# Ahmed Rizawan

AI Platform Engineer / LLMOps

Bangkok, Thailand | +66 65 651 3577 | ahm.rizawan@gmail.com | arizawan.com | linkedin.com/in/rizrizawan | github.com/arizawan

## SUMMARY

AI platform engineer with 15+ years in software and infrastructure. I run LLM workloads in production and make them fast, cheap and reliable: model routing with fallback, response caching, RAG over large datasets, pgvector and MCP servers. Results: an LLM gateway cut response time from 2.7 s to 11 ms with caching; a compliance agent cut screening from 4 hours to 45 seconds at about $0.03 per check; AI translation for 600+ users at under $0.02 of logged spend. Backed by real infrastructure depth: Terraform, Kubernetes, AWS, Google Cloud, CI/CD and observability, with deep SQL across PostgreSQL and MySQL. Based in Bangkok, available immediately.

## SKILLS

**AI platform:** LiteLLM (routing, fallback, retries, circuit-breaking, caching), Ollama, Claude, OpenAI and Groq APIs, pgvector, RAG, AI agents, tool calling, MCP servers, cost and latency optimisation
**Data:** SQL, PostgreSQL, pgvector, MySQL, MongoDB, Redis, Supabase; data modelling and query optimisation
**Cloud and containers:** AWS, Google Cloud, Cloudflare, Hetzner, Vercel; Docker, Kubernetes (GKE), Helm, Linux
**Infrastructure as code:** Terraform, Ansible
**CI/CD and developer experience:** GitLab CI, GitHub Actions, Git, smoke-gated deploys, over-the-air releases, internal tooling
**Observability and operations:** Prometheus, Grafana, Datadog, Sentry, New Relic, Logstash, health checks, on-call, incident response, runbooks, postmortems, SLOs
**Languages and frameworks:** Python (FastAPI), Bash, Node.js, TypeScript, PHP (Laravel); Next.js, React, Expo
**Ways of working:** Agile, cross-functional teams, code review, documentation

## SELECTED PROJECTS

**Local-first LLM gateway** (homelab): LiteLLM routes Ollama models first with cloud fallback, retries, circuit-breaking and response caching (2.7 s to 11 ms); MCP gateway; tool calling benchmarked across local models on an RTX 4060.

**Compliance screening agent** (KYC/AML): Claude and RAG agent over 1.7M+ sanctions records; screening dropped from 4 hours to 45 seconds, at about $0.03 per check versus $2-5 for enterprise tooling.

**MyMuayThai** (co-founder and technical lead, 2025 - present): live booking platform in Thailand with AI chat translation. 600+ users, 6 languages, 4,291 cached translations, logged AI spend under $0.02 since Feb 2026 (Groq).

**Taptober and Nabro** (Norway, 2026): run production operations for a multi-tenant league platform (FastAPI, Expo, Hetzner) and a quiz app live on the App Store (Next.js, Supabase, Vercel). Shipped the Oslo Oktoberfest 2026 launch with over-the-air releases and smoke-gated deploys; 130-test suite covering tenant isolation and RBAC; GDPR export and erasure in production; pgvector AI translation layer; runbooks, incident reports and postmortems.

**vidlizer** (github.com/arizawan/vidlizer): open-source Python CLI and MCP server that turns video, images or PDFs into structured JSON with local or cloud LLMs. Sole maintainer.

## PROFESSIONAL EXPERIENCE

**Full Stack PHP Developer, Iglu**
*08/2024 - Present | Bangkok, Thailand (hybrid)*

- Develop and maintain APIs and microservices, and keep systems stable with backup solutions.
- Create automated software testing procedures and apply DevOps practices to the development workflow.

**Full Stack Engineer and DevOps, Moo Gruppen AS**
*08/2020 - 09/2025 | Oslo, Norway (remote)*

- Automated infrastructure deployments with infrastructure as code and CI/CD pipelines.
- Developed integration microservices that sync millions of records between systems; integrated OpenAI-based AI functions and XRPL blockchain into them.
- Built a production RAG sales assistant for the WooCommerce store: sub-500 ms responses, five switchable LLM providers.
- Ran code quality and security analysis with SonarQube as part of technical due diligence.
- Monitored services and servers; optimised web shop speed with caching and network optimisations.

**Full Stack Engineer (part-time), Nomadory**
*03/2020 - 09/2020 | Remote*

- Managed AWS servers, built the backend admin, deployed to staging and live, designed scalable APIs and reviewed code for security.

**Site Reliability Engineer, Qwil**
*10/2019 - 04/2020 | Remote*

- Built Terraform for infrastructure management and ran a Kubernetes cluster on Google Kubernetes Engine.
- Automated deployments with GitLab CI; weekly deployments to staging and dev; managed cloud network security, security patching and on-call duty.

**Senior Full Stack Developer, Patch Homes**
*05/2016 - 01/2020*

- Built CI pipelines in GitLab CI and managed a Docker cluster on AWS; designed APIs, wrote unit tests for testing and automation, and reviewed code.

**Senior DevOps and Product Designer, HungryNaki**
*09/2013 - 12/2018*

- Maintained and scaled all servers and AWS infrastructure for a food delivery platform; maintained deployments; built a custom e-commerce API for the local market.

**Earlier roles**

- Full Stack Developer (part-time), Kyra and Vir (01/2018 - 04/2019); Head of IT, Rebate Networks / OScom (01/2011 - 09/2012)

## EDUCATION

**Diploma in Engineering, Computer Engineering**, SAIC Institute of Management and Technology, Bangladesh | Graduated 2011