

MUHAMMAD MUSTAFA

+92 317 8761858 | mustafadev486@gmail.com | linkedin.com/in/muhammad-mustafa | github.com/muhammadmustafa18

EDUCATION

FAST NUCES Karachi | *CGPA: 3.44 / 4.0*

Karachi, Pakistan

Bachelor of Science in Computer Science

Aug 2023 – Jun 2027

- Dean's List Awardee — 2 consecutive semesters
- Relevant Coursework: Algorithms, Database Systems, Computer Networks, Operating Systems, Compiler Construction, Artificial Intelligence, Deep Learning

EXPERIENCE

Backend Engineer (Contract) — Zolvr.ai

Aug 2026 – Oct 2026

Remote | Contract

- Developed core backend services for a production platform as a contract engineer, owning API design and third-party integrations end to end
- Integrated **Stripe** (payments, webhooks) and **Shopify** with idempotent webhook handling and retry logic
- Designed and shipped an **MCP server** exposing platform capabilities as tools for AI agents
- Load- and rate-limit-tested critical endpoints with **k6** and fixed the bottlenecks found; added **Redis** caching on hot read paths and served assets via **BunnyCDN** to cut latency and database load

AI Engineering / Agentic AI Instructor — AI Season

Jul 2026 – Present

Karachi, Pakistan | Remote | Part-time

- Design and deliver 3 sessions on agentic AI engineering for the AI Season cohort: LangGraph orchestration, Harness Engineering with design patterns, and MCP architecture
- Teach the backend layer of AI systems end to end — API design, SSE streaming, tool-calling protocols, persistence, and deployment — using shipped production projects as the running examples
- Review the wider curriculum (RAG, vector databases, memory layers, AI security, deployment with Nginx / Certbot / Postgres) and author the capstone rubric: an end-to-end agent with MCP integration, vector memory, and a live deployable URL

IT Intern — United Bank Limited (UBL)

Jun 2026 – Present

Karachi, Pakistan | On-site

- Embedded with the enterprise data and MLOps teams on a **Teradata** warehouse and **Informatica** ETL pipelines, tracing data from source ingestion through staging, transformation, and load, including reconciliation and data-quality checkpoints
- Followed the ML lifecycle on **Kubeflow Pipelines** over **Kubernetes** — warehouse joins, feature selection, training, and versioned deployment — and studied INT4 / INT8 quantization for on-prem inference
- Rebuilt the observed patterns independently on a local PostgreSQL and Python / Pandas stack to turn exposure into hands-on working knowledge

Full Stack Developer Intern — The Bridge of Hopes

Jul 2026 – Present

Karachi, Pakistan | Remote

- Build and extend a production Android application in Kotlin with clean-architecture separation between UI, domain, and data layers
- Integrated on-device inference models with async workflows that keep the UI thread free, optimizing memory, battery, and background execution on resource-constrained devices

PROJECTS

Proteus | *FastAPI, SQLAlchemy, ChromaDB, MCP SDK, OAuth2, SSE, Docker*

2026

- Multi-service FastAPI backend powering a self-hosted multi-provider AI workspace (OpenAI, Groq, Ollama, any OpenAI-compatible endpoint) with SSE-streamed responses
- MCP *client* layer supporting stdio and HTTP transports plus OAuth2 for remote servers — one-click connections to Notion, GitHub, and custom tool servers
- Vector-backed persistent memory via ChromaDB scoped per user session; agent mode with tool calling (file I/O, bash, Python execution, web search) and a goal/verifier feedback loop with automatic retry

Recluze | *Tauri 2, FastAPI, Baileys, LangGraph, SQLite, .msi Packaging*

2026

- Open-source desktop app that turns any WhatsApp number into a brand-voiced AI responder — WhatsApp login, LLM key, knowledge base, live in under 5 minutes
- Local-first architecture with multi-provider LLM routing (Anthropic / OpenAI / Groq), SQLite persistence, and full conversation memory — no data leaves the device
- FastAPI sidecar handles on-device inference; shipped as a single .msi installer for non-technical users

Redbus Karachi | *Next.js, TypeScript, Leaflet.js, Dijkstra's Algorithm, GeoJSON*

Aug 2025

- Converted raw Google Earth geographic data into GeoJSON and implemented a Dijkstra shortest-path routing service for city-scale transit queries
- Served 1,600+ unique users with sub-second route response across the Karachi network
- Co-led a 2-person team, owning the routing backend while the partner owned the map UI layer

TECHNICAL SKILLS

Languages: Python, TypeScript, JavaScript (ES6+), Java, Kotlin, SQL, Bash/Shell, C++

Backend / APIs: FastAPI, Node.js, Spring Boot, Hibernate/JPA, REST API Design, OAuth2 / JWT, SSE, WebSockets, Webhooks, Rate Limiting, Caching, Stripe, Shopify, OpenAPI / Swagger, Pydantic, Middleware, Pub/Sub Patterns

Databases / Storage: PostgreSQL, SQLite, Redis (Upstash), Teradata, Informatica (ETL), NeonDB, Supabase, ChromaDB, pgvector

Infra / DevOps / Testing: Docker, Docker Compose, Kubernetes, Kubeflow Pipelines, Git / GitHub Actions, Nginx, BunnyCDN, k6 (Load Testing), Vercel, cloudflared, Tailscale Funnel, Cron Jobs, OpenTelemetry, Langfuse, Structured Logging, Trace IDs

Data / BI: Power BI, Tableau, Pandas, NumPy, ETL/ELT, Data Warehousing, Data Validation, Reconciliation

AI / ML: LangGraph, LangChain, CrewAI, MCP (Client & Server), OpenAI / Anthropic / Groq SDKs, RAG, Vector Retrieval, Multi-Agent Systems, Prompt Engineering, Quantization (INT4 / INT8)

Frontend: React, Next.js (App Router), Tailwind CSS

CERTIFICATIONS & AWARDS

Winner — Design and Development Hackathon, FAST NUCES Karachi (DevDay 2026)

Winner — Web Development Hackathon, Sir Syed University (SMEC 2026)

Winner — App Development Hackathon, NED University (AIC 2026)

LeetCode — Active DSA problem solver | leetcode.com/muhammad.mustafa