

Charitha Sree Veluru

USA | Open to Relocation | +1 (857) 313-4385 | charithasreeveluru@gmail.com | [LinkedIn](#) | [GitHub](#) | [Portfolio](#)

PROFESSIONAL SUMMARY

AI/ML Engineer with 3+ years of experience building production-grade Generative AI, Multi-Agent AI, RAG pipelines, NLP and Machine Learning solutions using LangChain, LangGraph, PyTorch, Hugging Face and PostgreSQL/PGVector. Experienced in deploying scalable AI systems on AWS with Docker, Terraform and CI/CD, delivering high-performance semantic search, LLM evaluation, AI observability and end-to-end ML pipelines. Independent researcher in LLM hallucination detection with expertise in representation learning, hidden state analysis, statistical evaluation and production AI systems

CORE SKILLS

AI/LLM & Agent Frameworks: LangChain, LangGraph, LlamaIndex, CrewAI, OpenAI/Anthropic APIs, MCP, RAG Pipelines, Prompt Engineering, Function/Tool Calling

Memory & State: Conversation memory, long-term memory design, state persistence, checkpointing

Multi-Agent Systems: Task orchestration, agent delegation, hierarchical workflows

Observability & Eval: LangSmith, Langfuse, agent tracing, hallucination detection, benchmarking

Agent Reliability: Guardrails, prompt injection defense, input/output validation

ML/DL & Fine-Tuning: PyTorch, TensorFlow, PEFT/LoRA, 4-bit Quantization (BitsAndBytes), Hugging Face Accelerate, SHAP, NLP

Databases & Data Engineering: PostgreSQL, MongoDB, Redis, Elasticsearch, Amazon Redshift, ETL/ELT Pipelines, SQL Optimization

Cloud & MLOps: AWS (SageMaker, Lambda, ECS, S3, IAM), GCP (Vertex AI, BigQuery), Docker, Kubernetes, Terraform, MLflow, GitHub Actions

PROFESSIONAL EXPERIENCE

Agentic AI Engineer

December 2025 - Present

Humanitarians AI (Boston, MA)

- Developed a multi-agent RAG platform using LangChain, LangGraph and PGVector, delivering semantic search over 100K+ documents with 99.8% retrieval accuracy.
- Designed and implemented stateful multi-agent workflows using LangGraph, incorporating tool calling, conditional routing and multi-step reasoning to process 10K+ daily user queries with sub-2-second response latency.
- Developed Sentence Transformers-based embedding pipelines for 100K+ research papers, reducing content generation time by 95% and supporting 25K+ active users.
- Implemented LLM observability and evaluation with LangSmith, including agent tracing, prompt evaluation and hallucination monitoring, while building a Next.js analytics dashboard that improved user satisfaction to a 4.9/5 NPS.
- Containerized and deployed AI services on AWS ECS using Docker, Terraform and GitHub Actions CI/CD, achieving 98% service availability, reducing infrastructure costs by 40% and enabling automated deployments.

Data Scientist

January 2025 - May 2025

ClinicMind (Florida)

- Developed a GPU-accelerated claim denial prediction model using XGBoost (RAPIDS cuML) and SHAP on 500K+ healthcare records, achieving 91% accuracy and reducing claim denials by 18%.
- Built a real-time anomaly detection system using TensorFlow/Keras Autoencoders, reducing billing errors by 25% and saving \$120K annually.
- Designed and validated ML evaluation frameworks using Precision, Recall, F1-Score, AUROC, bootstrap confidence intervals and A/B testing across multiple healthcare datasets.
- Developed an NLP-based claim classification pipeline using text classification and statistical feature engineering to automate insurance claim routing, reducing processing turnaround time from 3 days to same-day resolution.
- Developed Airflow ETL pipelines and AWS QuickSight dashboards for 10+ insurance providers, reducing data processing latency by 50%.

Data Engineer

July 2021 - July 2023

Wipro Limited (India)

- Built and maintained scalable data pipelines for a medical equipment client, integrating procurement, inventory and supply chain data into PostgreSQL, eliminating manual workflows and reducing order delays by 40%.
- Designed and implemented ETL pipelines using Python and SQL to consolidate data from multiple enterprise systems,

- reducing inter-service data coupling by 60% and enabling unified cross-functional reporting.
- Engineered a real-time inventory synchronization platform using Redis and RabbitMQ across multiple warehouses, eliminating oversell incidents (0%) and reducing inventory reconciliation time from 3 hours to 10 minutes.
- Optimized high-volume SQL queries on a 10K+ SKU product catalog by integrating Elasticsearch, replacing PostgreSQL LIKE searches with sub-200 ms full-text search and improving query performance by 30x.
- Automated CI/CD pipelines using GitHub Actions, Docker and AWS ECS, reducing deployment time from 4 hours to under 18 minutes while achieving a 100% deployment success rate.
- Implemented role-based access control (RBAC) using JWT, OAuth 2.0 and Redis session management across three enterprise applications, resulting in zero security vulnerabilities during penetration testing.

Research Intern

July 2019 - August 2019

Bharat Electronics Limited (BEL) (India)

- Performed EDA using Python (Pandas, NumPy, SciPy) and SQL, identifying supply chain bottlenecks and reducing manual reporting by 40%.
- Researched and developed predictive demand forecasting models using regression analysis and time-series forecasting techniques to support procurement planning, delivering data-driven insights for senior leadership decision-making.
- Developed a React.js, Node.js and PostgreSQL analytics dashboard, reducing procurement report generation time from 2 days to 4 hours.

PROJECTS

AbroadIQ - Multi-Agent RAG Recommendation Platform

Python, LangGraph, Groq API, PGVector, Django, Docker, BeautifulSoup

- Built a dual-agent RAG architecture using Groq API with qwen-qwq-32b and llama-3.1-8b models, implementing PGVector semantic search with BM25 hybrid retrieval and cross-encoder reranking, achieving sub-2s response latency.
- Engineered embedding pipelines using Sentence Transformers (BAAI/bge-small-en-v1.5), converting raw content to 384-dimensional vectors enabling real-time semantic search across 100+ sources in under 500ms.
- Designed multi-agent orchestration with LangGraph, implementing tool-calling agents with conditional routing, query decomposition and quality-scored response generation.
- Integrated Django backend with PostgreSQL/PGVector, LangGraph agents and Docker containerization, delivering personalized recommendations with budget estimates.

RESEARCH

Unsupervised Hallucination Detection in Large Language Models — Research Paper

PyTorch, HuggingFace Transformers, BitsAndBytes, Scikit-learn, SciPy, PCA, Mahalanobis Distance

- Built an unsupervised hallucination detector using PCA and Mahalanobis distance, achieving 0.956 AUROC on HaluEval QA and outperforming SelfCheckGPT by 59 points.
- Engineered a multi-layer probing pipeline across 6 functionally-motivated transformer layers, discovering that mid-layer representations (25–50% depth) outperform final-layer probing used by all prior work.
- Conducted cross-domain experiments on 7 datasets, showing hidden state geometry detects LLM confabulations but not human-crafted distractors.
- Validated method across 4 LLM architectures (Llama-3-8B, Mistral-7B, Qwen2.5-7B, Phi-3-mini), demonstrating architecture independent hallucination signal in transformer representations.
- Implemented 4-bit quantized model inference using BitsAndBytes on GPU, enabling large-scale experimentation on consumer hardware (T4/A100).
- Proposed novel unsupervised LLM verification frameworks (LSJ & PHSV) based on hidden state consistency, in collaboration with an Adobe Software Development Engineer.

CERTIFICATIONS

- AWS Certified Solutions Architect - Associate | Amazon Web Services

EDUCATION

Northeastern University, Boston, December 2025

Master of Science in Information Systems

PES University, Bangalore, India, May 2021

Bachelor of Technology in Computer Science and Engineering