TEJUS PATURU

Boulder, CO · +1 (720) 209-1106 · tejusp08.us@gmail.com
LinkedIn · GitHub · Portfolio

EDUCATION

University of Colorado Boulder — Boulder, CO
Master of Science in Computer Science (Specialization in AI) | *GPA: 4.0* — Aug 2024 May 2026

National Institute of Technology Puducherry — Puducherry, India
Bachelor of Technology in Computer Science Engineering | *GPA: 8.5/10* — Dec 2020 May 2024

TECHNICAL SKILLS

- **Languages:** Python, C++, SQL, JavaScript, Bash.
- **AI & GenAI:** LangGraph, LangChain, PyTorch, TensorFlow, OpenAI API, Hugging Face, FAISS, RAG, Ollama, Llama-3 (SLM), Prompt Engineering, OpenCV.
- **MLOps & Cloud:** Docker, Kubernetes, AWS (SageMaker, Lambda, S3), MLflow, Prometheus, Grafana, GitHub Actions, CI/CD.
- **Data & Web:** FastAPI, PostgreSQL, MongoDB, Pandas, NumPy, Apache Airflow, Streamlit.

WORK EXPERIENCE

The Sprouting Company — Remote, USA
Machine Learning Intern — May 2025 Aug 2025

- **Executed** day-zero benchmarking of **GPT-5 vs. GPT-4o**, authoring a strategic technical report on reasoning capabilities and token economics that defined the companys roadmap for **Agentic workflows**.
- **Designed** a cost-optimized hybrid inference architecture, validating that routing complex reasoning to **GPT-5** while offloading routine tasks to **Small Language Models (SLMs)** on AWS would reduce token costs by **40%**.
- **Engineered** a production-grade computer vision pipeline to detect fungal contamination, fine-tuning **EfficientNetV2** on noisy real-world data to achieve **80% precision** while minimizing false positive rates.
- **Developed** automated **LLM evaluation pipelines** using Python and **LangSmith** to monitor model hallucination rates and drift, ensuring reliable performance for biological data analysis.

Treosoft IT Solutions — Bengaluru, India
Machine Learning Engineer — Jan 2024 June 2024

- **Architected** a feature extraction pipeline for high-dimensional sequence data, utilizing **PCA** to compress 400-D vectors, enabling efficient multi-label classification without information loss.
- **Optimized** inference throughput by **65%**, replacing complex ensemble models with a streamlined **KNN-based classifier** that achieved state-of-the-art accuracy on independent test sets while reducing computational overhead.
- **Developed** a Retrieval-Augmented Generation (RAG) system using **LangChain** and **FAISS** to query internal system datasets, improving technical information retrieval speed by **50%** for engineering teams.

Ford Motor Private Limited — Chennai, India
Software Intern — May 2023 July 2023

- **Architected** automated data extraction pipelines for ALM systems using **Python** and **SQL**, slashing processing time by **99% (from 2 hours to <1 minute)** and enabling real-time project tracking for 20+ cross-functional teams.
- **Modernized** legacy reporting workflows by migrating Excel logic to **Pandas**, eliminating manual data entry bottlenecks and improving dashboard update frequency by **97%**.
- **Built** an internal document retrieval bot using **OpenAI Embeddings** and vector search, allowing teams to instantly query internal documentation and significantly reducing onboarding time.
- **Designed** interactive data dashboards connected to automated pipelines, providing leadership with live visibility into key operational KPIs.

Treosoft IT Solutions — Bengaluru, India
Data Scientist — May 2022 May 2023

- **Deployed** a real-time recommendation engine processing **100K+ transactions**, utilizing **Market Basket Analysis** to identify purchasing patterns and drive a **15% lift in net sales**.
- **Engineered** a predictive upselling pipeline using **XGBoost**, integrating ML-driven insights directly into the sales platform via **REST APIs** to optimize conversion rates through personalized product suggestions.
- **Developed** an NLP text classification system using **NLTK** and **TF-IDF** to process unstructured customer feedback, automating sentiment analysis and identifying critical product improvement areas.
- **Built** scalable data preprocessing workflows to handle diverse data inputs, ensuring high-quality feature engineering for downstream machine learning models.

PROJECTS

Stateful Agentic Personal Trainer | *Tech: LangGraph, FastAPI, Docker, K8s, Ollama, PostgreSQL*

- **Architected** a multi-agent orchestration system using **LangGraph**, implementing a "safety critique loop" where autonomous Trainer and Physiotherapist agents collaborate to generate injury-safe workout plans with automated revision cycles.
- **Engineered** a hybrid RAG architecture supporting dual LLM backends (**Ollama/Mistral + OpenAI**), enabling natural language parsing of workout logs while retrieving historical user context for personalized, adaptive coaching.
- **Refactored** the MVP into a production-grade microservice with **FastAPI** and **PostgreSQL**, implementing **JWT authentication** and containerizing the system with **Docker/Kubernetes** for scalable cloud deployment.

Multi-Agent RAG System | *Tech: LangChain, FAISS, GPT-4, Cohere Rerank*

- **Developed** a collaborative Multi-Agent RAG system where specialized agents (Researcher, Critic, Writer) coordinate via a shared scratchpad to decompose complex queries, improving answer accuracy on technical documents by **35%**.
- **Implemented** a hybrid retrieval strategy combining semantic search (**FAISS**) with keyword search (BM25) and **Cohere Rerank**, ensuring precise context retrieval even for obscure domain-specific terminology.
- **Designed** an intelligent routing layer that dynamically selects between cached responses, vector search, or web search based on query ambiguity, optimizing system latency.

LLM-Ops Evaluation & Observability Pipeline | *Tech: MLflow, Prometheus, Grafana, LangSmith*

- **Architected** an end-to-end LLM Ops pipeline to monitor production model performance, integrating **LangSmith** to trace agent execution chains and pinpoint latency bottlenecks at the granular span level.
- **Developed** a "Golden Dataset" evaluation framework using **MLflow**, automating the daily scoring of model outputs against ground-truth answers to detect semantic drift and regression.
- **Deployed** real-time monitoring dashboards in **Grafana** (backed by **Prometheus**) to visualize token usage, cost-per-query, and error rates.

PUBLICATIONS

- **Resume Classification Using ML Techniques** IEEE, 2023
- **Potato Disease Classification Using Diverse Feature Extraction Methods and Machine Learning Models** Springer, 2023
- **Text Feedback Classification using Machine Learning Techniques** IEEE, 2023
- **Investigating Feature Extraction for Effective Lung Disease Detection Using Chest X-Ray Images** Springer, 2024