

SYED MOHAMMAD ABBAS

Full-Stack AI Engineer • Backend Systems Engineer • RAG / Automation Developer

Email: syed.abbas.ons@gmail.com
Location: Karachi, Pakistan
Portfolio: thesyedabbas.xyz
GitHub: github.com/SyedAbbas-CM
LinkedIn: linkedin.com/in/syed-mohammad-abbas
Upwork: upwork.com/freelancers/~0116a4b80466a0625f
Phone: +923318774222
Live System: dashboard.eternalconquests.com

SUMMARY

Full-Stack AI Engineer and Backend Systems Engineer with experience building enterprise RAG applications, LLM orchestration layers, multi-tenant SaaS platforms, FastAPI/NestJS backends, PostgreSQL/Redis infrastructure, scraping pipelines, and production automation systems. Strong background in Python, TypeScript, FastAPI, NestJS, PostgreSQL, Redis, Docker, FAISS, ChromaDB, PyTorch, LangChain, Azure, AWS, Unity, and Godot. Comfortable working directly with clients and stakeholders to turn unclear requirements into technical milestones, explain tradeoffs, debug production issues, and deliver usable systems.

TARGET ROLES

AI Engineer • Backend Engineer • Full-Stack Engineer • RAG Engineer • LLM Application Developer • Automation Engineer • Python Developer • FastAPI Developer • NestJS Developer

TECHNICAL SKILLS

Languages: Python, TypeScript, Kotlin, Java, GDScript, C#, C++, SQL, Bash
LLM / RAG: LangChain, FAISS, ChromaDB, BM25, embeddings, vector search, hybrid search, cross-encoder reranking, prompt orchestration, multi-agent workflows
Backend: FastAPI, NestJS, Ktor, Django REST, REST APIs, PostgreSQL, Redis, Docker, webhooks, HMAC validation, API integrations, multi-tenant architecture
Cloud / Data: Azure, Azure Flexible Server, Azure Databricks, Azure Service Bus, AWS, PySpark, SQL Server, Data Lake pipelines
ML / Audio AI: PyTorch, Faster-Whisper, Qwen3-TTS, Qwen3-ASR, forced alignment, WER scoring, FFmpeg, PoseNet
Graphics / Simulation: Unity 3D, Godot 4.6, C#, GDScript, GLSL/GDShader, WebGL, Three.js, Structure-of-Arrays, pathfinding
Delivery: Client communication, stakeholder updates, requirements breakdown, technical scoping, milestone planning, demo delivery, production debugging

PROFESSIONAL EXPERIENCE

Cloud Primero

Full-Stack AI & Systems Backend Engineer

Karachi, Pakistan
Aug 2024 – Present

Project: Enterprise RAG Platform

Stack: FastAPI, FAISS, BM25, Cross-Encoder Reranking, LLM Orchestration

AI research and consulting assistant that retrieved internal knowledge, generated grounded answers, streamed responses, and automated multi-step document workflows.

- Developed a multi-LLM orchestration layer for an enterprise RAG platform serving **500+ concurrent users** with **99.2% uptime** across containerized production deployments.
- Reduced p95 response latency by **73%** from **2.3s to 620ms** through async FastAPI streaming, connection pooling, request-path optimization, and backend execution profiling.
- Implemented hybrid **FAISS + BM25** retrieval with cross-encoder reranking, improving grounded answer quality with **91% retrieval accuracy** and **94% context retention**.
- Built *PhaseGraph V2*, a DAG-based planner-retriever-writer-reviewer workflow engine for multi-step consulting document generation and review automation.
- Worked with product and client-facing teams to clarify AI workflow requirements, explain retrieval and latency trade-offs, and convert ambiguous feature requests into deliverable backend milestones.

Project: Multi-Tenant Dental SaaS Backend

Stack: NestJS, PostgreSQL, Redis, Azure Flexible Server

Production SaaS backend for tenant-isolated dental operations, claims processing, payment workflows, clinical charting, and Unity-based visualization.

- Architected a database-per-tenant PostgreSQL isolation model on Azure Flexible Server across **22 modules**, **46 domain entities**, and approximately **68,700 LoC**.
- Designed Redis-backed runtime tenant resolution using `dbConfig:<tenantCode>`, automated idle-connection cleanup, and semaphore-based request controls to protect shared infrastructure under load.

- Built a custom NestJS `EntityManagerForkInterceptor` to bind request-scoped `em.fork()` transaction contexts, resolving connection pool exhaustion during concurrent tenant traffic.
- Programmed a Vektis MZ301 v1.3 claims serialization engine with fixed-width 310-character line validation, integer-cent conversion, BSN, AGB provider type, BIG, and UZOVI rule checks.
- Built an 8-stage PayT dunning and invoice webhook pipeline using raw-body HMAC-SHA256 validation, `timingSafeEqual()`, and partial-unique PostgreSQL indexes to prevent duplicate payment processing.
- Engineered tooth-level clinical observation APIs mapping Universal identifiers across surfaces and condition states, feeding chronological color updates into a Unity 3D dental visualization client.

Project: Truffle POS Systems, Canada

Stack: Ktor, Netty, ESC/POS, TCP 9100

POS printer integration for receipt parsing, Epson ESC/POS behavior, network discovery, socket transport, and hardware communication debugging.

- Reverse-engineered Epson ESC/POS and EPSONq network discovery behavior for Truffle POS systems in Canada using Netty/Ktor socket architectures over TCP port 9100.
- Built a print-buffer parsing layer with custom Trie and finite-state token extraction, achieving **95%+ field extraction accuracy** from unformatted receipt and printer payloads.
- Diagnosed inconsistent hardware communication behavior through real printer testing, protocol tracing, and structured documentation of integration assumptions and edge cases.
- Communicated protocol limitations, test findings, and implementation tradeoffs to stakeholders so integration decisions could be made without deep hardware/protocol knowledge.

Climate Culture AI

Full-Stack AI Engineer

Remote / Contract
Feb 2024 – Jun 2024

Project: Recommendation Engine

Stack: FastAPI, BGE Embeddings, LightGBM, MMR

On-device personalization, embedding retrieval, ranking, diversification, and telemetry-driven recommendation weighting.

- Built an on-device personalization and recommendation engine combining BGE embeddings, LightGBM ranking, MMR diversification, and telemetry-driven vector weighting inside a containerized FastAPI ecosystem.
- Converted product-level recommendation goals into ranking features, embedding retrieval flows, diversification controls, and feedback-driven personalization logic.

Active Mentor

Software Developer, Computer Vision & Runtime Systems

Karachi, Pakistan
Nov 2023 – Feb 2024

Project: Kinematic Computer Vision Integration

Stack: TensorFlow, PoseNet, Kotlin / Android Runtime

Pose estimation, tablet runtime integration, posture overlays, and automated exercise motion feedback.

- Integrated TensorFlow PoseNet Thunder and Lightning profiles into tablet runtimes for real-time posture overlays, kinematic node tracking, and automated rep-counting feedback.
- Built motion-feedback overlays for live exercise correction by connecting pose landmarks to canvas rendering, repetition detection, and user-facing posture guidance.
- Tested model/runtime behavior on constrained tablet hardware and tuned implementation paths for real-time visual feedback.

ImproData

Data Engineer

Karachi, Pakistan
Jul 2023 – Nov 2023

Project: Enterprise ETL Pipelines

Stack: Azure Databricks, PySpark, SQL Server

Data Lake ingestion, PySpark normalization, dashboard refresh optimization, and SQL tracking views.

- Developed high-throughput Azure Databricks and PySpark extraction pipelines for unstructured assets stored in Azure Data Lake, with indexed SQL Server tracking views to improve dashboard refresh cycles.

Mindstorm Studios

Software Engineering Intern

Karachi, Pakistan
Jun 2023 – Aug 2023

Project: Unity Mobile Game Jam

Stack: Unity, C#

Gameplay implementation, team coordination, mobile optimization, and deadline-driven delivery.

- Led a junior engineering cohort during a 2-month Unity mobile game jam, coordinating gameplay implementation, optimization passes, and delivery ahead of evaluation deadlines.

SELECTED PROJECTS

Project: Solo Game Engine: Cross-Platform Action-Roguelike + RTS Hybrid

Stack: Godot 4.6, GDScript, GLSL

Independent game/simulation framework combining RTS-scale entity simulation, procedural generation, networking, shaders, and editor tooling.

- Built a **76,569 LoC** Godot engine framework exported across **6 deployment builds**, demonstrating large-scale solo architecture across gameplay systems, runtime managers, content pipelines, and tooling.
- Implemented a data-oriented RTS simulation using Structure-of-Arrays layouts across approximately **30 packed attribute arrays**, supporting **2,000 active instances** without per-unit GameObject allocation.
- Developed an 80-px spatial hash grid to convert broad proximity scans into near-constant-time localized queries for combat, targeting, and unit interaction systems.
- Built BFS flow-field pathfinding on a 32-px matrix, with staggered AI update windows for retargeting, flow recomputation, formation damage counters, and rear-impact modifiers.
- Programmed host-authoritative co-op networking over ENet with UPnP port discovery, split broadcast intervals, compressed snapshot deltas, and client-side visual ghost objects.
- Created a procedural content pipeline for random-walk dungeon tracks, cellular-automata caverns, Voronoi mine-carving nodes, and density-biased frontier expansion.
- Authored custom GLSL/GDShader effects for world-space liquids, dynamic player-relative wakes, portal swirls, and internal performance telemetry monitors.
- Designed the project as a production-style codebase with modular managers, reusable resource definitions, profiling tools, editor automation, and a custom **World Baker** pipeline generating **1,944 compressed chunk files**.

Project: Project AI Audio Lab: Automated Voice Cloning & Audio Production Pipeline

Stack: Python, PyTorch, Faster-Whisper, Qwen3, FFmpeg

Local speech generation, ASR quality control, forced alignment, resumable batch execution, and audio mastering pipeline.

- Built a local speech generation and audio production stack using Qwen3-TTS, Qwen3-ASR, Qwen3-ForcedAligner, Faster-Whisper, and FFmpeg for long-form reference-conditioned narration.
- Engineered a resumable batch-processing system using TSV job manifests, per-job JSON state tracking, subprocess PID monitoring, timeout enforcement, and stalled-job recovery scripts.
- Developed ASR-based QA tooling that transcribes generated chunks and calculates Word Error Rate, character similarity, mismatch diagnostics, and pronunciation compliance.
- Built forced-alignment and text-normalization workflows for sub-word timestamps, acronyms, technical symbols, proper nouns, and repeated long-form narration consistency.
- Created FFmpeg mastering chains for crossfades, edge trims, loudness balancing, and room-tone masking to reduce generated-audio seam artifacts.
- Profiled and resolved local inference failure modes including Apple Silicon stalls, prompt instability, reference-audio mismatch, and model noise-floor fluctuations.

Project: Distributed Product Intelligence Engine & Scraper Fleet

Stack: Node.js, Python, Raspberry Pi, Jetson Nano, Edge Nodes

Edge-node scraping fleet, document ingestion, rate-limit handling, clustering, and recurring market signal detection.

- Built a distributed collection fleet across 4 physical edge boards, including Raspberry Pi, Jetson Nano, and Windows To Go nodes, running scraping pipelines that ingested **40,000+ public documents**.
- Managed rate limits through OAuth failure traps and deployed clustering pipelines using Jaccard matrices and TF-IDF vectors to group recurring industry indicators.

EDUCATION

Bachelor of Computer Science

FAST-NUCES

GPA: 3.2 Advanced Level (A-Levels) & Ordinary Level (O-Levels)

Beaconhouse School System *A-Levels: 3 As, 1 B • O-Levels: 7 As, 2 Bs* **Languages:** English, Urdu • **Certification /**

Assessment: English C1 Advanced

Graduation: 2024