Muhammad Kashif

AI Engineer | LLM Applications | Agentic AI | RAG Systems

+92-312-1502390 | kashifmayar7@gmail.com | linkedin.com/in/muhammad-kashif-867631224 | github.com/kashifmayar

Professional Summary

AI Engineer with experience building LLM powered applications, Agentic AI systems, RAG pipelines, and backend services. Skilled in Python, FastAPI, LangChain, OpenAI, Claude, Ollama, vector databases, prompt engineering, and production ready AI applications.

Technical Skills

Programming: Python, SQL, C++, JavaScript, Node.js, MATLAB
LLM & Generative AI: OpenAI API, Claude API, Ollama, LLM Applications, Prompt Engineering, Function Calling, Tool Calling, Structured Outputs, AI Agents, Agentic Workflows, RAG Pipelines, Embeddings, Semantic Search
AI Frameworks: LangChain, Hugging Face Transformers, PyTorch, TensorFlow, Scikit-learn, Keras
Vector Databases: FAISS, Pinecone, ChromaDB
Backend & APIs: FastAPI, Flask, REST APIs, Microservices, API Integration, Pydantic, Node.js
Databases & Cloud: PostgreSQL, Supabase, SQL, AWS
MLOps & Deployment: Docker, Git, GitHub, CI/CD, Model Deployment, Logging, Monitoring, Pipeline Automation
Data & NLP: NumPy, Pandas, ETL Pipelines, Data Cleaning, NLP, Sentiment Analysis, Intent Detection, Document Parsing, Summarization
Speech & Audio AI: Speech-to-Text, Text-to-Speech, ElevenLabs API, Audio AI Integrations
Computer Vision: YOLOv8, CNNs, OpenCV, Transfer Learning, Medical Image Classification

Experience

Markhor Systems Islamabad, Pakistan
AI Engineer *Feb 2026 – June 2026*

- Built LLM-powered workflows integrating OpenAI, Claude, Ollama, and external APIs for business automation.
- Developed RAG systems using embeddings and VectorDB(FAISS/Pinecone/ChromaDB) for semantic retrieval and Q&A.
- Designed FastAPI-based backend services and REST APIs for production AI applications.
- Used Docker, AWS, Git, and CI/CD to deploy and maintain scalable AI solutions.

Projects

Agentic RAG Chatbot System *FastAPI, LangChain, FAISS, Ollama*

- Built an Agentic RAG chatbot capable of retrieving information from custom knowledge bases and generating context aware answers.
- Implemented document ingestion, chunking, embedding generation, semantic retrieval, prompt engineering and LLM response generation.
- Integrated tool using workflows where the LLM could interact with external APIs and structured backend logic.
- Used FAISS for local vector search and optimized retrieval quality through chunking and prompt refinement.

AI Document Parsing and Automation Pipeline *FastAPI, Ollama, Supabase, CI/CD*

- Developed an AI document processing pipeline to extract, chunk, summarize, and organize documents automatically.
- Built backend APIs using FastAPI and integrated Ollama based LLM inference for document understanding and summarization.
- Used Supabase for data storage and workflow management, enabling searchable and reusable document intelligence.
- Applied pipeline automation and deployment practices to improve repeatability and maintainability.

RoohAI – Emotional Digital Twin *Python, LLM APIs, NLP, Sentiment Analysis*

- Worked on an AI powered system involving intent detection, sentiment analysis, and LLM-based response generation.
- Integrated speech and language capabilities including STT, TTS, and ElevenLabs API for improved user interaction.
- Applied prompt engineering and API orchestration to support natural, context aware conversations.

FabrIQ – Fabric Defect Detection System *Python, PyTorch, YOLOv8*

- Built a real time computer vision system using YOLO to detect and classify fabric defects for textile quality inspection.
- Handled model training, dataset preparation, evaluation, and system testing.
- Designed the solution to support automated inspection and reduce manual quality checking effort in textile production.

Education

FAST National University of Computer and Emerging Sciences Islamabad, Pakistan
B.S. Artificial Intelligence *Aug 2022 – Jun 2026*
Tufail Shaheed Army College Mardan, Pakistan
F.Sc. Pre-Engineering *2019 – 2021*

Certifications

- Machine Learning Specialization – DeepLearning.AI and Stanford University
- Google Prompting Essentials Specialization – Google
- Google IT Automation with Python – Google
- Introduction to Generative AI – Google Cloud
- Programming for Everybody – University of Michigan