Hasnain Ali

AI/ML Engineer

codingwithhasnain@gmail.com || https://www.linkedin.com/in/hasnainali3/ || +923135085477 || Pakistan

Profile

A results-driven AI/ML Engineer with over 3 years of experience specializing in building and deploying sophisticated AI agents, Large Language Models (LLMs), and Retrieval-Augmented Generation (RAG) systems. Expertise in fine-tuning LLMs (LoRA, QLoRA) in multi-GPU environments for high-profile clients. Proven ability to architect complex, agentic workflows with LangGraph and deliver scalable, production-ready AI solutions that solve critical business problems.

Skills & Technologies

- **AI Agents & LLMs:** Agentic AI, LangGraph, Retrieval-Augmented Generation (RAG), LLM Fine-Tuning (LoRA, QLoRA), LangChain, Transformers, Hugging Face, OpenAI SDK, Gemini API
- **ML & Data Science:** PyTorch, TensorFlow, Scikit-learn, NLP (spaCy, NLTK), OCR, Computer Vision
- **Data Stores & Vector DBs:** PostgreSQL, MongoDB, Vector DBs (ChromaDB, Pinecone), Graph Databases (Neo4j)
- **Cloud & MLOps:** AWS (SageMaker, Bedrock, S3, EC2 GPU Instances), GCP, CI/CD, Docker, Multi-GPU Environments
- **Backend & Data:** Python (Django, Flask), RESTful APIs
- **Tools & Collaboration:** Git, GitHub, Jira, React.js

Experience

Software Engineer | DEVSINC | OCT 2024 – PRESENT

- Lead engineer on the **"Smart Advocate"** project for a US-based client, developing a state-of-the-art document intelligence system on client-provisioned multi-GPU virtual machines.
- Fine-tuned a custom Large Language Model using **LoRA** for specialized legal information extraction, achieving superior performance on domain-specific tasks.
- Engineered complex, **cyclical agentic workflows using LangGraph** to automate chronological data analysis and construct dynamic knowledge graphs in Neo4j.
- Architected and optimized enterprise-grade **Retrieval-Augmented Generation (RAG)** systems, improving information retrieval accuracy by 40% for core NLP applications.

Associate Software Engineer | AMROOD LABS | MAR 2024 – OCT 2024

- Developed and optimized the primary Django backend for a mobile application, integrating PostgreSQL and AWS S3 for efficient, scalable data storage and retrieval.
- Integrated OpenAI models to automate contract analysis and response generation, cutting manual processing time for legal documents by 60%.

Founder & Lead Engineer | 2ndPlace *(Jan 2023 - Feb 2024)*

- Designed and deployed an AI-powered search system that integrated geolocation and user preferences, reducing average search time by 80%.
- Built the full-stack application on a Django and React stack and managed its deployment via a

CI/CD pipeline to AWS.

Projects

Smart Advocate

Tech Stack: Python, LangGraph, LoRA, PyTorch, OpenAI, MCP, OCR, Docker, Local GPU Instances, VMs, TFS

- **Custom LLM Fine-Tuning:** Led the end-to-end fine-tuning of a specialized LLM on a multi-GPU server to master the domain of legal document analysis. This involved curating a high-quality dataset and implementing LoRA for efficient training, resulting in a model that adeptly extracts complex entities and clauses.
- **Agentic Workflow for Document Processing:** Designed and implemented a multi-agent system using LangGraph with GuardRails where distinct agents handled OCR, document classification, and information extraction. This modular, stateful approach enabled the creation of detailed event chronologies from unstructured text with high accuracy.
- **Token-Efficient Classification:** Engineered a novel classification system that categorized a high volume of documents across dozens of classes using a minimal number of tokens, significantly reducing API costs and processing latency.

ConvoPilot | Custom RAG Chatbot Framework

Tech Stack: Python, LangChain, PostgreSQL, Vector DB (Pinecone), OpenAI API, Gemini API

- Developed a comprehensive framework for building and evaluating Retrieval-Augmented Generation (**RAG**) systems, focusing on user-controlled performance and security.
- Designed a scalable, multi-tenant architecture to ensure data isolation and optimized the retrieval pipeline, significantly reducing search latency.
- Implemented a control interface for A/B testing different LLMs (OpenAI, Gemini) and vectorization models, allowing for systematic evaluation of cost, speed, and accuracy trade-offs.
- Integrated NLP modules for **PII** detection to **redact sensitive data** and intent detection to route queries efficiently.

EDUCATION

Bachelors in Computer Science

Comsats University Islamabad (CUI) | CGPA — 3.33