

Madhav Makwana

Software Engineer | LLM Applications · RAG Systems · Agentic AI | DTU 2026

📞 (+91) 9717891589 ✉ madhavmakwana.mm@gmail.com 🌐 madhavmakwana.dev 🐙 GitHub in LinkedIn

Education

Delhi Technological University

2022 – 2026

B.Tech in Engineering Physics (Electronics), Minor in Computer Science

Experience

Appsoleut Games

June 2025 – Aug 2025

Data Science Intern

Gurugram

- Analyzed **10K+ gameplay sessions** to map early-player behavior patterns (retries, collisions, session duration) and identify high-churn business segments for product intervention.
- Trained a logistic regression classifier on engineered gameplay features, achieving **AUC > 0.80** on hold-out data.
- Delivered a **3-tier risk segmentation model** (low/medium/high) to guide onboarding strategy and retention decisions.

Projects

LLM Gateway — Secure, Cost-Aware Routing Proxy for LLM APIs

[🐙 GitHub](#)

FastAPI, Redis, FAISS, Docker, AWS EC2, Prometheus, Grafana, pytest

- Built an OpenAI-compatible gateway with circuit-breaker-based routing across Groq and OpenAI providers, automatically failing over on timeouts and rate limits.
- Implemented prompt injection/jailbreak detection and audit logging using embedding similarity matching.
- Added dual-layer caching with Redis exact-match and FAISS semantic similarity matching for near-duplicate prompts, reducing redundant LLM calls and inference cost.
- Developed per-API-key token/cost tracking with hard budget-cap enforcement to prevent runaway spend; deployed as a multi-container Docker Compose stack to AWS EC2 with Prometheus observability and mocked tests.

ResearchMind — Multi-Agent AI Research System

[🔗 Live Demo](#) | [🐙 GitHub](#)

Python, FastAPI, LangChain, Groq API, Tavily API, BS4, JWT, Prometheus, Docker, Streamlit

- Architected a **4-step LLM orchestration pipeline** integrating external web APIs (Tavily, BeautifulSoup) with an LLM backend for grounded, real-time report generation.
- Engineered a production-grade FastAPI backend with JWT auth, modular APIs, and JSON logging with trace IDs.
- Designed a Critic Chain self-evaluation loop scoring reports on accuracy and depth, reducing hallucination risk.
- Monitored pipeline with **Prometheus metrics** (per-step histograms, error counters), async test suite, deployed on Render.

YTLens — RAG-Powered YouTube Q&A System

[🔗 Live Demo](#) | [🐙 GitHub](#)

FastAPI, LangChain, FAISS, HuggingFace, JWT, SQLAlchemy, Docker, Streamlit

- Built an end-to-end **RAG pipeline** with multi-source transcript ingestion, structured chunking, **BGE embeddings**, and FAISS vector store persistence — effectively an ETL flow for unstructured media data.
- Replaced similarity search with **MMR retrieval** (k=5, fetch_k=20) to diversify context and improve answer grounding.
- Implemented a **3-layer transcript fetching strategy** (direct → proxy → manual fallback) ensuring zero hard failures.
- Hardened backend with bcrypt-hashed JWT auth, SQLAlchemy persistence, and shipped Dockerized service on Render.

Skills

Languages: Python, C++, SQL

CS Fundamentals: Data Structures & Algorithms, OOP, DBMS, Operating Systems

Backend: FastAPI, REST APIs, AsyncIO, Pydantic, SQLAlchemy, JWT Auth, OAuth2, Middleware, Logging

GenAI & LLM Systems: LangChain, RAG Pipelines, Semantic Caching, OpenAI API, Groq API, Agentic AI

Databases & Caching: MySQL, Redis, SQLite, FAISS, ChromaDB

DevOps & Infra: Docker, AWS EC2, Prometheus, Grafana, Git, GitHub, Render, Locust

ML: scikit-learn, TensorFlow, Keras, HuggingFace

Tools: Streamlit, Postman, Linux

Achievements

- Ranked among **top 5%** in NPTEL Introduction to Database (Q4 toppers list) [\[CERTIFICATE\]](#) [🔗](#)
- Awarded **Proficiency in Computer Science** in Class 12 [\[CERTIFICATE\]](#) [🔗](#)
- SOF Olympiads (top ~5% globally, qualified Level 2): **IEO Delhi Rank 19** [\[CERT\]](#) [🔗](#); **IMO Int. Rank 4033** [\[CERT\]](#) [🔗](#); **NSO Int. Rank 5065** [\[CERT\]](#) [🔗](#)