

PROFESSIONAL SUMMARY

AI Engineer with 3 years of Python-driven backend experience, specializing in the end-to-end design, evaluation, and deployment of agentic AI systems and Retrieval-Augmented Generation (RAG) pipelines. Adept at bridging the gap between cutting-edge Large Language Models (LLMs) and reliable enterprise architecture. Proven expertise in architecting vector-database-backed retrieval layers to eliminate model hallucinations and ground AI outputs in highly accurate, domain-specific data. Strong track record of translating complex AI capabilities into scalable, task-specific autonomous agents, seamlessly integrated into production environments utilizing robust microservice architectures with FastAPI, PostgreSQL, and Docker.

WORK EXPERIENCE

AB {Ark} Solutions <i>Associate AI Engineer</i>	Lahore, Pakistan Nov 2025 – Present
● Agentic AI Systems: Engineered autonomous multi-agent pipelines using LangGraph to replace manual multi-step operations, reducing business workflow execution time by 20%. ● RAG Architecture: Architected production RAG systems with hybrid search across Pinecone and Weaviate vector databases, optimizing chunking and embedding pipelines to achieve retrieval accuracy. ● LLM & API Integration: Integrated enterprise LLM APIs with asynchronous FastAPI backends and cloud database infrastructure, supporting daily inference requests with under 500ms response latency. ● Production Reliability: Implemented automated prompt evaluation, token caching, and latency monitoring in production using langfuse, lowering API compute costs by 25%.	
Turing <i>Software Developer</i>	Pakistan (Remote) Aug 2024 – Apr 2025
● LLM Evaluation: Evaluated code generation models across 500+ complex Python and JavaScript tasks, assessing algorithmic correctness, time complexity, and syntax reliability. ● Test Case Design: Developed 1,000+ edge-case test suites to probe error handling and logical edge cases, uncovering critical failure modes in reasoning logic. ● Model Feedback: Delivered structured human-in-the-loop (RLHF) annotations and ranking data, contributing to a 15% improvement in model response accuracy and instruction-following consistency.	
Viztera Solutions <i>Software Developer</i>	Lahore, Pakistan Jul 2023 – Jul 2024
● Feature Development: Built and deployed scalable backend services in Python across the complete software lifecycle, delivering 6+ core platform features ahead of sprint deadlines. ● Backend Architecture: Refactored monolithic endpoints into modular, OOP-compliant RESTful APIs and indexed database queries, improving API throughput by 35%. ● Code Quality: Collaborated with the team to identify and resolve bugs, improving code quality and application performance.	

TECHNICAL SKILLS

Languages: Python, SQL, JavaScript
Generative & Agentic AI: Generative AI (GenAI), agentic AI, RAG pipelines, LangChain, LangGraph, LLM evaluation, Hermes AI agents, OpenClaw
ML & Data: PyTorch, data science, embeddings, vector databases (Weaviate)
Backend & Infrastructure: FastAPI, REST API integration, PostgreSQL, SQL, Docker, Git, cloud database infrastructure
Practices: workflow automation, rapid prototyping, production monitoring, object-oriented design

EDUCATION

Information Technology University of the Punjab <i>MS Data Science</i>	Lahore, Pakistan 2024 – 2026
University of Engineering and Technology <i>BS Mechatronics and Control Engineering</i>	Lahore, Pakistan 2018 – 2022

LANGUAGES

- **Languages:** Urdu (native) · English (B2 — CEFR: Listening B2, Reading B2, Writing B1, Speaking B2)