for user-uploaded code reducing latency by **60%**.
- Integrated multiple agents with a **ChromaDB** vector store, combining **Similarity and Full-Text Search** to improve retrieval accuracy by **95%** and enable more precise queries of RDK documentation and codebase.
- Architected persistent **user session management** and chat history storage using **MongoDB**, enabling seamless retrieval of **past interactions** and supporting **context-aware** responses.
- Integrated a **Neo4j Knowledge Graph** to map and analyze **RDK component dependencies**, accelerating dependency mapping by **40%** and enabling precise identification of inter-component relationships.
- Leveraged **Selenium**-based **web scraping** to automate extraction of patches from **Gerrit** and integrated the results into **Jira** tickets, increasing tracking efficiency by **70%** for over **500 code changes**.
- Deployed the end-to-end application on an **AWS EC2** instance using **Docker** containers, automating **CI/CD workflows** on **GitHub Actions** and achieving a **65% reduction** in deployment with **99.9% uptime**.

Tableau
- Developed **50+** interactive **RDK dashboards** in **Tableau**, including clone, code, and contribution **metrics portals**, boosting insights by **75%** and significantly improving data accessibility and efficiency for **stakeholders**.
- Leveraged advanced Tableau techniques to create visually appealing **charts, graphs, and maps**, representing **complex data** more effectively and enhancing **stakeholder** understanding by **50%**.
- Implemented innovative **data integration** with **Tableau Prep Builder** for **ETL** process like cleansing and **transformation, cutting data cleaning time** by **85%** and improving loading efficiency into **MySQL**.

CERTIFICATES

- Fundamentals of Deep Learning by NVIDIA
- AWS Academy Machine Learning Foundations