

AHMAD RISHI

Senior AI / LLM Engineer · GenAI & Agentic Systems · AI Team Lead

Generative AI · RAG · Agentic AI (LangGraph, MCP) · LLMOps · MLOps · Python · AWS

Dubai, UAE (open to remote) | ahmadrishi@gmail.com | linkedin.com/in/ahmad-rishi | github.com/ahmadrishi

PROFESSIONAL SUMMARY

Senior AI/LLM engineer and current AI Team Lead — 4 years across freelance and full-time software engineering, ~2 years building production Generative AI, promoted from Software Engineer to AI Team Lead in ~16 months. Initiated and led the **Agentic Development Board**, an internal AI platform (FastAPI + Next.js) that drives a business request through spec → plan → code → tests → docs as **LangGraph multi-agent workflows** behind human approval gates; personally built its agent runtime — the dual-engine LLM abstraction, subprocess management, Atlassian auth layer and self-healing generation loop. Also initiated **CodeHarbor**, an agentic LLM security scanner, and built production RAG over an enterprise Jira/Confluence corpus that beat the incumbent tool **240–30–13** in a blind pairwise evaluation. Deep in LLMOps and evaluation (self-hosted LiteLLM gateway, Langfuse tracing, LLM-as-judge), NLP/machine learning, and secure AI delivery on AWS.

TECHNICAL SKILLS

Generative AI & LLMs: RAG (Retrieval-Augmented Generation), agentic & multi-agent systems, prompt engineering, structured outputs, tool/function calling, Model Context Protocol (MCP), LLM-as-judge, NLP (Natural Language Processing)

Frameworks & Orchestration: LangChain, LangGraph, LangSmith, n8n, Claude Code SDK, Cursor CLI, Hugging Face Transformers

Models & Providers: Claude (Anthropic API), GPT / OpenAI API, Gemini, DeepSeek, Llama; embeddings (text-embedding-3, ada-002)

RAG & Vector DBs: Pinecone, Milvus, ChromaDB, semantic chunking, hybrid/semantic search, reranking, citations, embeddings

LLMOps & Evaluation: LiteLLM (gateway/routing), Langfuse, MLOps, evaluation harnesses, cost optimization, observability/tracing, prompt logging

Backend & Full-Stack: FastAPI, Flask, REST APIs, async/asyncio, microservices, Kafka, Celery, Redis, SSE/WebSockets, PostgreSQL, MySQL, MongoDB; Next.js, React, Angular, TypeScript, NgRx, Zustand, TanStack Query, Tailwind

Cloud & DevOps: AWS (EC2, Auto Scaling, RDS, ALB, ECR, ACM, IAM, SSM, EKS, Bedrock), Docker, Kubernetes, Terraform, GitLab CI/CD, GitHub Actions

ML / CV & Security: Machine Learning, PyTorch, TensorFlow, OpenCV, EasyOCR, DE-GAN, model deployment; LLM security (prompt-injection defense, OWASP LLM Top 10), secrets scanning, secure code review; Playwright E2E automation

Programming Languages: Python (Expert), SQL, TypeScript/JavaScript, PHP

Spoken Languages: Arabic (native), English (fluent)

PROFESSIONAL EXPERIENCE

Maids.cc — AI Team Lead | Dubai, UAE (Remote) | May 2025 – Present

Lead a 4–6 engineer AI team; own the agentic-platform roadmap and org-wide AI tooling across business units, and mentor engineers and new-grad hires to ship production features.

- **Initiated and led the Agentic Development Board** — an internal AI platform on AWS that drives a business request through **spec → plan → code → tests → documentation** as **LangGraph multi-agent workflows** with human approval gates between phases. Wrote the repo's first commit and remained maintainer across ~8 contributors: **350 of 1,003 human commits**, merging 94 feature branches to master. **~116K LOC** across a FastAPI backend and a Next.js 15 / React 19 / TypeScript frontend.
- Built the platform's **LLM provider abstraction** — one interface over two incompatible coding-agent runtimes (a forked CLI binary and an in-process SDK), swappable by a single Parameter Store value — including the **factory (100% mine)**, the **Claude Code SDK execution path (85%)**, and the **subprocess layer that makes forking a 600-second child safe (684 of 686 lines)**.
- Designed **16 workflow types** as a persisted phase machine over a **46-table MySQL schema**, deliberately without a LangGraph checkpointer so one store answers "where is this task"; wrote the **natural-language → security-validated read-only API** workflow and its **self-healing validate→repair loop** (semantic + syntax back-edges, 96–100% mine), and the Kafka **event publisher and SSE streaming path** that surfaces multi-minute agent runs to the browser.
- Authored the **multi-model agent framework** — **16 model-specialized agent definitions** (discovery, analysis, backend/frontend implementer, code-reviewer, impact-analyzer, test-writer and E2E variants) packaged as drop-in agent kits, **38K lines, sole author** — with every convention derived by measuring the real repositories rather than asserted; plus the platform's **Atlassian OAuth 2.0 + JWT session layer end-to-end (487 of 488 lines)**, the auth foundation every integration on the platform authenticates against.
- Operated the platform's **AWS deployment** — VPC, autoscaling EC2 behind an ALB with ACM TLS, RDS MySQL, self-hosted Kafka (KRaft), ECR and SSM-managed secrets, provisioned with Terraform (44 resources across 26 AWS types) and released via containerised CI/CD.
- **Initiated CodeHarbor** and led its scan engine — an agentic LLM security scanner (FastAPI + LangGraph + Claude Code SDK) that runs a coding agent inside a **sandboxed per-execution clone**, scans GitLab diffs on a schedule against a **9-check security checklist** — public-webhook-to-internal-ERP backdoors, prompt-injection/excessive-agency, hardcoded secrets, injection-prone DB operations — and makes it **prove every finding**, with Pinecone similarity de-duplication and severity-based email + dashboard alerts. Largest human contributor: **207 commits, ~25K insertions**.
- Stood up the team's **LLMOps backbone** — a self-hosted **LiteLLM gateway** fronting every major provider behind one endpoint with per-key budgets and routing, and **Langfuse** (ClickHouse) tracing — then **migrated production traffic off Claude 3.5 Sonnet to a candidate ~12x cheaper per token**, gated on an LLM-as-judge replay harness over real conversations rather than on price alone.
- Standardized org-wide LLM tooling (Cursor, LangGraph, Langfuse) and built the adoption-analytics stack behind it: a custom VS Code/Cursor extension (diff engine) feeding a FastAPI/MySQL service and an **Angular 20 + NgRx** dashboard, plus a **queue-backed prompt-logging service** (Redis Streams, a 300 req/s rate ceiling, bounded queue shedding to 503, XCLAIM dead-worker recovery, dead-letter stream, idempotent batch upserts).
- Built and shipped a **WhatsApp AI agent pilot on n8n** — 22 workflows / 357 nodes with Twilio inbound orchestration, a 5-tool agent, and Redis conversation state (debounce list, durable attribute hash, 10-turn memory) including a **stale-reply guard** that drops an in-flight reply when a newer message arrives; **load-tested it to destruction** (30,300 requests ramped 26→940/s) and identified a hard ceiling of **~65 sustained req/s** with 70.7% timeouts beyond it — successful throughput stayed flat regardless of arrival rate, proving a saturated worker pool rather than a network limit. Also authored the fleet-sync and backup tooling covering **9 n8n instances / 5,558 workflows**.

Maids.cc — AI Engineer | Dubai, UAE (Remote) | Sep 2024 – May 2025

- Built a **production Jira/Confluence knowledge-management RAG** (LangChain + LangGraph over **Pinecone**, dual knowledge bases; migrated from ChromaDB) indexing **30–60K Jira issues across an enumeration reaching 40 projects**, answering with numbered source citations generated by *construction* so a citation cannot reference an unretrieved document. **Blind pairwise evaluation on 283 real questions: preferred over the incumbent tool 240–30–13 by an LLM judge and 246–34–3 by a human domain expert, the two agreeing 91.5% of the time**. Async Flask + Celery, CI-deployed, with Redis-cached rolling conversation summarisation, a DB-backed feedback loop and latency analytics.

- Shipped a **second production RAG** over the company's OpenAPI/Swagger specs using a **custom OpenAPI-aware semantic chunker** (endpoint/schema/tag chunks with \$ref resolution) and a LangGraph ReAct retrieval agent with text-embedding-3-large.
- Designed a **multi-provider, multi-judge LLM evaluation framework** integrating OpenAI, Anthropic, Gemini and DeepSeek — 7 evaluation arms replaying real production conversations against the Claude incumbent, with deliberate reciprocal cross-judging and a self-preference probe — and used it to clear the migration to a candidate ~12× cheaper per token. Key finding: **swapping the judge moved the pass rate 37.4 points on byte-identical outputs** (82.7% → 45.3%) against a sampling error of ±9–11, and **37.6% of judge verdicts failed to parse** — so judge choice and output reliability, not the candidate model, dominated the result.
- Built an automated **code-security workflow** (GitLab webhooks + LLMs) reviewing every commit diff across 15 repositories, emitting structured severity-ranked findings to email + an audit log for remediation.
- Delivered a **passport CV/OCR pipeline** — OpenCV (CLAHE contrast, denoise, perspective de-skew), EasyOCR, and PassportEye MRZ check-digit validation, benchmarked against DE-GAN/GFPGAN/NAFNet deep-enhancers — with an automated MRZ-confidence + image-clarity + alignment QA gate that cut the manual photo-review step for flagged passports.
- Built an **event-driven OCR microservice** (FastAPI + Kafka, Regula DocReader, Gemini image enhancement) with idempotent processing, dead-letter recovery, and Prometheus metrics; plus a fastText **language-detection API** covering 200+ languages and an 85-language LLM quality-evaluation study.

Maids.cc — Software Engineer | Dubai, UAE (Remote) | Jan 2024 – Sep 2024

- Built the Maids.cc **Ziwo telephony integration suite**: a production CTI **screen-pop** linking inbound/outbound calls to agent ERP profiles via a deterministic E.164 number-normalisation funnel, and a live **missed-calls analytics dashboard** over the Ziwo CDR API with adaptive time-window bisection around the 5,000-record export cap.
- Engineered a **real-time waiting-agents queue counter** (FastAPI, HTTP/2, uvloop) on the IVR's critical path feeding Ziwo smart-routing, and reverse-engineered Ziwo's storage API to snapshot **417 production call-flow/IVR configurations** that had no export path.
- **Owned and gated the ERP's Playwright + TypeScript E2E test framework** as reviewer-integrator — merged every change into main across the library and its consumer template (16 Angular Material component wrappers handling CDK overlays and virtual scroll, 53 helper utilities, a parameterized-SQL database-verification client, 232 tests), drove it to a scoped release on a self-hosted private npm registry, and caught defects in review including a credential-leak gap in the ignore rules.
- Containerised and self-hosted two internal operations dashboards — moving one off managed serverless onto Docker Compose with a deployment runbook, hardening the image build so build-time secrets never reach an image layer, and acting as merge gate on the trust-boundary rewrite that moved every privileged database call server-side.

ADDITIONAL EXPERIENCE

Anjz AI — Co-Founder & CTO (concurrent early-stage venture) | Dubai, UAE | Sep 2024 – Apr 2025

- Co-founded an early-stage startup and led all technology; built an **AI-powered legal document-management platform** (AWS + LangChain) that automated review and cut document-processing time significantly.
- Delivered a multilingual **OCR + LLM extraction pipeline** for legal documents.

Freelance Developer — Web Automation & Full-Stack | Remote | Jul 2022 – Mar 2024

- Built Python web scrapers (Scrapy) and browser-automation workflows (Playwright) for SMM platform clients, and delivered PHP/Laravel client websites with admin panels, REST APIs, and third-party integrations (payments, social, analytics).

SELECTED PROJECTS

- **Agentic Development Board** — Internal AI development-lifecycle platform I initiated. FastAPI · Next.js 15 · LangGraph · Kafka · MySQL · Terraform/AWS. 16 workflow types, 46-table schema, human approval gates; ~116K LOC.

- **CodeHarbor** — Agentic LLM security scanner I initiated. FastAPI · LangGraph · Claude Code SDK · Pinecone. Sandboxed agentic diff-scanning of GitLab repos against a 9-check checklist with evidence-gated findings.
- **protoframe** — AI UI-design platform I initiated and led. FastAPI · Celery · Angular 21 · LangGraph · PostgreSQL · Redis · MinIO. Generates production React/Tailwind designs; includes a deny-by-default CSS allow-list and an offline A/B eval harness for design quality.

EDUCATION

Bachelor of Science in Information Technology — Syrian Private University | 2019 – Jan 2024

CERTIFICATIONS & COURSES

- Associate AI Engineer for Developers — DataCamp | Jan 2025
- Developing AI Applications — DataCamp | Nov 2024