

Muhammad Annus Shabbir

AI Engineer | Machine Learning Engineer | LLM/RAG Backend Engineer

Lahore, Pakistan | annus.0191@gmail.com

linkedin.com/in/muhammad-annus-shabbir

SUMMARY

AI Engineer with 5+ years of production software engineering experience and an MS in Artificial Intelligence from LUMS. Specialized in LLM applications, retrieval-augmented generation, agentic AI, ML/NLP pipelines, deep learning, model optimization, and scalable backend systems. Built production RAG platforms, government-scale AI services, LLM evaluation systems, and edge-optimized LLM deployments using FastAPI, Elasticsearch, PostgreSQL, PyTorch, Hugging Face, LangGraph, MLflow, AWS, and modern backend infrastructure.

TECHNICAL SKILLS

AI/ML: Machine Learning, Deep Learning, Natural Language Processing, Retrieval-Augmented Generation, Agentic AI, ReAct Agents, Prompt Engineering, LLM Evaluation, Model Context Protocol, Efficient LLMs, Diffusion Models, Flow Matching

Frameworks and Models: PyTorch, Hugging Face, LangGraph, LangChain, MLflow, LangSmith, scikit-learn, Sentence-Transformers, Whisper, WhisperX, FAISS, ChromaDB, llama.cpp, llama-cpp-python, Mistral, LLaMA, RoBERTa, HuBERT

Backend Engineering: Python, FastAPI, Django, Node.js, Express.js, NestJS, TypeScript, JavaScript, REST APIs, SSE Streaming, Webhooks, Async Processing, Monoliths, Monorepos, Event-Driven Systems

Data and Search: PostgreSQL, pgvector, Elasticsearch, BM25, dense kNN, Reciprocal Rank Fusion, Prisma ORM, Redis, BullMQ, Informix SQL, Embeddings, Content Hashing, Data Enrichment

Cloud and DevOps: AWS EC2, RDS, S3, Lambda, CloudFront, IAM, Secrets Manager, Docker, GitHub Actions, Nginx, Gunicorn, Cloudflare, Heroku, GCP, Sentry, Structured Logging, Readiness Checks

Frontend and Tools: Next.js, React.js, Angular, Redux, Tailwind CSS, Chakra UI, HTML5, CSS3, Postman, SOAP UI, Jira, Asana, Linear, Slack, Figma, Mixpanel, Google Tag Manager

EDUCATION

Lahore University of Management Sciences (LUMS)

Sep 2024 – Jun 2026

MS in Artificial Intelligence, CGPA: 3.90/4.00

Lahore, Pakistan

- Focus: LLMs, deep learning, computer vision, NLP, efficient AI, agentic AI systems, and AI alignment.
- Selected MS projects: LLM optimization for edge deployment, multi-agent game builder, diffusion/flow matching reward alignment, call center agent evaluation, and MCP-based cyber-physical AI control.

University of Engineering and Technology Lahore (UET)

Sep 2016 – Aug 2020

BSc Electrical Engineering, CGPA: 3.595/4.00

Lahore, Pakistan

- Final Year Project: Built a single-sensor Non-Intrusive Load Monitoring system for appliance-level energy disaggregation using machine learning classification.

Beaconhouse School System

Sep 2014 – Jun 2016

*A-Level, Grades: 3 A*s*

Lahore, Pakistan

Beaconhouse School System

Sep 2011 – Jun 2014

*O-Level, Grades: 3 A*s, 4 As, 1 B*

Lahore, Pakistan

EXPERIENCE

Phaedra Solutions

Oct 2025 – Present

Senior AI Developer

Lahore, Pakistan

- Designed and implemented **AskTT AI Service**, a production **FastAPI RAG backend** for the MyTT Government Portal with hybrid retrieval, BM25, dense kNN, Reciprocal Rank Fusion in Elasticsearch, OpenAI-powered answers, numbered citations, KB-only and general-answering modes, and context-aware query rewriting for follow-up questions.
- Delivered end-user AI features on one backend service: streaming chat over **Server-Sent Events**, AI-assisted search with an LLM-generated overview and ranked results, source/citation surfacing, conversation APIs, and shared conversation state across text and voice workflows.
- Built voice chat using **Whisper speech-to-text, RAG answer generation, and optional text-to-speech**, with optional S3 persistence for voice assets and reusable conversation state across modalities.
- Engineered the AI data and operations layer with PostgreSQL and Prisma for conversations/messages, idempotent knowledge-base ingestion, embeddings, content hashing, background starter-prompt generation, Redis-ready caching, dependency readiness checks, structured logging, and optional LangSmith tracing for LLM observability.
- Developed an LLM-based recruitment matching engine for the Trinidad and Tobago Government, evaluating **400,000+ applicant-vacancy pairs** using weighted scoring models, prompt engineering, LLM-powered seniority inference, and structured skill/experience/domain-fit extraction.
- Built explainable scoring outputs for hiring workflows, including skill overlap, experience fit, seniority gap penalties, domain match, rationale summaries, and transparent matching reasons for ministries and agencies.
- Optimized high-volume LLM inference through asynchronous parallel execution, intelligent caching, response compression, checkpointed batch processing, and failure recovery, reaching **30–50 real-time matches/sec** while

reducing token cost and latency.

Esketchers

Senior Software Engineer

Nov 2021 – Oct 2025

Lahore, Pakistan

- Built an internal **Django RAG proposal generation system** using hybrid semantic and keyword search over company portfolios, pgvector embeddings, AI-generated summaries, and multi-format document processing for PDF/DOCX sources.
- Deployed the RAG system on **AWS EC2 and RDS PostgreSQL/pgvector** with Nginx, Gunicorn, and CI/CD pipelines, improving availability, maintainability, and reuse across proposal workflows.
- Developed and maintained Mentalhealth Match using **Next.js 14, Angular SSR, Node.js, Express, PostgreSQL, Prisma, Redis/BullMQ, Cloudflare, Mixpanel, and ActiveCampaign**, supporting **30,000+ therapists** and high-volume seeker traffic.
- Improved SEO, crawlability, and web performance through Next.js App Router migration, Angular server-side rendering, CDN caching, Cloudflare optimization, and reporting database migration from Heroku to Amazon RDS; achieved **85+ Google PageSpeed** on key pages.
- Implemented Redis/BullMQ queues for purge-cache and bulk image import tasks, improving responsiveness and reliability of background operations.
- Integrated analytics and CRM automation using Google Tag Manager, Google Sheets, Mixpanel, and ActiveCampaign to improve tracking, reporting, and engagement funnels.

i2c Inc.

Software Engineer

Jul 2020 – Nov 2021

Lahore, Pakistan

- Supported production digital payment systems across global clients, analyzing APIs, frontend behavior, Informix SQL data, ETL jobs, logs, and transaction flows for Visa, Mastercard, Discover, and UnionPay integrations.
- Certified payment APIs using Postman and SOAP UI, created detailed test plans and cases, coordinated fixes through Jira, and validated critical releases with E2E, smoke, negative, and beta testing.
- Troubleshoot client-facing issues by investigating application logs, data flows, database records, and transaction processing behavior in PCI-domain payment systems.

SELECTED AI AND MACHINE LEARNING PROJECTS

LLM Optimization for Edge Deployment | *PyTorch, Hugging Face, llama.cpp, CUDA, GGUF*

2025

- Compressed **LLaMA-3.1-8B from 15GB to 3.95GB** using Taylor-based structured pruning, SmoothQuant outlier suppression, and Fisher Information-guided mixed-precision quantization, reducing model size by **73.6%**.
- Reduced MLP parameters from 8.03B to 6.34B with 30% structured pruning and used Q4/Q5/Q6 mixed-precision quantization with sensitivity-aware bitwidth allocation across 32 transformer layers.
- Achieved **273% decode speedup** on CPU, improving generation from 1.16 to 4.33 tokens/sec using imatrix calibration on WikiText-2; deployed on Raspberry Pi 4 with an approximately 4.2GB memory footprint.

Multi-Agent Dino Game Builder | *LangGraph, ReAct Agents, Databricks, MLflow*

2026

- Built an agentic AI system that autonomously generated a playable Chrome Dino-style game through a multi-agent workflow with Architect, Engineer, QA, Scorer, and Human-in-the-Loop agents.
- Designed LangGraph orchestration with conditional routing, persistent state, memory checkpoints, and iterative feedback loops to refine game code across multiple development cycles.
- Implemented tool-using ReAct agents for code generation, execution, debugging, fallback recovery, and QA validation of gameplay features including jumping, ducking, obstacles, scoring, and high-score tracking.
- Added LLMops monitoring with MLflow to track agent latency, groundedness, execution quality, and reasoning performance across iterations.

Inference-Time Reward Alignment for Diffusion and Flow Matching | *PyTorch, Diffusion Models, Flow Matching, VLMs*

2026

- Researched sparse vs. dense inference-time reward alignment for diffusion and flow matching models, focusing on reward hacking, credit-assignment mismatch, high-noise denoising steps, and compute-quality tradeoffs.
- Proposed **ODE-Dense Inference-Time Alignment (ODE-DITA)** using deterministic n-step probability-flow ODE rollouts, Sequential Monte Carlo resampling, dense gain weighting, and a VLM semantic corrector in the hybrid variant.
- Evaluated 50 prompts across 3 seeds and showed sparse best-of-K selection improved PickScore by **+0.46**, while dense SMC did not significantly outperform vanilla under matched compute, revealing a practical training-inference gap in dense reward guidance.

Custom LLM System for Call Center Agent Evaluation | *Whisper, FAISS, Mistral 7B, Gradio*

2025

- Engineered an end-to-end evaluation pipeline using Whisper/WhisperX transcription, speaker diarization, RoBERTa sentiment analysis, HuBERT tonal emotion detection, and a FAISS-backed RAG layer over evaluation rulebooks.
- Hosted a locally quantized Mistral 7B model through llama-cpp-python and Gradio to extract KPIs such as empathy, tone, compliance, problem resolution, and coaching feedback from call transcripts.

MCP Cyber-Physical AI Controller | *FastMCP, Raspberry Pi, GPIO, Blockchain, Google Calendar*

2026

- Built an MCP server exposing **16 tools** across P2P energy-trading blockchain operations, Raspberry Pi GPIO fan control, and Google Calendar automation, enabling natural-language system control through Claude Desktop.
- Integrated blockchain tools for participant registration, trade submission, phase advancement, round/status queries,

and smart-contract interaction while abstracting transaction details behind agent tools.

- Implemented Raspberry Pi relay control for a physical fan with sub-100ms response, OAuth-based Google Calendar scheduling with Google Meet creation, and Tailscale remote access for distributed Raspberry Pi nodes.

AUTO KPI: AI Pipeline for Automated Call Center QA | *WhisperX, RoBERTa, HuBERT, LangChain, FAISS, Mistral* 2025

- Designed a multimodal call-center QA pipeline combining speech transcription, diarization, sentiment analysis, tonal emotion detection, rulebook retrieval, LLM scoring, and qualitative feedback generation.
- Built a scalable evaluation framework for high-volume call data, enabling objective agent assessment and customer-experience insights.

Agentic LLMs with Reflection and Tool Integration | *Llama 3, Groq API, HTTP Tools, IPython* 2025

- Implemented agentic reasoning patterns in Llama 3 with iterative self-reflection and external tool interaction for more reliable outputs.
- Built workflows using Groq API, HTTP integration, and IPython to demonstrate practical tool-using LLM behavior.

Car Price Prediction Model | *Python, scikit-learn, Regression, L2 Regularization* 2024

- Developed linear regression models with and without L2 regularization on a real-world PakWheels dataset, including preprocessing, baseline benchmarking, and comparative evaluation.

Non-Intrusive Load Monitoring | *Machine Learning, Signal Processing, Classification* 2020

- Built a single-sensor NILM pipeline to identify electrical appliances from unique load signatures using machine learning classification and end-to-end signal acquisition to appliance-level prediction.

CERTIFICATIONS

- Retrieval Augmented Generation, DeepLearning.AI, Aug 2025
- Neural Networks and Deep Learning, DeepLearning.AI, Jul 2025
- AI For Everyone, DeepLearning.AI, Jul 2025
- HTML, CSS, and JavaScript for Web Developers, Johns Hopkins University, Nov 2021

VOLUNTEER WORK AND LEADERSHIP

Shaukat Khanum Memorial Cancer Hospital Lahore, Pakistan

- Designed Arduino-based science projects, including LED pattern systems and a Bluetooth-controlled car, for patient learning activities.
- Explained technical concepts and science projects to diverse patient groups in an accessible, hands-on format.

UET IET On Campus Society, Vice President Lahore, Pakistan

- Organized the All Pakistan International Conference on Electrical Engineering at UET Lahore.
- Represented IET UET at Lahore Science Mela, featuring 50+ projects from across Pakistan, and organized IET's Present Around The World competition.