Ayush Kumar Malik

🌐 https://ayushmalik.dev ✉ ayushkumarmalik10@gmail.com | in 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

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

- Built **real-time tennis match analysis system** using **Llama 3.3-70B** via Venice.ai API, generating professional head-to-head predictions with streaming responses.
- Developed **web scraping infrastructure** using **Playwright** headless browser with anti-detection measures (user-agent rotation, heading-based navigation), extracting H2H stats from matchstat.com.
- Implemented **TTL caching layer** with thread-safe operations, order-independent keys, and automatic eviction, reducing redundant scraping by caching H2H data for 2 hours.
- Deployed **FastAPI** backend with **Server-Sent Events (SSE)** for real-time streaming, **Docker** containerization, and **Caddy** reverse proxy handling HTTPS termination.

Riverside Global LLC — **Jun 2025 – Dec 2025**
AI Engineer — *Hampton, IL (Remote)*

- Architected production **RAG system** with 5-stage pipeline: query routing, reformulation, **hybrid retrieval** (BM25 + semantic), cross-encoder reranking, and GPT-4 generation with hallucination checking, reducing document research time by 60%.
- Built **hybrid search engine** combining Sentence-BERT embeddings with BM25 keyword matching using **Reciprocal Rank Fusion (RRF)**, achieving 94% retrieval relevance on 10,000+ environmental documents.
- Developed **LLM-powered data extraction** pipeline using GPT-4 function calling with custom JSON schemas, achieving 95% accuracy and reducing manual extraction from 3 hours to 15 minutes per document.
- Implemented document classification system for permits, water quality reports, and EPA notices with automated validation checks and human-in-the-loop review for compliance workflows.

Projects

CodePilot: Multi-Agent AI Coding System — [GitHub] | [Live Demo]

- Architected **multi-agent orchestration system** with 4 specialized agents (Planner, Coder, Reviewer, Explorer) using **Claude Sonnet 4.5** and function calling for autonomous code generation.
- Designed **hybrid retrieval engine** combining BM25 keyword search with semantic embeddings using **Reciprocal Rank Fusion (RRF)**, improving code retrieval precision by 25% over single-method approaches.
- Implemented **token-efficient context tools** (file outlines, code chunk extraction) achieving **40x reduction** in token consumption vs. naive full-file reads.
- Built iterative **feedback loop** between Coder and Reviewer agents with state machine orchestration; deployed on **HuggingFace Spaces** with real-time agent visualization.

ML-Monitor: Real-Time Fraud Detection MLOps Platform — [GitHub] | [Live Demo]

- Built production **MLOps platform** achieving **sub-100ms inference latency** (5x faster than industry standard 500ms) for real-time fraud detection using **XGBoost** with feature engineering and model optimization.
- Implemented **automated model retraining pipeline** with drift detection using KL divergence and PSI metrics, triggering retraining when feature distributions shift beyond thresholds.
- Deployed scalable **FastAPI** inference service handling **10K+ predictions/sec** with request batching, async processing, and horizontal scaling via load balancing.
- Integrated **Grafana + Prometheus** observability stack for real-time monitoring of prediction latency, model performance metrics, and system health dashboards.

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

- Developed **intelligent LLM routing system** using fine-tuned **DistilBERT** classifier achieving **97% routing accuracy** at 50ms latency, routing 70% queries to GPT-3.5 and 30% to GPT-4 based on complexity.
- Achieved **60% cost reduction** vs 100% GPT-4 baseline by optimizing model selection through query complexity classification without sacrificing response quality.
- Implemented **hybrid cache strategy** combining exact-match **Redis** cache with semantic similarity search via **Qdrant** vector database, reducing cache false positives from 15% to ¡2% through threshold tuning.
- Built end-to-end pipeline including dataset labeling (1,000+ queries), model training with cross-validation, and production deployment with **Streamlit** web interface.

Technical Skills

Languages: Python, SQL
ML Frameworks: PyTorch, HuggingFace Transformers, Sentence-Transformers
LLM & RAG: GPT-4 API, Claude API, Function Calling, Hybrid Search (BM25 + Semantic), Cross-Encoder Reranking (ms-marco-MiniLM), Prompt Engineering
Vector Databases: ChromaDB, FAISS
Infrastructure: AWS (EC2, S3), Docker, FastAPI, PostgreSQL, pdfplumber