Kevin Hernandez

Senior AI Full Stack Engineer

kevin@stackera.space | (517) 300-6537 | Lansing, MI, 48906 | linkedin.com/in/kevin-hernandez-99a679407 | U.S. Citizen

PERSONAL SUMMARY

Senior AI & Full Stack Engineer specializing in LLM based platforms for real-time voice, text, document processing, and autonomous workflows. Experienced in cloud-native development across Azure, AWS, and GCP, with hands-on expertise in low/no-code tools including Vellum AI, Zapier, Make, n8n, and Pipedream. Proven ability to collaborate with clinical and operations teams to deliver scalable, reliable AI solutions that improve documentation quality and reduce manual effort.

EXPERIENCE

LILT AI, *Senior AI Full Stack Engineer* 05/2023 – 12/2025 | San Francisco

- Led development of a HIPAA-compliant AI platform on AWS (EKS, Amazon Bedrock, SageMaker) with RBAC and end-to-end audit logging, enabling real-time voice, text, and document workflows that scaled to 500+ clinicians and care coordinators in the first year.
- Engineered a low latency document processing pipeline (AWS Textract + Amazon Bedrock) with RAG based EHR context injection and MCP driven tool orchestration, sustaining 93%+ extraction accuracy and increasing downstream clinical report completeness by ~28%.
- Designed intelligent workflow automation using AWS Lambda and Step Functions, leveraging RAG pipelines over structured and unstructured clinical data to generate OASIS reports, care summaries, and compliance documentation reducing processing cycles by ~35%.
- Deployed autonomous agentic systems with MCP compliant tool interfaces for care plan routing, escalation handling, and automated patient follow ups, orchestrated via AWS Lambda and LLMs, cutting manual entry errors by 30% and improving clinical operational efficiency.
- Scaled FastAPI microservices on EKS with containerized inference workloads via Amazon ECR, enforcing HIPAA aligned network policies, PHI data isolation, and RBAC at the service mesh level, achieving 99.8% uptime and reducing cloud spend by ~20% through autoscaling.
- Implemented continuous LLM evaluation and observability pipelines with RAG relevance scoring, hallucination detection, and PHI redaction monitoring enabling proactive regression detection and compliance-safe performance mitigation across production environments.

DeepHow, *Full Stack Engineer* 02/2021 – 04/2023 | Detroit

- Built and maintained customer-facing web application using Next.js with TypeScript and GraphQL, serving 50k+ monthly users across manufacturing and industrial clients for workforce training, knowledge capture, and operational reporting.
- Developed core FastAPI backend services and GraphQL APIs powering DeepHow's AI knowledge engine, improving knowledge transfer workflows by ~30% and reducing customer support tickets by ~18%.
- Containerized and deployed DeepHow's platform on GCP (GKE, Cloud Run, Cloud Storage) using Docker and GitHub Actions CI/CD pipelines, improving deployment reliability and reducing infrastructure costs by ~20%.
- Built internal data pipelines and reporting dashboards using Next.js and FastAPI, enabling clients to track workforce knowledge usage and cutting analytics turnaround from days to hours.
- Mentored junior engineers on TypeScript, FastAPI, and Docker best practices and collaborated with product managers to improve sprint planning and reduce post-release defects by ~25%.

KODE Labs, *Backend Engineer* 06/2018 – 12/2020 | Detroit

- Built backend services and REST APIs using Django and Django REST Framework to support smart building and real estate portfolio management features, accelerating feature delivery by ~25%.
- Developed APIs with Firebase authentication, validation, and error handling, improving API reliability and reducing error incidents by ~20%.
- Designed and optimized PostgreSQL database schemas for IoT device and real estate data, improving query performance by ~35%.
- Integrated Kafka to stream real-time IoT sensor data from smart buildings, enabling live tracking of energy usage, occupancy, and facility operations reducing data latency by ~40%.
- Deployed backend services on Azure (App Service, Functions, Event Hubs) using Docker, achieving 99.9% uptime and reducing infrastructure costs.

Sage Solutions Group, *Frontend Developer* 10/2016 – 05/2018 | Livonia

- Built a React single-page application with MUI components, integrating FastAPI and MongoDB to streamline HR services and recruiting workflows for outsourcing clients, which accelerated candidate processing and reduced onboarding time.
- Developed product features including candidate tracking, job posting management, and employee onboarding interfaces, reducing manual HR workload for clients.
- Participated in Agile ceremonies, code reviews, and CI workflows with Docker, improving sprint velocity by ~15%.

EDUCATION

Bachelor of Science, *University of Michigan* 04/2016 | Ann Arbor, MI
Computer Science

TECHNICAL SKILLS

Languages
JavaScript (ES6+), TypeScript, Python, Go, Java, C#

Cloud & DevOps
AWS, GCP, Azure, Docker, GitHub Actions, Jenkins, Vercel, CI/CD

Workflow Automation
Vellum AI, Zapier, Make, n8n, Pipedream, Microsoft Power Automate

Frontend
React, Vue.js, Next.js, Angular, Tailwind, MUI, SCSS

AI/ML
OpenAI, Claude, Gemini, DeepSeek, Sora, Midjourney, Stable Diffusion, Flux, Veo, STT/TTS, Runpod, Lambda Labs

AI based No/Low-Code Platforms
Lovable, Bolt.new, v0, Replit Agent, Builder, Base44, Bubble

Backend
Node.js, ExpressJS, NestJS, Django, Flask, FastAPI, REST, GraphQL, Laravel

AI Agents
LangChain, LangGraph, LangSmith, RAG, MCP, Multimodal AI, Multi-agent AI, Deepgram, Vapi, Whisper, AssemblyAI

Databases
PostgreSQL, MySQL, MongoDB, DynamoDB, Firebase, Supabase, Redis

AI Coding Assistant
Claude Code, Cursor AI, Google Antigravity, Github Copilot