Rachana Shree H S

AI/ML Engineer (GenAI & Agentic Systems)

+91 9113638597 | shreerachanaa1010@gmail.com | LinkedIn | GitHub | Neetcode | Bengaluru, India

SUMMARY

AI/ML Engineer with 2.3 years of experience building production Agentic AI, RAG, and LLM-based systems. Strong background in hybrid retrieval, model evaluation, FastAPI, ASP.NET Core, and Azure cloud platforms. KSCST-funded researcher and IISc Bangalore Research & Innovation Program 2nd Place winner.

WORK EXPERIENCE

Associate AI/ML Engineer – Gainwell Technologies, Bengaluru *Jan 2025 – Present*

- Built **agentic workflows** (LangChain, LangGraph, FastAPI, Azure OpenAI, Claude) for healthcare guideline analysis and grounded response generation, serving 40+ users, 500+ requests/week.
- Implemented **RAG over Azure AI Search and Cosmos DB Gremlin** for relationship-aware retrieval; added **PostgreSQL-backed LangGraph checkpointing** for fault recovery and human-in-the-loop approvals.
- Automated RFP response with a **CrewAI + Azure Document Intelligence** pipeline for requirement extraction and draft generation, cutting first-draft time from 8 to 2 hours.
- Built a **claim fraud detection pipeline** (Azure ML, MLflow, Scikit-Learn) with tracked experiments and **ONNX** serving for real-time and batch inference; reached 0.93 AUC / 82% precision at 70% recall on 250K claims.
- Developed LLM-powered engineering analytics on Azure DevOps APIs and Bronze/Silver/Gold claims-reconciliation workflows, reducing manual reconciliation effort by 40%.

Software Engineering Intern – Furnishka *Sep 2024 – Jan 2025*

- Built and optimized React features for an e-commerce platform, improving page load time by 30%, working directly with the founder on requirements.

PROJECTS

Sentinel – Agentic Financial Research & Regulatory Compliance Engine GitHub | Live
LangGraph, QLoRA, Qwen 2.5, BAAI/bge-reranker-base, RAGAS, FAISS, PostgreSQL, FastAPI, Langfuse, Docker, AWS

- Fine-tuned Qwen-2.5-7B Instruct with 4-bit QLoRA on 2,174 domain examples; achieved 0.996 Macro F1 and 99.6% accuracy on 271 held-out test claims (95% CI 98.0–99.9%) with a +39.5% relative lift over GPT-4o-mini, 9x cost reduction ($1.00 vs $9.00/1k queries), and 612ms p50 latency.
- Engineered observability, evaluation, and QLoRA fine-tuning pipelines for Qwen-2.5-7B with Langfuse, enhancing retrieval quality, financial reasoning, and production-grade reliability of AI-generated research reports.

Legacy Modernization Copilot – AI-Powered .NET Refactoring GitHub | Live
C#, Python, Roslyn, Gemini, RAG(Hybrid search), React, SQLite, Render, Github Actions

- Architected a .NET modernization copilot using Roslyn, hybrid RAG, and Gemini, automating identification of 18+ legacy code anti-patterns across repository-scale codebases.
- Improved retrieval quality by 32% over semantic search using repository-aware hybrid RAG, while achieving a 78% verification pass rate for AI-generated refactors across 40 legacy .NET files.

Kisaan Sampurna – Multimodal Agri-Advisory & Market Intelligence Google Play Store
LLMs (Gemini 2.5, Qwen-2.5), LangGraph, QLoRA, Neo4j, GraphQL, FastAPI, XGBoost, PostgreSQL, AWS

- Built a multimodal agri-advisory platform using LangGraph, Neo4j, and FastAPI, integrating a 3,140-node crop-disease knowledge graph and market intelligence engine.
- Fine-tuned Qwen-2.5 on 178K+ training pairs and developed ensemble forecasting models across 6,273 mandis, achieving 9.19% validation MAPE and supporting 50,000+ installs with 3,500+ subscribers.

TECHNICAL SKILLS

GenAI / LLMs	Agentic AI, RAG, Hybrid Retrieval, Reranking, Prompt Engineering, QLoRA Fine-Tuning, Model Evaluation, Responsible AI, Azure OpenAI, Gemini, Claude, Qwen
AI Frameworks	LangChain, LangGraph, CrewAI, FAISS, BM25, PyTorch, Scikit-Learn, TensorFlow
MLOps	Azure ML, MLflow, ONNX, Experiment Tracking, Model Registry, Model Serving, Langfuse
Data & Search	Azure AI Search, PostgreSQL, Cosmos DB (Gremlin), Neo4j, SQLite, Redis, GraphQL, Pandas
Backend & Cloud	Python, C#, SQL, FastAPI, ASP.NET Core, Microservices, Docker, Kubernetes, AWS, Azure, GitHub Actions
Frontend	React, TypeScript, JavaScript, Vite

EDUCATION

B.E., M. S. Ramaiah Institute of Technology (CGPA 9.32/10) 2021 – 2025

CERTIFICATIONS

Building with the Claude API — *Anthropic Academy* Generative AI Fundamentals — *Databricks, Oct 2026*