

Mohammad Abdullah

AI Engineer | Agentic Systems, LLM Training & Applied ML

• +92 300 436 3509 • abdullah.shahid4315@gmail.com • Lahore, Pakistan • [LinkedIn](#) • [Github](#) • [Portfolio](#)

Summary

AI Engineer with 4+ years building production LLM and backend systems across FinTech, HealthTech and E-commerce. Currently training and evaluating multimodal agents for on-device task execution across first-party Android and iOS applications at Google (via Turing), while shipping client-facing agentic products end to end. Works in the space between model behaviour and system reliability: SFT and RLHF pipelines, evaluation harnesses, RAG, and the backend architecture that keeps them running once the demo is over.

Skills

LLMs & Agents: Agentic AI • Multi-agent workflows • Agent orchestration • Tool use • Supervised fine-tuning (SFT) • RLHF • Evaluation harnesses • Benchmarking • Prompt engineering • Structured tool-call outputs • Prompt caching • Anthropic Claude • OpenAI GPT • Llama • Hugging Face Transformers • LangChain • LangGraph • LlamaIndex • LangSmith

RAG & Retrieval: RAG • Hybrid retrieval • Pinecone • FAISS • Chroma • Sentence Transformers • Embedding pipelines • Retrieval evaluation

Backend: Python • FastAPI • Flask • Django • TypeScript • NestJS • REST APIs • WebSockets • Celery • Event-driven architecture • PostgreSQL • MongoDB • Redis • OAuth 2.0 • JWT • RBAC • Rust • Go

Frontend: React • Next.js • Tailwind • TanStack Query • Redux Toolkit

Cloud & DevOps: AWS (Lambda, S3, RDS Proxy, API Gateway, DynamoDB, EventBridge, SQS, CloudWatch) • GCP • Docker • GitHub Actions • CI/CD • Observability • n8n workflow orchestration

Experience

Senior Software Engineer | Code District | Lahore, Pakistan

Jun 2026 - Present

- Sole engineer on an AI lecture-generation platform that writes slide decks in each teacher's own voice, owning the full stack from database schema through AI workflow orchestration, billing and CI/CD.
- Built a style-fingerprinting pipeline that extracts a measurable writing profile from a teacher's uploaded lecture materials across weighted structural and narrative metrics, then scores generated decks for consistency against that profile.
- Designed a staged generation pipeline (outline, detailed content, personalisation, speaker notes) with per-stage isolation, so cost and failure are attributable to a single step rather than the whole run.
- Replaced hand-edited prompts with a version-controlled prompt pipeline across 11 AI workflows: each prompt lives as a file injected at publish time, and CI diffs changed workflows and republishes only those with per-environment credentials.
- Forced every model call to respond through a single named tool call rather than free text, so output arrives as structured data the backend can trust without re-parsing.
- Tuned Anthropic prompt caching (breakpoint placement, cache TTL, sampling count and model choice) against measured cost and score-drift from real dev/UAT executions, rather than caching by default.
- Delivered the full stack: PostgreSQL/TypeORM schema and migrations, NestJS services on AWS Lambda behind RDS Proxy, Next.js frontend, Stripe subscription billing on server-verified webhooks, OAuth-authorized Google Slides export, and an admin dashboard for monitoring generation runs, accounts and billing health.

LLM Engineer (Contract) | Turing (Client: Google) | Remote, US

Jun 2025 - Aug 2026

- Trained and fine-tuned LLMs for agentic task execution on first-party Android and iOS applications, enabling AI-driven automation of complex multi-step mobile workflows.
- Developed and curated high-quality training datasets for mobile agentic tasks, covering UI navigation, multi-step task completion and in-app action sequencing.

- Designed evaluation frameworks and benchmarks measuring agent task success rate, reliability and performance across diverse mobile environments.
- Applied supervised fine-tuning (SFT) and RLHF to iteratively improve LLM decision-making for on-device agentic scenarios.
- Built structured prompting strategies with explicit constraints and evaluation tests, improving reasoning quality and output consistency in LLM workflows.
- Validated, optimised and debugged large-scale production AI codebases in Python, Rust and Go, working across inference logic and performance-critical components.
- Partnered with product and engineering teams to align model capabilities with real first-party application requirements, delivering in Agile/Scrum cycles.

Machine Learning Engineer | InfoStack | Lahore, Pakistan

Mar 2025 - Sep 2025

- Built and launched an internal AI product from scratch, leading system design, backend workflows and the critical path components, including synchronisation across multiple data sources.
- Built RAG pipelines with hybrid retrieval over Pinecone and FAISS for semantic search across enterprise document sets, cutting retrieval latency by roughly 30% while improving relevance.
- Fine-tuned and deployed domain-specific language models (GPT, BERT, Llama) with Hugging Face Transformers, PyTorch and LangChain for search, summarisation and content generation.
- Applied prompt engineering and few-shot/fine-tuning techniques to improve contextual alignment on enterprise use cases, measured against internal benchmark sets.
- Stood up MLOps workflows with MLflow, Docker and Git-based CI/CD for reproducible training, model versioning and production monitoring.

Python Developer | Stampa Solutions | Lahore, Pakistan

Mar 2022 - Feb 2025

- Initiated and led the complete backend development of an Android TV application, designing the system architecture, technical documentation and backend workflows from scratch.
- Engineered a microservices architecture on AWS with Flask, implementing JWT-based authorisation for secure user access and improving authentication reliability by 30%.
- Integrated Stripe for payment processing, enabling seamless subscription management and achieving a 25% faster transaction flow compared to the initial prototype.
- Designed and implemented a Firebase-based synchronisation system for multi-location content delivery, reducing sync latency by 40% and ensuring real-time updates across 50+ devices.
- Developed and deployed an AI-powered chatbot integrated with Pinecone, Gemini and OpenAI embeddings, improving contextual query resolution accuracy by 35%.
- Led and mentored a backend team of 5 developers, managing architecture decisions, code reviews and delivery timelines, increasing the on-time delivery rate to 95%.
- Created backend documentation and API specifications, enabling cross-functional collaboration and reducing onboarding time for new developers by 20%.

Projects

Clinical Conversation AI | REST API, AWS Lambda, Cognito, DynamoDB, Anthropic Claude, Socket.io

- Conversational system for healthcare providers that captures doctor-patient interactions, applies speech-to-text with medical term extraction, and generates structured clinical notes (SOAP/BIRP) integrated securely with DynamoDB and downstream EHR systems.

Education

University of the Punjab | Bachelor of Business and Information Technology (BBIT) | 2019 - 2023

Lahore, Pakistan

Certifications

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

Advanced RAG with Vector Databases and Retrievers | IBM - 2025