

GADI COHEN

AI • Staff-Level Full-Stack Engineer • 0→1

San Francisco, CA 94110 • gadicohen93@gmail.com • +1 858 352 8260

www.gadicohen.xyz

PROFESSIONAL SUMMARY

Staff-level engineer building production-grade AI systems. I take ambiguous product requirements and design end-to-end AI platforms, balancing reliability, safety, velocity from early architectural decisions through production.

TECHNICAL SKILLS

AI & LLM Systems: RAG, agent orchestration, Claude Code, OpenAI, Langfuse, LangGraph, Deepgram, ElevenLabs

Backend & Orchestration: Node.js, Python, Temporal, Celery, Postgres, EdgeDB/Gel, GraphQL, Flask, REST

Frontend & UX: React, Next.js, TypeScript, React Native / Expo, Tailwind

DevOps & Infra: Kubernetes, Docker, ElasticSearch, Helm, AWS, Github Actions

PROFESSIONAL EXPERIENCE

GRANTED HEALTH

San Francisco, CA / New York, NY

Staff Software Engineer, AI Agent Systems (Founding team)

JANUARY 2025 - PRESENT

Early engineer at seed-stage health-tech startup leading the development of our agent system to help users resolve complex medical billing issues. Staff-level owner across agent architecture, retrieval systems, infra, and evaluation.

- Designed and led the greenfield Case Automation system, a multi-agent platform that powers all AI-driven user-facing workflows, modeling long-running, asynchronous processes with explicit state, human-in-the-loop approvals, retries, and escalation paths, translating LLM reasoning into auditable actions.
- Redesigned and shipped the agent's core context engine — replacing a monolithic prompt pipeline with a modular, real-time assembly layer that dynamically scopes context per task, supports streaming agent workflows, and enabled the team to independently develop and test new automation capabilities.
- Built a production RAG system end-to-end (OCR ingestion, chunking, metadata modeling, scoped retrieval, answer synthesis), taking it from prototype to a core customer-facing feature, with multi-document support, dynamic prompt assembly, and automatic escalation for ambiguous queries.
- Established evaluation and observability foundations for AI workflows, including online evals, workflow-level metrics, tracing, A/B comparisons, and regression detection
- Led the Durable Execution architecture for AI actions (Temporal), defining idempotency boundaries and activity models and translating the design into initial production implementation.
- Drove quarterly technical roadmapping, authoring prioritized architecture proposals spanning task orchestration, domain expert feedback loops, and evaluation infrastructure.

ASSORT HEALTH

San Francisco, CA

Founding Software Engineer

OCTOBER 2023 - April 2024

Most senior engineer on the team, led the development of an LLM agent-powered contact center platform for healthcare, collaborating with the founding team to shape product and technological directions.

Selected Project Experience:

- Proposed, built, and shipped a critical feature that detects a user's initial intent to route them to the correct flow or contact, increasing overall automation rate for our voice AI from 70 to 85 percent

- Led overall infrastructure efforts, including scoping out deployment of agents that interface with ElevenLabs, OpenAI, and Deepgram to support thousands of concurrent LLM-driven patient conversations

BLACKBIRD.AI

San Francisco, CA / New York, NY

Senior Software Engineer

SEPTEMBER 2022 - October 2023

A central member of the engineering team, leading the design and implementation of high-impact AI projects to solve disinformation problems. Handpicked by leadership to design, build, demo high-impact, time-sensitive prototypes.

Drove several system scalability and UX efforts across the wider engineering team.

Selected Project Experience

- Single-handedly designed + implemented an LLM integration prototype in < 1 month in early 2023; it helped secure substantial funding, demonstrated expertise in LLMOps, rapid productionisation, early LLM UX
- Led a critical scalability project, resulting in a >50% improvement in network visualization generation speed by improving helm deploys, autoscaling, migrating data from Neo4J to AWS Athena, completely refactoring Flask app integrations with Postgres, Celery task queue, Elasticsearch, etc.

General Responsibilities

- Developing and maintaining scalable APIs and services, integrating diverse data sources and user inputs
- Driving UX enhancements, leading React routing initiatives and middleware improvements, boosting interactivity and page responsiveness by double-digit seconds.

SISENSE

San Francisco, CA

Software Engineer II

January 2019 - April 2022

Designed, coded, shipped, and helped maintain backend + frontend for several business-critical data analytics features on a 3-4 person team.

Selected Project Experience

- New Python/R computing platform. Helped design the architecture for the system's components and coded large parts of the feature, released in January 2022
- Architected RBAC permissioning and audit system for Periscope Data, enhancing security protocols and reinforcing system integrity

pMD

San Francisco, CA

Software Engineer

June 2016 - December 2018

Full Stack Development: Designed and implemented new features for scalable, real-time, secure systems catering to thousand of users on a server-client stack using Linux/Java/MySQL/JSP/JS

Mobile App Leadership: Managed Android development and architecture for clinical software app, maintaining, updating and publishing different app versions for thousands of Android users

Integrations: Developed HL7, answering service interfaces between pMD and external clinical systems

Partnerships: Built relationships with our customers at all levels during sales, implementation, support, and business development

EDUCATION

BROWN UNIVERSITY

Providence, Rhode Island

MsC and BA in Computer Science

2011-2016

OTHER INTERESTS

Language learning (currently self-studying Levantine Arabic); writing.