

MUHAMMAD ARSALAN KAMRAN

AI Engineer | Full-Stack AI Developer | LLM, RAG, AI Agents | Python, FastAPI, Django

Peshawar, Pakistan (Open to Remote) | arsalankamran80@gmail.com | +92 335 9100337 | [LinkedIn](#) | [GitHub](#)

PROFESSIONAL SUMMARY

AI Engineer with 3+ years of experience building and deploying production AI systems: LLM applications, RAG pipelines, AI agents, text-to-SQL, voice AI, document AI and machine learning models, on scalable Python backends (FastAPI, Django). Currently a Full-Stack AI Engineer on a HIPAA-compliant healthcare AI platform for a US company. Has taken AI SaaS products and MVPs from idea to production in healthcare, fintech and legal tech, including React/Next.js frontends and AWS/GCP/Azure deployment.

TECHNICAL SKILLS

- **AI / LLM Engineering:** Generative AI, LLM application development, AI agents, agentic workflows, multi-agent systems, RAG, text-to-SQL, prompt engineering, function/tool calling, structured outputs, LLM evaluation (LLM-as-judge, golden datasets), MCP, voice AI, document AI, OCR
- **AI Frameworks & Models:** LangChain, LangGraph, DSPy, CrewAI, OpenAI (GPT-4o, GPT-5, Responses API), Azure OpenAI, Anthropic Claude, Google Gemini, Vertex AI, Llama, Groq, Deepgram, ElevenLabs, Replicate
- **Vector Databases & Search:** pgvector, Pinecone, ChromaDB, Qdrant, embeddings, semantic search
- **Machine Learning & Computer Vision:** PyTorch, TensorFlow, Keras, scikit-learn, CNN, RNN, LSTM, Transformers, YOLO, ONNX Runtime, Stable Diffusion, ControlNet, OpenCV, MediaPipe
- **Backend:** Python, FastAPI, Django, Django REST Framework, Flask, Node.js, Express, Celery, Redis, REST APIs, WebSockets, SSE streaming, microservices, multi-tenant SaaS, OAuth 2.0, JWT, RBAC
- **Databases & Data Engineering:** PostgreSQL, MySQL, MongoDB, BigQuery, Firestore, DynamoDB, Supabase, Databricks, PySpark, FHIR/HL7
- **Frontend & Mobile:** TypeScript, JavaScript, React, Next.js, Vite, Tailwind CSS, shadcn/ui, Zustand, TanStack Query, Capacitor (iOS, Android)
- **Cloud, DevOps & Deployment:** AWS (EC2, ECS, ECR, Elastic Beanstalk, Lambda, API Gateway, S3, RDS, Amplify, CloudFront, SQS, SES, EventBridge), GCP (Cloud Run, Cloud Build, Vertex AI, IAM), Azure, Docker, Docker Compose, Terraform, GitHub Actions, CI/CD, Nginx, Gunicorn, Vercel, Firebase, RunPod
- **Integrations:** Stripe, Plaid, QuickBooks, Gmail and Google Drive APIs, Microsoft Graph, Slack, Twilio, Shopify
- **Product & Leadership:** AI SaaS development, MVP development, requirement gathering, technical scoping, client communication, sprint planning, team leadership, Agile, HIPAA compliance

PROFESSIONAL EXPERIENCE

Full-Stack AI Engineer | Advanced Management USA, LLC (AdvancedACO)

Remote | Jan 2026 – Present

- Build production AI features for CareCompass AI, a HIPAA-compliant population health SaaS for US Accountable Care Organizations (ACOs), across a FastAPI AI service and a Next.js/React/TypeScript app.
- Designed and built a LangGraph text-to-SQL agent on BigQuery (LLM table selection, RBAC filter injection, period detection, ReAct SQL with sqlglot validation) that answers questions when the primary data API fails.
- Built an LLM evaluation regression gate in GitHub Actions with a 25-question golden dataset, baseline metrics, a version-comparison scoreboard and telemetry, run on every pull request.
- Drove major parts of the 485 AI Document Review pipeline, which extracts, analyzes and summarizes home-health plans of care from scanned faxes with Claude, Gemini and GPT-5, applying 19 clinical decision rules, page-orientation correction and 100% patient-identity matching before analysis.
- Re-architected document processing from long synchronous requests into an async job system (Firestore job store, cron worker, checkpoints, retries with backoff, capacity lock) with batch upload of up to 20 documents.
- Managed multi-provider LLM migrations and fallbacks across about 10 AI routes (Gemini 2.5 to 3.x, Claude Sonnet 4.5 to Claude 5, Azure GPT-5 fallback on quota errors) and built an LLM-as-judge evaluation pipeline with a developer UI.
- Designed service-to-service security for HIPAA data: HMAC-SHA256 signed principal envelopes, fail-closed tenant scoping, PHI disclosure masking, BigQuery cost caps, rate limiting, least-privilege GCP IAM and secret scanning in CI.
- Built Databricks/PySpark pipelines moving FHIR/HL7 data from BigQuery into PostgreSQL for care-plan flags.

Lead AI & Full-Stack Engineer (Part-Time) | Clentro

Remote | Dec 2025 – Oct 2026

- Led delivery of AI SaaS products and MVPs for US and Australian clients: requirement discovery, scoping and proposals, architecture, sprint planning, team task allocation and client communication through release.

- **BlackGrid (AI CFO for small businesses):** Bootstrapped the Django REST backend (about 54% of its Python code) and built the AI CFO on the OpenAI Responses API with tool calling, cited web search and memory, grounded in finance engines.
- Integrated Plaid, QuickBooks, Gmail, Stripe and push notifications with Celery and Redis; deployed with Docker Compose on AWS EC2, RDS PostgreSQL and S3; shipped iOS and Android apps from the React 19 web app with Capacitor.
- **TradieMate (SaaS for electrical contractors):** Designed a provider-agnostic LLM layer (Gemini, OpenAI, Anthropic), AI quote review and a job-level RAG chat on pgvector with SSE streaming, plus Gmail ingestion and Celery on AWS SQS.
- **Locaro.ai (AI location analysis):** Migrated the AI engine to Django REST and built a two-stage LLM pipeline (Claude Haiku classifies competitors, Claude Sonnet writes a schema-validated report); deployed on AWS Elastic Beanstalk.
- **Globugg (location-based social video):** Built the Django REST backend from scratch (Firebase auth, WebSockets, push notifications, GPT-4o-mini video tagging) and helped move it to AWS Lambda and Terraform.

AI Backend Engineer (Contract) | Boost Panda

Remote | Dec 2025 – Feb 2026

- Built end-to-end AI document generation for a legal research platform: FastAPI context aggregation, GPT-4o drafts streamed over SSE into a React editor, PDF export and a guided multi-turn DSPy agent.
- Improved the RAG chat with rolling conversation summarization, inline citations grounded only in retrieved sources, and paginated MongoDB aggregation APIs over state statutes and regulations.
- Wrote most of a Python/Selenium WhatsApp Web data-extraction pipeline with media capture and incremental sync, and added DSPy natural-language-to-Gmail query translation to a FastAPI gateway for 11 SaaS integrations.

AI Engineer | Ikonik

Islamabad, Pakistan | Sep 2025 – Feb 2026

- Sole developer of the FreshPass AI backend for a salon booking SaaS: a FastAPI service (about 22K lines of Python, 42 endpoints) delivering 6 AI products, integrated with the Laravel platform for billing.
- Built an AI hairstyle try-on pipeline (MediaPipe, PyTorch face-shape classifier, DeepFace, GPT-4o vision) rendering 3 haircuts in 4 views; v1 ran the Stable-Hair diffusion model (Stable Diffusion, ControlNet) on RunPod GPUs.
- Engineered a real-time AI voice receptionist on the Deepgram Voice Agent API over WebSockets, with 10 function-calling tools to book, check and cancel appointments and buy subscriptions by voice.
- Built multi-tenant RAG assistants on ChromaDB with per-business PDF upload and streaming responses, plus Redis async jobs, retries with exponential backoff, rate limiting and AI-credit debits with automatic rollback.
- Designed and built the Dev Portal in about 2 weeks (Django REST, PostgreSQL, JWT/RBAC, React/TypeScript), with a Gemini 2.5 Flash agent that reads a project brief and ranks the best-fit engineers for staffing.
- Shipped continuous ElevenLabs audiobook narration for the ChattyBook client platform, and converted YOLO11 to ONNX with a custom ONNX Runtime inference CLI and latency benchmark.

Team Lead, Backend/AI Developer | CortechSols

Islamabad, Pakistan | Aug 2024 – Sep 2025 (promoted from Backend/AI Engineer, Nov 2024)

- Led a team of 10+ backend, frontend and AI developers delivering 12+ client projects and closed 3 client deals.
- Architected Django REST and FastAPI microservices serving 10K+ daily requests with Celery, PostgreSQL and Redis; built RAG systems on Qdrant, Pinecone and ChromaDB and AI chatbots with OpenAI, Claude, Gemini and Llama.
- Deployed to AWS (EC2, S3, RDS, Lambda, Amplify) and GCP with Docker Compose, Nginx and Unicorn, ran GPU inference on RunPod, and set up CI/CD with GitHub Actions.

Machine Learning Engineer | Ayass BioScience, LLC

Remote (Frisco, Texas) | Sep 2023 – Nov 2024

- Built CNN/RNN cancer classification models (98% accuracy) on 100K+ medical records, leading to 3 clinical deployments.

SELECTED PROJECTS

- **CareCompass AI** (carecompassai.com): HIPAA-compliant ACO platform with AI document review and text-to-SQL analytics.
- **FreshPass AI** (getfreshpass.com): AI try-on, voice receptionist and RAG assistants for salons. FastAPI, GPT-4o, Deepgram.
- **BlackGrid AI CFO** (getblackgrid.com): AI CFO for small businesses, web and mobile. Django, OpenAI, Plaid, QuickBooks.
- **TradieMate** ([live demo](#)): Quoting and job management SaaS for electrical contractors with RAG chat and AI quote review.
- **Mackinac AI Research Platform** (research.mackinac.org): Legal research assistant with grounded RAG and AI drafting.
- **VisuAlly** (published in IJIST journal): Real-time object detection and text-to-speech app for visually impaired users.

EDUCATION

BSc Computer Systems Engineering | University of Engineering and Technology, Peshawar | 2024 | CGPA 3.80/4.00