

Eitan Poplinger.

CTO · AI & Cloud Solutions Architect

Cape Coral, FL · eitanpop@gmail.com · github.com/eitanpop · eitan.ai

I design and build production AI systems, with a deep niche in healthcare. Over the past several years I've architected and personally built multi-tenant clinical AI platforms on AWS end to end — data ingestion through grounded, citation-backed LLM answers — spanning RAG and hybrid retrieval, reranking, prompt and cost optimization, PHI de-identification, and HL7/FHIR/CCDA interoperability. Behind that sits twenty years of software and cloud architecture: CTO and CIO seats, dev management, and hands-on delivery across healthcare, telecom, logistics, and more.

SKILLS

AI & GenAI: RAG (hybrid retrieval, RRF, reranking, MMR) · AWS Bedrock (Claude, Titan, Cohere) · Langchain · LLM agents · Prompt engineering & optimization · LLM evaluation / ground-truth grading · Embeddings & vector search (OpenSearch KNN/HNSW, Pinecone) · OCR & document extraction · LLM cost optimization

Healthcare Data: FHIR · HL7 · CCDA · Epic EHR integration (OAuth2 / FHIR APIs) · Wearables & health time-series (Amazon Timestream) · Carequality / TECCA · PHI de-identification · Clinical data provenance & grounding

Cloud & Infra (AWS-first): Step Functions · Lambda · ECS Fargate · DynamoDB single-table · Timestream · OpenSearch · S3 / SQS / SNS / EventBridge · Cognito / Auth0 · CDK (TypeScript) · Terraform · Multi-tenant SaaS architecture · Azure

Languages: Python · TypeScript / JavaScript · C# · SQL

Architecture: Event-driven & serverless · Microservices · CQRS / Event Sourcing · DDD · Clean / Hexagonal

Full-stack & Integration: React · React Native · GraphQL · WebRTC · OAuth2 · .NET

EXPERIENCE

CTO — Clarity

2020 – Present

- Architected and built a multi-tenant healthcare RAG platform on AWS: a five-stage Step Functions pipeline (ingestion → canonicalization → vectorization → LLM enrichment → delivery) with per-tenant state machines, vector indexes, and API domains.
- Designed the retrieval stack: hybrid BM25 + KNN search over OpenSearch fused with weighted Reciprocal Rank Fusion, Cohere Rerank v3.5 cross-encoder reranking, and MMR diversity selection — plus exact and semantic near-duplicate suppression at index time.
- Built end-to-end data provenance: every LLM answer resolves to deterministic evidence atoms (source document, XPath/page, quote), so clinical answers are grounded and auditable back to the original ingested file.
- Engineered healthcare data ingestion and interoperability: an HL7 ↔ FHIR canonicalization change feed, a containerized C-CDAs ↔ FHIR conversion service on autoscaling ECS Fargate behind a private VPC Link, PDF extraction with AWS Textract (forms/tables), and Carequality/TECCA patient discovery and document retrieval.
- Built Epic EHR integration: OAuth2 authorization-code flow with CSRF state validation, refresh-token lifecycle management in DynamoDB (TTL + streams), an authenticated FHIR API proxy, and event-driven patient queries over EventBridge.
- Drove LLM cost optimization: model tiering on Bedrock (routing lightweight passes to cheaper models), rerank-skip heuristics, token budgeting, batched embeddings, and a usage-metering/billing subsystem with idempotent events and versioned rate cards.
- Built prompt optimization workflows: S3-hosted Jinja2 prompt templates hot-swappable without redeploy, multi-pass citation repair, and an evaluation harness with ground-truth grading.
- Built PHI protection and clinical NLP services: reversible PII de-identification (tokenized redaction with secure re-identification) and clinical entity extraction covering conditions, medications, and PHI.
- Built a wearables health-data platform: webhook-driven ingestion from ~30 wearable and health-device brands (Garmin, Fitbit, Apple, Oura, Whoop, CGMs) into Amazon Timestream for time-series analytics, with LLM-agent insights built on

Langchain.

- Designed the multi-tenant identity platform: a FastAPI/PostgreSQL service mapping external IdP tokens to stable internal principals with per-tenant roles, DB-backed JWKS/RS256 verification, Auth0 provisioning with saga-style compensation, and internal-only service auth over VPC Lattice with SigV4.
- Owned the cloud architecture end to end: AWS CDK (TypeScript), Python Lambdas and FastAPI on Fargate, DynamoDB single-table design, Timestream, S3, SQS/SNS/EventBridge alerting, and Cognito/Auth0 authorization.

Python · AWS Bedrock · OpenSearch · Step Functions · CDK · Langchain · AdalFlow · FHIR/HL7

Owner / Consulting Architect — Auvul Co.

2018 – Present

- Consulting practice delivering cloud architecture across industries — serverless TypeScript/NodeJS big-data backends (Kinesis, DynamoDB, GraphQL) for robotics company Tailos; EPIC EHR and vendor integrations over HL7/FHIR with CQRS and event sourcing for Enquire Solutions.
- Built full AWS products end-to-end: Python serverless backends with Cognito/OAuth2, React and React Native frontends, and WebRTC real-time audio/video (Airplate); modernized legacy Java systems into Python Lambdas (Celink).

TypeScript · Python · C# · AWS · GraphQL · Serverless

Dev Manager — Arrow Electronics

2016 – 2018

- Directed multiple agile teams; partnered with tech leadership on migrating the on-premises stack to Azure.

Azure · C# · .NET

CIO — 012 Global

2015 – 2016

- Headed technology strategy and operations for CRM, web APIs, and live telephony, leading a multinational team.

.NET · SQL Server

Sr. Consultant — Avanade (Microsoft / Accenture)

2013 – 2015

- Built and managed systems for Microsoft Lync / Skype for Business, leading an offshore team.

C# · .NET

CTO · Software Developer — UFX · eWave

2006 – 2013

- CTO of a Tel Aviv trading platform (UFX, 2011–2013); full-stack .NET development at eWave (2006–2011).

C# · .NET