

Jhorlin De Armas

Principal AI Engineer, Agent Infrastructure

Orlando, FL · Remote · jhorlin@gmail.com · jhorlin@skillfaber.com · jhorlin.com · skillfaber.com · linkedin.com/in/jhorlin · [AWS ML Blog \(PDIQ\)](#)

HIGHLIGHTS

- I build the harness around LLMs: the agentic runtime that turns a model into a working system (tool-use loops, streaming, context management, capability detection, metering, and evals), on AWS in TypeScript and Go.
- Co-author of the AWS Machine Learning Blog post on PDIQ, PDI's enterprise RAG system, where answer approval measured 60% to 79%.
- Eval-driven: benchmarked a hybrid retriever at 86% qrels-hit@k on SciFact with regression floors in CI; every claim here is measured, not asserted.
- 20 years shipping software; the last decade architecting AI and serverless systems on AWS, and mentoring teams to build with AI-native tooling.

EXPERIENCE

AI Platform Architect · PDI Technologies

Mar 2024 – Present

- Architected PDIQ, PDI's multi-tenant AI assistant platform on AWS: composable assistants built from knowledge bases, models, guardrails, and agent tools, streaming chat via CloudFront and Lambda Function URLs. Co-authored the AWS Machine Learning Blog post on its RAG architecture.
- Designed the 'one pipeline, many crawlers' ingestion architecture for a company grown through 35+ acquisitions: containerized crawlers (Confluence, SharePoint, Jira, ServiceNow, Azure DevOps) go to the data instead of centralizing it, and summary-prepended embeddings raised answer approval from 60% to 79%.
- Built PDIQ's agent-tool framework, independently converging on the pattern MCP would standardize and shipping it within days of the spec's announcement: each tool a versioned Lambda with a describe/execute contract, a DynamoDB registry, per-tool credentials, and cross-account support so customers expose their own Lambdas as agent tools.
- Mounted PDIQ onto Salesforce case pages for the support org via a thin adapter and an RS256 JWT SSO into Cognito with JIT provisioning, selected over a third-party vendor and Salesforce's native AI. With past resolutions ingested as knowledge, average case resolution fell from four days to one.
- Hired to explore what AI could do for PDI; as the proof of concept showed viability it became PDIQ, the support org's daily tool, embedded in Salesforce as a Lightning component. To sustain what it had grown into, stood up and trained an offshore team of 8 new-graduate engineers: fundamentals first, then AI-augmented development with Claude Code.

Founder & Principal Engineer · [Skillfaber \(Independent\)](#)

Jan 2026 – Present

- Founded and solo-built Skillfaber, a production multi-tenant AI agent platform on Amazon Bedrock (~270 Lambdas, fully serverless): compose a role from a model, guardrails, knowledge, and tools, then reach it from web, Slack, Teams, email, a public API, or agent-to-agent delegation.
- Built the agent runtime: a checkpoint-and-continue loop that chains 15-minute Lambdas into hours-long turns (~50 generations), with loop state in S3, primary-to-backup model failover, and context management that compacts a conversation past the model's context window without losing its session state, its id, or its in-flight tool calls.
- Solved multi-model support with feature detection and polyfill: a probe harness tests every model across Bedrock, OpenRouter, and Meta where provider capability APIs fall short. A model lacking vision is told so in-prompt and handed reader tools; where emulation would fabricate, the harness fails loud rather than guessing.
- Created 'recipes': TypeScript tools authored in-product and run in a QuickJS sandbox with brokered, allow-listed access to tools, MCPs, and KMS-held secrets. Codifying one payroll customer's repeated agent work cut a 25-million-token, 45-minute run to ~100k tokens and two minutes.
- Made AI spend a control surface: per-tenant budget ceilings gate turn admission before the first model call (warn at 80%, deny at 100%), every turn lands with eleven attribution dimensions down to org, team, user, and role, Athena partition-projection reporting turns them into usage anyone can query, and billable tokens land on a Stripe usage meter.

- Integrated 128 vendor services (Salesforce, Workday, NetSuite, SAP) as MCP tools behind a Bedrock AgentCore Gateway with KMS-encrypted credentials resolved user to role to team, and shipped six live channels (email, Slack, Discord, Teams, Messenger, Instagram) over one FIFO ingress keyed by conversation.

Co-founder & Director of Software Development · Kazzcade

Jan 2017 – Feb 2024

- Built sfSync in-house after Salesforce quoted ~\$250k a year for a five-minute-at-best sync: one extracted metadata file drives code generation for the whole pipeline (Liquibase, Go ORM, GraphQL, 107 Apex triggers), and changes stream out of Salesforce as they happen, nearly free to run.

Senior Software Engineer · Under Armour

Jan 2016 – Nov 2016

- Full-stack engineer on Endless Aisle, which let any Under Armour store find a product across every inventory source and ship it to the customer free and expedited. Rolled out to every US retail store.

Sr. Architect · Riptide Software

Mar 2013 – Jan 2016

- Architected Waypoints, an in-application training overlay: React components injected onto the client's own product that explain each on-screen element, capture interactions as xAPI statements, and test mastery inline. Later sold to General Dynamics.

Sr. Software Engineer · NCR Corporation

Sep 2010 – Mar 2013

- Built Delta's airline check-in for web and cross-platform mobile (Knockout MVVM, PhoneGap, Spring MVC), deployed across Delta's airport network; a client-side state machine kept the check-in business rules in one place.

Software Developer · Toptech Systems

Sep 2007 – Sep 2010

- Modernized real-time control software for fuel-terminal automation, a safety-critical domain: ported QNX C systems to Linux C++ and built a client/server bill-of-lading reconciliation tool.

Software Developer · Highwinds Software

Sep 2006 – Sep 2007

- Optimized the CDN cache index, a 130 GB on-disk data structure in multithreaded C++: the tuning doubled storage capacity and cut drive-lookup time by three seconds without slowing responses.

PROJECTS

Notitia — serverless RAG — A serverless-first, multi-tenant RAG service in Go for clients with large corpora. Designed as a drop-in knowledge layer for Skillfaber. Retrieval quality is measured, not assumed.

SKILLS

Agent Infrastructure: Agent runtimes / tool-use loops, Agentic systems & AI agents, Large Language Models (LLM) & generative AI, Multi-agent orchestration & sub-agents, MCP / tool calling, Prompt & context engineering, Model capability detection & polyfill, Guardrails & responsible AI, Evals & benchmarking (LLM-as-judge), Durable execution & agent state machines, Context-window management, Inference cost & latency optimization, Usage metering, per-tenant cost attribution & enforced budget ceilings

Models & Providers: Amazon Bedrock, OpenRouter, Anthropic Claude, Amazon Nova, Meta Llama, Cohere (embed + rerank), Multi-model routing & failover, Prompt caching

Data & Retrieval: Retrieval-Augmented Generation (RAG), Hybrid retrieval & reranking, Chunking & semantic search, Conversational analytics / text-to-SQL (Trino, Superset), pgvector, OpenSearch, Aurora PostgreSQL, DynamoDB

Identity & Security: Enterprise SSO (Okta OIDC, SAML, OAuth 2.0), RFC 8693 token exchange / OBO, JWT (RS256) / JWT-bearer, Multi-tenant isolation & RBAC, KMS encryption & audit logging

Cloud, Infra & Delivery: Infrastructure as Code (SST/Pulumi, CDK, CloudFormation), Serverless (Lambda, Step Functions), Event-driven (EventBridge + Pipes, SQS/SNS, Kafka), Microservices & APIs (AppSync, gRPC, REST), IoT Core (MQTT), Containers (Docker, Kubernetes, Fargate/ECS), CI/CD (CodePipeline, GitHub Actions), Observability (X-Ray, CloudWatch, RUM, Datadog), Cost optimization / FinOps

Languages: TypeScript, Go, JavaScript / Node.js, C#, C / C++, Java, SQL, Apex

EDUCATION

BS in Computer Science, University of Central Florida (2006)