NAGARJUNA PENDEKANTI

Ph: +1(940)843-8699 |Email: nagu07799@gmail.com | GitHub

SUMMARY

AI Engineer with 4+ years building real-time financial AI and **multi-agent** systems. Delivered fraud detection at scale across 30M+ transactions per day, agentic platforms using **LangChain**, **LangGraph**, and **MCP**, and low-latency model serving on AWS. Strong in **PyTorch** and TensorFlow, PySpark and end-to-end MLOps with MLflow, Docker, Kubernetes, Terraform, and CI/CD, with governance-ready explainability and drift monitoring.

PROFESSIONAL EXPERIENCE

AI Engineer: Procal Technologies (Client: Wells Fargo) **June 2025 - Present**

- Designed and implemented a **multi-agent AI platform** using **LangChain** and **LangGraph** to automate **trading**, **risk management**, **compliance**, and **customer analytics** for large-scale banking operations.
- Architected **specialized agent workflows**, enabling collaborative decision-making via **Trading Agent**, **Risk Agent**, **Compliance Agent**, and **Customer Intelligence Agent** modules, mirroring expert advisory teams.
- Integrated **Model Context Protocol (MCP)** for seamless agent access to **financial APIs**, **regulatory databases**, **live market feeds**, **credit scoring tools**, and **AML screening systems**, supporting dynamic data-driven decisions.
- Developed **agent memory systems** combining real-time conversational context with **persistent knowledge bases**, ensuring continuity and personalization across **1,000+ user sessions**.
- Engineered **inter-agent communication protocols** for **distributed cognition**, enabling agents to collaboratively solve complex **cross-domain banking problems**.

ML Engineer: Tata Consultancy Services (Client: Bank of America) **Aug 2021 - Aug 2023**

- Built **real-time fraud detection models** for **30M+ daily transactions**, improving **precision to 98%** and reducing **false positives by 35%**.
- Built **ensemble fraud models** combining **gradient boosting (XGBoost)** and **deep neural networks**, achieving **94% precision** and **87% recall** while reducing **false positives by 52%**.
- Engineered a **streaming feature store** with **PySpark**, delivering **200+ behavioral and merchant features**.
- Deployed **PyTorch** and **TensorFlow** models on **AWS SageMaker** with **autoscaling**, maintaining **99.9% uptime** across peak loads.
- Accelerated inference using **quantization** and **pruning**, cutting **latency by 30–45%** and lowering **compute cost by 40%**.
- Implemented end-to-end **ML CI/CD** with **GitHub Actions**, **Docker**, **Terraform**, and **Helm**, reducing deployment cycles from days to **under three hours**.
- Set up **MLflow** experiment tracking and a **model registry** with approval gates for **risk and compliance**, ensuring full **lineage** and **auditability**.
- Delivered **FastAPI** gateways and **TensorFlow Serving** endpoints, plus **Kafka** and **REST** contracts that enabled **real-time case management** and **cut investigation time by 45%**.
- Introduced **SHAP-based explainability** and **transaction-level reason codes**, meeting **governance requirements** and supporting **regulator** and **internal audit** reviews.

University of North Texas: Research Assistant **Jan 2024 – May 2025**

- Designed and implemented an **Agentic RAG** chatbot enabling **24/7 support** for UNT students on **library access**, **scholarships**, **campus services**, and **academic queries**.
- Engineered **agentic workflows** integrating specialized sub-agents (**Library Info Agent**, **Scholarship Advisor**, **Campus Services Agent**) for **domain-specific reasoning** and **multi-turn dialogues**, mirroring human-like guidance for diverse student needs.
- Leveraged **open-source LLMs** (**Llama 3**, **Mistral**) and **vector databases** (**FAISS**, **Chroma**) for **semantic search** and rapid retrieval from diverse university data sources.
- Built **scalable, user-friendly web interface** with **contextual chat history**, **document uploads**, and **interactive Q&A** tailored for **real-time student engagement**.
- Used **external LLMs as judges** for pipeline validation, benchmarking RAG chatbot outputs on **answer correctness**, **hallucination rate**, and **user satisfaction**.
- Built **data ingestion and ETL pipelines** processing **PDFs**, **CSVs**, **web pages**, and **FAQs**, automating updates to keep the **knowledge base** current for all student-related topics.

PROJECTS

- Developed a **scalable large language model inference platform** using PyTorch, FastAPI, and AWS Bedrock. Focused on optimizing latency and throughput via **distributed serving, caching, and model parallelism**. Integrated with **Kubernetes and Docker** for containerized deployment and MLOps automation, achieving significant performance and cost improvements.
- Built an **AI-powered assistant** to automate engineering workflows like code review summarization, documentation generation, and knowledge retrieval. Designed a **cloud-native API platform** using Python, LangChain, and AWS Lambda that handled thousands of daily requests with sub-second latency, boosting developer productivity and accelerating GenAI adoption internally.

SKILLS

- **Programming:** Python, C++, Java, Bash
- **ML and DL:** PyTorch, TensorFlow, Keras, XGBoost, ONNX, ONNX Runtime, scikit-learn, SHAP
- **GenAI and Agents:** LLMs GPT Llama Claude, LangChain, LangGraph, Model Context Protocol, Agentic RAG
- **Data and Streaming**: PySpark, Spark, Kafka, Airflow, feature stores, data contracts
- **Serving and APIs:** FastAPI, Flask, REST, AsyncIO, Redis, batching, parallel execution
- **Cloud and Infrastructure on AWS:** SageMaker, Bedrock, Lambda, EKS, ECS, API Gateway, DynamoDB, S3, ECR.
- **MLOps and Deployment:** Docker, Kubernetes, Terraform, MLflow, GitHub Actions,

EDUCATION

M.S. Data Science **Aug 2023-May 2025**
University of North Texas **GPA 4.0/4.0**