Muhammad Shayan Khan

Location: Pakistan (Open to Remote)
Email: devvshayan@gmail.com | **Phone:** +923358042133
Links: LinkedIn | GitHub

PROFESSIONAL SUMMARY

Innovative AI Developer with expertise spanning full-stack development and generative AI technologies. Specialized in building agentic AI solutions, Large Language Model (LLM) applications, and Retrieval-Augmented Generation (RAG) systems that deliver measurable business impact. Demonstrated success in reducing processing time by 35% and increasing user engagement by 28% through AI-powered features. Passionate about leveraging cutting-edge AI to create intelligent applications while maintaining excellence in technical implementation and user experience.

WORK EXPERIENCE

AI Engineer | Sim Engine (Digital Human Agent Platform) (US-REMOTE)

03/2025 - Present

- Architected and developed a complex agentic platform managing autonomous AI personas.
- Spearheaded the transition to Model Context Protocol (MCP) for tool integration, enabling seamless interaction with external services (Threads, Networks, Groups).
- Designed a robust Redis-backed queue system for real-time task management and task scheduling.

- Implemented a sophisticated RAG architecture using Pinecone and LangChain for persistent persona memory and knowledge retrieval.
- Implemented Sub-Agent Registry Pattern (Factory Pattern) to standardize agent creation and scalability.
- Developed comprehensive authentication system with HS256-signed tokens for Perceptron, User, and Service levels.
- Engineered observability pipeline with Sentry for production monitoring and MongoDB tracing for agent execution analysis.
- Built 'AgentTurn' context container to encapsulate multi-modal interaction states (execution, simulation, sensory).
- Designed Collective Session Service using Redis to manage multi-perceptron thread contexts with sliding expiration.
- Integrated multimodal LLM support (GPT-5, Gemini-3-pro) and FFmpeg for multimedia processing.

Full Stack AI Engineer | Power Nap (Istanbul, Türkiye · Remote)

Apr 2025 - Present · 1 yr 1 mo

- Developed an AI-powered Instagram grid generator, transforming brand assets and design inputs into cohesive 12-post social media grids.
- Engineered the full-stack system with React, TypeScript, FastAPI, and Google Gemini, incorporating AI image generation, background removal, and caption generation.
- Implemented object-aware logo placement using YOLOv10 and ONNX Runtime to enhance layout quality and prevent subject occlusion.
- Designed asynchronous processing for image-intensive tasks and created an interactive UI for content generation, customization, and export.
- This product significantly reduces manual effort for small teams and business owners in creating consistent branded marketing content.

Full Stack AI Developer | CarZoomo (Minnesota, United States – Remote)

01/2025 – 03/2025

- Engineered a natural language search engine using RAG and vector embeddings, enabling conversational queries (e.g., “family cars under $50k”) that increased search engagement by 35% and reduced bounce rates by 22%.
- Developed an AI-powered content generation system that automatically created engaging vehicle descriptions from technical specifications, reducing listing creation time by 40% and improving readability scores by 28%.
- Implemented automated email distribution system for leads, increasing open rates by 25% through personalized AI-generated content.
- Integrated AI models with backend systems using JavaScript, Node.js, and MongoDB.

Gen Eng Team Member | PIAIC (Islamabad, Pakistan)

07/2023 – Present

- Researched and applied emerging AI technologies, staying current with industry best practices.
- Delivered projects on time and within budget, maintaining a 100% client satisfaction rate through clear communication and quality deliverables.

EDUCATION AND TRAINING

Bachelor of Computer Science

Virtual University of Pakistan | *10/2023 – Present*

Higher Secondary School Certificate in Computer Science

F.G Sir Syed College, Rawalpindi | *03/2021 – 03/2023*

PERSONAL SKILLS

Languages

- **Urdu:** Native
- **English:** Proficient (C1/C2)
- **Hindi:** Expert (C2)
- **Punjabi:** Proficient (B2)
- **German:** Beginner (A1)

Technical Skills

- **Agentic AI:** LLMs, Retrieval-Augmented Generation (RAG), OpenAI Agents SDK, LangGraph, LangChain, LangSmith, CrewAI, stateful workflows with memory, tool use, and complex decision logic, Prompt Engineering, NLP.
- **RAG Pipelines:** Pinecone, FAISS, LangChain for grounded, accurate retrieval.
- **Full-Stack AI Applications:** Next.js/React, TypeScript, Tailwind, Python, FastAPI, PostgreSQL, MongoDB, Redis.
- **Cloud & Tools:** Azure, Google Cloud, Docker, Git, GitHub, Agile Methodologies, AI Model Testing & Validation, MCP integrations.
- **AI Models & Platforms:** Claude Code, OpenClaw, OpenAI Codex.

ADDITIONAL INFORMATION

Selected AI Projects & Client Builds

- **Autonomous Franchise Operator AI (Upwork Client):** Building an AI for a multi-location franchise, centralizing intelligence to automate reporting, detect anomalies, explain business performance, and provide conversational AI leadership.
- **Trueears (Opensource):** Built Trueears, an open-source Tauri desktop AI dictation application using React/TypeScript and Rust that captures microphone

audio, transcribes it, applies context-aware LLM formatting and rewriting based on the active application rules set by the user, and inserts the result back into desktop tools such as editors, chat apps, email clients, and notes apps. Implemented native OS automation for global shortcuts, active-window detection, clipboard/paste workflows, app-specific profiles, select-to-transform editing, structured logging, Google OAuth, and licensing/payment support.

- **Local-First AI Audio Cleanup App (Opensource):** Built an open-source local-first desktop AI audio-cleanup application using a React/Tauri frontend and a bundled Python FastAPI inference engine to denoise audio extracted from audio and video files on-device. Implemented one-time model setup, asynchronous job orchestration, cancellation, saved-output management, waveform-based before/after comparison, and Rust sidecar startup/runtime hardening, enabling users to validate cleanup quality without uploading media to the cloud.
- **Newsletter Agent (Upwork Client):** Automated agent ingesting research articles to publish newsletters using Python, OpenAI Agents SDK, Playwright, and Google Docs API.

Certifications

- **Embedding Models: From Architecture to Implementation** – DeepLearning.AI
- **AI For Everyone** – DeepLearning.AI ($^{11}/_{2023}$)
- **Generative AI for Everyone** – DeepLearning.AI ($^{11}/_{2023}$)
- **Programming for Everybody** – University of Michigan ($^{11}/_{2023}$)
- **Introduction to Generative AI** – Google Cloud Training Online ($^{10}/_{2023}$)