

Abuzar Zulfiqar

AI/ML Engineer | AI Agents, RAG & LLM Automation | Azure & AWS

Lahore, Pakistan · **Open to relocation** | +92 301 0202022 | abuzarzulfiqar18@gmail.com
abuzar18.github.io | linkedin.com/in/abuzar-zulfiqar | github.com/abuzar18 | Upwork

30× faster	**25+**	**−35%**	**60–70%**
p50: 126 s → 4 s on 1.24M records	production RAG assistants shipped	LLM API cost (−40% latency)	less drafting time via AI agents

Summary

AI/ML engineer with **3+ years** taking Generative AI from prototype to production for enterprise knowledge, financial research and healthcare-insurance workflows, now delivering independently for startups and SMBs. Builds RAG assistants, agentic document workflows and LLM data pipelines with Python, FastAPI, LangGraph, Azure OpenAI and AWS, and focuses on what keeps AI reliable and affordable in production: evaluation, grounding, PII redaction, idempotency, observability and cost control.

Core Skills

Generative AI & Agents: LLMs (GPT-4, Azure OpenAI, Llama), LangGraph, LangChain, RAG, embeddings, tool calling, structured outputs, prompt engineering, human-in-the-loop agents, text-to-SQL, n8n

Backend & Data: Python, FastAPI, Pydantic, REST APIs, async processing, PostgreSQL, PgBouncer, Redis, SQL, Pandas, Pinecone, FAISS, ChromaDB

Cloud & DevOps: AWS (Lambda, SQS, S3, API Gateway, RDS, CloudWatch, CloudFormation), Azure, Docker, Kubernetes, Terraform, GitHub Actions, Azure DevOps, CI/CD

LLMOps / MLOps: Evaluation harnesses, grounding checks, schema validation, prompt/model versioning, drift monitoring, release gates, PHI/PII redaction, audit logs, latency & token-cost telemetry

ML & Deep Learning: PyTorch, TensorFlow, Hugging Face Transformers, scikit-learn, NLP, text classification, information extraction, RNNs, SHAP, LIME

Experience

Freelance AI Engineer – AI Agents, RAG & Automation
Aug 2026 – Present
Self-employed (Upwork & direct clients) — *Remote*

- Build AI agents, RAG assistants and workflow automations for startups and SMBs, scoping each engagement from prototype to a measurable, production-ready deployment.
- Packaged a reusable delivery standard (policy gates, human approval, idempotency, audit logging, testing and CI/CD) as open-source projects, led by a citation-grounded Llama RAG system with an evaluation harness (see Open Source).

Software Engineer (AI/ML)
Aug 2025 – Jul 2026
Rolustech — *Lahore, Pakistan (Hybrid)*

- Cut p50 career-matching time from ∼**126 s to 4 s (∼30×)** across **1.24M account records** at 100 concurrent requests by replacing sequential scans with normalized, indexed PostgreSQL queries and event-driven API Gateway, Lambda and SQS processing with S3 payload offloading.
- Owned LLM research and enrichment pipelines for executive biographies, investment intelligence and earnings-call analysis: search orchestration, content extraction and source-aware synthesis into schema-validated JSON/HTML synced to the CRM.
- Designed meeting-intelligence automation: FastAPI services schedule meeting bots, ingest recording and transcript webhooks into S3, generate Azure OpenAI summaries and action items, and sync results to **Salesforce**.
- Hardened pipelines with authentication, signed requests, idempotency, deduplication, retries with exponential backoff and replay-safe recovery; managed database load with PgBouncer and worker concurrency limits.
- Ran CI/CD for tests, migrations, Terraform infrastructure and deployments; added structured logging, CloudWatch monitoring and per-request LLM token, cost and latency telemetry (including corrected reasoning-token billing).

Stack: Python, FastAPI, Azure OpenAI, PostgreSQL, PgBouncer, AWS Lambda/SQS/S3/API Gateway, Terraform, Salesforce

Associate AI Engineer
Jul 2023 – Aug 2025
OptimusFox (joined as AI Intern, Jun – Jul 2023) — *Lahore, Pakistan*

- Led delivery of **25+ policy-focused RAG assistants** (GPT-4, Llama, LangChain, Pinecone) with automated ingestion, ETL, embedding and index sync on Kubernetes, reducing manual knowledge-refresh work by ∼**90%**.

- Reduced inference latency by ∼**40%** and LLM API costs by **35%** through prompt routing and Redis caching.
- Built an **agentic insurance-appeal workflow** in LangGraph: document ingestion, fact extraction, RAG over policy clauses and citation-grounded drafting with Azure OpenAI; schema-validated DOCX/PDF output with e-signature handoff, PHI/PII redaction, audit logs and prompt/model versioning. Cut drafting time by **60–70%**.
- Developed **SAM BOT**, a conversational BI service (FastAPI, Azure OpenAI, PandasAI) that turns natural-language questions over spreadsheets and databases into tables, charts and executive summaries; encoding business rules into text-to-SQL prompts cut manual analysis time by ∼**60%**.
- Contributed to architecture reviews, technical planning, production troubleshooting and mentoring for customer-facing AI services; also ran R&D, researching and prototyping new GenAI approaches and handing validated work to the delivery team.

Stack: Python, LangChain, LangGraph, GPT-4, Llama, Azure OpenAI, Pinecone, Redis, PandasAI, FastAPI, Kubernetes

Open Source

- **Llama Agentic RAG** (flagship): self-correcting, citation-grounded RAG on local Llama with relevance grading, query rewrite, groundedness checks and an LLM-as-judge eval harness; FastAPI service and MCP server. *LangGraph, LangChain, Ollama, FastAPI, Docker, CI.*
- **Agentic AI Control Plane**: safe-agent architecture with policy gates, human approval, idempotent execution and typed audit events. *Python, FastAPI, Docker, CI/CD.*
- **AWS MLOps Reference Platform**: release promotion gates, drift measurement and infrastructure as code. *FastAPI, CloudFormation, Docker, CI/CD.*
- **Production AI Systems Blueprint**: citation-verified AI workflows with health checks, tests, observability and CI/CD. *FastAPI, Docker, GitHub Actions.*

Research Leadership & Teaching

Lead Researcher (Part-time)
Oct 2019 – Jun 2023
COLAB, FAST-NUCES

- Led a **30-member** applied AI research team across NLP and ML: set technical roadmaps, reviewed architectures, standardized reproducible experiments and mentored researchers on evaluation, code quality and deployment.
- Built a Wi-Fi human-activity-recognition system (RNN on Wi-Fi signals) with edge inference on Raspberry Pi for healthcare monitoring.

Teaching Assistant – Digital Logic Design
Jan 2021 – Jun 2021
FAST-NUCES

- Designed labs, assignments and workshops for **80+ students**, contributing to a ∼**15% improvement** in average assessment performance.

Education

BS Computer Science, FAST National University of Computer and Emerging Sciences (FAST-NUCES) **2019 – 2023**

Certifications

Microsoft Certified: Azure AI Engineer Associate (2024) | Hugging Face NLP Course (2023) | Databricks Lakehouse Fundamentals (2023) | Samsung Innovation Campus, Artificial Intelligence (2022)

Recognition & Community

- Ranked **top 30 of 200+ teams** at the ICPC programming contest hosted by GIKI (2022).
- Organized and led **three LLM and MLOps workshops** for **150+ participants** at FAST-NUCES.