Zaid Rahim

Lahore, Pakistan | +92 343 9863954 | zaidrahim162@gmail.com
https://linkedin.com/in/zaid-rahim-zaid | https://github.com/Zaid-Rahim

Professional Summary

AI/ML and Software Engineer with hands-on experience building end-to-end AI applications, RAG pipelines, agentic workflows, LLM integrations, and Python backend services. Built a 5,000+ document legal RAG system with FAISS, LangChain, LangGraph, OpenAI/Gemini APIs, and a FastAPI inference backend. Also experienced in LLM fine-tuning with LoRA/QLoRA, semantic search, and model development with PyTorch. Strong software engineering foundation across Python, REST APIs, databases, Git, and full SDLC workflows.

Technical Skills

Programming & Backend: Python, OOP, FastAPI, REST APIs, Async Programming, API Integration
LLMs & Generative AI: Hugging Face Transformers, LLM Fine-Tuning, LoRA/QLoRA, PEFT, TRL, Prompt Engineering, OpenAI API, Gemini API, Claude API
RAG & AI Agents: RAG Pipelines, LangChain, LangGraph, Agentic Workflows, Semantic Search, FAISS, Sentence Transformers, Knowledge Graphs
ML & Deep Learning: PyTorch, NumPy, Pandas, scikit-learn, Neural Networks, CNNs, Transformer Architecture, Model Training & Evaluation
Databases: PostgreSQL, MySQL, MongoDB, Oracle
Tools & Deployment: Git, Docker, Linux (Ubuntu), Streamlit, Kaggle GPU Environments, Postman, VS Code

Professional Experience

Medic Patron LLP Islamabad, Pakistan
Backend Engineer Intern Jun 2023 – Aug 2023

- Designed backend architecture, database schemas, and REST API integrations for a medical portal using Python and FastAPI.
- Delivered features across the software development lifecycle while collaborating with cross-functional teams.
- Authored technical documentation for system architecture and API interactions to support reproducibility and team handoffs.

FAST National University of Computer and Emerging Sciences Islamabad, Pakistan
Teaching Assistant – Operating Systems & Intro to CS Feb 2024 – Jun 2024

- Guided 30+ undergraduate students through systems programming, debugging, and algorithmic problem-solving.
- Evaluated programming assignments and led exam preparation sessions, communicating technical concepts clearly.
- Coordinated with faculty stakeholders to manage student workflows and course activities.

Selected Projects

AI-Powered Legal Research Assistant 2025 – 2026
Python, LangChain, LangGraph, RAG, FAISS, FastAPI, LLM APIs

- Built an end-to-end RAG pipeline over 5,000+ legal documents using FAISS vector embeddings, semantic retrieval, knowledge graph integration, and grounded response generation.
- Designed multi-step agentic workflows with LangGraph and integrated OpenAI and Gemini APIs with engineered prompts for legal entity extraction, summarization, and complex reasoning.
- Fine-tuned a domain-specific Small Language Model using LoRA/QLoRA with custom tokenization strategies for regulatory question answering.
- Deployed the LLM inference stack as a FastAPI REST backend and maintained technical documentation across the system.

QLoRA Fine-Tuning: Islamic Finance Assistant 2026
Python, PyTorch, Hugging Face, QLoRA, PEFT, TRL, Kaggle

- Fine-tuned Qwen2.5-3B-Instruct on a custom Islamic Finance Q&A dataset using 4-bit BitsAndBytes NF4 quantization and LoRA adapters with PEFT r=16, alpha=32.
- Built a custom training pipeline using TRL SFTTrainer and a 5-step rolling-average checkpointing strategy to select adapter weights based on smoothed training loss.
- Curated structured JSON training data using Qwen chat templates with a 512-token maximum sequence length and published the dataset to Hugging Face.

LexisSearch – Semantic Legal Search Engine 2025
Python, Sentence Transformers, FAISS, NumPy, Streamlit

- Developed a semantic search application using all-mpnet-base-v2 embeddings and cosine similarity implemented from scratch in NumPy.
- Precomputed document embeddings with FAISS for fast retrieval and deployed the end-to-end application on Streamlit Cloud.

Fundamental ML Architectures From Scratch 2025
PyTorch, NumPy, Backpropagation

- Implemented Transformer architecture from scratch in PyTorch, including multi-head self-attention, positional encoding, LayerNorm, and residual connections.
- Built a 10.8M-parameter decoder-only GPT from scratch and a 2-layer neural network in NumPy with manual forward propagation, backpropagation, and gradient descent, achieving 96%+ accuracy on benchmarking data.

Education

Bachelor of Science in Computer Science Sep 2022 – Jun 2026
FAST National University of Computer and Emerging Sciences, Islamabad, Pakistan

Certifications & Leadership

Certifications: Hugging Face – Fundamentals of LLMs (2026); Hugging Face – Fine-tuning Language Models (2026)
Leadership: Led AI and software engineering teams of up to 10 members through requirement analysis, task delegation, milestone tracking, code reviews, and stakeholder presentations.
Activities: Zindagi Prize Entrepreneurship Program – Campus Round Co-organizer (2024); NUETEC University Tech Event – Volunteer (2023, 2024)