



MICHAEL GIFFORD-SANTOS

Product engineer — AI in production: agents, real-time voice, retrieval, distributed inference

hello@michaelgifford.com · +1 408.599.2275 · michaelgifford.com · github.com/mijka1 · linkedin.com/in/michaelgifford

Fifteen years full-stack, on teams and on my own. Engineer #8 at Center, growing with it past 100 people and leading onboarding for every new web engineer as the team scaled from two to three scrum teams; then founding engineer and first hire at a K-12 platform that reached six-figure weekly transaction volume on a team that never exceeded two engineers. The last two years I've built and operated production AI systems — inference, agent orchestration, real-time voice, mobile, billing, monitoring, incident response — with a co-founder on product and go-to-market, plus the operational layer that only appears once agents do sustained work: supervision, trust boundaries, cost, provenance, recovery. I run what I build.

WHAT I BRING

- **I ship AI into production, then operate it.** A real-time voice agent live in the App Store and Google Play, with the retrieval layer behind it — OCR-and-web ingestion into Qdrant, reranking and reciprocal-rank fusion, hybrid vector + full-text search over Postgres. The agent fleet that builds my own work runs on infrastructure I wrote: **Sigil** (open source) supervising the sessions, **Drydock** deciding what is allowed to run unattended.
- **I design the thing, too.** Six years as a working designer before engineering — and already automating it: a weather page a newsroom had assembled by hand for decades became a pull, a format and a drop into InDesign, in minutes. On every product below I owned the interface and the brand as well as the stack, which collapses the translation layer between the two.
- **I own money paths.** Two production payments systems in different domains, plus three years building corporate card and expense infrastructure. Collections, failed charges, chargebacks, refunds, entitlements, make-whole tooling.
- **I own things end to end — including after launch.** Apps live in the App Store and Google Play, with the runbooks, monitoring and written postmortems behind them. Two engineers carried Ordo through multiple raises on an architecture we kept deliberately small. Doing it without a support cast is how I know what every layer costs to run.

EXPERIENCE

WunderGuide (JNSQ Labs) — Co-founder / CTO · Aug 2025–present

Voice-first AI travel concierge, launched Hawai'i-first; live in the App Store and Google Play. Incubated inside Dodecki, spun out Aug 2025. Co-founded with a partner who leads go-to-market; we work the product together near-daily. Hands-on across engineering, product design, brand and print collateral, and infrastructure/ops.

- Production real-time voice agent. LiveKit concierge grounded in live itinerary, location, travel-time and weather context under a real-time latency budget — grounding freshness traded against turn latency explicitly.
- Built the reliability half that makes it a product rather than a demo: an authenticated RevenueCat → entitlement pipeline with server-side reconciliation, plus make-whole tooling for when money and entitlement disagree.
- Built the retrieval layer behind the concierge. PDFs, web pages, documents and scanned images (OCR) through chunking, dedupe and contextual enrichment into Qdrant; query expansion, reranking and reciprocal-rank fusion on the way out; exposed as agent tools so text and voice share one knowledge base. Plus hybrid search over deals and business profiles in Postgres — vector + full-text + rating + recency — so the agent can tell a strong semantic match from a weak one.
- One product across native, embed and web: an affiliate-attribution system (probabilistic fingerprint + install-referrer), a partner portal for hospitality operators, an embeddable widget, and a web client mirroring the React Native app.

Ordo — Founding Engineer (first hire) · Sep 2021–Oct 2024

K-12 food-service platform; Accel- and XYZ Venture Capital-backed. Joined at seed with one live school and one signed; sole engineer for roughly the first year, one of two after that, in a ~10-person company that grew to hundreds of schools across 15+ states.

- Delivered the MVP in under three months, replacing the founders' Bubble prototype with the platform the company onboarded every school after the first onto.
- Built and owned the payments path — Stripe at six-figure weekly transaction volume, including collections, failed charges, chargebacks and refunds.
- Ran infrastructure, deploys and on-call solo, and onboarded the second engineer when engineering doubled.
- Shipped an iPad check-in kiosk, then followed real service. Larger lunchrooms ran on paper, so roster- and homeroom-grouped PDF report packs became the deliverable instead — and a per-meal name-labeling step got cut once live operations showed it added labor with no service benefit.
- Built web, mobile and admin systems (React, React Native, Node.js, Hasura (GraphQL/Postgres), Stripe) and the support tooling a two-person team used to serve those hundreds of schools.
- The founders drove the business model — local caterers and vendors — and a strong ed-tech salesperson turned it into contracts. I built what the model needed to run, through two further raises totalling eight figures. Named at ordo.com/our-story; retained equity.

Dodecki Labs — Founder / Product Engineer / Designer · 2020–present

My product studio, founded 2020, set down for the Ordo years and picked back up full-time in Oct 2024. WunderGuide was incubated here before spinning out in 2025; GentlePrep, Span and Yapp are built and run here now. The AI systems above are operated from here, with runbooks, monitoring and written incident writeups across a Kubernetes / Docker / Proxmox estate — self-hosted on Linode and Cloudflare rather than a hyperscaler, at the size where that is cheaper to run and simpler to reason about.

- **I run a blend of coding agents in concert, not one.** Claude, Codex, Antigravity and others go on the same problem deliberately — consensus where it matters, different reads where it doesn't, and disagreement treated as a signal to go and check. Much of what I ship is code I directed, reviewed and debugged rather than typed: Sigil is a Go daemon I architected and operate without writing Go.
- **GentlePrep** — clinical-prep companion, iOS + Android, shipped and operated solo. A photographed prep sheet is parsed by a vision model into a structured instruction set, which a deterministic scheduler counts back from the appointment into the timeline the app runs on.

- **Ensembl** (Sep 2020–Sep 2021) — real-time meeting assistant on self-hosted Jitsi/WebRTC, co-founded with a business-side partner while I led engineering and owned full technical execution from first commit to production. React web, React Native iOS + Android, Firebase, Stripe. IP retained — and the real-time media foundation the voice-agent work later stood on.

Center — Senior Software Engineer, CenterCard · 2017–2020

Joined as software engineer #8; left at 100+ employees. Corporate card and expense platform where correctness and reconciliation were the product — the foundation I helped build was acquired by American Express in 2025, five years after my tenure.

- Built and owned major features end to end — async job processing, approval delegation, inbox, and custom workflow configuration, plus light backend work in Node on Lambda against DynamoDB.
- Led onboarding for every new web engineer as the web team scaled from 2 engineers to 3 scrum teams.
- Maintained the internal React component system; led cross-team code and design reviews.

Dodecki Software — Founder & CEO · 2012–2017

Order-ahead and pickup for Hawai'i restaurants — my first venture, and the one where I was the business half.

- **Founder and CEO** alongside a technical co-founder who led engineering: I raised capital, recruited the team, designed the product, and ran merchant onboarding and daily ops.
- **And I built everything around the app** — site, infrastructure, integrations and the systems the business ran on. My co-founder owned the app itself.
- Wound down after five years — the venture that taught me the most, and the one whose people I kept: several remain friends, including my WunderGuide co-founder, a fellow **Founder Institute** alum.

TECHNICAL

AI/ML systems — LLM agent orchestration · multi-model workflows (Claude, Codex, Antigravity, others in concert) · RAG and hybrid retrieval (Qdrant, reranking, RRF) · MCP · Anthropic API / Claude Agent SDK · MLX inference, continuous batching, on-device · agent trust boundaries · cost governance **Real-time** — LiveKit / WebRTC voice agents · self-hosted media servers (Jitsi, mediasoup) · TURN/NAT traversal · latency budgeting · live context grounding — in production since 2020 **Languages** — Python · TypeScript / JavaScript · SQL · shell **Platform & ops** — FastAPI · Node (Express, Fastify) · REST / GraphQL · React / React Native / Expo · auth and applied crypto (OAuth, JWT minting, entitlements, Ed25519 identity, HMAC signing, AES) · QUIC · Docker · Kubernetes (k3s) · Proxmox · CI/CD · Linux administration · Linode / Akamai · Cloudflare (DNS, TLS, CDN) · AWS SES · networking and firewalls (OPNsense, VLANs) · S3-compatible storage · self-hosting continuously since the 1990s **Data & serving** — Postgres · MySQL · SQLite · MongoDB · Supabase (self-hosted) · Firebase (RTDB, Firestore) · Caddy · nginx **Mobile** — iOS + Android · React Native / Expo · SwiftUI · StoreKit / RevenueCat · App Store and Play release ops

SELECTED SYSTEMS

Sigil — agent-session infrastructure · Go · **open source (MIT)**, mijkal.github.io/sigil Daemon that turns tmux into a supervised process manager for long-running AI agent sessions: observable, interruptible and auto-restoring after host crashes — with sigil-web, a browser client that attaches to any session from any device over HTTPS, so the fleet is reachable from wherever I am. Built so the work is not tied to the machine I am sitting at: sessions live on a server, and I drop into the same running one from my workstation, a laptop, or a phone. Closing the lid stops nothing.

Drydock — multi-agent operations platform · React / FastAPI / Postgres Operator dashboard coordinating coding-agent work across a ~16-project portfolio. It runs real work today — my job search among it — under supervised delegation and human gating; full autonomy is the direction, not the current state.

- HMAC-signed hook callbacks and provenance tainting — context from untrusted sources routes the work onto the human-approval path instead of running unattended.
- Human-gated proposal model, with a per-task usage ledger and rate-limit ceilings — agent cost control as a first-class part of the platform, not an afterthought.
- Built to survive host failure: agent sessions auto-restore; runaway workers are bounded by concurrency caps, host-pressure gating and per-host jails.

Mycellm — distributed LLM inference network · Python / QUIC · **open source (Apache-2.0)**, PyPI, iOS app · mycellm.ai Aiming at a BitTorrent-style protocol for inference: pool heterogeneous hardware, contribute capacity, draw on it. What runs today is a QUIC overlay with NAT-friendly reverse connections — seeders dial out and the gateway routes inference back down the live session, because hole-punching does not survive symmetric NAT. Public, private and federated networks are all first-class, and a node serves local weights or brokers any compatible API behind one interface.

- **Made continuous batching the default MLX path**, driving `mlx-lm`'s BatchGenerator directly rather than vendoring an existing wrapper (prior art: oMLX, Apache-2.0, credited in NOTICE) — then measured it: ~4.3× aggregate throughput at 32 concurrent streams, single-stream latency unchanged, reproduced three months later across two MLX releases on a host that was serving other work (M1 16GB, Qwen3-1.7B 4-bit; ~1.9× at 16 streams on a 30B MoE — dense models batch better).
- On-device iOS inference in a shipped app; routing and capability announcement across heterogeneous nodes.

EDUCATION

- BA Advertising — **University of Oklahoma**, 2006
- Graphic Design (AA equivalent) — **University of Hertfordshire**, UK, 2004
- **Founder Institute** (startup accelerator) — Honolulu, 2013