

MUHAMMAD JUNAID

Full Stack & AI Engineer | Agentic AI · Voice Agents · RAG · LLMOps

itxjunaid22@gmail.com · +92 341 8914317 · [Portfolio](#) · [GitHub](#) · [LinkedIn](#) · [Kaggle](#) · [Coursera](#)

PROFILE

Full Stack & AI Engineer who builds production-grade agentic AI systems end to end — from LangGraph and CrewAI multi-agent pipelines and real-time voice agents down to the FastAPI backends and Next.js frontends that serve them. Comfortable across the whole LLM lifecycle: RAG and Agentic RAG, LoRA/QLoRA fine-tuning, evaluation, model observability, and LLMops/MLOps deployment. Currently a Full Stack AI Developer at Voltek US, working as a voice agents specialist and applied AI researcher on agent runtimes such as OpenClaw and Hermes Agent.

CORE SKILLS

Languages: Python, C/C++, JavaScript, TypeScript, SQL, HTML, CSS

Generative AI & Agents: LangChain, LangGraph, CrewAI, LlamaIndex, OpenClaw, Hermes Agent, Agentic Workflows, Multi-Agent Orchestration, Model Context Protocol (MCP), Tool/Function Calling, Prompt Engineering, Structured Outputs

Voice AI: Real-time voice agents, Whisper STT, Kokoro / edge-TTS, Silero VAD barge-in, WebSocket audio streaming, low-latency turn-taking

RAG & Retrieval: Agentic RAG, hybrid search, re-ranking, chunking strategies, embeddings (bge, fastembed); Qdrant, FAISS, ChromaDB, Pinecone

LLMOps & Model Observability: LangSmith tracing, evaluation harnesses & LLM-as-judge, prompt and model versioning, guardrails, hallucination & drift monitoring, token-cost and latency optimisation

MLOps: MLflow, Weights & Biases, Docker, CI/CD with GitHub Actions, model registry & versioning, scalable inference deployment, production monitoring

ML & Deep Learning: PyTorch, TensorFlow, Keras, Scikit-learn, Hugging Face Transformers, CNNs / ResNet50, transfer learning, Grad-CAM explainability, LoRA & QLoRA fine-tuning, NLP, Computer Vision

Full Stack & Backend: React, Next.js, FastAPI, REST APIs, WebSockets, Server-Sent Events, Node.js, PostgreSQL, MongoDB, SQLite, Tailwind CSS

Platforms & Tools: Git/GitHub, Linux, Docker, Gradio, Streamlit, Hugging Face Spaces, Groq, OpenRouter, Ollama, Tavily, Flutter (on-device AI)

Soft Skills: Leadership, team collaboration, analytical thinking, problem solving, technical communication

WORK EXPERIENCE

Full Stack AI Developer — Voltek US *May 2026 – Present*

Voice Agents Specialist · AI Researcher

- Own AI products end to end — FastAPI services, Next.js interfaces, and the LLM and agent layers behind them — taking features from prototype to production.
- Build real-time voice agents with streaming speech-to-text, text-to-speech and VAD-based barge-in, tuning for sub-second turn-taking and natural conversational flow.
- Architect agentic workflows and multi-agent orchestration with LangGraph, CrewAI and MCP tooling to automate multi-step business processes.
- Research and benchmark emerging agent runtimes including OpenClaw and Hermes Agent, prototyping self-hosted, memory-persistent agents for internal automation.
- Establish LLMOps practice: tracing and model observability, evaluation suites, prompt versioning, guardrails, and token-cost and latency optimisation across deployed models.

AI / ML Engineer — Xact Mind *Mar 2023 – Sep 2025*

- Optimised ML models for predictive analytics and NLP, lifting accuracy through advanced algorithms and feature engineering.
- Integrated AI features into web applications with Python and TensorFlow, improving automation and operational efficiency.
- Built an agentic AI chatbot using LLMs, prompt engineering and tool integration to automate user queries and internal workflows.
- Containerised and served models with Docker and FastAPI, adding monitoring for latency, throughput and model drift.

Kaggle Competition Contributor — Kaggle *Apr 2023 – Present*

- Compete in AI/ML competitions and publish public notebooks, focusing on model efficiency, reproducibility and documentation quality.

AI/ML Study Group Lead — University of Kotli *Jan 2023 – Present*

- Coordinate a student AI/ML community, mentoring members and running hands-on sessions with real datasets.

PROJECTS

RepoRadio — Voice AI Codebase Companion *Python · FastAPI · Groq*

- Paste any GitHub repo and talk to an AI radio host that has read the entire codebase — live, out loud, interruptible mid-word.
- Real-time voice pipeline: server-side Silero VAD barge-in → Whisper STT → RAG → Llama 3.3 70B on Groq → Kokoro/edge TTS, streamed gaplessly over WebSockets.
- Answers grounded in real source via ChromaDB + bge-small retrieval, with a per-commit version archive that broadcasts spoken changelogs; runs at zero cost on free tiers.

Multi-Agent Research Assistant — Agentic RAG + Web Search *Apr 2026 – Present*

- Production-grade research assistant built with FastAPI, LangGraph, Qdrant and Next.js: upload PDFs or URLs and get cited, evidence-backed answers.
- LangGraph pipeline of Planner → Researcher → Reasoning agents wired to a Qdrant RAG tool and Tavily web search, feeding a generator that returns answers with sources.
- Pluggable LLM backends (OpenRouter / Claude / Ollama) and Qdrant modes (memory / local / cloud), SSE streaming, LangSmith monitoring, fully Dockerised stack.

TrailGuard AI — Offline On-Device Survival Assistant · Demo *Flutter · Gemma · LiteRT-LM*

- Fully offline hiking and survival companion running Gemma on-device through Google's LiteRT-LM engine — no connectivity or API keys required.
- Multimodal photo analysis of plants, injuries and terrain, SOS first-aid decision trees, GPS trail tracking and cached OpenStreetMap offline maps.
- Hands-free voice in/out with speech-to-text and text-to-speech, plus a medical profile (blood type, allergies, contacts) for context-aware answers. Built for a Kaggle AI + Outdoor Safety hackathon.

YouTube Content Producer — Multi-Agent AI System · Kaggle writeup *Nov 2025*

- Multi-agent system using Python, LLM APIs and prompt orchestration to automate topic research, trend analysis, script writing and SEO optimisation for YouTube content.
- Real-time streaming workflows and session management with FastAPI and Server-Sent Events, plus JSON/PDF/CSV export pipelines.
- Responsive web interface with robust error handling and a scalable architecture; submitted as the Kaggle Agents Intensive capstone project.

Medical Q&A Assistant — RAG + LoRA Fine-Tuning *Sep 2024 – Jan 2025*

- Production-ready medical Q&A system built on TinyLlama with QLoRA fine-tuning for efficient, domain-accurate responses.
- FAISS-backed RAG retrieval for fast evidence search, keeping answers grounded in medical literature.
- High-performance REST API with a clean Gradio web interface.

Brain Tumor Detection Web App — Explainable Medical AI *TensorFlow · FastAPI*

- ResNet50 deep-learning system that classifies brain tumours from MRI scans, with Grad-CAM heatmaps that visually explain every prediction.
- FastAPI backend with doctor authentication and medical-license verification, patient records, scan history and dashboard analytics over SQLite.
- Automated PDF medical report generation and a responsive Tailwind CSS interface built for clinical use.

EDUCATION

University of Kotli — BS Software Engineering, Azad Kashmir *Dec 2022 – Dec 2026*

Relevant coursework: Machine Learning, Artificial Intelligence, Data Science

Muhammdan College — A-Levels: Mathematics (A), Physics (A), Computer Science (A), Hajira *Aug 2021 – Nov 2022*

ACHIEVEMENTS & CERTIFICATIONS

- Kaggle Competitor — active competitor building and publishing ML notebooks on real-world, data-driven problems.

- [Coursera Specializations \(ML, DL, AI Agents\)](#) — industry-focused certifications in Machine Learning, Deep Learning and AI Agent development.
- **Hackathons** — Kaggle and hackathon certificates and achievement badges, including TrailGuard AI for the Kaggle AI + Outdoor Safety Hackathon.
- **Software house internships & university FYP** (Aug 2024 – Present) — hands-on full-stack and AI systems development in professional and academic settings.

INTERESTS

Chess · Cricket · PC gaming · Hackathons · Travelling · Reading (fiction & non-fiction)