

Muhammad Abdullah

AI Engineer | Agentic AI, LLM Systems & RAG

• +92 311 840 0589 • abdullah.muhammad.4315@gmail.com • Lahore, Pakistan • [LinkedIn](#) • [Github](#)

Summary

AI engineer with **7+ years** of software engineering, building and shipping agentic AI, RAG and voice-agent systems to production. Currently working on LLM workflows, **RLHF and evaluation** for Google via Turing. Has built **LangGraph and LiveKit voice agents**, fine-tuned and quantized models for CPU inference, and RAG pipelines on **Pinecone and FAISS**, backed by Python, FastAPI and AWS engineering that keeps these systems reliable in production.

Skills

LLMs, Agents & RAG: Agentic AI • Multi-agent Workflows • Tool Use • Prompt Engineering • SFT • RLHF • Fine-tuning (LoRA) • Evaluation Harnesses • Anthropic Claude • OpenAI GPT • AWS Bedrock • LangChain • LangGraph • LlamaIndex • LangSmith • Hugging Face Transformers • Voice Agents • RAG • Hybrid Retrieval • Pinecone • FAISS • Chroma

ML & Data: PyTorch • TensorFlow • Scikit-learn • XGBoost • SpaCy • ONNX Runtime • MLflow • Pandas • NLP • Computer Vision

Backend & Cloud: Python • FastAPI • Flask • Django • NestJS • TypeScript • Rust • Go • REST APIs • WebSockets • PostgreSQL • MongoDB • DynamoDB • Redis • AWS (Lambda, S3, SageMaker, API Gateway, SQS) • GCP (Cloud Run) • Docker • CI/CD • n8n • Zapier

Experience

AI Engineer | Turing | Remote, United States

Feb 2025 - Present

Embedded as an AI Engineer for Turing's Client (Google), working on large-scale production AI systems involving LLM workflows, prompt engineering, RLHF, and performance-critical backend components using Python, Rust, and Go.

- Embedded as an AI Engineer for Google (via Turing), validating, optimizing, and debugging large-scale **production AI codebases** to improve model correctness, performance, and reliability.
- Designed **prompt engineering strategies** for LLM-based workflows, creating structured prompts, constraints, and **evaluation tests** to improve reasoning quality and output consistency.
- Implemented prompt iteration and **RLHF workflows** to align model behavior with human feedback and domain-specific requirements.
- Contributed to model quality and infrastructure improvements using **Python, Rust, and Go**, working across inference logic and performance-critical components.
- Collaborated in Agile (Scrum/Kanban) teams with cross-functional stakeholders to deliver **production-ready AI systems** on tight timelines.

AI Engineer | Nexxt.ai | Remote, United States

Jul 2023 - Jan 2025

Architected and maintained secure, scalable AI-powered platforms across Fintech, E-commerce and Real Estate sectors with a strong emphasis on compliance, real-time processing, and service resilience.

- Built an **AI-powered voice agent** for businesses to handle customer queries and manage bookings, supporting **20+ concurrent calls** with **sub-1000 ms** end-to-end latency, using asynchronous call handling, real-time speech processing, and backend API integrations for availability checks and scheduling.
- Built a Formula 1 race outcome prediction system using **Random Forest**, training on **50+ years** of qualifying and sprint data to estimate top-5 finish probabilities, and deployed real-time inference via event-driven AWS services for low-latency predictions.
- Developed a cryptocurrency market trend classifier using technical indicators (RSI, MACD, Bollinger Bands, EMA) and **XGBoost**, backtesting on multi-year OHLC data to generate bullish/bearish signals that improved risk-adjusted trade selection compared to rule-based strategies.
- Designed an NLP-based resume parsing pipeline using Transformer models and **SpaCy NER** to extract structured candidate attributes from **1,000+ resumes**, reducing manual screening effort and enabling downstream candidate ranking and search automation.

- Implemented an image-based product search and recommendation system using **CNN feature extraction** (VGG16, ResNet50) and cosine similarity, enabling visual catalog search and fallback recommendations to increase product discoverability.
- Productionized ML systems using **FastAPI and AWS SageMaker**, implementing model versioning, automated retraining triggers, logging, and inference monitoring, with CI/CD workflows to promote models safely from experimentation to production.

Machine Learning Engineer | Ascendia AI | Remote, Global

Jan 2022 - Jun 2023

Developed AI-driven recruitment platforms combining real-time voice AI, RAG-based semantic search, and event-driven backend architectures, delivering low-latency, production-grade intelligent systems.

- Architected real-time **Voice AI interview system** using LiveKit Agent Framework and WebRTC for bidirectional audio streaming, implementing LangChain orchestration pipelines with **sub-500ms latency** through optimized audio chunking and streaming transcription (Deepgram).
- Designed **event-driven microservices** architecture for large-scale job scraping and candidate matching, orchestrating parallel crawlers across LinkedIn, Indeed, and Dice using AWS Lambda, S3, and DynamoDB, processing **20K+ jobs monthly** with CronJob-triggered workflows.
- Built **RAG-based semantic matching engine** combining fine-tuned LED-base-16384 for long-document summarization with OpenAI embeddings (text-embedding-3-large) and Pinecone VectorDB, achieving **3x faster job sourcing** and 85%+ match accuracy through hybrid retrieval and LLM reranking.
- Engineered adaptive conversational AI flows using **LangGraph state machines** for multi-turn interviews with context-aware follow-up logic, integrating sentiment analysis, tone detection, and engagement scoring via real-time audio feature extraction and transformer-based classification.
- Optimized custom fine-tuned LED model for production deployment using **8-bit quantization and ONNX Runtime**, enabling CPU-only inference with fp16 mixed-precision training and gradient accumulation, reducing VRAM requirements by **70%** while maintaining summarization quality.
- Built AI phone outreach system using **Twilio Voice APIs** with TTS/STT integration, implementing asynchronous call orchestration, automated scheduling logic, and 2-way SMS screening workflows, reducing candidate response time by **80%**.

Software Engineer | Infostack | Pakistan

Feb 2019 - Dec 2021

Worked on backend engineering for healthcare and financial platforms, building secure APIs, event-driven services, and data pipelines while gaining early exposure to automation and data-centric workflows.

- Built backend services using **Flask and FastAPI** for healthcare and financial platforms, implementing **JWT and OAuth 2.0** authentication with RBAC, blocking malicious requests through rate limiting and token validation mechanisms.
- Built and maintained RESTful APIs and event-driven workflows using **AWS Lambda and webhooks**, integrating **Stripe** payment APIs with idempotency, retry mechanisms, and failure notifications to ensure high transaction reliability.
- Improved system performance by implementing **Redis caching**, query optimization, and database indexing, reducing API response times by **~40%** and increasing throughput under concurrent load.

Projects

Platinum Detailing Voice Agent | Retell AI, FastAPI, Google Calendar API, Urrable CRM, Cloud Run | 2026

- Built a Retell AI voice agent that answers inbound calls, routes by intent and books appointments from **live Google Calendar availability** across detailers, with 15-minute slots and booked or past times filtered out.

TeacherAide: AI Slide Generation | Claude, n8n, Next.js, NestJS, AWS Lambda, PostgreSQL, Stripe | 2026

- Built a Claude-powered generator that extracts a teacher's style fingerprint from uploaded lectures and produces decks in four stages across **11 version-controlled n8n AI workflows**.

Education

University of the Punjab | Bachelor of Science in Computer Science | 2015 - 2019

Lahore, Pakistan

Certifications

IBM RAG and Agentic AI Specialization | IBM (Coursera) - 2025

Advanced RAG with Vector Databases and Retrievers | IBM - 2025