VIKRANTH REDDIMASU

College Park, MD 20740, USA, Earth | vikranthreddimasu@gmail.com | LinkedIn | GitHub

SUMMARY

AI/ML Engineer and full-stack developer pursuing an MS DS at the University of Maryland, with hands-on experience building production-grade agentic AI systems, Gen AI applications, and data-driven platforms. Proven ability to design and deploy multi-agent architectures, conversational AI, and end-to-end applications using Python, LangGraph, FastAPI, and React. Passionate about applying emerging AI technologies to transform wealth management and financial services through innovative, client-centric solutions.

EDUCATION

University of Maryland - College Park **Aug 2024 - May 2026**
Masters, Data Science *College Park*

- **GPA:** 3.85
- **Coursework:** Deep Learning, Advance Machine Learning, Natural Language Processing

TECHNICAL PROJECTS

WEALTHAGENT: MULTI-AGENT AI SYSTEM FOR FINANCIAL ANALYTICS & TCA | GitHub

- Engineered a production-grade multi-agent AI system using LangGraph and Anthropic Claude, orchestrating 3 specialized agents with LLM-based intent classification to route financial queries and deliver real-time advisory responses across portfolio management and trade analytics workflows.
- Implemented a TCA module computing execution quality metrics including slippage, market impact, and VWAP deviation on historical trade data, extended with an ML-based counterparty and algorithm recommendation engine that identifies optimal execution strategies, analyzing historical trade performance and market conditions, enabling end-to-end pre-trade and post-trade decision augmentation.
- Engineered real-time WebSocket-based token streaming with live agent status indicators and persistent chat history, alongside a Text2SQL interface that converts natural language queries into structured PostgreSQL commands using LLM function calling with schema-aware prompting, delivering a low-latency, conversational financial analytics experience.
- Architected a full-stack application with FastAPI, React, PostgreSQL, and Docker featuring 12+ REST endpoints, Pydantic v2 validation at all API boundaries, and 85%+ test coverage across a 64-test pytest suite covering TCA computations and multi-agent orchestration pipelines.

DISTRIBUTED ML TRAINING FRAMEWORK FOR APPLE SILICON MACS FROM SCRATCH | GitHub

- Engineered a distributed ML training framework enabling data-parallel PyTorch training across up to 18 Apple Silicon Macs via Thunderbolt 4, with automatic workload balancing based on GPU core count and memory
- Implemented intelligent gradient compression pipeline (Top-K sparsification + FP16) achieving ~20x bandwidth reduction for distributed synchronization within Thunderbolt's bandwidth constraint
- Designed WeightedDistributedSampler that intelligently splits training data proportional to each node's compute capacity maximizing cluster utilization and preventing bottlenecks on lower-powered nodes
- Built end-to-end training infrastructure including checkpoint management, gradient accumulation for simulating larger batch sizes, and containerized dependencies, demonstrating production-grade DevOps practices

OFFLINE NOTEBOOK LM: RAG ASSISTANT WITH VECTOR INTELLIGENCE & LLM ORCHESTRATION | GitHub

- Architected and deployed a production-grade offline-first RAG application (Electron + React + FastAPI + Python) with a PGVector-compatible ChromaDB vector store, supporting local document ingestion, embedding generation, and LLM-powered retrieval without cloud dependency.
- Designed a modular agentic retrieval workflow where a routing agent classifies query intent to invoke either summary-level or chunk-level search across a 2-stage pipeline, achieving 2-3x faster query times with reduced memory overhead and consistent accuracy on multi-document notebooks.
- Implemented intelligent LLM backend selection with prompt optimization that benchmarks model performance on domain-specific queries and applies knowledge distillation techniques to compress larger model outputs into efficient Phi-3 and Mistral inference pipelines, reducing latency for RAM-constrained deployments.
- Built a robust multi-format document ingestion pipeline supporting 7+ file types with adaptive chunking strategies, sentence-transformers embeddings, and ChromaDB vector storage, achieving ~366 chunks/second throughput with consistent retrieval accuracy across document sizes and formats.

PRODUCTION-READY GAN FOR MNIST DIGIT SYNTHESIS WITH HF DEPLOYMENT | GitHub

- Designed and implemented a fully functional Generative Adversarial Network with 1.49M+ generator parameters and 1.46M+ discriminator parameters using PyTorch
- Built a production-ready Gradio web app with comprehensive error handling, structured logging, full type hints, and input validation. Implemented intelligent device detection with automatic GPU acceleration
- Deployed the trained GAN model to HF Spaces with seamless CI/CD integration, and interactive live demo to end users. Optimized model serialization & containerized dependencies for deployment across multiple cloud environments
- Implemented training techniques like LeakyReLU activation and Batch Normalization for generator stability. Showcased full-stack ML expertise with deep learning, PyTorch and software engineering practices(PEP 8, SRP, DRY)

TECHNICAL SKILLS

- **Programming**: Python, SQL, R, Data Structures, Algorithms, Object-Oriented programming, Software Development, Git, Docker
- **AI & ML**: Generative AI, Agentic AI, Large Language Models (Claude, OpenAI, Llama), RAG, Text2SQL, LLM Fine-tuning & Distillation, Distributed Training (DDP, DeepSpeed), Neural Network Architecture, Transformers, NLP
- **Framework**: LangGraph, LangChain, FastAPI, LlamaIndex, PyTorch, TensorFlow, APIs, React, REST APIs, WebSockets, Pydantic
- **Databases & Cloud**: PostgreSQL, ChromaDB (Vector DB), PGVector, Docker, CI/CD, Cloud Deployment (HuggingFace Spaces), GitHub Actions
- **Data & Visualization**: Pandas, Numpy, Plotly, Streamlit, yfinance, Jupyter

LEADERSHIP & STARTUP EXPERIENCE

- Invited to Stanford University for Silicon Valley Meet-up of University Innovation Fellows in March 2023
- Editorial Board Member of Stanford UIF's Change Forward Journal 3rd edition
- Hosted Design Thinking workshops at multiple universities that impacted over 5000+ students
- Co-Founded Young Web Solutions, a student-led startup providing web development and UI/UX design services