

SHIVAM RATNANI

+1 (978) 799-7813 • shiv@ratnani.org • <https://www.linkedin.com/in/shivamratnani/> • <https://ratnani.org>

SUMMARY

Full-stack Forward Deployed Engineer with 2+ years building production systems at scale, from payment infrastructure processing \$10M+ monthly to AI-powered clinical tools serving 100+ physicians. I help startups move fast without breaking things by architecting AI cloud-native solutions that handle real user load and real compliance constraints.

EDUCATION

University of Wisconsin-Madison

Bachelor of Science (BS) - Computer Science & Data Science (Double Major) | Graduated: May 2025

WORK EXPERIENCE

CollectWise, Inc. – Forward Deployed Engineer – Manhattan, NY

Jan 2026 – Present

- FDE on founding team owning end-to-end client integrations across 20+ enterprise accounts, generating \$150K–\$200K MRR, working directly with client CEOs from scoping through production launch.
- Built and maintained payment processing integrations handling more than \$10M+ in live consumer debt payments monthly; diagnosed and resolved production payment failures to protect collections revenue.
- Architected SIP trunking and caller ID reputation infrastructure (Jambonz, HAProxy, Five9, Livekit) supporting 500K–1M AI voice calls monthly; evaluated attestation vendors to preserve answer rates.
- Implemented FDCPA/TCPA/Reg F compliance infrastructure, including validation letters and SMS consent systems.
- Resolved production incidents spanning payment processors, SFTP webhook CPU saturation, and dialer throughput; built CloudWatch/Grafana observability across AWS (EC2, RDS, Lambda) and Fly.io processing 10–50K logs/hr.

Expert AI – Lead Software Engineer – Farmington Hills, MI

May 2025 – Nov 2025

- Rebuilt the platform across API, web, and mobile, delivering an agentic SOAP-note pipeline with parallelized sub-agents and a six-tier priority queue; achieved 92%+ structured-note acceptance with one-click Athena import.
- Drove revenue capture by cutting failed EHR uploads from 15% to <3% for 100+ physicians, preventing lost charges.
- Migrated Flask to FastAPI & refactored to asyncio, doubling throughput, reducing p90 latency from 600ms to 200ms.
- Designed a central document aggregator injected into each SOAP subagent, reducing retries by 97%.
- Launched a MedicalGPT product (Python API + Next.js UI) with 6 clinician modes including lab analysis, medical record analysis, differential diagnosis, treatment planning, guideline lookup, and general Q&A.
- Fine-tuned MedGemma 27B-it on Vertex AI using thousands of (no PHI) examples; achieved ≤6s TTFT.
- Built CI/CD pipeline on GCP Cloud Run with dev & prod containers, reducing buggy deployments from 29% to 4%.
- Operated as de facto tech lead: managed contractors to deliver the web revamp on schedule.

Rightworks – Software Engineer Intern – Nashua, NH

Jul 2024 – Aug 2024

- Resolved 21 critical bugs in AccessHub's AngularJS web app, improving stability for 300,000+ active users
- Developed Windows AccessHub C# application and monitored API performance using DataDog.

ActiveLinks – Co-Founder, Lead Developer – Madison, WI

Dec 2024 – Nov 2025

- Developed iOS application in Swift/SwiftUI serving 500+ beta users with real-time data synchronization.

EduTools – Co-Founder, Lead Developer – Madison, WI

May 2024 – Jan 2025

- Founded and developed EduTools, a secure educational platform with OCR analysis and GPT assistance, enabling 1,000+ users to build collaborative knowledge bases from academic materials.
 - Architected dual-LLM system utilizing Mistral-7B for content analysis and specialized models (LLama3, Gemma2, Qwen2.5-coder, Claude 3.5 Sonnet) for targeted educational solutions
 - Built cross-platform desktop application using Electron, React.js frontend, and Node.js backend; deployed on Azure with microservices architecture for scalable content processing.
 - Migrated from AWS to Azure, implementing cost-saving measures with Azure App Services, Blob Storage, CosmosDB, and Azure Functions
-

Projects:

Stock Volatility Analyzer: Python/React AWS application analyzing 19,000+ stocks with automated screening

Connect 4 AI: Python game using Minimax algorithm with alpha-beta pruning and Pygame GUI

Skills:

Languages: Python, JavaScript, TypeScript, Java, C#, Swift, Kotlin, SQL, R, Bash/Linux

Backend & APIs: FastAPI, Django, Flask, Node.js, Express.js, REST APIs, GraphQL, WebSockets

Frontend: React.js, Next.js, AngularJS, SwiftUI, React Native

Voice AI & Telephony: LiveKit, Retell AI, Vapi, Jambonz, Five9, ElevenLabs, SIP trunking, HAProxy

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

Cloud/DevOps: AWS (EC2, S3, Lambda, RDS, API Gateway, CloudWatch), GCP (Cloud Run, Cloud Build, Compute Engine, Vertex AI), Azure (App Services, Functions, Blob Storage, CosmosDB), Fly.io, Docker, Kubernetes, GitHub Actions

AI/ML: LangChain, LangGraph, TensorFlow, Keras, scikit-learn, pandas, NumPy, LLM fine-tuning (Vertex AI), Prompt Engineering

Tools & Observability: Git, OAuth, JWT, Nginx, Gunicorn, pytest, Jest, SonarQube, DataDog, Grafana, Ansible, Jira, Confluence