

LAIBA KHAWAR

AI Engineer | LLM & NLP Systems | Backend & Data Engineering

Islamabad, Pakistan | +92 305 3335757 | laibakhawar7@gmail.com

linkedin.com/in/laiba-k | github.com/LaibaKhawar

PROFESSIONAL SUMMARY

AI Engineer building and deploying production LLM and NLP systems, including retrieval-augmented generation (RAG), automated model evaluation, graph-based memory, and semantic search. Takes models from integration and fine-tuning through to scalable Python services, data pipelines, and cloud deployment.

TECHNICAL SKILLS

AI & NLP: LLM pipelines, RAG, semantic search, fine-tuning, summarization, concept extraction, natural language inference (NLI), LLM-as-a-judge evaluation, multilingual NLP

Frameworks & Models: LangChain, Hugging Face, PyTorch, Sentence Transformers, FAISS, Ollama, Mistral, LLaMA, Scikit-learn

Backend & APIs: Python, FastAPI, Flask, REST APIs, Streamlit, ETL pipelines, Pandas, NumPy, Spark

Databases: PostgreSQL, pgvector, SQL Server, MySQL, MongoDB, Neo4j, Memgraph

Cloud & DevOps: Azure AI, AWS (EC2, S3, Lambda, Glue, Redshift, API Gateway), Docker, Git, GitHub

PROFESSIONAL EXPERIENCE

AI Engineer | SysReforms International

Feb 2026 – Present | Islamabad, Pakistan

- Build and optimize PVS-LLM, a production AI/NLP platform that converts unstructured veterinary and public-health reports into structured, validated insights.
- Develop production LLM pipelines for summarization, question answering, concept extraction, semantic mapping, and learning-needs analysis using Mistral, Ollama, Azure AI, Sentence Transformers, and fine-tuned models.
- Design automated model-evaluation and QA frameworks combining semantic similarity, NLI, BERT-based metrics, and LLM-as-a-judge scoring, and improve throughput via batching, parallelism, model preloading, and resumable checkpointed workflows.
- Build multilingual FastAPI and Streamlit applications for conversational AI, evaluation, reporting, and human review, backed by PostgreSQL, SQL Server, Docker, and cloud-hosted inference.

AI Backend Engineer | Dataspecc

Jan 2025 – Jan 2026 | Islamabad, Pakistan

- Built Hyperengage's Customer Memory Graph, an AI infrastructure layer using Neo4j and Memgraph to model accounts, contacts, commitments, risks, deals, and time-bounded customer relationships.
- Implemented hybrid retrieval combining pgvector semantic search with graph traversal, and developed LLM-powered services that generated and maintained contextual memory summaries for customer accounts.
- Built data pipelines integrating HubSpot, Salesforce, Segment, Intercom, and Zendesk into a unified graph model, and contributed backend functionality to the Signals Engine and Copilot Q&A for real-time pattern detection and natural-language querying.

Teaching Assistant – NLP and Database Systems | FAST-NUCES

Feb 2025 – Jun 2025 | Islamabad, Pakistan

- Supported lectures, lab sessions, and assessment for undergraduate NLP and Database Systems courses.

Machine Learning Intern | Devsort

Jun 2024 – Aug 2024 | Islamabad, Pakistan

- Contributed to development, testing, and deployment of ML solutions across data collection, cleaning, preprocessing, exploratory analysis, model development, and evaluation.
- Used Python, Pandas, NumPy, Scikit-learn, Matplotlib, and Git to surface patterns and data-quality issues and communicate results through visualizations and reports.

PROJECTS

Hybrid Rule-Based and LLM-Assisted EDI Mapping System (Final Year Project) – Python, LLaMA 3.1. Mapped Tradacoms, X12, and HL7 formats using rule-based logic for standard cases, LLM assistance for complex mappings, and automated syntax and consistency validation.

Conversational Retrieval-Augmented Generation Chatbot – Python, LLaMA, FAISS, LangChain. Built chunking, embeddings, vector search, reranking, conversation memory, and query rewriting for grounded multi-turn question answering.

AI Agent for Research Paper Summarization and QA – Python, LLM Agents, RAG. Built a tool-calling agent for PDF parsing, semantic search, structured summarization, and retrieval-augmented QA.

Computer Vision: Zero-Shot Recognition, Captioning, and Real-Time Detection – PyTorch, Hugging Face, CLIP, OpenCV. Built CLIP-based zero-shot recognition with transformer captioning, plus CNN face-mask and emotion models optimized for low-latency webcam inference.

EDUCATION & CERTIFICATIONS

BS Data Science | FAST-NUCES, Islamabad – 2021–2025. Dean's List: Spring 2023, Fall 2024

A Levels (Engineering) | Beaconhouse Margalla Campus, Islamabad

Certifications: LangChain: Chat with Your Data | AWS Web Application Builder | AWS Data Pipeline Builder | DataCamp: Understanding Data Visualization | DataCamp: Intro to Data Science in Python