Matt Vegas

Principal/Staff Software Engineer | AI/LLM Systems (RAG, Agents) | Full-Stack (React/Next.js/Node/TypeScript)

Colorado, USA (Remote/Hybrid) | linkedin.com/in/mattvegasonline | inference-stack.com | Contact: LinkedIn message

---

Summary — Full-stack engineer and AI solutions architect with 20+ years delivering enterprise-grade software platforms and GenAI systems. Hands-on builder who pairs clean architecture with security-conscious design, strong DX, and measurable outcomes.

Core strengths

- LLM application development: RAG, embeddings, vector search, prompt flows, tool/function calling, and agent workflows (LangChain, LangGraph).
- Modern web platforms: React, Next.js, TypeScript; API design and integration (REST, GraphQL); real-time UX patterns (WebSockets).
- Scalable architecture: multi-tenant SaaS patterns, RBAC/least-privilege access, performance budgets, and observability-minded engineering.
- Quality and DX: strict typing, test strategy, CI quality bars, component systems, and accessibility (WCAG 2.1).
- Cloud delivery: AWS (Lambda, EC2, S3, API Gateway, Cognito), Azure; Vercel/Firebase; pragmatic DevOps and release discipline.

Technical keywords

GenAI: OpenAI (GPT-4/4o), Anthropic (Claude), LangChain, LangGraph, Pinecone, Weaviate, vector search, embeddings, RAG. Frontend: React, Next.js, Angular, Tailwind, Storybook. Backend: Node.js, REST, GraphQL, WebSockets, serverless. Cloud: AWS, Azure, Firebase, Vercel. Practices: Clean Architecture, SOLID, CI/CD, testing, performance, accessibility.

Selected impact

- Warner Bros/Microsoft entertainment portal: reached 1M+ users in the first month; improved retention ~40%; reduced page load times ~30%; expanded access ~15% via WCAG improvements.
- SaaS client delivery (React/TypeScript): reduced bug reports ~35% post-release via strict typing and higher-confidence test coverage; accelerated feature rollout ~25%; improved API response reliability ~40%.
- Healthcare GenAI platform (Soluna AI): designed a HIPAA-aligned, privacy-first architecture with multi-agent orchestration, retrieval pipelines, and FHIR-ready APIs.

Recent roles

Inference Stack — Principal Engineer / AI Solutions Architect (consulting + product) | 2022–Present

The Smyth Group — Senior React Developer (contract) | 2025

Microsoft & Warner Bros (via CMG) — Senior Frontend React Developer (contract) | 2021–2022

TalentReef — Frontend React & UX Functional Lead | 2019–2021

Full resume, references, and deeper project detail available upon request.