

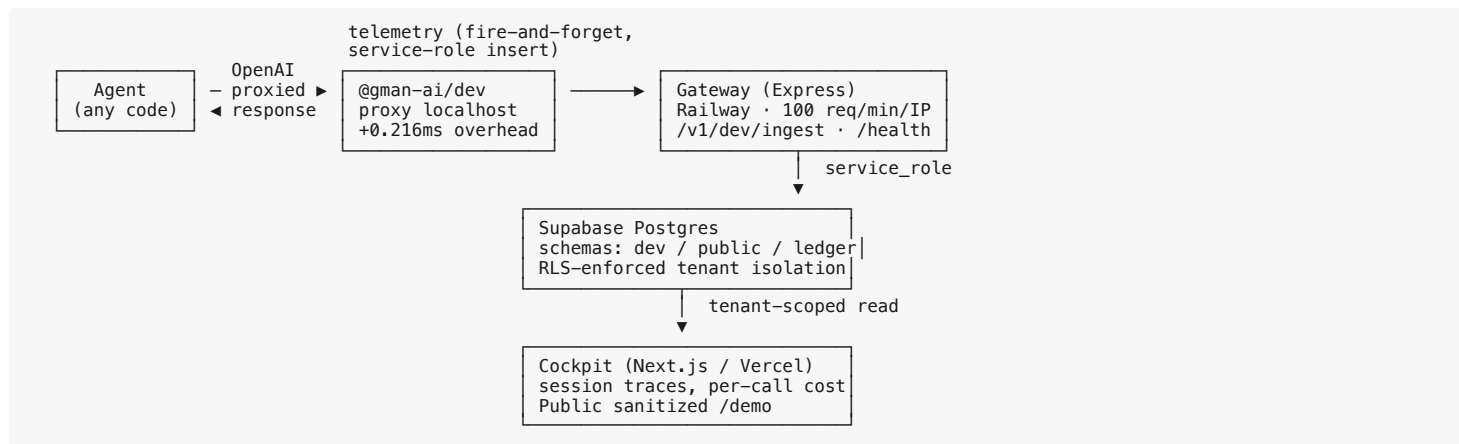
Uday Kumar Gandhasiri: Project Portfolio

Senior engineer · 9 years · production AI/ML + real-time data systems

Portfolio: udayx.com · GitHub: github.com/gman-ai · Live demo: getmyagentnow.com/demo

GMAN: AI AGENT OBSERVABILITY PLATFORM

Published npm SDK (@gman-ai/dev, 9 releases). Multi-tenant gateway with live cockpit and public sanitized demo. The SDK runs as a transparent local proxy: set OPENAI_BASE_URL to localhost:9000 and every LLM call gets intercepted with no code changes in the consuming agent.



- Server-derived tenant_id via Postgres RPC + RLS. Wire envelope has no tenant field, never trusted from client.
- Cross-tenant data leakage blocked at database layer (RLS-enforced across dev / public / ledger schemas).
- Measured local proxy overhead: **+0.216 ms** (async telemetry, non-blocking).
- Rate-limited 100 req/min per IP at gateway. SLOs tracked on ingest p50/p99 and gateway uptime.
- Public sanitized demo: live agent traces, per-call cost breakdown, inspectable prompts and responses at getmyagentnow.com/demo.

Stack: Next.js, React, TypeScript, Node.js, Express, Postgres (Supabase), Docker, Railway, Vercel, GitHub Actions, npm publishing.

TRIAGE-AGENT: AGENT HARNESS WITH VALIDATION FRAMEWORK

End-to-end agent harness built in three weeks: planning, tier-ordered retrieval, validation, replay. Working repo at github.com/gman-ai/triage-agent, 179 tests passing.

- Multi-step reasoning agent (model-agnostic; Sonnet 4.6 reference client) with bounded tool use, retry-with-feedback, and failsafe termination.
- Plan-gated tier-ordered retrieval decides what enters the context window. Source pointers and a bundle hash flow into an append-only audit ledger for cross-session replay without raw payload retention.
- 3-pass citation-grounded validator: schema, citation existence, citation support. Pass 3 walks the cited field_path in each retrieval and verifies the value matches the model's expected_value, catching hallucinated values, not just hallucinated citations.
- Eval harness with a 30-case gold set and a 12-category adversarial set (prompt injection, stale data, schema drift, cross-tenant isolation). **SUT exact-match 0.933, adversarial pass rate 1.000, ECE 0.085**. Blended cost ~\$0.015 per case.

Stack: Python 3.12, FastAPI, Pydantic, Anthropic SDK, structured retrievals, uv, pytest, GitHub Actions.

TRADING SYSTEMS: THREE COORDINATED REAL-TIME PROJECTS

Single 200 GB Hetzner Linux VM. IBKR paper trading. Cross-process coupling via shared read-only data; no message bus, no socket, no HTTP server in the trading lane.

paisamaker: Dual-lane IBKR execution engine. Two `clientId`s in separate processes so a stuck market-data subscription cannot poison order flow. Mode state machine (`DATA_ONLY` → `RECONNECT_WARMUP` → `FULLY_LIVE`) blocks new entries during recovery. 20+ pre- and post-trade safety gates.

gex-advisor: Multi-ticker real-time signal evaluator. 73-feature pipeline per tick on a 30s loop. Statistical anomaly detection on time-series data. New signals ship as shadow alerts first, then promoted to production.

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. **p99 write latency 2 ms**. Stdlib only, no frameworks; deliberate latency-audit choice.

Stack: Python 3.11, `asyncio`, `ib_insync`, `pandas`, `NumPy`, `scikit-learn`, `aiohttp`, `WebSocket`, `protobuf`, `zstandard`, `Docker`, `Linux`, `systemd`.

PRODUCTION AI AT CURRENT ORGANIZATION (SOLE SENIOR IC)

Six production AI systems serving the organization, all running on a platform I architected.

- **Natural-language analytics platform** on Azure OpenAI GPT-4o: RAG retrieval with chain-of-thought prompting across 2.6M+ vectorized records, function-calling conversational agent, vector semantic search, automated validation pipelines.
- **Real-time edge AI on computer vision** (KiwiVision occupancy monitoring, over 99% accuracy, live since January 2026).
- **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.
- **Data warehouse** (SQL Server): 30+ automated ETL pipelines integrating 10+ operational source systems. The data fabric every downstream application and AI system reads from.
- **AI policy and governance framework** authored. AI strategy and platform roadmap presented to the Board of Directors (April 2026).

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

SOCIAL-AGENT: AUTONOMOUS CONTENT PIPELINE

TypeScript agent on a GitHub Actions cron. GPT-4o relevance scoring, Ed25519-verified Discord human-in-loop enforced at the network boundary (cron holds no posting credentials). Every LLM call traced through the GMAN proxy. 30-day window: 84 runs, 20 posts, ~\$0.29 total.

Live observability at getmyagentnow.com/demo · Open source at github.com/gman-ai · Portfolio at udayx.com