

ADITYA KAKADE

Mumbai, India | +91 9518538113 | adityaakakade@gmail.com | linkedin.com/in/adityaakakade
[Github](#) | [Leetcode](#) | [Portfolio](#)

SUMMARY

AI Engineer Intern building production Generative AI and LLM systems in **Python** and **FastAPI**: **RAG** pipelines with **embeddings** and **vector databases**, tool-using **AI agents** (LangGraph, MCP), **prompt engineering**, **REST APIs**, and Microsoft 365 automation, with quality measured through programmatic evals and expert-labelled benchmarks.

EXPERIENCE

Zycus Infotech Pvt Ltd.

Jun 2026 – Present

AI Engineer Intern

Mumbai, India

- Built **User Hub**, an internal access gateway (**Python**, **FastAPI**, **OAuth 2.0**) giving employees secure access to **100+ products** across client tenants, with user, client, and project mappings that grant access automatically, cutting manual configuration time by **40%**.
- Built **Zyra**, a **Microsoft Teams** IT self-service chatbot (**Copilot Studio**, **Power Automate**, **Entra ID**) automating **8 Microsoft 365 admin workflows**, including mailbox and distribution list management and end-to-end employee **onboarding and offboarding**.
- Built a **natural-language query layer** (**LangGraph** agent with **MCP** tools and structured prompts) over tenant, product, and workforce services, replacing manual navigation across internal systems.
- Developed the **FastAPI** backend and **Docker** deployment for **ITIS**, an internal issue-resolution portal, automating ticket routing and speeding up resolution by **35%**.

PROJECTS

LLM Strategy Agent with Eval Harness | *Python, FastAPI, React, Numba, Ollama*

[GitHub](#)

- Built a **tool-using LLM agent** that turns plain-English trading ideas into strategy code, with iteratively engineered prompts, static checks, a **subprocess sandbox**, a decline tool for impossible requests, and **rate-limit-aware failover** across providers.
- Designed a **39-case eval harness** that grades code on hidden data against reference implementations and detects look-ahead by perturbing future bars; **gpt-oss-120b passed 92%** vs. 28% for local Qwen2.5-Coder 7B.
- Built the underlying **event-driven backtesting engine** with a **Numba** fast path (**25,000 parameter combinations in under a second**) and **GitHub Actions CI** that verifies the eval grader.

MedRAG: Medical Document QA System (RAG) | *Python, FastAPI, LangChain, pgvector, React, Docker*

[GitHub](#)

- Built a full-stack **RAG** system answering medical questions grounded only in uploaded PDF, DOCX, and TXT files, with document preprocessing and chunking, **pgvector** (HNSW) search, **Tesseract OCR**, and streamed answers with cited sources.
- Evaluated on **PubMedQA** (1,000 expert-labelled questions): **98.0% top-1** and **99.4% top-4** retrieval accuracy, **76.9%** answer agreement with experts (human baseline 78%), 1.5 s median latency.
- Added a Gemini-to-Groq **LLM fallback chain** and a tuned similarity cutoff that rejects off-topic queries before any LLM call, wrongly refusing only **2 of 900** valid questions.

Adaptive LLM Inference Gateway | *Python, FastAPI, Redis, FAISS, Docker*

[GitHub](#)

- Building an OpenAI-compatible **LLM inference gateway** for my thesis on adaptive cost-quality-latency routing across open-weight LLM fleets, with request normalization so equivalent requests share one cache key.
- Implemented a two-tier cache: **SHA-256 exact match in Redis** backed by a **semantic cache** (BGE-small embeddings, **FAISS**) whose similarity scores are calibrated into hit probabilities with **isotonic regression**.
- (In progress) Adding **vLLM**-served Qwen2.5/Llama replicas, a GPU-time cost model, circuit breakers, and a **LinUCB** router.

TECHNICAL SKILLS

Languages: Python, SQL, C++, TypeScript

AI/ML & LLM: LangChain, LangGraph, sentence-transformers, Ollama, MCP, OpenAI, Anthropic, Gemini, and Groq APIs

Generative AI & NLP: Prompt Engineering, RAG (hybrid retrieval), Embeddings, Vector Databases, NLP, AI Agents, LLM Evaluation, Semantic Caching, LLM Routing and Fallback

Microsoft 365 & Automation: Copilot Studio, Power Automate, Microsoft Teams, Entra ID (Azure AD), Exchange admin workflows

Backend & Frameworks: FastAPI, REST APIs, asyncio, Pydantic, SQLAlchemy, Pandas, Numba, React, pytest

Databases & Infrastructure: PostgreSQL (pgvector), Redis, ChromaDB, FAISS, Docker, AWS (S3), Git, GitHub, GitHub Actions

Core CS: Data Structures & Algorithms, Object-Oriented Programming, REST API Design, System Design

EDUCATION

Vellore Institute of Technology (VIT), Bhopal

2022 – 2027 (Expected)

Integrated M.Tech in Artificial Intelligence; CGPA: 7.76