VENKATESH SHANMUGAM

New York, NY | (703) 216-2540 | svenkatesh.js@gmail.com
linkedin.com/in/svenkatesh-js | github.com/Venkat-Git98 | venkatjs.netlify.app

SUMMARY

First AI engineer at Aithon (NYC, AI for B2B sales teams), promoted to Senior. Built the agents behind **lead generation**, its most adopted product, turning 10+ sources into millions of data points a month across about **10,000 companies**. Onboarded 6 customers, deployed and monitored **12+ production services**, cut about $8K a month with a self-hosted model, and own the LLM infrastructure behind about **42.7M calls** in 30 days. Nearly 5 years in production software, including automation for 14+ telecom client teams.

PROFESSIONAL EXPERIENCE

AITHON AI — New York, NY
Senior AI Engineer — Oct 2025 - Present

- Built **lead generation** into the most adopted product, covering **about 10,000 companies**: LangGraph agents fuse 10+ sources, including 612K job postings, into millions of data points a month, with evals gating every prompt and model change.
- Engineered a **research agent** that profiles each company on its own from its filings and web presence, searching **917K documents** with RAG and citing every claim; retrieval changes ship only after replaying a fixed question set.
- Designed a **CRM agent** that works through every deal and account in each customer's own Salesforce or HubSpot, thousands per customer, keeping a self-updating memory that tells sellers where to sell next; models are compared blind before release.
- **Onboarded 6 customers** onto the production system, found why users weren't adopting and shipped the fixes with them.
- **Deployed 12+ production services**, from LangGraph agents to dbt and Databricks data pipelines, on AWS (ECS, SageMaker) with Terraform and CI/CD, with monitoring on every service and each customer's data isolated.
- **Cut about $8K a month** by moving document parsing from AWS Textract to a self-hosted open-source model on a GPU endpoint that scales to zero, raising financial-table cells captured from near zero to 82%.
- Architected the **LLM infrastructure** every Aithon service uses to reach models: a gateway routing about **42.7M calls** across 17 services in 30 days over five providers, with every call traced and priced.

ScriptChain Health — Washington, DC
Machine Learning Engineer — May 2024 - Sep 2025

- Architected and built the **agentic AI platform** (LangGraph, multi-agent) that writes personalized nutrition and exercise plans for metabolic-syndrome patients, with NVIDIA NeMo Guardrails keeping every plan within AHA and U.S. Dietary Guidelines.
- Predicted 30-day hospital readmissions at about **78% ROC-AUC** with a hybrid model (graph attention plus a time-aware LLM).
- Reduced **preprocessing time 99%** on a 500M+ record dataset by reworking the AWS pipeline (S3, Glue, Athena), and halved GPU memory use, saving about $6K a year.

Tata Consultancy Services — Chennai, India
Software Engineer — Apr 2021 - Aug 2023

- Owned automation for **14+ telecom client teams** end to end: gathered requirements, mapped their processes, found where automation would save them time, then built, deployed and drove its adoption as their technical-to-business liaison.
- Shipped **12+ RPA bots** on client CRM, billing and network systems, cutting **infrastructure costs 40%** and turnaround about 30%.

TECHNICAL SKILLS

Agentic AI: Multi-step and multi-agent systems (LangGraph), tool calling, MCP, agent memory, guardrails, human-in-the-loop review
Generative AI: LLMs, prompt engineering, RAG, hybrid search, reranking, LoRA fine-tuning, PyTorch, inference, evals, testing
Full-stack: Python, TypeScript/JavaScript, Go, SQL, FastAPI, REST APIs, React/Next.js frontends, AI coding agents (Claude Code)
Backend & data: PostgreSQL, Redis, Elasticsearch, vector databases, dbt, Databricks, data pipelines, distributed systems
Cloud & security: AWS (ECS, SageMaker, Bedrock), GCP, Azure, Kubernetes, Docker, Terraform, CI/CD, observability, security

PROJECTS

Between-Visit: agentic care manager | *LangGraph, Claude, FastAPI, Next.js, PostgreSQL* — between-visit-agent.netlify.app

- Built a **13-tool LangGraph agent** that runs 90-day care programs behind code-enforced safety rules and clinician approval; in evals on live model runs it escalated **214/214 red-flag replies** (keyword baseline: 121/214) with 0 unsafe writes in 44/44 fault runs.

Agentic Graph-RAG for Building Codes | *LangGraph, Neo4j, LangSmith* — vabuildingcode.netlify.app

- Created **five LangGraph agents** that answer questions over the Virginia building code with Graph-RAG, evaluated in LangSmith.

EDUCATION

The George Washington University — May 2025
Master of Science in Computer Science (GPA: 3.88/4.0)

SRM University — May 2020
Bachelor of Technology in Computer Science (GPA: 3.5/4.0)