

Summary

Hands-on agentic AI engineer and technology leader with 25+ years building production software, cloud platforms, and engineering organizations from inception through Series C. Builds stateful AI agents, custom orchestration, retrieval-augmented generation (RAG), persistent memory, context engineering, tool integrations, and self-hosted model inference. Combines strong software engineering and distributed-systems fundamentals with production AI delivery and executive leadership in HIPAA-regulated healthcare. Grew multiple engineering organizations from zero, including a 60+ person organization that supported growth from \$0 to \$20M ARR through Series B.

Skills

Agentic AI	Stateful agents, custom orchestration, multi-agent workflows, planning, tool use, retries, session recovery, human oversight
RAG / Context Models	Embeddings, vector and hybrid retrieval, grounding, persistent memory, retrieval-aware context assembly, MCP Anthropic Claude, OpenAI-compatible APIs, open-weight models, llama.cpp, local GPU inference, structured extraction, conversational AI
Engineering	Python, FastAPI, TypeScript, Node.js, React, Angular, GraphQL, REST APIs, distributed systems
Cloud / Data	AWS (ECS, RDS, CloudFormation, Glue, SageMaker), Kubernetes, OpenSearch, Elasticsearch, MongoDB, Redis, PostgreSQL, RabbitMQ, Docker
Production	LLM observability and tracing, audit logs, testing, CI/CD, Infrastructure as Code, DevSecOps, HIPAA and PHI-aware systems

Experience

Trackable	Remote
CTO	July 2022 – Present
- Built the engineering organization and complete SaaS platform as the first technical hire for a bootstrapped healthcare startup, scaling the product to 400+ dental practices - Developed a patent-pending matching algorithm that correlates patient records with inbound leads through multidimensional heuristics - Designed and built production agentic AI systems for call intelligence, voice appointment booking, and SMS engagement using stateful workflows, tool calls, location-specific prompts, and retrieval-aware customer context - Built a RAG-based persistent memory system with semantic retrieval and runtime context assembly, exposed through Model Context Protocol (MCP) interfaces for reliable, token-efficient agent grounding - Implemented production observability and traceability for AI conversations, recording prompts, history, responses, tool calls, token usage, and latency for debugging and auditability - Rewrote transcription and structured-data extraction pipelines to run self-hosted open-weight models with GPU-accelerated inference - Led an AI engineering transformation that introduced agent-assisted development across the organization, enabling product and non-engineering staff to implement production UI - Built reusable multi-agent development workflows with Claude Code, custom MCP servers, skills, commands, and subagents; established practices for agent use across product and engineering - Architected hybrid infrastructure across AWS ECS and bare-metal Kubernetes with OpenSearch analytics, UTM attribution, and CI/CD for multiple daily production deployments; led repatriation to self-hosted Kubernetes, cutting costs approximately 30%	
Okkanti	Madison, WI
Fractional CTO	October 2025 – Present
Provide executive technical leadership for a perinatal care platform connecting providers with expectant families	

Janus Health Technologies

Madison, WI

CTO

June 2020 – July 2022

- Grew engineering from zero to approximately 60 across application development, data science, data engineering, and automation as the first technical hire, supporting growth from \$0 to \$20M ARR through Seed, Series A, and Series B with Series C underway at departure
- Built a security-first AWS hub-and-spoke architecture using CloudFormation, ECS, Kubernetes, RDS, Elasticsearch, and Glue, with CI/CD for multiple daily deployments of systems handling regulated healthcare data
- Deployed a claims-denial prediction pipeline using Glue ETL, SageMaker, and Label Studio
- Led development of the Angular, Node.js, and GraphQL web platform

SwanLeap

Madison, WI

VP of Engineering

February 2019 – June 2020

- Led approximately 40 people across Engineering and Product and established workflows for design, estimation, prioritization, and delivery
- Migrated a legacy Meteor platform to a React-based AWS microservices architecture, cutting infrastructure costs approximately 40%
- Rebuilt Jenkins CI for parallel test automation and multiple daily production deployments; established test-driven development, linting, and structured code review practices

OrgSpan / Interactive Intelligence / Genesys

Durham, NC and Madison, WI

Development Group Manager, Genesys

April 2014 – February 2019

- Led approximately 20 engineers building configuration, onboarding, and billing platforms for PureCloud, a high-availability SaaS omnichannel contact center
- Built a virtual reality contact center experience using Unity, three.js, and WebRTC, demonstrated as a keynote to 2,000+ attendees at Interactions 2016

Senior Software Engineer, OrgSpan

April 2013 – April 2014

- Architected a real-time XMPP chat platform using Node.js, Redis, MongoDB, and WebRTC that supported 1M+ simultaneous connections

Rally Software

Raleigh, NC

Senior Software Engineer

April 2011 – April 2013

- Served as a full-stack developer and sole DevOps engineer for agilezen.com across a 20-node AWS infrastructure
- Built a real-time project management application using Socket.IO, RabbitMQ, and a custom client-side MVC framework in CoffeeScript
- Designed a zero-downtime continuous deployment system

TDS Telecom

Madison, WI

Software Engineer

November 2008 – April 2011

- Won the Inspiring Excellence award for a VoIP automation tool that cut processing time from 4 hours to 15 minutes

Network Specialist

November 2002 – November 2008

Projects

Accordyx

accordyx.com

Conference Management Platform

- Multi-tenant platform for managing academic conferences from proposal submission through peer review, acceptance, and scheduling
- Configurable proposals, blind review, revision workflows, moderator management, and automated email communication
- Next.js and TypeScript platform backed by Elasticsearch, Redis, BullMQ, S3, and SES

Audoken

audoken.com

Personalized Audio News Platform

- Personalized audio-news platform for discovering and listening to important independent journalism and major news
- Multi-source story clustering as a major-news consensus signal, with full source provenance
- Privacy-first, self-hosted models for summarization, classification, embeddings, and speech
- Personalized OpenSearch ranking designed to avoid similarity-driven filter bubbles, with one React Native application for iOS, Android, and web

Orchestrel

github.com/rbrcurtis/orchestrel

Multi-Project AI Coding Agent Orchestration Platform

- Kanban-based platform with custom orchestration for long-running, stateful AI coding agents across multiple projects
- Managed agent sessions with tool execution, real-time event streaming, provider and model routing, session recovery, context compaction, and automated lifecycle transitions

Local LLM Inference Gateway

Python / FastAPI / Ray / llama.cpp

OpenAI-Compatible Multi-GPU Model Platform

- Production-style inference gateway for open-weight chat, embedding, and classification models across NVIDIA CUDA and AMD ROCm GPUs
- Built model-aware routing, lazy loading, health monitoring, streaming responses, process recovery, and OpenAI-compatible APIs using Python, FastAPI, Ray, and llama.cpp

Education

University of Wisconsin–Madison — Bachelor of Science in Computer Science, January 2009

Patents

Usage of Virtual Reality in a Contact Center Environment — Pending, filed November 9, 2016