Ayush Kumar Malik

https://ayushkm.com ayushkumarmalik10@gmail.com | linkedin.com/in/ayush67

Education

Indiana University Bloomington — **Bloomington, IN**
MS in Computer Science, GPA: 3.9/4.0 — *Graduated: 2025*

Shiv Nadar University, Delhi NCR — **Greater Noida, India**
BSc in Computer Science, GPA: 3.8/4.0 — *Graduated: May 2020*

Work Experience

Radical Squares — **Jan 2026 – Present**
GenAI Engineer — *Remote*

- Developed **production GenAI application** integrating **OpenAI GPT-4o APIs** with **LangChain** for intelligent data processing, implementing prompt engineering and token optimization strategies reducing API costs by 94.6%.
- Built **full-stack AI-powered platform** with **React frontend** and **FastAPI backend**, integrating multiple LLMs (GPT-4, GPT-4o-mini) for automated SQL generation and natural language processing.
- Implemented **RAG pipeline** with automated metadata extraction from **PostgreSQL, MySQL, SQL Server** databases, using vector embeddings for semantic search and intelligent field mapping across heterogeneous data sources.
- Deployed **microservices architecture** using **Docker** containers with Redis caching layer, SQLAlchemy ORM for database management, and REST APIs handling 40+ endpoints for enterprise ETL workflows.

Brckt (Peristyle Labs) — **Dec 2024 – Present**
AI/ML Engineer — *Indianapolis, IN (Remote)*

- Built **real-time GenAI application** using **Llama 3.3-70B LLM** with streaming responses via **Server-Sent Events (SSE)**, implementing prompt optimization for sports analytics predictions.
- Developed **scalable backend API** with **FastAPI** and async processing, implementing caching strategies and thread-safe operations for high-throughput data processing.
- Containerized application using **Docker** with multi-stage builds, deployed with **Caddy** reverse proxy for production-grade HTTPS termination and load balancing.

Riverside Global LLC — **Jun 2025 – Dec 2025**
GenAI Developer — *Hampton, IL (Remote)*

- Architected enterprise **RAG system** using **GPT-4 API** with **LangChain** orchestration, implementing 5-stage pipeline with hybrid retrieval combining BM25 and semantic search using vector databases.
- Developed **production-ready APIs** with **FastAPI** for document processing, implementing function calling with OpenAI APIs and custom JSON schemas for structured data extraction.
- Built **vector search infrastructure** using **ChromaDB** and **FAISS** for 10,000+ documents, achieving 94% retrieval accuracy with cross-encoder reranking and Reciprocal Rank Fusion.
- Deployed on **AWS EC2** with S3 storage, implementing monitoring and logging for production GenAI applications processing environmental compliance documents.

Projects

CodePilot: Multi-Agent GenAI System with LangChain [GitHub] | [Live Demo]

- Architected **multi-agent GenAI system** using **Claude API** and **LangChain** for autonomous code generation, implementing function calling and prompt chaining for complex task orchestration.
- Developed **hybrid RAG engine** combining keyword and semantic search with **vector databases**, achieving 40x token reduction through intelligent context management and chunk extraction.
- Built **React-based frontend** with real-time agent visualization, deployed on **HuggingFace Spaces** with containerized backend services handling concurrent LLM requests.

Cascade: Intelligent LLM Router with Semantic Caching [GitHub] | [Live Demo]

- Developed **intelligent routing system** for **OpenAI GPT APIs** using fine-tuned DistilBERT, optimizing model selection between GPT-3.5 and GPT-4 based on query complexity analysis.
- Implemented **hybrid caching strategy** with **Redis** for exact matches and **Qdrant vector database** for semantic similarity, reducing API costs by 60% while maintaining response quality.
- Built end-to-end **ML pipeline** with model training, API integration, and production deployment using **Docker** and **Streamlit** frontend for real-time inference.

ML-Monitor: Production MLOps Platform with Real-Time Monitoring [GitHub] | [Live Demo]

- Built production **ML deployment platform** with **FastAPI** backend handling 10K+ predictions/sec, implementing model versioning and automated retraining pipelines.
- Integrated **Grafana + Prometheus** for real-time monitoring of model performance, drift detection, and system health metrics with custom dashboards.
- Deployed using **Docker Compose** with horizontal scaling, implementing CI/CD pipeline for automated testing and deployment of ML models.

Technical Skills

Languages: Python, SQL, JavaScript
GenAI/LLM: OpenAI GPT-4/GPT-3.5 APIs, Claude API, Llama, LangChain, Prompt Engineering, Function Calling
RAG & Vector DB: ChromaDB, FAISS, Qdrant, Hybrid Search (BM25 + Semantic), Embeddings, Reranking
Backend: FastAPI, Django, REST APIs, Microservices, SQLAlchemy, Redis
Frontend: React, HTML/CSS, Streamlit
Databases: PostgreSQL, MySQL, SQL Server, Redis (NoSQL)
Cloud & DevOps: AWS (EC2, S3), Docker, CI/CD, Grafana, Prometheus
ML/AI: PyTorch, HuggingFace Transformers, Fine-tuning, XGBoost