YAZAN KITTANEH

Senior Full Stack & AI Engineer

hi@yazan.io · (312) 785-4427 · yazan.io · github.com/yazankittaneh

AI ENGINEERING

Specification & Constraint Design
Agent Orchestration
Multi-Agent Decomposition
Evaluation Harnesses
Failure Pattern Recognition
Context Architecture
Cost & Token Economics
On-Prem LLM Deployment
Fine-Tuning (Qwen3 VL 8B)
Trust & Security Design
RAG / Retrieval Pipelines

TECH STACK

LANGUAGES JS · TS · Python · Go · Kotlin · Bash

FRONTEND React · Next.js · Svelte · Vue · Angular

BACKEND Node · NestJS · FastAPI · Django · GraphQL

DATA PostgreSQL · MongoDB · Redis

INFRA Docker · K8s · GCP · AWS · Azure · Proxmox

TESTING Jest · Cypress · Playwright · Selenium

EDUCATION

Grinnell College
B.A. Computer Science
May 2017

PROJECTS

OpenSpar
Open-source comparative AI analysis platform. Next.js.

Who Owns What
Property ownership data, Chicago market. Django backend, React frontend.

Jail.app
Whisper model rewrite for serverless infra. React dashboard, Django API.

All projects at github.com/yazankittaneh

EXPERIENCE

SurePayroll — *Sep 2025 – Present*

Senior Full Stack & AI Engineer Remote

- **Multi-Agent System:** Orchestrated a unified agent layer across Sierra AI and Intercom with tightly scoped instructions, explicit output constraints, and defined fallback paths—preventing specification drift that produces confident but wrong responses in production. Full integration test coverage before any response reaches a live user.
- **Context Architecture:** Structured in-session payroll state into retrievable context pills surfacing the right data per interaction type—keeping prompts lean and model responses accurate without inflating token usage. Built a React node-tree UI highlighter enabling agents to guide users through multi-page flows with zero re-engineering on site changes.
- **ML Recommendation Engine:** Built a self-evaluating onboarding recommendation system for pay types and schedules. User overrides feed back into the next inference cycle, creating a closed feedback loop that improves quality over time without manual retraining.

TechMade — *Jan 2021 – Aug 2025*

Senior Full Stack & AI Engineer Remote

- **On-Prem LLM Deployment:** Modeled cost-per-token economics across OpenAI 4o API vs. on-prem, projecting $10k/month in savings. Fine-tuned Qwen3 VL 8B with explicit classification schemas and output constraints, deployed across a 12-node Apple M4 cluster, and owned every layer from inference to UI.
- **Failure Monitoring:** Built failure detection across the model pipeline—context degradation, silent classification errors, data drift—with real-time transaction flagging as the evaluation layer. System processed 100k+ CSV rows and ~3,000 PDFs per session without a silent failure reaching production.
- **Legacy App Revamp:** Led 3-person team migrating React 16 to Next.js + React 19. Traced 75% of excess API calls and 50% of error logs to specification drift in legacy Redux before rewriting anything. Result: 50% fewer errors, 75% fewer API calls.
- **App Consolidation:** Architected NestJS microservices unifying proprietary databases into MongoDB. Delegated discrete data contracts to 3 independent remote teams (US, Mexico, Colombia)—each operating on their own spec without blocking others. Projected 50% SaaS cost savings.
- **Architecture Overhaul:** Refactored monolithic React/Node app to Next.js in 3 months by identifying the failure patterns driving 80% of regressions before restructuring. Result: 200ms load improvement, 10x error reduction.
- **NFT Tool Launch:** Led 4-person team launching a trading aggregator that hit 60,000 users in two months. Decomposed into independently deployable services with clear ownership boundaries. Stack: React, Next.js, Redis, Supabase, NestJS.

Subletinn — *Jan 2019 – Dec 2021*

Full Stack Developer Chicago, IL

- **Real Estate CRM:** Built a property management CRM in React, Node, and MongoDB. Structured lease, payment, and property data into machine-searchable schemas for N8N automation agents to act on reliably. Secured $100k in seed funding.

Outcome Health (PatientPoint) — *Oct 2017 – Nov 2018*

Android / Full Stack Developer Chicago, IL

- **Android Deployment:** Configured a Zygote-stage APK for ~20,000 wallboard devices with a trust verification layer—pre-flight checks before any APK pull or install—eliminating silent failures across the entire fleet.