

ASHISH NEGI

AI Engineer | Agentic AI & GenAI | RAG & LLM Systems | LLMOps | Responsible AI

Pune, India | +91 8628038851 | ashishnegi1496@gmail.com | linkedin.com/in/ashishranbirnegi | github.com/AshishNegi1496

SUMMARY

AI Engineer with 3.5+ years of experience building and productionizing GenAI, Agentic AI, and ML systems. Designed agent architectures using LangGraph and MCP and built grounded RAG pipelines using PGVector, hybrid retrieval, and reranking over 250,000+ documents. Reduced measured hallucination rate by 25% through Ragas and LangSmith evaluation. Experienced in LLM application security, LLMOps, Responsible AI, Python, FastAPI, and classical ML/deep learning.

TECHNICAL SKILLS

- **GenAI & Agentic AI:** LangChain, LangGraph, MCP, RAG, Graph RAG, Multi-Agent Systems, Prompt Engineering, Tool Calling, LoRA Fine-Tuning, OpenAI, Hugging Face
- **Retrieval & Evaluation:** PGVector, FAISS, Hybrid Search, Cross-Encoder Reranking, Ragas, LangSmith, Golden Dataset Evaluation
- **Machine Learning & Deep Learning:** PyTorch, TensorFlow, LSTM, Time-Series Forecasting, YOLO, Computer Vision, Feature Engineering
- **Responsible AI:** Prompt-Injection Controls, PII Handling, AI Guardrails, Hallucination Mitigation, Traceability
- **Backend & Data:** Python, FastAPI, SQL, PostgreSQL, REST APIs, Microservices, ETL Pipelines, Spring Boot, React
- **MLOps & Cloud:** Docker, Kubernetes, GitLab CI/CD, AWS, Prometheus, Grafana, Model & Prompt Versioning

PROFESSIONAL EXPERIENCE

AI Engineer | C-DAC, Pune Apr 2023 – Present

Museums of India – GenAI & AI Agent Platform

- Designed and productionized Agentic AI architecture (agents, tools, orchestration, memory) using LangGraph and MCP, enabling multi-step retrieval across 250,000+ historical documents.
- Built end-to-end grounded RAG pipeline (ingestion, chunking, embeddings, dense + sparse retrieval on PGVector, cross-encoder reranking), improving retrieval relevance by 35%.
- Implemented LLMOps practices with Ragas and LangSmith (golden-dataset checks, tracing, prompt versioning), reducing measured hallucination rate by 25%.
- Fine-tuned Llama-3-8B with LoRA for domain vocabulary and tool-calling; embedded Responsible AI controls including prompt-injection mitigation, PII handling, and response guardrails.
- Developed FastAPI services to operationalize model outputs and mentored 4 interns on RAG, Graph RAG, and ETL pipelines.

eForest – Government Digital Platform

- Architected and delivered the platform from 0→1 with the engineering manager, using Spring Boot + React microservices, supporting 3,000+ active users and 5,000+ daily transactions.
- Designed JWT-based authentication and RBAC authorization across 6 roles, enforcing role-aware access to platform capabilities and APIs; implemented Redis-based IP rate limiting and caching for frequently accessed queries to improve performance and resilience.
- Built Prometheus/Grafana observability and automated GitLab CI/CD, reducing release cycles from 2–3 days to under 2 hours and cutting incident response time by 35%.
- Phase 2 — Architecting a multi-modal conversational AI platform with multi-tenancy, tenant-aware ACLs, network isolation, and tool-level URL whitelisting for secure enterprise AI access.
- Designed the AI retrieval pipeline using BGE-Large embeddings + pgvector, cross-encoder reranking, and Llama 8B for grounded conversational responses, with retrieval and generation components designed for tenant isolation and controlled tool access.

River Sensor Forecasting & Geospatial Object Detection

- Built LSTM time-series models for river sensor forecasting, handling missing values, outliers, and multi-station data.
- Trained YOLO object detection models on high-resolution GeoTIFF satellite imagery with tiling, georeferencing, and augmentation pipelines.
- Owned full model lifecycle: preprocessing, feature engineering, hyperparameter tuning, and evaluation (RMSE/MAE, mAP/IoU).

ADDITIONAL PROJECTS

Secure Multi-Tenant RAG Platform (In Progress) | github.com/AshishNegi1496/secure-rag

- Designing multi-tenant RAG architecture with tenant isolation, RBAC/ABAC, PII handling, and prompt-injection mitigation (OWASP Top 10 for LLM Applications). Deployed on AWS EC2 with automated security checks.

CERTIFICATIONS & EDUCATION

- IBM RAG and Agentic AI Professional Certificate – Coursera
- DevOps, DataOps, MLOps – Duke University (Coursera)
- PG Diploma in Advanced Computing – CDAC (2022–2023)
- B.Tech, Electronics & Communication Engineering – HP Technical University (2013–2017)