Afnan PK

Agentic AI Engineer

GenAI Engineer | LangGraph, RAG, Multi-Agent Systems | Production AI, not prototypes · $25/hr · Available

Kerala, India · linkedin.com/in/afnan-pk

SKILLS

Python, FastAPI, LangChain/LangGraph, RAG pipeline design, vector databases (Qdrant; familiar with ChromaDB/Pinecone architecture), prompt engineering, multi-agent orchestration, hybrid vector retrieval, semantic caching, AWS, Docker, CI/CD, REST API integration, production observability and cost optimization, PostgreSQL, Redis, BullMQ

ABOUT

I build AI systems that hold up under real traffic, not demos that work once. Most recently I architected a multi-tenant agentic platform at Wiral AI: LangGraph-based agent orchestration, RAG pipelines with Qdrant, async processing with Redis/BullMQ, and observability that cut model costs by 40-50% while serving 300-400 daily live customer interactions. I care about the parts most people skip: evaluation, fallback logic, structured validation, and what actually breaks at scale. Backend-first, Python and Node.js, comfortable owning a system end to end from prototype to production.

EXPERIENCE

Agentic AI Engineer · Wiral AI

Jul 2025 – Present

Doha, Qatar · full time

- Engineered a production AI engagement platform for WhatsApp and web channels, supporting paying customers
- and approximately 300–400 daily customer interactions across sales and support workflows.
- Helped shape a multi-tenant AI platform with isolated knowledge bases, prompts, agent behavior, and runtime
- configurations for multiple client environments.
- Orchestrated LangGraph-based multi-agent workflows for enquiry handling, booking, and back-office operations,
- while reducing user-facing response latency by moving non-critical actions to post-response processing.
- Enhanced the RAG stack through knowledge-base ingestion, web content extraction, hybrid vector retrieval, and
- tenant-level data isolation using Qdrant.
- Established asynchronous backend services with Node.js, Redis, and BullMQ for reliable webhook handling,
- retries, and queue-based processing.
- Connected Chatwoot and Evolution API workflows to automate WhatsApp follow-ups, lead capture, and booking
- journeys.
- Introduced monitoring and diagnostics for conversation replay, retrieval tracing, token usage, and cost visibility,
- contributing to a 40–50% reduction in model cost through caching and model-selection optimizations.

Backend Engineer · WebMavericks Softcoders

Jan 2025 – Jul 2025

Remote · full time

Created and maintained Apache Airflow ETL workflows integrating data from 50+ external platforms across marketing, e-commerce, and business systems.

• Engineered Python-based ingestion and transformation pipelines using Airflow, AWS S3, and Redshift for analytics and reporting workloads.

• Maintained containerized data processing services using Docker to ensure scalable and repeatable pipeline execution.

• Delivered REST APIs within a Django/PostgreSQL operations platform, enabling secure multi-role workflows through role-based access control.

Software Engineer · Bridgeon Solutions

Mar 2024 – Jan 2025

Kerala, India · full time

- Contributed to full-stack application development using Django backend services and React-based frontend
- interfaces.
- Integrated third-party APIs and internal services to streamline application workflows and improve data exchange.
- Resolved production bugs and optimized CI/GitHub workflows to improve system stability and release reliability.

EDUCATION

INDIRA GANDHI NATIONAL OPEN UNIVERSITY · Bachelor's, Computer Science

Jun 2024 – Jun 2027

Delhi, India

PROJECTS

Wiral - AI Sales Agent

wiral.ai

LangGraph · Node.js · Redis · BullMQ · Qdrant · Chatwoot · Evolution API

Architected and delivered a production conversational AI platform supporting business interactions across WhatsApp and web channels. The system replaced a limited single-tenant chatbot with a scalable multi-tenant, multi-agent architecture designed for reliability and operational visibility.

• Architected a multi-tenant AI platform enabling each client to manage independent prompts, knowledge bases, and agent configurations without redeployment.

• Structured a LangGraph-based multi-agent orchestration layer coordinating enquiry handling, booking workflows, and operational assistance.

• Built a production knowledge retrieval system with hybrid vector search, document processing pipelines, and tenant-isolated indexes using Qdrant.

• Designed an asynchronous webhook processing architecture using Express, Redis, and BullMQ to reliably process Chatwoot events with retries and queue-based workers.

• Integrated WhatsApp automation through Evolution API, enabling AI-driven customer interaction and automated follow-up flows.

• Introduced semantic caching and operational diagnostics to analyze retrieval behavior, latency, token usage, and model cost efficiency.

• Created internal operations tooling for knowledge-base management, tenant configuration, and AI system debugging in production.

Enterprise ETL Pipeline (Apache Airflow + AWS)

Confidential(Company Product)

Python · Apache Airflow · AWS S3 · Redshift · Docker

Designed and implemented ETL workflows to collect, transform, and load data from multiple external systems into a centralized analytics environment. The solution supported automated reporting and downstream business intelligence use cases.

• Built Apache Airflow DAGs to orchestrate data extraction, transformation, and loading from 50+ external platforms.

• Developed Python-based ingestion pipelines for automated and scheduled data movement into AWS-backed storage and warehouse systems.

• Used AWS S3 and Redshift as part of a centralized analytics architecture for reporting and data consolidation.

• Containerized pipeline components with Docker to improve portability, deployment consistency, and execution reliability.

• Structured workflows to support maintainability, repeatable scheduling, and scalable processing across multiple sources.