Chad Avalon

AI Product Engineer / Founding Engineer | Agentic Systems, LLM Observability, Product Metrics
San Francisco Bay Area · 503-801-1095 · avalon.chad@gmail.com
LinkedIn: linkedin.com/in/chad-avalon | GitHub: github.com/cavaloni | Portfolio: dev.chadavalon.com

TECHNICAL SKILLS

Proficient: TypeScript, React, Redux, Next.js, Zustand, shadcn, SCSS, Node.js, Python, Django, MongoDB, Prisma, Cypress, CI/CD, Docker, RunPod, Vercel, AWS (EC2, S3, CloudWatch, SQS), GCP (Compute Engine, Cloud Storage, Cloud Functions, BigQuery), RabbitMQ

Exposure: Redis, PostgreSQL, Supabase, GraphQL, Kubernetes, Terraform, vLLM, HuggingFace, LangChain, Pinecone, WebSockets, Server-Sent Events (SSE), Tool Calling, Multi-Model Routing, LLM Observability (Langfuse, Arize Phoenix)

PROFESSIONAL EXPERIENCE

Routly | Founding Engineer | Remote **Jan 2025 – Present**

- Built a carbon-aware LLM routing system optimizing across carbon intensity, latency, cost, and model quality, reducing average workload carbon intensity by ~40% per session via intelligent multi-region routing.
- Designed a production streaming inference platform using Node.js, Redis, vLLM, and WebSockets/SSE with health checks and automatic failover (8.5% fallback rate, zero data loss).
- Implemented end-to-end observability for routing decisions, latency (p50/p95), throughput, error rates, and per-request carbon accounting with historical persistence and dashboards.

Independent AI Product Engineer (Agentic Systems) | | Remote **Feb 2025 – Present**

- **SomniScope — Agentic CPAP Analytics System**: Built an agentic AI analysis system for CPAP data where the LLM plans and explains results while deterministic tools compute exact analytics, preventing hallucinated metrics and enabling auditability.
- Engineered for high-frequency time-series data (25Hz CPAP signals; ~21.6M samples/month) using streaming ingestion, tiered storage (SQLite + DuckDB + Parquet), and columnar analytics for interactive performance.
- Integrated RAG over personal sleep journals to correlate subjective notes with objective CPAP metrics; added evidence artifacts, provenance tracking, and production observability (Langfuse) with cost and latency KPIs.
- **WhatsCal**: Built a tool-calling conversational scheduling assistant over WhatsApp/SMS for dental practices, handling booking, rescheduling, and FAQs with safety guardrails and human handoff.
- Implemented AI-driven broadcast optimization (send-time, wave sizing, channel selection) with production observability (Arize Phoenix), improving fill rates and reducing time-to-fill.

Somnology | Lead Frontend Engineer | Bay Area **Jun 2023 – Feb 2025**

- Led development of a physician-facing sleep apnea management platform using React, TypeScript, and Node.js, integrating data from 7 wearable device types.
- Implemented LLM-assisted clinical search (RAG) over patient and device data to support faster, reviewable clinical insight.
- Built CPAP compliance visualizations that reduced patient onboarding time by 35% and automated ingestion pipelines reducing data entry errors by 60%+.

Zulily | Software Engineer | Seattle **Dec 2019 – Nov 2022**

- Built a vendor onboarding portal in React/Redux that reduced product listing time by 150%.
- Improved AWS-based logging and error visibility across 4 teams, reducing mean time to resolution by 70%.
- Led standardization of a shared React component library across 5 teams.

Syndio | Software Engineer | Santa Cruz **Oct 2017 – Jun 2019**

- Implemented a pay equity calculator contributing ~$100K in new sales within three months.
- Built legal analytics dashboards cutting review time by 40%.

Montaia Global | Frontend Engineer (Contract) | **2016 – 2018**

- Built an event and lodging management platform in React/Node.js that increased event page engagement by 300%.

EDUCATION

M.S. Computer Science – Sofia University, Palo Alto, CA — 2016