Hieu Pham

San Francisco, CA | 619-315-7238 | hmpham317443@gmail.com | LinkedIn | GitHub | Portfolio

SUMMARY

Product-focused software engineer building production AI systems in TypeScript and Python. Shipped an LLM agent serving live iMessage conversations for a consumer dating product, real-time voice workflows on a Medicare platform serving 500+ agents and 50,000+ daily queries, and an AI interview platform used across 2,500+ sessions. Experienced owning products end to end, from distributed infrastructure and evaluation systems to user-facing features.

TECHNICAL SKILLS

Languages: TypeScript, Python, Go, SQL

Product Engineering: React, Next.js, React Native, Node.js, Bun, NestJS, FastAPI, REST APIs, WebSockets

AI / LLM: OpenAI (GPT-5), Anthropic Claude, LangGraph, LangChain, RAG (Retrieval-Augmented Generation), LLM Evals (Braintrust, LangSmith), Pinecone, LiveKit, ElevenLabs

Infrastructure & Data: GCP (BigQuery, Pub/Sub, GKE), Docker, Kubernetes, Terraform, Restate, MongoDB, PostgreSQL, Redis, BullMQ, RabbitMQ, CI/CD (GitHub Actions)

WORK EXPERIENCE

Ditto - AI Matchmaking Over iMessage — San Francisco, CA

Member of Technical Staff (Software Engineer) — *May 2026 - Present*

- Built and now own SendIntent, the multi-channel delivery layer for iMessage, SMS, and email that preserves message ordering across 100+ phone lines and 3 providers during 400+ message-per-minute bursts, using Redis leases, epoch fencing, and durable per-line BullMQ dispatch queues.
- Drove chatbot quality and latency as owner: built a release-gating harness of 6 Braintrust scorers (3 deterministic, 3 LLM-judge) to prevent quality regressions, identified the AI stage as 85-90% of response latency, and shipped skill pre-activation and no-reasoning fast paths.
- Designed and shipped end-to-end location-based matching, adopted across Summer 2026 matches for real-world date spots, by building the location picking and saving flow, H3 geo-indexed intersection matching, and an LLM venue-selection agent.

CheckRx - Medicare Intelligence Platform — San Diego, CA

Founding Software Engineer — *Mar 2024 - May 2026*

- Built CMS Medicare data pipelines that handle inconsistent schemas and missing records, powering plan recommendations for 500+ insurance agents and 50,000+ daily queries within a HIPAA-compliant platform.
- Cut plan lookup time from 10 seconds to 200ms through query optimization and Redis caching, letting agents handle 3x more client calls per day.
- Developed real-time voice agents with LiveKit and ElevenLabs that handle plan comparisons and eligibility questions, eliminating 10-minute manual lookup workflows.

Mocktalk - AI-powered Mock Interview Platform — San Diego, CA

Founding Software Engineer — *May 2023 - Jun 2024*

- Owned the React frontend and built a real-time code editor with WebSocket-based context injection, enabling AI feedback as candidates typed. Platform served 300+ users across 2,500+ sessions.
- Established Jest and Cypress coverage for critical user flows, supporting daily releases with no Sev-1 regressions over 3 months.
- Built a multimodal AI interviewer using GPT-4 Turbo and Azure Speech, synchronizing live transcripts with code-editor state to generate context-aware feedback.

PROJECTS

Frontforge

Frontend Scaffolding CLI App

- Built a Go CLI with Charmbracelet that scaffolds production-ready React, TypeScript, and Tailwind projects.
- Handles dependency pinning, routing setup, and test scaffolding through guided Quick and Custom workflows.

EDUCATION

University of California, San Diego — San Diego, CA

Bachelor of Science in Computer Engineering — *Sep 2020 - Jun 2024*