

/ SUMMARY

Engineer building production agentic systems — tool-calling loops over typed tools, MCP (Model Context Protocol) servers, RAG architectures, and evaluation harnesses in Python/FastAPI and TypeScript. Shipped production-quality AI prototypes and integrations solo; made four customers' on-premise products run as SaaS on OCI, architecture through production cutover. Build before documenting, think in reusable patterns, and leave behind reference architectures and working code rather than one-offs. A decade at Oracle in AI infrastructure and AI cloud, working directly with founders, CTOs, and ISV partner engineers.

8	6	1st	2×
PRODUCTION SaaS SHIPPED SOLO	SELF-AUTHORED REPOS	PLACE, AUGHACKS AI HACKATHON AT MIT	OCI CLOUD ARCHITECT

/ AI ENGINEERING — FOCUS AREA: AGENTIC

- **Agentic systems:** a tool-calling loop over 15+ typed tools with parallel specialist subagents, agentic pipelines with subagent fan-out, and agent-graph and agent-loop frameworks handling graph construction, state, memory, scheduling, and failure recovery.
- **Tool calling, MCP and developer platforms:** define the technical best practices, tool contracts, and agent-operability standards other engineers' agents build against — authoring MCP (Model Context Protocol) servers exposing typed tool access to internal, enterprise, and third-party APIs over stdio transport, with OAuth2 token refresh.
- **RAG architectures and evaluation:** chunking, embeddings, vector search, retrieval evaluation — with the evaluation harnesses around them: scored regression suites, tracing, and prompt observability, so agent changes are measured, not guessed — behind human-in-the-loop approval gates.
- **LLM APIs, model serving and multi-model orchestration:** build against the OpenAI and Anthropic APIs plus Google, xAI and OpenRouter; route across frontier and self-hosted models by task and cost, with model serving on self-hosted inference runtimes for 24/7 agent workloads at no marginal token cost.
- **Reference architectures and reusable artifacts:** spec-driven development, reference architectures, deployment patterns, technical writing and developer documentation, and public demos — eight shipped products at live URLs.

/ SELECTED TECHNICAL PROJECTS

Building since 1996. Three decades hands-on across systems, full-stack web, distributed inference, and agentic AI — the projects below are the current, documented work.

Venus — autonomous agentic wedding planner | sole author

Vercel AI SDK & AI Gateway · Next.js · TypeScript · Tavily · Upstash KV · svix · Resend

- Architected a tool-calling loop over 15+ typed tools with an instruction-file persona layer and parallel specialist subagents researching four workstreams concurrently — model calls routed through Vercel AI Gateway, state in Upstash KV, imagery via Tavily.
- Closed the loop end to end: autonomous outbound email via Resend, an svix-signature-verified inbound webhook, model-classified reply understanding, bounce and complaint handling, and cron-driven follow-up sweeps — runs unattended for months, escalating to the human only on decisions.

Agentic outreach service | sole author

Python · FastAPI · Uvicorn · WebSockets · Playwright (headless Chromium) · Ollama · Gemini

- Built a closed-loop outreach agent as a FastAPI/Uvicorn service driving headless Chromium via Playwright, with a WebSocket-backed human-in-the-loop control plane, JSONL run telemetry, and fuzzy entity resolution — dual-provider inference (hosted Gemini, fully local Ollama) behind a single provider interface, so hosted and self-hosted models are swappable at zero call-site cost.

Self-custody AI home lab — distributed local inference | sole architect

exo · RDMA over Thunderbolt 5 · Ollama · MLX · Qwen · Apple Silicon · Hermes · OpenClaw

- Distributed systems in practice: a two-node Apple Silicon cluster interconnected over RDMA on Thunderbolt 5, serving a single sharded model as one address space — benchmarked at roughly a 99% inter-node latency reduction versus IP networking, with exo pooling both machines' unified memory: inference optimization at the interconnect layer.
- Run production agent graphs and multi-step agent loops on that fabric with Hermes and OpenClaw — model selection and routing, context and memory management, scheduling, failure recovery — with Qwen served locally through exo/Ollama/MLX, a data-sovereign serving architecture with no data leaving the network.

Forward-deployed cloud engagements — architecture and POC work in customer-owned environments

AWS · Amazon Aurora · Kubernetes · GraphQL · Firebase · Netlify

- Customer-facing engineering inside customer-owned accounts and repositories: stood up cloud-native architecture on AWS with Amazon Aurora, Kubernetes for container orchestration, and GraphQL API layers — then ran technical product discovery and proof-of-concept (POC) validation in the customer's own environment, producing the reference architecture their team built against.

AugHacks 2025 at MIT — 1st Place (LONG track), AI hackathon

Python · C/C++ · Lab Streaming Layer · NeuroLM · Jupyter · OpenBCI

- Built a real-time EEG and machine-learning brain-computer interface in a weekend, team of three: an OpenBCI headset streaming 6-channel EEG over Lab Streaming Layer into a NeuroLM neural network for attention scoring, unsupervised EEG clustering in Python/Jupyter, and a C/C++ wearable-camera trigger.

✓ TECHNICAL SKILLS

AI & agents: agentic systems, tool calling / function calling, MCP (Model Context Protocol) servers, RAG architectures, evaluation harnesses and LLM evals, LLM APIs (Anthropic, OpenAI, Google, xAI, OpenRouter), agent orchestration and multi-model orchestration, human-in-the-loop (HITL) gates, agent-graph and agent-loop frameworks (Hermes, OpenClaw, Vercel AI SDK and AI Gateway, hand-rolled typed-tool loops), Tavily search API, tracing and prompt observability

Inference & GPU: model serving on self-hosted inference runtimes (exo, Ollama, MLX), inference optimization at the interconnect layer, RDMA over Thunderbolt 5 interconnect, unified-memory pooling and single-model sharding across nodes, open-weight models (Qwen, Llama), GPU workloads — sizing and architecture (NVIDIA A100 80GB, H100, HPC and SuperCluster), PyTorch, multimodal (speech, image retrieval, EEG/BCI ML)

Languages & backend: Python, FastAPI, Uvicorn, TypeScript, JavaScript (Node.js, ESM), C/C++, SQL, Bash, REST and GraphQL APIs, WebSockets, distributed systems, Next.js, React, scheduled jobs

Cloud & infrastructure: Kubernetes, Docker container registry and runtime workflows, cloud-native development, AWS (Aurora, Bedrock), Google Cloud / GCP (Firebase, Gemini), Microsoft Azure (multicloud interconnect), Oracle Cloud Infrastructure (OKE, API Gateway, Functions, Resource Manager/Terraform), deployment automation and repeatable deployment patterns, infrastructure-as-code, GitHub Actions CI, Git/GitHub (forks, PRs, code review), open source

Integrations & auth: reusable integration patterns, OAuth2 with automatic token refresh, webhooks (svix signature-verified inbound, Stripe), SDKs (Vercel AI SDK, Stripe server SDK, self-authored MCP tool surfaces), enterprise APIs and partner integrations — four third-party commerce APIs normalized behind one interface, developer platforms, passkeys (FIDO2/WebAuthn), IAM, Playwright automation

Data & retrieval: retrieval systems and vector databases (Pinecone, OpenSearch) with reranking (Cohere Rerank), chunking, embeddings, vector search, PostgreSQL (Neon, Amazon Aurora), Redis (Upstash KV), data pipelines (JSONL/CSV ingestion, run telemetry, scheduled sync)

✓ PLATFORM & PARTNER EXPERIENCE

Senior Field Account Executive — Oracle Cloud Infrastructure (OCI)

2020 – Apr 2026

Oracle · New England & Boston–Cambridge — field territory (enterprise, ISV & AI-native accounts)

- Turned a one-off partner migration into a reference architecture and repeatable deployment pattern reused across four enterprise customers — each customer's on-premise licensed product taken to SaaS on OCI end to end, architecture and POC through production cutover and multi-year commitment.
- Partner engineering and customer-facing engineering with founders, CTOs, and ISV partner engineers from seed stage to enterprise: architecture reviews, live technical demos, and POC scoping run inside partner and customer engineering teams, plus ISV, systems integrator (SI), and technology-partner co-development through cloud-marketplace listings. Opened a Boston–Cambridge startup and AI-native ISV patch (20 net-new logos, cold outbound), sizing and specifying GPU workloads on NVIDIA A100 80GB and H100 capacity, including HPC and SuperCluster configurations.
- Left behind reusable artifacts rather than one-offs: an AI prospecting framework adopted by a 300-person organization, and technical writing at scale — co-authored Oracle's "Why OCI" GTM-for-AI manifesto (2023), adopted across a 1,500-person field organization.
- Grew the territory to \$5M self-sourced net-new ARR, expanded landed accounts 4× to \$20M territory ARR, and closed the region's largest-ever deal at \$10M TCV — 150%+ quota six consecutive years, President's Club 2023 and 2024.

Field Account Executive — Oracle Cloud Infrastructure (OCI)

2018 – 2020

Oracle · Vermont, New Hampshire, Maine, Rhode Island — four-state greenfield territory

- Opened OCI's four-state New England territory from a zero install base, self-sourcing 100% of pipeline and landing the region's first net-new OCI logos — discovery, architecture reviews, and POC scoping run solo as the first Oracle sales rep to earn the Oracle Cloud Architect Associate certification (2018). **Promoted to Senior Field Account Executive in 2020.**

Account Executive — Oracle Database (License) 2017 – 2018 · Oracle, Research Triangle Park, NC — sold Oracle Database and its options (RAC, partitioning, Advanced Security, in-memory, multitenant) into enterprise accounts; promoted from inside sales to field AE.

Inside Sales Representative — Oracle Hardware 2016 – 2017 · Oracle, New York, NY — sold Oracle enterprise hardware (x86, SPARC, Exadata engineered systems, storage) into the New York metro; promoted to Account Executive within a year.

Earlier — AT&T, Sales Representative · LA Fitness, General Manager 2012 – 2016 · AT&T (Walpole, MA, 2014–2016): top regional performer, closed the store's largest single transaction. LA Fitness (Walpole & Stoughton, MA, 2012–2014): ran two clubs' membership sales billing **\$33,000/month**, leading a **7-person team** with hiring, coaching, and P&L.

Education: The University of Massachusetts Amherst · **Certifications:** 2× Oracle Cloud Architect Associate (2018, 2021) · Anthropic — Claude with Amazon Bedrock (2025) · **Community:** Techstars · MassChallenge · Massachusetts AI Coalition