Contact

sawairaasghar05@gmail.com

www.linkedin.com/in/sawaira-goraya-ai (LinkedIn)

Top Skills

Retrieval-Augmented Generation (RAG)

Agentic AI / AI Agents

LangChain & LangGraph

Certifications

Fundamentals of Prompt Engineering Course: ChatGPT Prompt Engineering for Developers

Data Handling and Processing Course: Preprocessing Unstructured Data for LLM Apps

Introduction to API Wrappers Course: Getting Started with Generative AI API Specialization

Multi AI Agent Systems with crewAI

Build Autonomous AI Agents From Scratch With Python

Sawaira Asghar

Agentic AI | Python, RAG, LangGraph, LangChain, AWS | Building Production Multi-Agent Systems with Optimized LLM Performance

Punjab, Pakistan

Summary

Hi, I'm Sawaira Asghar – Production AI Engineer & RAG Specialist

Currently building battle-tested, multi-agent AI systems at Mid-Chain Technologies that serve 10,000+ requests daily with <300ms latency.

What I do every single day:
‣ Design & deploy end-to-end RAG pipelines (FAISS, Pinecone, pgvector + Sentence-Transformers + Cohere reranker) → 50–65% higher retrieval accuracy & drastic hallucination reduction
‣ Architect scalable agentic workflows using LangGraph & LangChain that autonomously reason, retrieve, and act across tools
‣ Ship production microservices with Python, FastAPI, Django, Docker, AWS Lambda, ECS, Bedrock, SageMaker & OpenSearch
‣ Optimize Llama 3 70B, Mixtral, Claude 3.5, and GPT-4o inference using vLLM, TensorRT-LLM, and quantization → cut costs 40–60% while keeping quality high
‣ Own prompt engineering, chain-of-thought, and self-correction loops that consistently improve ROUGE/BERTScore by 18–25%
‣ Deliver real client impact: medical assistants, intelligent invoice OCR + automation, HR policy agents, resume parsers, and personalized recommendation engines

Fast career trajectory:
Intern → Full-time Junior AI Engineer → Promoted to AI Engineer in just 3 months (Oct 2024 – Jan 2025)

2025 Computer Science graduate from The Islamia University of Bahawalpur – graduated with hands-on experience and multiple shipped AI products.

Certifications & Continuous Learning:
‣ Advanced Retrieval-Augmented Generation (DeepLearning.AI)

- LangChain & LangGraph for Agentic AI
- Generative AI with LLMs (Coursera + AWS)
- Prompt Engineering & LLM Optimization

Currently obsessed with:
→ Multi-modal RAG | Agent memory & long-term reasoning | MLOps for LLMs | Cost-efficient inference at scale | Self-improving agents

Open to new challenges (remote or relocation) in 2025–2026 as:
AI Engineer | LLM Engineer | RAG Engineer | Generative AI Engineer | Machine Learning Engineer (L3–L5 equivalent)

If you're hiring for teams working on Gemini, Llama, Claude, Grok, DeepSeek, or any large-scale production AI system – let's talk.

I bring proven production experience, clean code, obsession with latency & cost, and the ability to ship reliable AI that real users depend on.

Let's connect and build the next generation of intelligent systems together

GitHub: github.com/Sawaira-0316
Portfolio & case studies: DM me for RAG cheat sheets, LangGraph templates, or live demos
Open to Work | Actively interviewing | Ready to relocate

#AIEngineer #RAG #LangGraph #LLMOps #GenerativeAI #AgenticAI #Python

---

Experience

Mid-Chain Technologies
1 year 3 months

AI Engineer
January 2025 - Present (1 year)
Remote

‣ Building production-grade multi-agent AI systems using LangGraph, LangGraph, Llama 3, GPT-4o, and Claude 3.5 that serve 10,000+ requests/day

‣ Designed and deployed end-to-end RAG pipelines (FAISS + Sentence-Transformers + Cohere reranker) → improved retrieval accuracy 50%+ and reduced hallucinations by 65% in medical & financial use cases

‣ Engineered scalable AI microservices in Python + Django + FastAPI on AWS (Lambda, ECS, Bedrock) handling 1,000+ concurrent requests with <300 ms p95 latency

‣ Led prompt engineering and chain-of-thought optimization strategies that cut token usage 40% while boosting answer quality (ROUGE/BERTScore ↑ 18–22%)

‣ Integrated and fine-tuned open-source models (Llama 3 70B, Mixtral, Mistral-7B) via Hugging Face, vLLM, and TensorRT-LLM for cost-efficient inference

‣ Delivered client-facing AI products:

• Multi-modal medical assistant (RAG + vision)

• Intelligent invoice automation (OCR + agentic workflow)

• Personalized recommender engine using embeddings + collaborative filtering

‣ Owned CI/CD for AI services (GitHub Actions + Docker + Terraform) and monitoring with LangSmith, Prometheus & Grafana

‣ Mentored 2 interns on LangChain/LangGraph best practices and production deployment

Technologies: Python, LangChain, LangGraph, LlamaIndex, FAISS, Pinecone, Sentence-Transformers, AWS (PostgreSQL + pgvector), AWS (Bedrock, Lambda, SageMaker, OpenSearch), Docker, FastAPI, Django, Git

AI Intern (

October 2024 - December 2024 (3 months)

Lahore, Punjab, Pakistan

‣ Built and shipped first-generation RAG prototypes and semantic search engines from scratch

‣ Developed classification, clustering, and embedding models that became core components of production systems

‣ Created REST APIs in Python + Django consumed by web and mobile frontends

‣ Collaborated directly with CTO and senior engineers on client deliverables that led to immediate full-time offer

Education

The Islamia University of Bahawalpur

Completed, Computer Science · (August 2021 - January 2025)