Rishabh Kumar

+91 9921557109 | irishabh.2311@gmail.com | linkedin.com/in/rishabh-kumar | github.com/Rishabh23112 | Pune, Maharashtra

Technical Skills

Languages & Frameworks: Java, Python, FastAPI, SQL

GenAI Stack: LLMs, RAG, LangChain, Prompt Engineering, Vector DBs (Qdrant, ChromaDB)

Projects

MentalHealth RAG API Nov 2025 – Dec 2025
FastAPI, Qdrant, Gemini API, MongoDB *(GitHub)*

- Developed a production-grade mental health API using **Retrieval Augmented Generation (RAG)** to ground responses in verified clinical PDF literature.
- Integrated **Qdrant** for high-dimensional embedding management and **MongoDB** for persistent session history, enabling long-context retention.
- Implemented a **Crisis Detection Protocol** that overrides LLM generation to trigger hard-coded safety responses when emergency intent is detected.

RepoChat Oct 2025 – Nov 2025
FastAPI, React.js, LangChain, ChromaDB, Gemini API *(GitHub | Live Demo)*

- Architected a full-stack RAG application that queries live GitHub repositories using **Gemini 2.5 Flash** and **ChromaDB** for code-aware vector retrieval.
- Engineered a recursive file ingestion pipeline to clone, parse, and embed repository structures, preserving file context for accurate semantic search.
- Reduced hallucination and improved answer relevance through optimized chunking + contextual windowing in the retrieval pipeline.

OpenRouter CLI Sept 2025 – Sept 2025
Python, OpenRouter API *(GitHub)*

- Developed a lightweight CLI client to query frontier models (GPT-4o, Claude 3.5, Gemini, Llama) via OpenRouter using direct HTTP requests for low-latency inference.
- Implemented **10+ flags** for model selection, verbose mode, token usage, glow output, file piping, and session persistence.
- Added **robust error-handling** for API failures, rate limits, and malformed inputs, ensuring stable developer workflows.

Experience

AI Research Intern Feb 2025 – Apr 2025
Planto.ai *Remote*

- Benchmarked Codium Copilot and other LLM coding assistants across 4 client environments, generating measurable ROI insights and actionable adoption recommendations.
- Automated research documentation and experiment workflows for a 10-member team, reducing manual overhead by 20% and improving evaluation consistency.
- Delivered 8+ technical briefs translating complex LLM performance data into clear, decision-ready insights for engineering and leadership stakeholders.

Education

Vellore Institute of Technology CGPA - 8.25/10.0
B-Tech in Computer Science and Engineering *Aug 2022 – Sept 2026*

Achievements & Extra Curricular Activities

- **Open Source Contributor**: Merged PR to **terminal-ai (100LinesOfAICode)** enhancing CLI stability with API validation, request timeouts, and automated retry logic.
- Ranked **Top 500** in the HackOnBlocks Web3 Hackathon, competing among 3,000+ participants.