

TEJA BABU MANDALOU

940-758-3312 tejamandalou22@gmail.com linkedin.com/in/teja-mandalou techwhizgenius.github.io

AI Engineer | LLMs, RAG & Agentic Systems

AI Engineer specializing in production LLM systems: retrieval-augmented generation, multimodal reasoning, and agentic orchestration with LangGraph, CrewAI, and MCP. I take models from Gemini and GPT-4 through evaluation (Ragas, LangSmith) to monitored inference on AWS and GCP, most recently deploying a HIPAA-compliant clinical AI system serving live production traffic.

Experience

Caretalk Health

Jan 2026 – Present

AI Engineer

Bethesda, MD

- Serve a fine-tuned domain LLM in production via vLLM on AWS EC2 for automated clinical report generation
- Orchestrate a LangGraph agentic workflow chaining multi-step validation and MCP-based tool routing
- Built an LLM evaluation harness with Ragas and LangSmith, cutting hallucination rate by 20%
- Built AWS Lambda ingestion pipelines integrating 10+ healthcare APIs, validated via Postman
- Integrated Zoho CRM data through REST APIs, removing manual entry as a source of error
- Enforced HIPAA-compliant access controls and audit logging across all AI infrastructure
- Partnered with clinical stakeholders to define evaluation criteria for report quality

Vosyn Inc.

Jun 2025 – Dec 2025

Applied AI Engineer

Chicago, USA

- Designed a multimodal RAG system on Gemini 2.5 and Vertex AI for cross-modal document retrieval
- Built an embedding evaluation pipeline across 10M+ document vectors, lifting retrieval accuracy 40%
- Instrumented RAG evaluation with Ragas and LangSmith, cutting hallucination rate by 30%
- Iterated on prompt and chain design in LangChain, improving response quality by 35%
- Led a 4-engineer team restructuring RAG ingestion architecture, raising throughput by 60%
- Ran A/B experiments across 5+ model and retrieval configurations, tracked in MLflow
- Containerized model-serving with Docker/Kubernetes, cutting deployment time by 92%

SEES Group

Jan 2024 – May 2025

Generative AI Engineer

Tennessee, USA

- Built an LLM-powered natural language interface for querying unified PM, EMR, RCM, and financial data
- Designed healthcare-tuned prompt templates, improving LLM query accuracy by 25%
- Orchestrated a CrewAI/AutoGen multi-agent system automating cross-system data reconciliation
- Architected the cloud data lake unifying 15+ source systems for LLM retrieval
- Maintained ETL pipelines feeding the GenAI application with 99.9% data integrity
- Established MDM and schema normalization so the LLM could reason across healthcare data
- Enforced HIPAA and SOC2 controls, including metadata tracking, across AI infrastructure

Tata Consultancy Services

Nov 2021 – Dec 2023

Machine Learning Engineer

Hyderabad, Telangana

- Deployed XGBoost and LSTM models to production on Azure ML Studio with CI/CD retraining
- Added a SHAP/LIME explainability layer, later extended into LLM evaluation work
- Optimized inference pipelines and monitoring, catching drift across 10+ deployed models
- Delivered ML-driven solutions supporting \$2M+ in Fortune 500 client engagements
- Translated churn and demand-forecasting problems into modeling requirements
- Established CI/CD and monitoring patterns later extended to generative AI deployments
- Supported model governance reviews, documenting assumptions for stakeholder sign-off

Larsen & Toubro Infotech

Aug 2020 – Nov 2021

Data Engineer

Hyderabad, India

- Built ETL pipelines in Python, PySpark, and Airflow processing 500GB+ daily from 15+ sources
- Designed dimensional data models and feature stores feeding 20+ ML applications
- Implemented data quality validation safeguarding 99.9% integrity across pipelines
- Tuned Azure Data Factory workflows for parallel processing, improving throughput 45%
- Built Kafka streaming pipelines for event-driven use cases at sub-second latency
- Managed Azure Blob Storage and SQL Database with geo-replication for high availability
- Redesigned warehouse schemas with star modeling, tripling analytics query performance

Skills

Languages: Python, SQL
LLMs & Generative AI: GPT-4, Claude, Gemini, LLaMA, Hugging Face Transformers, Fine-tuning (LoRA/QLoRA), Prompt Engineering
Agentic AI & Orchestration: LangGraph, CrewAI, AutoGen, LangChain, MCP (Model Context Protocol), Tool/Function Calling
RAG & Retrieval: Retrieval-Augmented Generation, Vector Search, Pinecone, Weaviate, FAISS, Embeddings, Semantic Search
Model Serving & Inference: vLLM, Docker, Kubernetes, Model Registry
Evaluation & Observability: Ragas, LangSmith, A/B Testing, Hallucination Analysis
Cloud AI Platforms: AWS (Bedrock, EC2, Lambda), GCP Vertex AI, Azure OpenAI & ML Studio
ML Foundations: PyTorch, TensorFlow, XGBoost
APIs: REST APIs, OpenAI API, Anthropic API
Compliance: HIPAA, SOC2

Education

University of North Texas <i>Master of Science in Data Science, GPA: 4.0</i> • Relevant Coursework: Large Language Models, NLP, Deep Learning, Cloud Computing	Jan 2024 – Dec 2025 <i>Denton, Texas</i>
Jawaharlal Nehru Technological University Hyderabad <i>Bachelor of Technology, GPA: 3.6</i> • Relevant Coursework: Machine Learning, Algorithms	Jun 2017 – Aug 2021 <i>Hyderabad, India</i>

Projects

Multimodal Clinical AI Assistant <i>PyTorch, BERT, ResNet, FastAPI, Cross-Attention</i> • Built a multimodal AI system fusing BioBERT clinical-text embeddings with ResNet-101 image features via cross-attention, reaching 91% diagnostic accuracy • Designed the inference service in FastAPI, exposing predictions and attention-based explanations through a production-style API • Built the preprocessing pipeline normalizing DICOM imaging and HL7 notes from 5 hospital systems into model-ready inputs, validated with clinicians across 5 specialties	September 2023
LLM-Augmented Market Forecasting Pipeline <i>Python, LSTM, BERT Sentiment, MLflow</i> • Combined an attention-based LSTM forecaster with BERT-derived news-sentiment signals to predict S&P 500 direction, reaching 87% directional accuracy • Engineered 20+ technical indicator features alongside NLP-extracted sentiment scores to strengthen model inputs • Served the model through a REST API with MLflow experiment tracking, supporting 10K+ daily predictions	March 2024

Leadership / Student Organizations

UNT Data Science Organization (2024 – Present): Ran hands-on sessions on LLM application development and agentic workflows for 30+ members; mentored a 5-student team building an applied AI project for a nonprofit partner
Association for Information Systems - AIS (2024 – Present): Led the technical build integrating LLM reasoning into the team’s AI-driven IT support submission, placing 2nd among 50+ teams

Awards

2nd Place - AIS AI-Driven IT Support Challenge <i>Recognized for building the team’s AI reasoning and automation components</i>	2025 <i>AIS Student Chapter</i>
Excellence in Innovation Award <i>Outstanding contribution to a technology innovation project</i>	2021 <i>T-Hub & Hiroshima Prefecture Japan</i>
UNT Welcome Center Appreciation Award <i>Data visualization excellence recognition</i>	2025 <i>UNT Welcome Center</i>