

Michael Gifford-Santos

Applied AI engineer — agent infrastructure, real-time voice, distributed inference **Seattle, WA** — relocating to the SF Bay Area · hello@michaeltgifford.com · +1 408.599.2275 · michaeltgifford.com · github.com/mijka1 · linkedin.com/in/michaeltgifford

Fifteen years full-stack, on teams and on my own. Engineer #8 at Center, growing with it past 100 people and taking the web team from two engineers to three scrum teams; then first engineering 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 been building and *operating* the AI systems most teams only prototype — inference, agent orchestration, real-time voice, mobile, billing, monitoring, incident response — with co-founders on product and go-to-market. 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. On every product below I owned the interface and the brand as well as the stack — which removes a hire and 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 ship alone and it holds up.** Apps live in the App Store and Google Play; ~15 deployed services with runbooks, monitoring and written postmortems. Two engineers sustained Ordo through multiple raises because of what we deliberately didn't build.
- **I've been both halves of a founding pair** — CEO/GTM at Dodecki v1 with a technical co-founder who shipped, the engineering half at Ensembl (Trip Terrano) and WunderGuide (Iqbal Ashraf).

SELECTED SYSTEMS

sigil — agent-session infrastructure · Go · **open source (MIT)**, github.com/mijka1/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 because operating a fleet of coding agents surfaced failure modes no local harness handles.

Drydock — multi-agent operations platform · React / FastAPI / Postgres Operator dashboard running an ~16-project portfolio of autonomous coding agents.

- 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.

mycel1m — distributed LLM inference network · Python / QUIC · **open source (Apache-2.0)**, PyPI, iOS app · mycel1m.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 — the BitTorrent parallel is the topology, not just the swarm. Hardware-agnostic: a node serves local weights or brokers any compatible API behind one interface.

- **Continuous-batching prototype: 4.29× throughput at concurrency 32**, single-stream latency unchanged — driving `m1x-lm's` BatchGenerator directly rather than vendoring an existing wrapper; prior art (oMLX, Apache-2.0) credited in NOTICE.
- On-device iOS inference in a shipped app; routing and capability announcement across heterogeneous nodes.

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 **Iqbal Ashraf**, 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.
- **Build 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 Software Engineer (first engineering 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 printed PDF report packs — sorted and grouped by roster and homeroom, matching how kitchens and front offices already worked — became the deliverable instead.
- **Cut a per-meal student-name labeling step** once live operations showed it added cost and 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; a deliberate “**v0.5**” **reporting stopgap** held in production nearly three years before its v1.
- Front end stayed mine end to end; the backend moved Node → Rails as engineering became a pair.
- **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 (Accel, then XYZ, cumulatively eight figures)** on a team that never exceeded two engineers. Named at **ordo.com/our-story**; retained equity.

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

Full-time since Oct 2024. The AI systems above are built and operated here — plus ~15 deployed services with runbooks, monitoring and written incident writeups across a k3s / Docker / Proxmox estate.

- **I run a blend of coding agents in concert, not one**. Claude, Codex and others are put on the same problem deliberately — to get consensus where it matters and genuinely different reads where it doesn't. Knowing which model to hand a problem to, and when disagreement between them is the signal, is its own skill.
- **GentlePrep** — clinical-prep companion, iOS + Android. Shipped and operated solo, store ops included — and **making sales on SEO/ASO alone, with no marketing spend**.
- **Ensembl** (Sep 2020–Sep 2021) — real-time meeting assistant on **self-hosted Jitsi/WebRTC**, co-founded with **Trip Terrano** on the business side 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 · acquired by American Express (Apr 2025)

Joined as software engineer #8; left at 100+ employees. Corporate card and expense platform where correctness and reconciliation were the product.

- Built and owned major features end to end: async jobs, delegation, inbox, custom workflows.
- **Grew the web team from 2 engineers to 3 scrum teams; led onboarding for every new web engineer** through that scaling.
- Maintained the internal React component system; led cross-team code and design reviews.

Dodecki Software — Founder & CEO · 2012–2017

Food ordering and delivery 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.
- Wound down — and it is why everything since ships with a runbook. **Founder Institute, 2013**.

TECHNICAL

AI/ML systems — LLM agent orchestration · multi-model workflows (Claude, Codex, 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 · shell **Platform** — FastAPI · React / React Native / Expo · QUIC · Docker · k3s · Proxmox · GitHub Actions · **self-hosting continuously since the 1990s** **Data & serving** — Postgres · MySQL · MongoDB · Firebase (RTDB, Firestore) · Caddy · nginx · Apache **Mobile** — iOS + Android · React Native / Expo · SwiftUI · StoreKit / RevenueCat · App Store and Play release operations

EDUCATION

BA, **University of Oklahoma** · Design, **University of Hertfordshire**, UK · **Founder Institute**, Honolulu, 2013