

Eric Allione

Full-Stack AI Agentic Architect

aigamma.com | ai-firehose.com | algonow.net | alohacarbon.org | codexproof.com | learnrust.ai
mathlimit.com | researchshop.org | selectsectors.com | spokenhistory.org | worldthought.com
[Bio page](#) | eric@aigamma.com | TS Clearance (Inactive)

AI architect and derivatives quant. Fractional CIO of a waste-to-energy company, and engineering lead on research systems at UCLA, UCI, and WWU. Peer-reviewed results at AgentNet 2026, a workshop of IEEE ICNP, and at The Digital Conference at King's College London: provenance for agent evidence, and AI-assisted oral history. Builds from raw code rather than inside a vendor's abstractions for all sectors.

Core Competencies

- **AI Architecture:** Agentic harnesses, MCP server design, BLAKE3 provenance layer, LLM orchestration, custom RAG pipelines, vector databases (Qdrant, Pinecone, Supabase/Postgres)
- **Systems Engineering:** Observability (OpenTelemetry, Prometheus, Grafana), latency optimization, evolving complex repositories
- **Backend & Analytics:** Rust, Python, JavaScript; full-stack architecture and risk analysis; revenue leak discovery and remediation

Technical Leadership & Consulting Experience

Full-Stack AI Architect

May 2025 - Present

Operate an independent AI infrastructure and research laboratory to develop production-grade core architectures, distributed telemetry pipelines with centralized analysis, and customized retrieval engines. Consult on infrastructure design, AI strategy, mathematical modeling, and enterprise solutions. First author on five CHI 2027 submissions and contributing co-author on three more, all due September 2026, with 50+ repositories held private pending double-blind review. Every build starts from a harness of finished prior systems rather than an empty repository, so a project begins at a running full stack and the time goes to architecture instead of scaffolding. Extreme development velocity achieved by dictation and proprietary automation harnesses: 2.4 million words per 6 months, mostly dictated directly into PowerShell.

Selected Production Infrastructure & Native Platforms (*all built without templates, site builders, or CMS*)

- **LLM Fleet Governance & Distributed Telemetry (mathlimit.com):** Established a centralized observability stack (OpenTelemetry, Prometheus, Grafana) to profile asynchronous traffic across 18 hand-coded domains (React, Rust, HTML). Leveraged this telemetry to engineer a serverless risk-desk platform that monitors live LLM spend, applies non-anticipating quantile limits, and bounds the dollar exposure of a spend excursion under stated metering and enforcement assumptions: a deployed floor rate of \$5 per five-minute bin, with an external \$200 per-period cap supplying the global dollar ceiling. Built a dynamic, self-releasing throttle ladder (P93, P97.7, P99.5 bands whose clamp depth and hold time trade off inversely) and deployed a Bill-at-Risk forecast, a VaR analogue that prices the probability of breaching the operator's cap, with policy analysis bounding a floor-rate spend flood to between \$12.90 and \$34.50 against those assumptions. Under a common false-alarm ceiling the controller detected 26 of 30 seeded six-hour twofold leaks, where all five baselines detected at most 7 of 30, at a median detection latency of 115 minutes.
- **Cryptographic Agentic Provenance (codexproof.com):** Engineered a full-stack Rust provenance platform, backed by four heterogeneous vector stores under one operator, Pinecone, Supabase pgvector and Qdrant among them, that deploys a keyless coordination gateway with Model Context Protocol (MCP) and A2A bindings to federate interrogations across a fleet of distinct trust domains. Implemented a just-in-time BLAKE3 hashing and Ed25519 signature layer that structurally guarantees data integrity and source when evidence crosses trust boundaries,

isolating cross-domain latency and integrity failures to the responsible store. That architecture is the AgentNet 2026 paper cited below, and its overhead is measured rather than asserted: about 15 microseconds to seal a node, 11 of them the Ed25519 signature, 27 to verify one, and 54 to authorize a certificate chain to its trust anchor, with structural overhead under 800 bytes per node, which leaves the retrieval hot path itself crypto-free. The site runs the verifier in the browser, so a visitor can tamper an evidence node and watch one trust domain fail while the others stay green.

- **Daily Parsing of AI Industry (ai-firehose.com):** A data-mining operation that surfaces the most salient YouTube videos along with data points on AI techniques, concepts, and tools, and shows how each ranks day to day by trend. Curates over 600 well-sourced glossary items with supporting illustrations in a Pinecone vector store. This site solves the problem every AI enthusiast faces: the daily overload of new information. The corpus is self-expiring and retains nothing older than about 90 days except for a hand-curated band that lasts 180 days. This way, cost stays flat by construction rather than by pruning. Published to npm as ai-firehose-mcp, which anyone runs with npx against the public read API: four tools, no repository, no key, and no local data.
- **Quant Platform (aigamma.com):** Published an MIT-licensed quantitative volatility-research codebase of 70+ live models across 10 mathematical families, including: rough volatility (rough Bergomi, RFSV, three-estimator Hurst triangulation), stochastic and local volatility (Heston, Bates, Dupire), Lévy jump models (Merton, Kou, variance gamma), SSVI/SVI surface calibration, and regime-switching, all recalibrated hourly on market days. Per-model agentic analysis flags each model's inefficiencies, anomalies, and patterns, funneled into a site-wide aggregate feed and companion browser extensions. Uses Supabase for retrieval.
- **Philosophy Knowledge Base (worldthought.com):** Features 2,200+ unique chatbots, one for every philosopher in the corpus and one for every pairing in the network, each with sharp topical guardrails, powered by a Pinecone corpus of primary sources that were mostly pulled from Project Gutenberg. Teaches the ideas of 150+ philosophers, each with educational AI tutors that yield academic-grade citations. Clicking any edge of the interactive network between two thinkers spawns an agentic summary comparing their philosophies, which opens into a focused chatbot that probes the hidden connections in that pairing. An interactive world map responds to a time-range slider to focus on specific periods.
- **Cross-Browser Core Extensions ([Chrome Store](https://chrome-store)):** Shipped Manifest V3 extension packages to the Chrome, Firefox, and Edge marketplaces using minimal-permission service workers, maximum user privacy, CORS-scoped fetches, and market-hours-gated CDN caching against custom Netlify Functions.
- **Agentic Lifecycle Tracing:** Wrapped non-deterministic AI agent loops and recursive tool-calling workflows in a manual OpenTelemetry instrumentation layer, capturing end-to-end execution spans to isolate and remove recursive loop regressions across the 2,300+ live chatbots running fleet-wide.
- **Rust Instruction Platform (learnrust.ai):** A site that teaches Rust, written in Rust from the server to the vector database, on the thesis that as agentic coding strengthens, a language's syntactic difficulty stops being a tax and becomes a moat: the compiler that is hard to satisfy by hand is a tireless verifier an agent can iterate against. Every fenced example and committed file compiles against the newest stable rustc in CI, weekly and on every push, and deliberately broken walkthroughs must fail with the exact documented error, so the homepage carries that result as a dated, falsifiable claim. 41 concepts across six tracks, corpus projection by power-iteration PCA written from scratch with no numerics dependency, and a hand-rolled MCP server on axum, public at learnrust.ai/mcp, exposing the corpus as five read-only tools.
- **Fleet Observability Spine:** Built a self-hosted OpenTelemetry collector taking metrics, traces, logs, and browser telemetry from every property to one bearer-authed endpoint, redacting prompt text at the collector so no application can leak it by accident, then forwarding a single egress to Grafana Cloud. Instrumented under OpenTelemetry semantic conventions so token and model data stay portable across backends, with cost treated as a fourth golden signal beside rate, errors, and duration, and cardinality discipline keeping prompt text and user identifiers out of metric labels entirely. Verified by sending a synthetic trace carrying a fake secret and confirming usage attributes present and prompt content absent. Runs on a 256 MB instance.
- **Per-Identity Access Gateway:** Deployed an authentication gateway in Rust on the principle that nobody ever sends anybody a credential: invitations travel, passwords are created where they live, and a reset is a fresh invitation rather than an email flow. Argon2id for human passwords and BLAKE3 for machine tokens, HMAC-signed httpOnly SameSite=Strict sessions, per-identity method and path scopes, and a reverse proxy that strips caller credentials and injects the upstream token, with one JSON audit line per request. Access control fails

closed, deliberately opposite to the fail-open posture of the enforcement layer elsewhere in the fleet, because the two protect different invariants. Verified live against 200, 403, and 401.

Client Case Studies (May 2025 - Present)

Spoken History Project, UCLA and Western Washington University (AI Engineering Lead and full-stack technical owner)

- **System Delivery** (spokenhistory.org): Took over a stalled, two-year-old heritage-archive platform with 15 prior contributors and rebuilt it from scratch single-handedly into a research system in coordination with the Smithsonian and Library of Congress. Developed all application code, the Firestore data layer, Pinecone retrieval, and a citation-bearing reference layer covering the people and the primary texts. Designed an interactive educational surface steered by cosine similarity. The corpus runs to roughly 16,000 SRT-anchored vectors across 140 interviews.
- **Ownership and Operations**: Created the repository and authored every commit in it. The system runs on my own domain, serverless Netlify deployment and fly.io workers; carrying the Firebase, Supabase, Pinecone, and Voyage AI accounts personally.
- **Data Cleansing**: Built an automated text-processing system that cleaned the entire historical transcript corpus of systemic typos, dictation artifacts, and semantic collisions in six days, then reconciled it line by line against the Library of Congress API. Found errors on the Library of Congress side and closed those gaps by inference. Added the timestamp resolution and VTT captioning which the national source archive lacks.
- **Publication Gate**: The original loop scored summaries at 80/80 and kept the best of three attempts even when none passed, so an unsupported summary could still ship. Replaced it with a gate that fails closed: both scorers must independently reach 90/90, the scoring is cross-family so one vendor's model grades another vendor's generator against the identical rubric rather than a family grading its own homework, a per-claim citation auditor blocks on an uncertain verdict and not only on a false one, and scorer disagreement routes to human review instead of publishing on a coin flip. Generation was grounded before it started by word-boundary-anchored matching against a schema-validated ground-truth corpus of 378 entries and 396 aliases, and a public machine-audit page documented the pipeline end to end.
- **Custom MCP Server Deployment**: Stood up a Model Context Protocol (MCP) server that makes timestamped VTT caption files the backbone of a near-zero-latency chaptered video player (2.5–4-minute segments with self-descriptive topics), enabling semantic context matching and automated thematic paraphrasing. Includes 12+ RAG features (streaming responses, jitter-free anchoring, explicit latency-vs-depth query routing) that together form an interactive learning surface. It runs publicly at mcp.spokenhistory.org/mcp with no login and no API key, exposing eight tools, three prompts and four resources, rate-limited per IP and scaled to zero on a 256 MB instance, logging request metadata and never querying text.
- **Technical Paper**: Wrote the team's full 21-page technical paper, start to finish. It accompanied Dr. O'Hara's presentation of the work at [The Digital Conference at King's College London](#), June 2026.
- **Method Transfer and Continuing Ownership**: Taught the universities' graduate students the agentic engineering practice behind the project, in weekly sessions across those five weeks. I remain the full-stack technical owner of the production system and continue to run it, and the principal investigator is extending the work into grant-funded projects in cultural heritage and community-based archives.

UCI, Department of Computer Science and Informatics (Engineering Lead, referred via UCLA Principal Investigator)

- **System Build**: Built the CODW workbench from scratch around 500 CODW files that only expensive commercial software could read, writing the format reader instead and converting a principal investigator's dissertation methodology and a decade of anonymized interview data into a React application, and bringing this proprietary methodology into digital life for the first time.
- **Analysis Performance**: Moved the aggregation into Postgres RPCs behind a per-session cache and cut first load to about 0.4 seconds. A second migration then precomputed per-code alpha and kappa, so the Ground Truth instrument reads 237 reliability rows instead of recomputing across all 12,509 per the original corpus structure.

- **Retrieval Evaluation:** Tuned RAG queries to sub-second response latency across dense datasets, on a two-backend adapter running pgvector with full-text search against Qdrant, with Voyage AI embedding and reranking. The choice between backends was settled by a gold-question evaluation harness rather than by preference, and retrieval stays citation-grounded: every answer traces to a transcript, a turn, and a timestamp.
- **Security & Guardrails (gated, private repository at leaderlogic.org):** Implemented data-confidentiality guardrails mirroring HIPAA standards to protect sensitive research interview data without slowing vector retrieval, on invite-only authentication with row-level security in Postgres. Data governance is enforced mechanically rather than by policy: a deny-by-default ignore list and a pre-commit guard stop any transcript on .docx, .pdf, .srt, .vtt, or secret from entering git at all.
- **Gated Agent Surface:** Deployed a bearer-authenticated MCP server at codw-mcp.fly.dev/mcp, exposing the workbench to agent clients under the same access model as the application itself.

International Travel Conglomerate (Revenue Systems Architect)

- **Revenue-Leak Forensics:** Recovered over a quarter of a million in annualized operating margins by exposing a click-hijacking system that was intercepting 99.1% of promised ad delivery from the advertising vendor consuming 77% of the client's budget. A 10-day data audit found it, and a forensic read of the vendor's public 10-K and 10-Q filings against its represented audience figures is what made the discrepancy provable rather than suspected. Eliminated the wasted spend and cut the overall vendor stack by 83%.
- **Infrastructure Cleanup:** Stripped a broken, dated web footprint of 5 redundant Google Tag Manager containers, 13 deprecated tracking scripts, and 48 conflicting third-party integrations. Transformed a C# ecommerce site (that was prone to shopping cart crashes) by directly sending patches and work orders to the offshore development team. Lifted PageSpeed desktop scores from failing to the high 90s. Manually configured ad campaigns to outperform the less precise, default AI-generated offerings.
- **Risk Management Systems:** Solved the currency exposure the company had flagged as unresolved at hire, delivering a mechanism for locking forward bookings eight to ten months out: debit verticals set opposite the yen forward rate for committed Japanese exposure, and a Chinese Treasury ETF position-sized from reported yuan revenue.
- **Global Governance:** Oversaw cross-border compliance, domain health, and IP reputation across offices in LA, NY, London, Shanghai, and Hong Kong, strengthening their application firewalls, and coordinating data integrations with major flag carriers and national tourism authorities.

Tri-City Men's Health and Granite Mountain Health, Prescott, AZ (interim HIPAA Security Officer, then full-time engineer)

- **Compliance Sprint:** Scoped, configured, and deployed a HIPAA-compliant operating environment in six weeks. Documented system boundaries, identified data-handling risks, mapped access control, wrote workstation hardening procedures and written operational guidelines for audit readiness, built security scorecards, and enforced clear separation between clinical and administrative data flows. Designed and delivered a staff security training program for both clinical and administrative teams.
- **Workflow Automation Library:** Delivered a library of 63 discrete operational workflows. Consulted on patient intake, intelligent scheduling, programmatic content generation, and retention tracking.
- **Margin Expansion:** After baseline controls and documentation frameworks were established, moved out of the security role into full-time engineering and expanded operating margins with an AI-driven marketing campaign. The practice signed a lease on an additional location on the strength of that growth.
- **Hybrid Inference Deployment:** Architected a split-inference framework for a practice whose protected data cannot leave the premises. Local LM Studio nodes handle anything carrying patient information; only de-identified secondary processing crosses an API boundary, with the boundary enforced in policy rather than by convention.

Aloha Carbon (Fractional CIO)

- Own the technology stack for a waste-to-energy company, setting AI and infrastructure strategy and advising the executive team on business, revenue, and technology decisions.
- **Team-wide AI coaching (ongoing):** Run AI coaching sessions to enable their team of executives, mechanical engineers, and civil engineers in hands-on AI techniques that enhance operations.

- **Production Website Rebuild:** Reverse-engineered a WordPress site into a secure, serverless stack with Google PageSpeed Insights scores on both desktop and mobile: 100/100 in Performance, Accessibility, Best Practices, and SEO.

Hedge Fund Launch (Agile Income Group LLC, Fractional CTO)

- **Infrastructure Launch:** Ran institutional infrastructure, DevOps pipelines, and quantitative research systems for a hedge-fund pre-launch, working with prime brokers and custodian banks to provision the launch scaffolding and logistics such as overnight distributed cash sweeps. When the seed funding fell through, released a substantially stronger public platform at aigamma.com.
- **Risk Playbook:** Migrated my own parameterized risk-management playbook into the fund operating framework, carrying 200 entries through triage and cutting roughly 20 to a working set of 180. Each entry sets a structure against a stated market condition, which is what turns a model output into a position.
- **Inference Engineering:** Hand-built a multi-GPU tensor-parallel inference pipeline optimized for bounded financial tasks and cutting execution costs while improving mathematical accuracy.
- **Engineering Governance:** Instituted repository controls, CI/CD checks, protected-branch protocols, and pull-request evaluation harnesses to turn loose research into repeatable R&D infrastructure.

Quantitative Finance Experience

Principal Architect (SpotGamma)

Jan 2023 - May 2025

- Designed the modern platform and its models from 2023 onward, directing all major product and quantitative analytical tool releases and modernizing older key-level software into a coherent methodology and an actionable set of models and tools. The resulting derivatives analytics are used daily by retail and institutional traders.
- Produced 585 consecutive market-close financial newsletters under real-time constraints and without oversight, plus a 700-page market-mechanics knowledge base. Owned product marketing, education, and ghostwriting end to end.
- Sat in engineering meetings as the primary Subject Matter Expert on market mechanics, derivatives modeling, product development, and strategic dynamics, carrying responsibility for the mathematical soundness of everything that shipped. Operated as the final escalation point for every technical email, across thousands of chains followed to resolution.
- Served as chief educator to thousands of clients, with over 10,000 direct answers in the customer training Discord plus live chat during releases and major events. Taught applied mathematics with both intuition and precision.

Independent Derivatives Trader (Quantitative Research)

Jan 2018 - Jan 2023

- Ran a self-directed research program spanning pricing theory in discrete and continuous time; stochastic, rough, and local volatility; jump-diffusion; and parametric smile calibration, built from hundreds of graduate texts and research papers.
- Converted the theory into parameterized risk-management playbooks, option-structure designs, and volatility models, then traded derivatives full-time as their empirical proving ground. The research base carried directly into the Principal Architect role at SpotGamma and now runs live as the quant platform at aigamma.com.

Defense & Aerospace Infrastructure Experience

Senior Systems Engineer 5 Defense Contract Management Agency (DCMA West)

Carson, CA · Apr 2017 - Jan 2018

- Headed the virtual estate for DCMA West as Lead Infrastructure Architect, steering its hardening, migration, and optimization across 22 countries.
- Drove virtual desktop infrastructure (VDI) modernizations, migrating thousands of active VMs from legacy clones to Horizon View Instant Clones on budget-constrained vSAN clusters.

- Charted the agency's first end-to-end network topology, gathering thousands of virtual machines across 20 global datacenters into an interactive mapping tool for department heads.
- Provided engineering direction, hands-on troubleshooting, and technical risk assessments for corporate Microsoft and VMware engineering teams during active production rollouts.

Virtual Server Admin / ISSO Headquarters Marine Forces Pacific (MARFORPAC)

HP Enterprise · Camp H.M. Smith; Aiea, HI · Mar 2016 - Apr 2017

- Formally appointed as the Information Systems Security Officer (ISSO) for the virtual server environment across MARFORPAC G-6 Headquarters Information Technology Services (HITS).
- Acted as the primary technical cybersecurity advisor reporting directly to the Branch Head; led STIG remediations and policy-compliance work alongside the base Cybersecurity department and Threat Red Teams.
- Managed international federal auditors through a two-week system review, earning a passing compliance score for the headquarters base.
- Restored VMware infrastructure by identifying and correcting a broken core upgrade sequence, then drafting the master SharePoint engineering documentation to redefine future enterprise upgrade matrices.
- Handled storage scaling, PKI certificate overhauls, and data backups, performing manual vSphere load-balancing and incident-response protocols during live cyber incidents and outages.

Publications & Conference Presentations

Allione, E., O'Hara, D. (2026). *Provenance-Native Retrieval: A Signed, Content-Addressed Thin Waist for Agent Evidence*. Peer-reviewed acceptance at AgentNet 2026, a workshop of IEEE ICNP; presenting in Tempe, AZ on October 5, 2026. Every retrieved evidence unit is named by the BLAKE3 hash of its canonical bytes, signed with Ed25519 under per-domain trust anchors, and linked to answers by content ID, so lineage and integrity form one self-certifying graph. A single verifier confirms answers drawn from four heterogeneous, mutually distrusting vector stores under one operator, localizes any tampered byte to the responsible node, and attributes slow retrievals to the store at fault, while the retrieval hot path stays crypto-free.

O'Hara, D., Sovelove, J., & Allione, E. (2026). *Time-Anchored Testimony: Architecture, Provenance, and Retrieval for an AI-Assisted Civil Rights Oral-History Archive*. Presented at The Digital Conference 2026, Department of Digital Humanities, King's College London. A UCLA and Western Washington University project, in consultation with the Library of Congress and the Smithsonian Institution. Technical report of record: ssrn.com/abstract=7134518

Education & Credentials

- **M.S. Cybersecurity**, Western Governors University, 2019. *Capstone Certificate of Excellence on "Data Center Prototyping."*
- **B.A. Philosophy (with Certificate in German)**, University of Hawai'i at Mānoa, 2012. *Golden Key International Honour Society.*

Certifications & Awards

- **Active Technical Certification:** VMware Certified Professional – Data Center Virtualization (VCP-DCV)
- **Historical Technical Certifications (Inactive):** CNDA (Certified Network Defense Architect), CHFI (Computer Hacking Forensic Investigator), CompTIA CASP+, CompTIA Linux+
- **Honors:** Recipient of the WGU Excellence Award in Cyberlaw and Governance
- **Industry Awards:** Selected as a global VMware vExpert for three consecutive years: 2016, 2017, and 2018