

ADVAIT RAKTATE

advrakte@gmail.com | +91 8433593640 | [LinkedIn](#) | [GitHub](#)

2026 B.Tech Graduate (CS – AI & Analytics) | 6 months production AI engineering experience

SUMMARY

AI Engineer specializing in agentic LLM systems and Retrieval-Augmented Generation (RAG), with production experience building multi agent orchestration, realtime conversational AI, and async backend services. Skilled in LangChain, LangGraph, FastAPI, Azure OpenAI, and hybrid search (keyword + vector) retrieval pipelines, with a focus on scalable architecture, reliability, and measurable engineering impact.

TECHNICAL SKILLS

Languages: Python, SQL, Java, C++, HTML, CSS, JavaScript

LLM & Agentic AI: LangChain, LangGraph, MCP (FastMCP), RAG, Vector Search, Azure AI Search, BM25, pgvector, Semantic Ranking, Reranking, Prompt Engineering, Multi Agent Orchestration, LoRA/QLoRA

Evaluation & Observability: RAGAS, LangSmith, LLM-as-a-Judge

Backend & Infra: FastAPI, Async Python, WebSockets, SSE, SQLAlchemy, Alembic, PostgreSQL, Redis, Celery, Docker, JWT/OAuth2, RBAC, REST APIs, CI/CD

ML/DL & Tools: PyTorch, TensorFlow, Scikit-learn, MLflow, NLP, PySpark

Core CS: Object-Oriented Programming (OOP), Data Structures & Algorithms, Multithreading, Operating Systems, DBMS, Git Fundamentals

Cloud: AWS, Azure OpenAI/Foundry, Azure AI Search, Docker, Kubernetes

EXPERIENCE

GenAI Intern | Arcitech.ai

Dec 2025 - May 2026

AGBrain — AI-Powered Employee Performance Evaluation Platform

- Developed a realtime conversational AI platform using the OpenAI Realtime API and WebSockets, reducing manual HR review effort by ~60% across performance evaluation cycles; used MCP (FastMCP) to expose organizational data and tools for agent tool calling, so the agent could dynamically retrieve information mid conversation.
- Redesigned the evaluation agent from direct OpenAI API calls to a LangGraph based multi agent architecture, replacing a single expanding prompt handler with 5 modular per-feature agents running in parallel for maintainability, with an OpenAI-to-Gemini fallback for provider outages.
- Designed async, parallel agent graphs to concurrently retrieve and merge employee data from Zoho tickets and PostgreSQL across 100+ employees, cutting response time by 5-6 seconds per evaluation.

BPOBox — Call Center Analytics Platform

- Created an async call ingestion pipeline using AssemblyAI for transcription and Gemini for sentiment analysis, compliance flagging, and summarization, without blocking downstream workflows.
- Designed an evaluation pipeline combining formula based KPI metrics across 8-10 KPIs with LLM-as-a-Judge scoring criteria, translating raw call transcripts into structured performance scores exposed through REST APIs that orchestrate retrieval and LLM calls with built-in observability.
- Managed data persistence with SQLAlchemy across 10-15 tables in relational databases and AWS object storage for call recordings, as one of two AI engineers owning BPOBox's AI processing pipeline; worked within a shared CI/CD pipeline across AGBrain and BPOBox with automated security checks and Alembic migration validation.

PROJECTS

DocuMind — Production RAG Pipeline ([docu-mind.tech](#)) ([Github](#))

Stack: LangChain, FastAPI, Azure OpenAI, pgvector, Cohere, OpenRouter

- Built DocuMind, a production RAG pipeline with hybrid search (pgvector dense + BM25 sparse + RRF fusion) and semantic reranking that grounds LLM responses in source documents with citation-based answers, reducing hallucinations and achieving 0.962 faithfulness and 0.966 answer relevancy on RAGAS evaluation.
- Implemented SSE streaming, session-isolated multi-user architecture, per-query cost tracking (\$0.00115/query), and LangSmith observability with full latency breakdown across three A/B pipeline variants.
- Deployed on Heroku with Cohere reranking API, eliminating a 500MB torch dependency from the production slug, and shipped a live document Q&A demo at docu-mind.tech.

AutoMed AI — Autonomous Multi-Agent Medical Image Classification Platform ([Github](#))

- Built an end-to-end multi agent pipeline with LangGraph, FastAPI, and PyTorch, automating dataset inspection, GAN-based data augmentation, and adaptive model training, with a Next.js dashboard streaming agent logs over WebSockets.

EDUCATION

B.Tech, Computer Science (AI & Analytics) | MIT ADT University, Pune

2022 - 2026

Coursework: Generative AI w/ LLMs (AWS), Agentic AI & RAG (IBM), Deep Learning - IIT Ropar (NPTEL), ML Specialization (Stanford)

ACHIEVEMENTS

- Placed Top 4 at UST's D3CODE Hackathon out of 20 thousand participants
- Co-authored "RASEED: AI-Powered Personal Finance Manager & Advisor" accepted at BRAICON 2026, MDIS Singapore.