Gowtham Venkata Sai Ram Maddala

Stony Brook, NY | (934) 246-9689 | gowthamvenkata.maddala@stonybrook.edu | Github | LinkedIn | Portfolio

EDUCATION

Stony Brook University, Stony Brook — New York
Master of Science in Data Science, ***GPA: 3.84/4*** — Aug 2024 - Dec 2025

International Institute of Information Technology, Bangalore — India
Advanced Certificate Programme in Data Science with Specialization in NLP, ***GPA: 3.8/4*** — Apr 2023 - Dec 2023

KL University, Hyderabad — India
Bachelor of Technology in Computer Science with specialization in Data Science, ***GPA: 9.02/10*** — Sept 2017 - May 2021

TECHNICAL SKILLS

Programming Languages: Python, C, C++, Java, Go, R Programming, MATLAB, HTML, CSS, JavaScript (JS), React, SQL

Tools and Platforms: Git, Docker, Kubeflow, Flask, FastAPI, Google Cloud Platform (GCP), Microsoft Azure, REST, VS Code

Frameworks and Libraries: TensorFlow, PyTorch, Keras, NumPy, Pandas, Scikit-learn, Matplotlib, Seaborn, XGBoost, CUDA, Spacy

Machine Learning: Predictive Modeling, GBDTs, Random Forests, Clustering, Decision Trees, Time Series Forecasting

NLP and GenAI: Transformers, Transfer Learning, Natural Language Generation (NLG), Training and Fine-tuning, LLMs

EXPERIENCE

Soroco — Bangalore, India
Software Engineer (Machine Learning) — Feb 2024 - Jul 2024

- Developed **4 Flask APIs** to generate flowgraphs of user activities based on screens used for the Workgraph product and trained YOLOv9 on annotated screenshots to detect interacted and non-interacted fields, enabling accurate data collection.
- Utilized the **Guidance library** to format **LLM** outputs, reducing post-processing needs, and optimized **Mistral-7B and LLaMA2** models using **Python**, cutting average inference time from **5 seconds to 0.8 seconds** without compromising quality.
- Created test cases using **PyTest** to validate API functionality and JSON outputs with Postman, achieving a **100%** test pass rate.
- Drove company revenue by **15%** through successful onboarding of **3 Fortune 500 clients** within a short timeframe of **4 months**.

Awone.ai — Client: Carelon Global Solutions — Hyderabad, India
Data Scientist — Apr 2023 - Feb 2024

- Developed a **RAG** pipeline in Python using **BioMedGPT-7B** to retrieve and generate accurate responses from a vector database of healthcare policies and patient records, ensuring relevance based on patient health conditions and policy applicability.
- Reduced model size and improved inference speed by quantizing BioMedGPT-7B from FP16 to INT4 using **QLoRA** on GPU, decreasing model size from **13.5GB to 4GB** and cutting inference time from over **60 seconds to 8 seconds**.
- Constructed specialized datasets for **DPO** training using advanced prompt engineering techniques with **Llama2**, tailored for healthcare policy queries and patient-specific scenarios.
- Achieved a Rouge score of **0.82** by fine-tuning BioMedGPT-7B using DPO, enhancing accuracy in patient-specific policy responses.

Ivy Comptech — Hyderabad, India
Software Engineer — Aug 2021 - Feb 2022

- Handled a high-volume transactional database with over **3 million** records as a key contributor to the wallet and payments team.
- Revamped **30** complex **SQL** queries, reducing execution time by **50%** and significantly boosting overall data pipeline performance.

Telescope (Voxlogic.inc) — Acquired by Meta — Sunnyvale, USA - Remote
Software Development Intern (AI Platform Team) — Jul 2020 - Dec 2020

- Architected a conversational search solution using Hugging Face's TAPAS model, enabling numerical question answering on tabular data with **97.45%** accuracy and integrated it into Slack with **TensorFlow** quantization for real-time responses.
- Played a key role in the development of Telescope, acquired by Meta for **$2.4 million** in 2021, by enabling conversational search capabilities and ensuring swift and accurate user interactions.

PROJECTS

DPO-Enhanced Qwen for Python Programming

- Fine-tuned the **Qwen2.5-3B** model using Direct Preference Optimization (DPO), achieving a **35.56%** improvement in Python programming and debugging tasks, supported by a curated dataset of **12,000** labeled examples.
- Enhanced model training through hyperparameter tuning and quantization techniques, and engineered preference-based optimization strategies to align large language model outputs with user-defined priorities, boosting AI-assisted programming workflows.

Delivery Time Estimation Using Neural Networks

- Analyzed key drivers of delivery time, such as total outstanding orders, hour of the day, and market dynamics, using Random Forest feature importance analysis, achieving an **MSE of 3.2** and **RMSE of 1.79** with the **Random Forest regressor**.
- Maximized predictive accuracy by fine-tuning **Neural Networks**, reducing error metrics to an MSE of **0.12** and RMSE of **0.34**, resulting in a more reliable delivery time forecasting system.