Sohaib Ali

Founder & CEO, Olympus Labs | Software & AI Systems Engineer

Lahore, Pakistan | US, EU and AU hours

OLYMPUS LABS
(SMC-PRIVATE) LIMITED

sohaibali999@gmail.com · +92 311 1714297 · github.com/SohaibAli9 · linkedin.com/in/sohaibali9 · olympuslabs.ink

01 SUMMARY

I build and operate 0-1 applied AI systems and the platforms around them for companies in the US, Europe, Israel and Australia: full-stack B2B platforms, AI security and inline DLP tooling, financial and CAD data pipelines, and the cloud architecture under them. Every model output is checked in code (deterministic validation, evaluation harnesses, human-approval review) before a person relies on it. Before Olympus Labs, I built and ran services on a 13-service Azure IoT platform handling 40,000+ events a second.

02 EXPERIENCE

Founder & CEO | Olympus Labs (SMC-Private) Limited | Remote

Sep 2026 to Present

Independent practice since Apr 2024, incorporated Oct 2026. 15+ client contracts, 100% Job Success.

AI Security and Inline DLP Engine | Cybersecurity SaaS, Israel — Nov 2026 to Present

- Tuned secret-detection and structured-PII validation rules in a JVM and Rust inline DLP scanner guarding enterprise AI agents.
- Built a single-pass JSON string unescaper with UTF-16 offset mapping and reproducible precision, recall and latency benchmarks.

B2B Acquisition Deal-Flow Platform | US search-fund community — Oct 2026 to Present

- Architecting and building a multi-phase Django and PostgreSQL platform with Circle headless SSO, nightly member sync, billing webhooks, role-gated access and service-provider directories.
- **3,800+ automated tests** run in CI/CD on every change before it deploys to staging.

Construction Plan-Reading Platform | Australian estimating SaaS — Jun 2026 to Present

- Designed, built and operate a production FastAPI API, RabbitMQ worker queues and a web review UI that read residential PDF plan sets for rooms, walls, glazing and floor areas.
- Matched **17 of 17 rooms** and **6 of 6 hand-measured areas** (0.96% scale error), with role-based auth, a human-approval correction loop, automated error diagnosis and regression-gated QA and production deploys with rollback scripts.

Engineering Drawing GPU Pipeline and AWS Operations | US energy AI company — Jun 2026 to Nov 2026

- Built an EC2 GPU batch pipeline that detects, OCRs and annotates text and symbols in **35,000 raster plant engineering drawings**.
- Audited live AWS spend and executed Terraform teardowns of idle environments to cut recurring infrastructure cost.

Financial Filings and Fund-Distribution Pipelines | Philippine financial services firm and US fintech — Jul 2026 to Nov 2026

- Built a deterministic-plus-LLM filings engine for **14 listed utilities**: source-page provenance, 5 accounting validation gates.
- Delivered a deterministic Python and PDF fund-distribution pipeline (1,287 estimate rows, 555 YTD rows, **131 automated validation checks**, 195 unit tests) that catches publisher typos before ingest.

Historical Corporate Annual Report Extraction | Faculty research, University of Tennessee, Knoxville — May 2026 to Jun 2026

- Built an OCR, schema-constrained LLM extraction and validation pipeline across **27,596 scanned annual reports** (1900 to 1945), with 27 structured fields per report, each tied to its source page.
- Derived a 278,816-row directors panel, a board-interlock network and an auditor-switching dataset (1,760 switches).

Insurance Claims (EOB) Extraction Pipeline | PatientStudio, US healthcare SaaS — Aug 2024 to May 2026

- Worked directly with CEO Derek Sanborn to architect and ship a Python and FastAPI service on Google Cloud Run that extracts line-item claims across 28 payer layouts.
- Processed **5,000+ claims at about 95% field accuracy**, cutting manual entry from 15 minutes to under 1 minute per claim.

Multilingual OCR and Quant Research Pipelines | Long-term independent research client — Sep 2024 to Present

- Built a React/TypeScript and FastAPI application on Google Cloud Run for Arabic, Farsi, Russian and Urdu OCR and translation with RTL/LTR handling. Rehired in 2026 for Python workflow automation on a quantitative research pipeline.

02 EXPERIENCE continued

Senior Software Engineer | Strategic Systems International (Quantis.ai, PocketFactory) | Lahore Jul 2023 to Sep 2026

Software Engineer, Jul 2023 to Jun 2025. Promoted to Senior Software Engineer, Jul 2025 to Sep 2026. Enterprise industrial IoT and real-time analytics platform for global beverage and manufacturing customers.

- Led the migration of 14 legacy .NET Framework services off 8 VMs to containerized .NET 8 on Azure, enabling horizontal autoscaling and **cutting infrastructure cost by 35%**.
- Built and maintained services in a 13-service event-driven pipeline streaming **40,000+ telemetry events a second** through Azure IoT Hub, Event Hubs and Service Bus into TimescaleDB at under 100 ms latency.
- Optimized live-dashboard analytics queries in TimescaleDB and PostgreSQL from 90 seconds to 2 seconds (**45x speedup**) for 50+ concurrent enterprise users.
- Designed and shipped a multi-tenant Document Management Service on Azure Data Lake with hierarchical permissions for 250+ enterprise users, **cutting storage cost by 60%** versus traditional blob storage.
- Resolved production-critical bottlenecks, including API thread exhaustion under 100+ concurrent connections, 5 GB caching memory leaks and Event Hub partition duplication, using Redis-based load balancing.
- Owned CI/CD and release management across 6 customer production instances, led code reviews, mentored engineers and built an agent-ready operations knowledge base mapping all 13 services, data stores and runbooks.

03 AI AGENTS, EVALUATIONS AND OPEN SOURCE

- **Production agent tooling:** Operate daily autonomous engineering and client workflows built on Claude Code, custom skills, hooks and MCP servers (Gmail, Drive, Trello, Upwork, RunPod), with explicit human-approval publish gates, untrusted-input isolation, a local Python multi-model routing proxy and scheduled triage agents.
- **Evaluation harnesses and red-teaming:** Audited ActuBench, a 200-item public actuarial LLM benchmark, by rebuilding its LLM judge. Hedged wrong answers passed 11 of 12 and prompt-injected wrong answers passed 9 of 12. Two instruction fixes drove both to **0 of 12** while correct answers kept passing. On a separate 82-case sales-outcome dataset, measured an LLM judge at AUC 0.47 (no better than chance) and replaced it with deterministic checks.
- **Self-hosted production infrastructure:** Deploy and maintain 10 self-authored applications on Linux using Docker Compose, PostgreSQL with pgvector, RabbitMQ, Cloudflare tunnels, self-hosted GitHub Actions CI and uptime monitoring.
- **gerberview:** Browser-native Gerber PCB viewer in TypeScript, with no backend or upload (github.com/SohaibAli9/gerberview).
- **axon:** Self-contained TypeScript edge agent for bridge VMs and air-gapped hosts (github.com/SohaibAli9/axon).
- **t3code:** Merged fix to pingdotgg/t3code that surfaces model refusals on early agent exit (PR #34621).

04 TECHNICAL SKILLS

Languages	Python, TypeScript, C#, SQL, Java / JVM, Rust (DLP rule tuning)
AI, LLMs and data	OpenAI, Anthropic Claude and Gemini APIs, vision models, agent harnesses, MCP, RAG, evals and red-teaming, Google Document AI, Azure Document Intelligence, AWS Textract, Mistral OCR, Pandas and ETL
Backend and frontend	FastAPI, Django, .NET 8 / ASP.NET Core, React, REST APIs, RabbitMQ, SignalR
Databases	PostgreSQL, TimescaleDB, pgvector, Redis, SQL Server
Cloud and operations	AWS (EC2 GPU, Terraform), Azure (IoT Hub, Event Hubs, Service Bus, Data Lake, AKS), Google Cloud Run, Docker, Kubernetes, GitHub Actions, Linux

05 EDUCATION AND LEADERSHIP

BS Computer Science | FAST NUCES, Islamabad 2019 to 2023

- **GPA 3.76 / 4.0**, Silver Medal, Summa Cum Laude, 7x Dean's List.
- Teaching Assistant across 6 CS courses over 12 semesters (2020 to 2023). Created the Buddy System peer mentorship program, pairing 20 seniors with 100 freshmen and **reducing the freshman course failure rate from 20% to 5%**.
- Final-year project: on-device Android driver monitoring system (drowsiness, seatbelt and Urdu voice-emotion models).