

Ajmal Ali

Senior Software Engineer, Applied AI and LLM Systems

Carol Stream, IL | ajmalsyedali@gmail.com | github.com/ajmalali
linkedin.com/in/ajmalsyedali

Summary

Senior software engineer with 6 years of production experience, now building Generative AI systems as an AI engineer at a technology consultancy. Builds agentic workflows on Google Gemini and Claude, semantic search with embeddings and hybrid retrieval, LLM document extraction, and evals that track precision, recall and cost. Background in machine learning clustering, data engineering and SQL. Works directly with clients and owns delivery from requirements to production code, infrastructure and tests.

Technical Skills

GenAI and agents: Google Gemini via Vertex AI, Anthropic Claude, OpenAI GPT, LiteLLM, prompt engineering, structured outputs with Pydantic, tool calling, agentic workflows, LangChain

Retrieval and search: RAG, semantic search, hybrid retrieval, embeddings with bge and nomic, vector similarity search, reciprocal rank fusion, LLM reranking, SQLite FTS5

Evals and AI observability: golden-set regression, LLM-as-judge, precision and recall, SWE-bench, token and cost tracking per session

Machine learning and data engineering: UMAP, HDBSCAN, Pandas, NumPy, ETL pipelines, SQL, PostgreSQL, InfluxDB time-series, MariaDB

Languages: Python, TypeScript, SQL, Bash shell scripting, Go, C/C++, Java

Cloud and delivery: Google Cloud with Vertex AI and Workload Identity Federation, AWS IoT Core, Lambda, CDK and RDS, Docker, GitHub Actions CI/CD

Backend and integrations: NestJS, Flask, Django, RabbitMQ, Redis, MongoDB, Stripe, Twilio, Slack Bolt, n8n, Zapier, webhooks

Experience

AKstudios | AI Engineer, technology consulting | Carol Stream, IL

Sep 2025 – Present

- Eliminated manual report and timesheet preparation by building an LLM document extraction pipeline that pulls structured fields from unstructured input, with deterministic code owning totals and dates so the model never does arithmetic.
- Converted a Fortune 100 manufacturer's requirements documents into 6 implementation-ready tickets for its fleet agreement module by building a pipeline that pulls requirements from multiple source documents and reconciles conflicts between them.
- Took Zaps, an EV charging station platform, from zero to production as the sole engineer, backed by 165 test files and 8 CI/CD pipelines, by building a NestJS and PostgreSQL API with RabbitMQ events, a React Native kiosk with Stripe Terminal payments, and a weekly job that forecasts disk and database growth.
- Enabled pilots at 3 institutions for Radiator Revivers, a smart radiator controller startup, by building multi-tenant Python gateway software on AWS IoT Core, Lambda and InfluxDB with per-tenant certificates, over-the-air updates and live-AWS integration tests.
- Traced 85% of missing sensor telemetry to gateway power loss across 1,398 gateway-days from 9 pilot gateways by writing Python analysis of sequence-number gaps and signal strength over 6 months of InfluxDB time-series data.

Penny Software | Senior Software Engineer, founding team | Riyadh, KSA

Nov 2020 – Aug 2025

- Shipped 500+ features over 4.5 years as a founding engineer on an enterprise procurement platform, including the marketplace used across 30+ enterprise deployments.
- Built an unsupervised machine learning clustering engine, UMAP dimensionality reduction with HDBSCAN, that groups purchase requests for the same vendor into one quotation request, aimed at unlocking bulk pricing and cutting off-contract spend.
- Sped up and stabilized releases by splitting a monolithic repo into Nx libraries with Nx Cloud CI/CD, Docker and GitHub Actions.

King Fahd University of Petroleum & Minerals | Research Assistant | Dhahran, KSA

Sep 2020 – Jan 2021

- Enabled campus COVID-19 outbreak scenario modeling by building an ETL pipeline that merges 3 data sources and feeds configurable parameters into GAMA agent-based simulations.

AmeriCloud Solutions | Software Engineer | Schaumburg, IL

Jun 2020 – Jul 2020

- Cut cell-site database generation for Nokia RF engineers from 15 minutes to seconds by replacing a manual process with a Flask and Pandas service.

AI Projects

- **SWE-bench coding agent, UIUC team project** | Python, Gemini 2.5 Flash, Vertex AI, LiteLLM. On a 40-task SWE-bench Verified subset, resolved 21 of 39 completed tasks, 53.8%, by extending mini-swe-agent with file read and write tools under a \$3 per-task cost cap and classifying failure modes such as hallucinated tool parameters.
- **swatter-bot** | Python, Slack Bolt, SQLite, Pydantic. Slack agent that turns bug reports into GitHub issues. Detects duplicates by retrieving then judging: FTS5 keyword search plus bge-small vector similarity merged by reciprocal rank fusion, then an LLM judge on the top 3. Handles model errors with schema validation and retry, and ships a golden-set eval for dedup precision and recall.
- **flux** | Python. Dropped a 4-session plan, audit and apply agent workflow from the default after measuring it at 2.8x to 2.9x the cost of a single session for identical acceptance scores across 3 benchmark runs, by building a harness that grades Claude Code workflows on held-out tests with the model held fixed.
- **codometer** | TypeScript, npm. Observability tool for coding agents that reads Claude Code, Codex CLI and opencode transcripts to track token cost per session and map which files agents change.

Education

University of Illinois Urbana-Champaign | Master of Computer Science | Expected Dec 2026

King Fahd University of Petroleum & Minerals | BS Software Engineering | 2020 | GPA 3.73/4.00