

Venkata Sai Rohit Bathi

San Francisco, CA • bathirohit@gmail.com • (941) 400-2951 • [LinkedIn](#) • [Portfolio](#) • [GitHub](#)

SUMMARY

Founding AI Engineer with experience in building and scaling production AI systems. Built a job-search platform to 1000+ users and deployed 10+ enterprise-grade AI agents for healthcare and recruiting. Specialized in Agentic Workflows (LangGraph), Voice AI latency optimization (<500ms), and RAG Architecture.

WORK EXPERIENCE

CellaNova Technologies | Founding AI Engineer | *May 2025 - Present*

- Built a **full-stack AI document-generation platform** using TypeScript, Next.js, Python, and FastAPI that converts complex RFPs and company evidence into citation-grounded proposal drafts, reducing Technical Approach development and saving **85%+** of writer time.
- Engineered a **PDF/DOCX ingestion** and OCR pipeline using Docling, OpenAI embeddings, and Pinecone, processing **280+** documents with a **94.3% clean extraction rate** across technical, compliance, personnel, and past-performance records.
- Designed a **3-stage evidence-matching**, drafting, and refinement pipeline with project-scoped retrieval, Pydantic-validated outputs, inline citations, staleness detection, and hard “NEEDS PROOF” guardrails; used LangSmith and a **60-question evaluation** set to improve **groundedness** to achieve **96.2%** claim traceability.
- Shipped a **secure reviewer-facing workflow** with source inspection, revision history, SSE streaming, and PDF export, backed by WorkOS authentication, RBAC, organization-scoped data isolation, encrypted storage, audit logging, and retry/fallback handling that achieved a **97.4% generation success rate**.

ApplyLoom (Founding Project) | Founding Engineer | *Aug 2024 - Dec 2024*

- Architected and shipped ApplyLoom end-to-end from **0 to 1**, built a production Next.js, Supabase auth platform scaling to **1,000+ active** users and Chrome Extension that semantically intercepts job applications across LinkedIn, Indeed and more.
- Built a real-time **mock interview** engine with AI Personas (Recruiter, HM, Final Panel) using an intent-based model router with **A/B-tested** routing; powered live audio/video sessions via **LiveKit** across 1,000+ user sessions with sub **300ms latency**.
- Integrated multi-layer caching stack (ElastiCache, Redis) for **semantics and context**, Fine tuned **Llama 3.3 70b** with **SFT, LoRA** on **GCP Model Garden** cutting LLM API costs by 90% and reducing **p95** latency to **2s** across production traffic.
- Engineered collaborative Rooms, a real-time multiuser job board enabled members to curate and track live postings.

Cypheryard | AI Software Developer | *Jan - Jun 2023*

- Developed a low-latency production **Voice AI** Agent using **Twilio** Media Streams and Amazon Transcribe; optimized the WebSocket pipeline to achieve **sub-800ms** turn-taking latency.
- Orchestrated an automated outbound outreach engine using **SQS** and **Lambda**, scaling candidate engagement by 5x without increasing recruiter headcount.

TECHNICAL SKILLS

Agentic AI and Eval: LangChain, LangGraph, LangSmith, A2A, Tool Calling, Prompt Engineering, RAGAS, LLM as Judge, Guardrails AI, mcp-use, Fine-tuning (PEFT, QLoRA, RLHF), vLLM, Pinecone, HuggingFace, VLMs, OpenAI API, Anthropic, LiveKit, Twilio, WebRTC

Full Stack: Python, C++, TypeScript, JavaScript, Node, FastAPI, GraphQL, Pydantic, React, Next.js, Tailwind CSS

ML and DL: PyTorch, TensorFlow, Keras, JAX, Scikit-Learn, XGBoost, A/B Testing, CoreML, TensorRT, ONNX, OpenCV, CUDA, YOLO

Infrastructure & cloud: AWS (EC2, ECS Fargate, Lambda, S3, DynamoDB, Cognito), GCP Vertex AI, Docker, Kubernetes, RTSP, GitHub Actions, Redis cache, Elastic Search, CI/CD, Vercel, JWT Auth

Data and MLOps: Pandas, NumPy, ETL, SQL, Supabase, PostgreSQL, Mongo DB, Neo4j, Vector DB, AWS Bedrock, MLflow

PROJECT EXPERIENCE

Founder | Atimuss Flow | *Dec 2024*

- Built a local first personal **voice agent** that lives entirely **on device**, no data leaving your machine. **AES encryption** ensures your voice data stays private by design.
- Engineered **sub-500ms** speech-to-speech latency with raise-to-wake invocation and global hotkey **STT** injecting transcribed text directly at the active cursor position across any application.
- Shipped a **desktop agent** that lives on device, capturing voice at **159 WPM** average (4x faster than typing) operates at the system level with zero friction and end-to-end encrypted local processing.

EDUCATION

Arizona State University | **Master of Science in Computer Science** | 3.6 GPA | *May 2025*

- Relevant Coursework: Artificial Intelligence, Statistical Machine Learning, Multimedia and Web databases.

Vellore Institute of Technology | **Bachelor of Technology in Computer Science** | 8.54/10 | *May 2023*