AMMAN FAISAL

+92 312 9815068 · Islamabad, PK · khawajaamman@gmail.com · linkedin.com/in/amman-faisal · github.com/Stradok · khawaja-amman-faisal.vercel.app

Professional Summary

Fresh graduate AI Engineer specializing in LLM application development, multi-agent orchestration, and production RAG systems. Experienced building fully offline agentic pipelines via local LLM inference (Ollama), fine-tuning open-source models with QLoRA, and delivering end-to-end AI systems across security, automation, and data intelligence domains.

Technical Skills

AI & LLM: LangChain, LangGraph, Ollama, Hugging Face Transformers, PyTorch, QLoRA Fine-Tuning, Vector DBs, Sentence-Transformers

Languages & Tools: Python, C++, SQL | FastAPI, Next.js, Git, Linux, OpenAI API, Anthropic API

Experience

AI Engineer *2025 – 2026*
Cyberlytics Lab Islamabad, PK

- Served as lead architecture designer for CODE-AI, a context-aware vulnerability detection pipeline combining semantic CVE search with dual-model validation.
- Owned the system end to end: vector search design, dual Ollama model orchestration, and the self-healing repair loop that auto-generates and patches vulnerable code.

GenAI Intern *June 2024 – August 2024*
KDD Lab, FAST NUCES Islamabad, PK

- Fine-tuned Mistral-7B via QLoRA to automate deterministic script generation from natural language queries.
- Engineered data verification and prompt engineering pipelines; achieved **15% structural accuracy improvement**.

Selected Projects

JANOS: Joint Autonomous Multi-Agent Neural System — *Ollama, LangGraph, ChromaDB, FastAPI, Whisper STT, Python*

- Architected a fully offline 14-agent system spanning 7 capability tiers for long-horizon autonomous task execution.
- Built a ChromaDB-backed self-improving RAG pipeline ingesting failure patterns for autonomous tool-call recovery.
- Implemented Whisper STT/TTS, PyAutoGUI PC automation, and wake-word activation across **29 stable modules**.

CODE-AI: Context-Aware Vulnerability Detection Pipeline — *Python, FastAPI, ChromaDB, Sentence-Transformers, Ollama, Next.js*

- Built a 6-stage local RAG pipeline over **200k+ NVD records** via Sentence-Transformers; zero external API exposure.
- Designed dual-model validation (DeepSeek-R1 + LLaMA-3.1) with self-verifying patch generation requiring fixes to clear the full pipeline.
- Delivered a full-stack Next.js/Monaco Editor application with real-time Server-Sent Events streaming.

Dual-GAN Multimodal Emotion Synthesis via CLIP-Guided Representation Learning — *PyTorch, CLIP, GANs, Computer Vision, NLP*

- Developed a bidirectional multimodal framework capable of translating facial expression images into natural language emotion descriptions and generating facial expressions from textual prompts.
- Designed two adversarial GAN architectures operating as inverse mappings and aligned through a shared CLIP embedding space for cross-modal representation learning.
- Leveraged contrastive vision-language embeddings to improve semantic consistency between generated images and textual emotion descriptions.
- Evaluated multimodal synthesis quality using reconstruction accuracy, adversarial loss convergence, and CLIP-based semantic similarity metrics.

Code Buddy: Local LLM Coding Assistant — *Python, Ollama, OpenRouter, LangGraph, AST Parsing*

- Built a local-first coding assistant routing between Ollama and OpenRouter, enabling code generation without mandatory cloud API calls.
- Implemented AST-based code context extraction to supply precise structural information to LLM generation prompts via a LangGraph workflow.

TEXT PANDAS: Tabular RAG Engine — *Python, Hugging Face Hub, Pandas, FastAPI*

- Published a custom text-to-pandas dataset on Hugging Face Hub mapping natural language queries to pandas operations.
- Solved tabular RAG token-bloat by injecting targeted structural arrays rather than raw spreadsheets into the context window.

Education

Bachelor of Science in Data Science *Expected June 2026*
FAST National University of Computer and Emerging Sciences Islamabad, PK