

Ram Sai Sri Surya Abothula

+91-7989503377 | surya.abothula@gmail.com | LinkedIn | GitHub | Portfolio | Rajahmundry, India

PROFILE

AI/LLM Engineer with 1+ year building and deploying production NLP/AI systems: LLM inference, semantic search, embeddings, candidate matching, and AI-driven ranking. Hands-on across self-hosted LLMs on NVIDIA DGX Spark (vLLM, Docker), a zero-downtime MongoDB → Qdrant migration, and recruiter-facing sourcing and ranking pipelines – from experimentation through production deployment.

TECHNICAL SKILLS

- AI/ML:** LLMs, NLP, Deep Learning, Agentic Systems, Transformers, Hugging Face, PyTorch, Embedding Models, Semantic Search, RAG, Prompt Engineering, Model Evaluation
- LLM Infrastructure:** vLLM, NVIDIA DGX Spark, Docker, Ollama, LangChain, LangGraph, Model Serving, Quantization, FastAPI, Linux
- AI Systems:** Qdrant, Elasticsearch, MongoDB, Neo4j, RabbitMQ, Vector Search, Retrieval Pipelines, AI/ML Pipelines
- Backend:** Python, FastAPI, REST APIs, Node.js, Express.js, Git/GitHub
- Frontend:** React.js, Next.js

EXPERIENCE

- SproutsAI**
AI/LLM Engineer (Intern → Full-Time)

June 2025 – Present
Remote, India

 - Benchmarked open-source LLMs (Qwen3.6-35B-A3B MoE, Qwen3.5, Gemma 4, Phi-4, Qwen2.5-72B-AWQ) against Gemini on resume parsing, candidate scoring, and email generation; shipped Qwen3.6-35B-A3B on NVIDIA DGX Spark (vLLM, Docker), ~39% faster than Qwen3.5 (~60% under concurrent load) and ~26% faster than Gemma 4.
 - Diagnosed inflated match scores on empty and sparse candidate profiles; consolidated scattered scoring logic into a single evidence-based, queue-driven (RabbitMQ) Match Parser pipeline, cutting false positives by 20% across real jobs and candidate profiles.
 - Expanded autosourcing beyond Elasticsearch-indexed ATS profiles to 8 new sources (Exa, Tavily, Apollo, Google, DuckDuckGo) and built a GitHub enrichment agent that infers skills with confidence scores from repo/commit evidence, cutting low-quality matches recruiters review by 40%.
 - Designed and built a feedback-driven candidate recommendation engine: multi-vector Qdrant retrieval (E5 embeddings across skills/experience/summary with Reciprocal Rank Fusion) over recruiter thumbs-up/down signals, re-ranked by Gemini 2.5 Flash with per-candidate scoring and natural-language reasoning.
 - Built a candidate filtering pipeline (4 rule-based filters, 3 LLM smart labels) across autosourcing and enrichment that rejects poor profiles early, cutting irrelevant profiles reaching recruiters by 30%+ and skipping downstream scoring.
 - Executed a zero-downtime production migration of vector search from MongoDB to Qdrant across 5 NLP repositories handling 10K+ daily queries, and deployed a FastAPI/Docker-based embedding service (Hugging Face SentenceTransformer) with backward-compatible validation.
- Care Automate**
Full-Stack Developer Intern

Dec 2024 – May 2025
Remote, India

 - Developed React.js frontend and Express.js backend with MongoDB; built and productionized Speech-to-Text and document extraction pipelines, integrating AI models into backend services while optimizing for real-world latency.

PROJECTS

- QEventGraph – Event-Graph Framework for Quantum Program Observability**
FastAPI, Neo4j, Qiskit, NetworkX, Next.js, React 19, Cytoscape.js

Github

 - Designed an event-graph observability framework for quantum programs (dual-relationship graph model, NEXT + QUBIT_DEP, 1000+ events, O(n) complexity) with a FastAPI + Neo4j backend and interactive React/Cytoscape.js frontend, including divergence analysis comparing noisy vs ideal executions via Qiskit Aer.
- Spendlytic AI – AI-Powered Finance Analyzer**
Next.js, Neon DB, Gemini AI, Clerk, ShadCN, Tailwind CSS

Github

 - Built an AI-powered finance analyzer using Gemini with retrieval-based context injection and prompt-grounded workflows to generate user-specific financial insights.

EDUCATION

- B.Tech in Computer Science and Engineering (Honors in AI & ML)**
Vishnu Institute of Technology (VITB), Bhimavaram, AP

2022 – 2026
CGPA: 9.11/10
- Intermediate (MPC)**
Sasi Junior College, Rajahmundry

2020 – 2022
93.8%

ACHIEVEMENTS & CERTIFICATIONS

- Star Performer Award (Employee of the Month)**, SproutsAI, Jul 2025 – recognized for contributions and ownership.
- Hackathons:** GDG HackATron 2025 – 4th of 100+ teams (NLP sports management); Smart India Hackathon Finalist (2023, 2024); DeMux Hackathon – Special Appreciation
- Certifications:** Deep Learning with PyTorch (IBM/edX); Programming with Generative AI (NPTEL); AWS Cloud Practitioner