Vinay Sekhar Reddy

AI Engineer
+1 9403543951 | drvinaysreddy@gmail.com | LinkedIn

Education

University Of North Texas Aug 2024 – May 2026
Master of Science in Computer Science — CGPA: 3.64/4.0 *Dallas, United States*

- Coursework: Machine Learning, Software Engineering, Artificial Intelligence, Computer Network, NLP research

Experience

Merrill Lynch Jan 2025 – Present
AI Engineer *Remote USA*

- Fine-tuned **LLaMA 3.2 3B (QLoRA)** on **2658** instruction pairs synthesized from unstructured company documents without labeled data to reduce hallucinations in domain-specific policy and product Q&A.
- Architected scalable production **multimodal RAG pipelines** over unstructured **enterprise documents** (100+ page PDFs, text, images) on **Azure AI Search** and **Azure AI Foundry** with **Azure OpenAI SDKs**, combining **(dense + BM25) indexing** and **semantic reranking** to lift retrieval precision by **35%** (Precision@10) and cut factual hallucinations by **40%.**
- Deployed and monitored AI models using **MLflow** and **SageMaker** Model Monitor, implementing data drift detection, prediction quality alerts, and automated retraining workflows.
- Compressed model memory from **12 GB** to **2 GB** via **4-bit quantization**, enabling low-cost training on **NVIDIA T4 GPUs**.
- Implemented prompt engineering strategies including **few-shot prompting, chain-of-thought, and structured output formatting** to improve LLM response quality for financial advisory and risk workflows.
- Implemented an **SFT pipeline** under a **15 GB VRAM** constraint using **TRL SFTTrainer, PEFT rank-16 LoRA adapters, Unsloth kernels, AdamW 8-bit optimizer**, and gradient accumulation for an effective batch size of 8.
- Collaborated with data scientists, financial analysts, and engineering teams in **Agile sprints** to scope, build, and deliver AI features supporting investment research and client advisory platforms.
- Achieved **39%** improvement in factual accuracy **(ROUGE-L: 0.14 to 0.54)** and **17% gain** in semantic alignment **(BERTScore: 0.69 to 0.88)** over the base LLaMA model, validated on 766 held-out domain-specific examples.

Canadian Western Bank Feb 2022 – Jul 2023
Software Engineer - AI/ML/Java *Hyderabad, India*

- Built and operated high-performance **REST APIs** and microservices in **Java** with Spring Boot, adding **pre and post-processors** for request validation, response formatting, and API error handling to improve **scalability** during peak banking hours, while meeting PCI-DSS, SOC2, and FFIEC requirements including data encryption and audit logging.
- Maintained **12+ dev, QA and UAT environments** and replicated new environments for sprint at client request and worked with QA and product teams to resolve **60+ critical UAT defects**, cutting reported bugs by **40%** and improving uptime by **25%**.
- Developed and trained credit risk, loan default, churn, and fraud detection models using **PyTorch**, **Scikit-learn**, **XGBoost**, and **LightGBM**, handling class imbalance with **SMOTE** and threshold tuning, on Python/SQL feature engineering pipelines built from banking transaction, credit and behavioral data.
- Automated feature extraction and batch scoring with **Airflow**; tracked experiments in **MLflow** and W&B; deployed models to **SageMaker** with auto-scaling, A/B testing, and Model Monitor; served inference via **FastAPI**.
- Delivered production-ready **UIs** with **JavaScript**, **React-based Visualizer**, and Tailwind CSS for **1500+** daily sessions, and fixed a UAT security **vulnerability** by removing session user IDs from PDF transaction URLs and enforcing server-side authorization.
- Participated in **Agile** ceremonies and collaborated with data engineers, business analysts, and risk teams to define model requirements and deliver AI solutions aligned with banking compliance requirements.

Weights & Biases Mar 2020 – Feb 2022
Machine Learning Engineer Intern *Hyderabad, India*

- Developed core ML experiment tracking and model versioning features for the Weights & Biases platform in Python, enabling data scientists to log, visualize, and compare training runs.

- Designed and implemented automated hyperparameter sweep workflows using W&B Sweeps, PyTorch, and TensorFlow, supporting grid search, random search, and Bayesian optimization.
- Built SDK integrations with PyTorch, TensorFlow, Scikit-learn, and Hugging Face to automatically capture training metrics, model artifacts, and dataset versions.
- Developed Python client libraries and REST API integrations for experiment tracking, model registry, and dataset versioning workflows.
- Created model evaluation visualizations, including confusion matrices, precision-recall curves, and custom metric dashboards, to help ML teams assess performance.
- Leveraged Docker, AWS, and CI/CD pipelines to deploy and test platform services across development and production environments.

Projects

Proj-H [Link] | *Python, FastAPI, OpenAI GPT-4, React, Node.js, Evals* Mar 2025

- Built an object-oriented Python/FastAPI inference layer that generates personalized health plans for 100+ users while using prompt caching to control OpenAI GPT-4 costs.
- Dogfooded the product to uncover a cultural-bias failure, then ran rigorous before/after evaluations on the same input set to validate the fix—now used as a release gate for all prompt changes.
- Delivered P95 latency of 250–300ms across 20K+ daily datapoints and load-tested the Oats coaching agent to sustain 100–150 QPS under production traffic.

AEC Multi Document Conflict Detector | *Python, LangChain, FAISS, RAG, Claude Vision LLM, FastAPI* Feb 2025

- Built a RAG pipeline with LangChain and FAISS to detect specification conflicts across multi-format architectural documents, cutting manual review effort by ∼70%.
- Integrated Claude Vision LLM for diagram reasoning to catch spatial and dimensional inconsistencies missed by text-only systems.
- Deployed as a FastAPI service with strict JSON schema validation that blocks malformed LLM outputs, and delivered production-ready conflict detection with full pytest coverage and low-latency retrieval for scalable AEC workflows.

Technical Skills

Languages & Frameworks: Python, SQL, Java, JavaScript, React, NoSQL, HTML/CSS, Node.js
Tools: Linux/Unix, Git, GitHub, Jupyter Notebook, Google Colab, Docker, RESTful APIs, Postman, Triton Server
AI/ML Frameworks: LLM Fine-tuning, RAG, Hugging Face Transformers, LLaMA, Ollama, Claude Code, PEFT, QLoRA, LangChain, LangGraph, OpenAI GPT-4, Prompt Engineering, PyTorch, TensorFlow, Scikit-learn, XGBoost, LightGBM
Vector Databases: Pinecone, FAISS, Chroma, Azure AI Search, Pgvector
MLOps: Weights & Biases, MLflow, AWS SageMaker, Kubeflow, Model Monitoring
Cloud - AWS: SageMaker, S3, Lambda, EC2, ECR, SageMaker Pipelines, SageMaker Model Monitor
Cloud - Azure: Azure OpenAI Service, Azure ML, Azure Functions, Blob Storage
APIs & Deploy: FastAPI, Flask, Docker, Kubernetes, REST APIs, GitHub Actions
Databases: PostgreSQL, MySQL, MongoDB, Redis, DynamoDB

Achievements

- **Co-authored an IEEE-published research paper** on Terahertz Communication for 6G Networks during M.S. at the University of North Texas, 2025. [Link]
- **Technical Writer** — Published an article on how large-scale platforms estimate unique Reel views using HyperLogLog, probabilistic algorithms, and distributed-systems concepts. [Link]
- **Open Source Contributor** — Identified and reported a UI bug in Font Awesome with detailed reproduction steps. [Link]
- **Winner, MLH Hackathon** — Built a multilingual AI chatbot using IBM Watson NLP, powered by Microsoft.