SAI VIVEKANAND REDDY

857-465-9966 | vangala.sa@northeastern.edu | Boston, MA | LinkedIn | GitHub | Portfolio Site

About: A problem solver bridging 3 years of professional experience in full-stack development with cutting-edge work in cloud infra and AI, aiming to consistently deliver solutions that improve performance & user experience

Education

Northeastern University, Boston, USA *September 2023 - August 2025*

Master of Science, Computer Software Engineering

- **Relevant Courses:** Networks & Cloud Computing, MLOps, Big Data & Intelligence Analytics

BITS Pilani, Pilani, India *August 2016 - June 2020*

Bachelor of Engineering, Electronics Engineering

- **Relevant Courses:** OOPs, AI, DBMS, Algorithms, Operating Systems, Neural Networks, Distributed Systems

Technical Skills

Languages & Web Development: Python, Java, Go, SQL, React, HTML5, CSS, Javascript, RESTful APIs
AI/ML: Gen AI, LangChain, LangGraph, NLP, TensorFlow, PyTorch, scikit-learn
Cloud & DevOps: AWS (S3, EC2, Lambda, Route 53, RDS, KMS), GCP, Docker, CI/CD
Frameworks & Tools: Spring, FastAPI, Spark, Linux, Airflow, Kubernetes, MongoDB, MySQL, PostgreSQL, Redis

Work Experience

Humanitarians AI (5 mos) Boston, USA

AI Engineer *Jan 2025 - May 2025*

- Architected multi-agent AI infrastructure using FastAPI and Python with custom communication protocol, enabling seamless message routing and autonomous task coordination between distributed AI agents
- Built a 4-tier architecture memory management system using ChromaDB for vector embeddings and OpenSearch
- Developed multi-modal document processing engine supporting 10+ file formats with Gemini Vision API integration, achieving 95%+ accuracy in OCR and table extraction from images

Puddl (1 yr 3 mos) Bangalore, India

Member of Technical Staff *June 2022 - August 2023*

- Engineered a client-side analytics platform for OpenAI API monitoring, architecting a secure system where user API keys are stored and processed exclusively in the browser to ensure 100% data privacy
- Created an intuitive dashboard using React, TypeScript, and data visualization libraries to translate complex API data (costs, tokens, requests) into actionable insights, including a feature for local currency conversion
- Implemented front-end logic to query OpenAI organization endpoints, enabling users to identify usage trends and patterns at a granular, user-by-user level for cost optimization

Blue Yonder (2 yrs) Hyderabad, India

Software Engineer *July 2020 - May 2022*

- Revamped the data-loading mechanism in the Demand Workbench by optimizing backend Java logic and Oracle SQL queries to selectively fetch only user-requested measures, reducing initial load times by an estimated 40%
- Enhanced user productivity by re-engineering the 'Save & Calc' feature into discrete Save, Calc, and Reset actions, using in-memory caching to allow planners to rapidly simulate multiple forecast scenarios without database writes
- Owned the end-to-end resolution of 70+ customer-reported bugs annually in a complex Spring MVC application, debugging across the full stack (Java, JSP, Oracle DB), implementing robust fixes with comprehensive JUnit tests
- Managed software maintenance across multiple legacy product versions, collaborating with support teams to replicate and resolve version-specific issues, ensuring continuous stability and support for a diverse global client base

Projects

Parallelizing Text-to-Image Generation (using Diffusion)

- Reduced preprocessing time by 48% for text and 33% for image embeddings by architecting and benchmarking parallel CPU (Joblib, Dask) and multi-GPU (torch.multiprocessing) pipelines
- Achieved 1.94x (text) and 1.5x (image) speedups by engineering a multi-GPU preprocessing pipeline with torch.multiprocessing and mixed-precision (autocast)
- Scaled U-Net model training on a high-performance cluster by implementing PyTorch's DDP with mixed-precision

RAG Application (with AI Agents)

- Designed a multi-agent (OpenAI API) research system using LangGraph, orchestrating RAG, Web Search, and Academic Paper retrieval agents to synthesize documents' information, web searches, and academic sources
- Automated document processing pipeline using Airflow, Docling & Pinecone for structured extraction and storage
- Built a secure platform using JWT authentication, featuring parallel agent execution and PDF export capabilities