Gowtham Venkata Sai Ram Maddala

+1(934) 246-9689 gowthamvenkata.maddala@stonybrook.edu LinkedIn GitHub

EDUCATION

Stony Brook University, Stony Brook, New York **Aug 2024 - Dec 2025**
*Master of Science in **Data Science**, Courses: Natural language Processing, Data Analysis, Probability*

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

Koneru Lakshmaiah Education Foundation - KL University, Hyderabad, India **Sept 2017 - May 2021**
Bachelor of Technology in Computer Science with specialization in Data Science *CGPA*: ***9.02/10***

TECHNICAL SKILLS

Programming Languages: C, C++, Java, **Python**, JavaScript, MATLAB, **MySQL**, MongoDB, R Programming
Frameworks and Libraries: **TensorFlow**, Keras, NumPy, Pandas, Scikit-learn, Matplotlib, Seaborn, **PyTorch**
Machine Learning: Predictive Modeling, GBDTs, Random Forests, Clustering, Time Series Forecasting
Natural Language Processing: Word Embeddings, Transformers, Text Processing, **LLMs - Mistral, LLaMA, GPT**

EXPERIENCE

Soroco **Feb 2024 - Jul 2024**
Software Engineer (Machine Learning) — **Azure DevOps, Python, Docker, Pytest, SQL, APIs** *Bangalore, India*

- Developed 4 Flask APIs to generate flowgraphs of user activities based on screens used for the Workgraph product.
- Leveraged **Guidance** library to format outputs from LLMs, significantly minimizing the need for post-processing.
- Streamlined performance evaluation between **Mistral-7B and LLaMA2** models, reducing average inference time from **5 seconds to 0.8 seconds** without compromising output quality.
- Designed the test cases using **PyTest**, ensuring smooth functionality of **APIs**, and validated JSON outputs using Postman.
- Boosted company **revenue by 15%** through successful onboarding of **3 Fortune 500 clients** in a timeframe of **4 months**.

Awone.ai — Client: Carelon Global Solutions **Apr 2023 - Feb 2024**
Data Scientist — **Python, LLMs, TensorFlow, Reinforcement Learning, Quantization** *Hyderabad, India*

- Developed a **RAG pipeline** on Kubeflow using BioMedGPT-7B for efficient response generation from vector databases.
- Optimized model size and improved inference speed by quantizing from FP16 to INT4 using **QLoRA on GPU**, reducing the model size from 13.5GB to 4GB and decreasing inference time from over **60 seconds to 8 seconds**.
- Constructed specialized datasets for the DPO trainer, leveraging advanced **prompt engineering** techniques with **Llama2**.
- Achieved a Rouge score of 0.82 by fine-tuning the BioMedGPT model using **Direct Preference Optimization (DPO)**.

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

- Managed a high-volume transactional database with over 3 million records as part of the wallet/payments team.
- Optimized 30 complex SQL queries, reducing execution time by **30%** and significantly boosting data pipeline performance.

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

- Architected a conversational search solution using the **TAPAS model** from Hugging Face, enabling numerical question answering on tabular data extracted through a custom web scraping pipeline, achieving 97.45% accuracy.
- Integrated the model into Slack, allowing users to input tabular data and receive real-time responses, and quantized the model using **TensorFlow** to optimize for speed, ensuring swift responses during conversational searches.
- Contributed to the development of Telescope, which was later acquired by Meta for **$2.4 million** in 2021.

PROJECTS

Direct Preference-Optimized Language Model for Advanced Reasoning and Debugging — GitHub

- Developed and fine-tuned the Qwen 2.5 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.
- Optimized model training through hyperparameter tuning and quantization techniques to improve performance.
- Engineered preference-based optimization strategies to align large language model outputs with user-defined priorities, enhancing AI-assisted programming workflows.