Rahul Talepa

India

+91 7014164372 • talepa.rahul6@gmail.com • LinkedIn • GitHub

AI Engineer with hands-on experience building production-grade LLM applications and Retrieval-Augmented Generation (RAG) pipelines using LangChain and LangGraph. Skilled in semantic search, vector databases, embeddings, prompt engineering, and tool calling, delivering accurate retrieval across large-scale datasets. Experienced in multimodal computer vision search, multi-agent architectures, and deploying scalable AI services with FastAPI and Docker.

Skills

- **Generative AI & LLMs:** LangChain, LangGraph, AI Agents, Agentic AI, Retrieval-Augmented Generation (RAG), Prompt Engineering, Tool/Function Calling, Structured Output Generation
- **LLM Systems:** Semantic Search, Vector Databases, Embedding Models, Hybrid Search, Metadata Filtering, Retrieval Evaluation, Document Processing
- **Programming & Backend:** Python, SQL, JavaScript, FastAPI, REST APIs
- **Computer Vision:** Transformers, Vision Transformers (ViT), Image Search, Similarity Matching
- **Tools & Platforms:** PostgreSQL, MongoDB, Docker, Git, GitHub, Linux, Jupyter Notebook

Professional Experience

Stuvio Digital — **India**

Junior AI Engineer — *Feb 2026 - Present*

- Engineered production-grade LLM applications and scalable RAG pipelines using LangChain, semantic search, vector databases, embeddings, prompt engineering, and tool calling to deliver accurate retrieval across lakhs of records
- Developed an AI-powered multimodal image search system using Transformer-based Computer Vision, vector similarity search, metadata filtering, and retrieval optimization
- Built automated data ingestion, preprocessing, metadata enrichment, indexing, and REST API integration pipelines using Python, FastAPI, and Docker
- Collaborated with backend teams to deploy scalable AI services to production

Independent Projects

AI Revenue Recovery Agent – Agentic B2B Invoice Recovery Platform — *GitHub*

- Built an agentic invoice-recovery workflow with LangGraph and Google Gemini, where the LLM recommends actions and a deterministic policy engine gates every execution — enforcing reminder caps, cooldowns, and forced escalation rules
- Engineered a FastAPI and PostgreSQL backend with an XGBoost risk-scoring model, Kafka event streaming, and Redis-based idempotency locks, backed by 120+ automated tests and CI/CD via GitHub Actions
- Integrated real third-party APIs (Google Gemini, Sarvam speech-to-text/text-to-speech, Resend email) behind swappable provider abstractions with automatic fallback, plus a Next.js and TypeScript dashboard for live case monitoring

Atelier – Research Intelligence Platform — *GitHub*

- Built a multi-agent research platform with LangGraph orchestrating a Director → Specialists → Evidence Analyst → Synthesizer pipeline that decomposes technical questions into investigation plans and produces cited reports
- Implemented LangGraph interrupt() for human-in-the-loop mid-research decisions, checkpointed session state via PostgreSQL, and real-time SSE streaming of phase progress to a Next.js frontend
- Director agent uses Gemini structured output with deterministic tool budgets by depth (quick/standard/deep) and rule-based fallback; source quality filtering and citation validation reject hallucinated references

Education

D Y Patil University — **Pune, India**

B.Tech, Computer Science Engineering — *2021-2025*

CGPA: 8.33