

RITHESH BEJJARAPU

Staff AI Engineer

Seattle, WA | rithesh.bejjarapu@gmail.com | (669) 333-1575 | github.com/rithesh-commits

PROFESSIONAL SUMMARY

Staff AI Engineer with 10+ years of experience across software engineering, data platforms, and production AI systems. Specializes in **local and open-source LLM inference**, RAG, **agentic systems**, enterprise knowledge platforms, and the data foundations required to operate AI reliably at scale. Proven experience building full-stack AI applications, hybrid local and cloud model architectures, and AI platforms that deliver **measurable business impact** through decision intelligence, automation, and cost optimization.

PROFESSIONAL EXPERIENCE

SLIDE | Staff AI / Data Engineer

Nov 2025 to Present

Remote

Data Foundation & Decision Intelligence

- Architected the company's governed data foundation on BigQuery, transforming fragmented operational, financial, and partner data into reusable models and data products powering executive analytics, AI applications, and agentic workflows.
- Built an evaluated AI data analyst that autonomously investigates financial anomalies, trends, and partner signals; surfaced **\$2.1M in revenue at risk** and generated opportunity signals associated with **75% of successful upsell outcomes** during the evaluated period.

Company Brain & Agentic Platform

- Designed a shared company brain that unifies structured and unstructured organizational knowledge into a common context layer for employees, AI assistants, and agents, with evaluation and recursive improvement loops that use failures and feedback to improve retrieval and response quality.
- Built an internal agentic platform connecting the company brain, data warehouse, and operational systems through MCP-based tools, enabling context-aware automation with RBAC-controlled access to sensitive data and actions.

Local Inference & Model Systems

- Designed a hybrid inference architecture spanning specialized SLMs, locally hosted open-source LLMs, and frontier cloud models, routing workloads based on task complexity, quality, privacy, latency, and cost.
- Deployed and operated **6+ open-source models** with vLLM and llama.cpp, routing **60%+ of eligible internal workloads** to local inference and reducing monthly external AI spend by **50%**, while preserving frontier models for workloads requiring higher reasoning capability.

AMAZON | Sr. System Development Engineer, AI & Security Automation

Sep 2022 to Oct 2025

New York, NY

AI-Assisted Security Assessment

- Architected an LLM-based security assessment system that evaluates third-party engagements against assessor runbooks, Amazon security policies, vulnerability intelligence, and internal security data, generating grounded initial assessments for human reviewers and reducing average assessment closure time **from 9 days to 5 days within four months**.
- Built an AI-powered customer intelligence system that aggregates OSINT, cybersecurity, fraud, vulnerability, and internal risk signals into structured risk profiles; evaluated **300+ companies** and surfaced **45 previously unidentified potential risk signals** for assessor investigation.

Assessment Platform & Automation

- Designed and built an AWS-based self-service assessment platform that dynamically assembles questionnaires, evidence requests, and evaluation workflows based on engagement context, replacing static assessment flows and improving the consistency and accuracy of security posture analysis.
- Designed middleware and orchestration services on AWS to connect the assessment platform with internal security systems, normalizing data and coordinating asynchronous workflows across multiple APIs and downstream services.

JPMORGAN CHASE & CO. | Senior Software Engineer

Nov 2021 to Sep 2022

Wilmington, DE

Automation Platform Engineering

- Built a governed low-code automation platform that enabled business teams to design, approve, deploy, and maintain workflows while securely accessing enterprise credentials, APIs, and internal systems through a controlled engineering layer.
- Expanded the platform to support **120+ reusable actions** and **50+ production workflows**, enabling self-service automation across business teams and saving **300+ engineering hours** previously spent on custom implementation and support.

SENIOR SOFTWARE CONSULTANT | Enterprise Automation & Platform Engineering

2016 to 2021

Clients: AMC Networks · Yale University · 7-Eleven · USAA

- **AMC Networks (2020 to 2021):** Built IT support and employee lifecycle automations across onboarding, offboarding, and service operations, reducing manual effort for the Service Desk by approximately **15 hours per month**.
- **Yale University (2019 to 2020):** Developed IT asset management workflows supporting healthcare and enterprise operations, and built a mobile experience enabling service desk agents to receive alerts and take operational actions from the field.
- **7-Eleven (2018 to 2019):** Built workflow and support automations for enterprise IT operations, improving request handling and reducing manual coordination across distributed teams.
- **USAA (2016 to 2018):** Worked within a core engineering team building and maintaining enterprise automation for access, provisioning, and internal engineering workflows used across corporate technology teams.

ENTREPRENEURIAL EXPERIENCE

SVARAM.AI | Founder, Side Venture

Oct 2025 to Present

New York, NY

- Founded an AI-assisted music studio for custom business anthems and promotional music, combining generative models with artist-in-the-loop creative direction to produce commercially usable tracks.
- Built AI-enabled production workflows for music generation, stem separation, vocal isolation, karaoke creation, and post-production, accelerating repetitive studio tasks while preserving human control over final creative quality.

Regionly | getregionly.com AI-powered competitive intelligence product that analyzes public business data and generates localized market insights and actionable recommendations.

EDUCATION

M.S. in Computer Science | Northwest Missouri State University | Missouri

2014 to 2015

B.E. in Computer Science | Vasavi College of Engineering | Hyderabad, India

2010 to 2014

CORE SKILLS

AI Engineering: Local and Open-Source LLMs, RAG, Agentic Systems, LangChain, LangGraph, Model Routing, LLM Evaluation, Prompt Engineering, Fine-tuning, Context Engineering, Embeddings, Vector Search, Guardrails, AI Observability, LLMOps, MCP

Inference & Model Systems: vLLM, llama.cpp, Ollama, Hugging Face, Small Language Models, Model Serving, Quantized Models, Local Inference, Hybrid Local and Cloud Architectures, Latency and Cost Optimization

Data & AI Platforms: BigQuery, Data Pipelines, Data Modeling, Enterprise Knowledge Systems, Structured and Unstructured Data, Retrieval Infrastructure, Vector Databases, Analytics Data Products, OpenSearch

Software & Cloud: Python, FastAPI, Java, JavaScript, REST APIs, AWS Bedrock, SageMaker, Lambda, API Gateway, OpenAI, Anthropic, SQL, NoSQL, PostgreSQL, Docker, CI/CD, Distributed Systems