Uday Kumar Gandhasiri

Senior engineer · 9 years · builds production AI/ML + real-time data systems at scale
Portland, OR · open to relocate or remote
Portfolio: udayx.com · udaygkumar33@gmail.com · github.com/gman-ai · linkedin.com/in/ukg3

SUMMARY

I design, build, and operate AI/ML platforms in production. I ship LLM systems, edge AI, and anomaly-detection models against real business problems. I built and open-sourced the tooling to monitor them.

HIGHLIGHTS

- Real-time edge AI on computer vision in production (KiwiVision occupancy monitoring, >99% accuracy, live since January 2026); statistical anomaly detection on time-series data (73-feature signal pipeline, gex-advisor). Transferable to industrial IoT, sensor data, route optimization, and predictive maintenance use cases.
- 6 production AI systems on Azure OpenAI frontier models (transformer-based LLMs): custom RAG with chain-of-thought prompting + function-calling across 2.6M+ vectorized records and 30+ ETL pipelines; transformer-based embeddings for semantic search; model evaluation against on-premise LLM baseline (Apple Silicon, local fine-tuning eval pathway); automated validation pipelines.
- Built GMAN: AI/ML observability platform with 9 npm SDK releases of `@gman-ai/dev`; intercepts every LLM call, retry, and dollar across consuming agents with no code changes in the agent; multi-tenant gateway with live public demo at getmyagentnow.com/demo. Solves the operational challenge of monitoring Generative AI in production. Open source on npm.
- Operate two production AI agents observed through GMAN: real-time 0DTE options trading agent (200 GB Hetzner VM, 429 req/30s ingest, 73-feature signal pipeline doing statistical anomaly detection on time-series, p99 2 ms write latency, 20+ safety gates) and autonomous content agent (GPT-4o commit scoring, Ed25519-verified Discord HITL approval, ~$0.29/30d LLM cost). End-to-end ownership: problem definition through design, implementation, deployment, and production monitoring.

EXPERIENCE

Multnomah Athletic Club · Portland, OR · 2018 – Present

Senior Engineer, Data + AI Platform · 2024 – Present

- Shipped 6 production AI systems across 40+ FastAPI endpoints: real-time edge AI on computer vision (KiwiVision, >99% accuracy, live since Jan 2026); natural-language analytics platform on Azure OpenAI frontier models with RAG retrieval across 2.6M+ vectorized records; function-calling conversational agent; vector semantic search; automated validation pipelines
- Built on-premise LLM capability on Apple Silicon: local inference + RAG over customer data on a privacy-preserving architecture (no PII leaves the building); evaluated against Azure OpenAI baselines using structured evaluation methodologies
- End-to-end delivery of the AI/ML platform: problem definition with business stakeholders, architecture, modular packaging, deployment via CI/CD, and production monitoring
- Drove architectural decisions and long-term technical strategy for the AI/ML platform; cross-functional partnership with executive leadership, product owners, and program directors to translate business requirements into scalable AI systems
- Project lead across AI and data platform initiatives for 5+ years: offshore engineering vendors, QA testers, and business analysts working under my direction on shipped projects; currently leading testing and validation staff on the Member 360 platform and automation validation programs

- Trusted technical advisor to Executive Leadership and Board of Directors; presented AI strategy and platform roadmap (April 2026); authored the organization's AI policy and governance framework
- Mentor engineers and consultants on AI engineering practices, code reviews, and AI-assisted tooling adoption; custodian of $80K annual AI POC budget; manage the GitHub organization, run quarterly security audits, lead vendor selection

Stack: Python, FastAPI, REST APIs, Azure OpenAI, pgvector, Postgres, Redis, Docker, Azure, Terraform, Next.js, React, TypeScript, GitHub Actions, OAuth 2.0, CI/CD

Applications Engineer | BI & Automations · Jul 2020 – Dec 2023

- Architected and operate the SQL Server data warehouse: 30+ automated ETL pipelines integrating 10+ operational source systems across the organization; the data fabric every downstream application and AI system reads from
- Designed query-optimization strategy across the warehouse: EXPLAIN-driven indexing, partitioned indexes on hot dimension tables, query cost analysis
- Built full-stack data applications (FastAPI + Next.js) for program directors across the organization; data visualization via Power BI and Tableau

Stack: Python, FastAPI, REST APIs, SQL Server, SSIS, Databricks, Postgres, Docker, Azure, Next.js, TypeScript, Power BI, Tableau, GitHub Actions

Data Analyst · Jun 2018 – Jul 2020

- Built the analytics foundations that later scaled into the warehouse: source-of-truth dashboards for program directors, integration patterns, automated workflows that replaced manual reporting across the organization
- Established the first data integrations with the organization's operational systems — the integration approach that later scaled to 10+ source systems

CVS Health · Remote · Jul 2021 – May 2022

Data Developer III *(Contract)*

FinOps analytics reporting to the FinOps Director.

- Built FinOps platform consolidating Azure, AWS, and GCP usage data for enterprise cost visibility
- Deployed containerized data services on Kubernetes with Docker; reporting infrastructure on Azure Databricks
- Power BI dashboards for cost attribution across business units

Robert Half · Portland, OR · 2018

Report Developer *(Contract)* · Jan 2018 – Jun 2018

- Built BI reporting and data-extraction infrastructure for a Portland enterprise client: cross-system reports across operational data sources, dashboard development, automated reporting workflows.

PERSONAL PROJECTS (PRODUCTION SYSTEMS)

GMAN: AI Agent Observability Platform · getmyagentnow.com

Multi-tenant AI/ML platform with published open-source npm SDK (`@gman-ai/dev`, 9 releases). The SDK runs as a transparent local proxy: set `OPENAI_BASE_URL` to `localhost:9000` and every LLM call gets captured (prompts, tokens, latency, cost) with no code changes in the consuming agent. Gateway forwards to OpenAI, ships telemetry to Supabase, cockpit renders sessions for audit. Built to solve the operational challenge of delivering and monitoring LLM-based cloud systems in production.

- Server-derived `tenant_id` via Postgres RPC + RLS; wire envelope has no tenant field, never trusted from client
- RLS-enforced row-level tenant isolation across dev / public / ledger schemas; cross-tenant data leakage blocked at the database layer

- Rate-limited at 100 req/min per IP at the gateway; service-level objectives tracked on ingest p50/p99 and gateway uptime
- Public sanitized demo at getmyagentnow.com/demo with live agent traces, per-call cost breakdowns, and inspectable prompts/responses

Stack: Next.js, React, TypeScript, Node.js, Express, REST APIs, Postgres (Supabase), Clerk (JWT / OAuth 2.0), Stripe Connect (OAuth), Webhooks, Docker, Railway, Vercel, GitHub Actions, CI/CD, npm publishing

Trading Systems: 0DTE options + gamma advisory

Three coordinated real-time projects on a single 200 GB Hetzner Linux VM, paper trading on IBKR. Cross-process coupling via shared read-only data; no message bus, no socket, no HTTP server in the trading lane.

paisamaker. Dual-lane execution engine. Two IBKR clientIds in separate processes (executor=1, chain/quote=2) so a stuck market-data subscription can't poison order flow. Mode state machine (DATA_ONLY -> RECONNECT_WARMUP -> FULLY_LIVE) blocks new entries during recovery. 20+ pre- and post-trade safety gates.

Stack: Python, asyncio, ib_insync, pytest, structlog, Docker, Linux, systemd, Prometheus-style metrics.

gex-advisor. Multi-ticker real-time signal evaluator. 73-feature pipeline per tick on a 30s loop. Four alert types (CASCADE_WATCH, CHARM_SQUEEZE, GAMMA_RECLAIM, GAMMA_RECLAIM_EXIT) with statistical validation — anomaly detection and statistical modeling on time-series data. New signals ship as shadow alerts first, then promoted to production.

Stack: Python, pandas, NumPy, scikit-learn, pytest, Discord webhooks, REST APIs, Linux, systemd.

market-data-fetcher. Real-time market-data ingest. 429 concurrent API requests every 30 seconds across 39 tickers, ~487k rows per RTH day. WebSocket-only in production since March 2026; REST scaffold retained for instant rollback. p99 write latency 2 ms. Stdlib only, no frameworks; deliberate latency-audit choice.

Stack: Python 3.11, asyncio, aiohttp, WebSocket, REST, protobuf, zstandard, Docker, Linux, systemd.

social-agent: Autonomous Content Pipeline · @gman_ai

TypeScript agent on a GitHub Actions cron (3×/day at 08:00, 14:00, 20:00 UTC). Reads commits across 8 of my repos, scores relevance via GPT-4o (2-stage gated at score 5/10 or higher), drafts a tweet, sends to Discord with Approve/Reject buttons. Every LLM call routes through the GMAN proxy with `X-GMAN-Step` headers; every step traced. The cron has no Twitter credentials; only a signed Discord interaction (Ed25519 verified) can authorize a post. Human-in-loop is enforced at the network boundary, not by convention.

- 30-day window: 84 cron runs, 81 successful, 24 drafts sent to Discord, 20 tweets posted to @gman_ai
- LLM cost: ~$0.29 / 30 days for the entire pipeline; ~$0.012 per draft amortized

Stack: TypeScript, Node.js, OpenAI API, Twitter API v2 (OAuth 1.0a), Discord interactions (Ed25519), Webhooks, REST APIs, GitHub Actions (CI/CD), Vercel

---

SKILLS

Languages: Python, TypeScript, JavaScript, SQL, Bash

AI / ML / LLM: AI Engineering, AI/ML Platform Engineering, Agentic AI, Large Language Models, frontier models, transformers, attention mechanisms, OpenAI (GPT-5, GPT-4o, o3, o1), Azure OpenAI, Anthropic Claude (Opus 4, Sonnet 4, Haiku 4), Google Gemini (2.5, 2.0 Pro, Flash), Meta Llama (4, 3.3), Microsoft Phi-4, Mistral, DeepSeek V3, LangChain, LangGraph, LlamaIndex, Claude Agent SDK, OpenAI Agents SDK, Model Context Protocol (MCP), PyTorch, TensorFlow, Hugging Face, scikit-learn, vLLM, Ollama, llama.cpp, RAG, chain-of-thought prompting, prompt engineering, function-calling, agent orchestration, guardrails, model evaluation, fine-tuning, GPU optimization, MLOps, LLMOps, embeddings, semantic search, vector databases, anomaly detection, statistical modeling, time-series analysis, computer vision (edge deployment), pgvector, ChromaDB

Data: Postgres, Snowflake, Databricks, Spark, PySpark, dbt, Airflow, Redis, JupyterLab, real-time streaming, query optimization (EXPLAIN, indexing strategy), data pipelines for GenAI (cleaning, chunking, embeddings), data visualization (Power BI, Tableau, Plotly, matplotlib, seaborn)

Web / App: Next.js, React, Tailwind, FastAPI, Node, Express, REST APIs, WebSocket

Infra / Cloud: AWS, Azure, GCP, Linux, Docker, Kubernetes, Terraform, GitHub Actions, CI/CD, Cloudflare, Vercel

Observability: Datadog, Grafana, Prometheus, OpenTelemetry, Langfuse, LangSmith, Helicone, Phoenix (Arize), GMAN (own product), SLO/SLI design, production monitoring

Methods: Agile, Scrum, Kanban

EDUCATION

Master of Science, Information Technology · Valparaiso University · 2016 – 2017 (Honor Society)