Tanmay Sutar

Atlanta, GA, USA

+1 (404) 423-9768 | tanmay21@gatech.edu | https://www.linkedin.com/in/tanmaysutar2109

https://github.com/Tanny2109

Summary

Recent graduate with strong software engineering experience, having developed scalable systems and production-ready applications using Python, React, and JavaScript. Experienced in collaborating across multidisciplinary teams to translate complex challenges into efficient, customer-focused solutions. Excited to contribute technical expertise and a proactive problem-solving mindset to drive impactful innovation in financial operations.

EDUCATION

Georgia Institute of Technology **2023 - 2025**
Master's, Computer Science - ML Specialisation *Atlanta, GA, USA*

IIT Guwahati **2019 - 2023**
Bachelor of Technology, BME *India*

EXPERIENCE

Social Wellbeing Lab - Georgia Tech **Jan 2025 - May 2025**
ML Engineer *Atlanta, USA*

- Collaborated with researchers, designers, and product stakeholders across GeorgiaTech, GSU, and UC Berkeley to translate exploratory LLM research into scalable software features aligning with product ownership values.
- Designed and architected a multi-GPU training pipeline for LLaMA-3.1-8B-Instruct using structured algorithms and data structures, achieving a 0.98 ATS score based on judge evaluations.
- Developed a distributed inference system with dynamic GPU load balancing across Nvidia H200 GPUs using async batch processing, processing over 10K conversations and achieving a 4x throughput improvement.
- Implemented production-ready AutoGen integration with multi-threaded conversation simulation, incorporating fault-tolerant error handling and GPU resource optimization for 24/7 deployment.

Bio-MIBLab Georgia Tech **Aug 2024 - Dec 2024**
ML Researcher - 3D Computer Vision *Atlanta, USA*

- Collaborated with medical researchers and cross-functional teams to convert innovative 3D computer vision findings into robust Python modules that support clinical and product design objectives.
- Engineered a 3D brain tumor segmentation pipeline using 3D U-Net and VMNet architectures, achieving a Dice coefficient of 0.83 on the BraTS dataset with custom data augmentation strategies.
- Optimized compute efficiency by implementing distributed PyTorch training with Luigi workflow orchestration on CUDA 12 infrastructure, improving performance by 70% and reducing training time from 48 to 14 hours.

Providence Global Centre **May 2022 - Jul 2022**
SWE Intern *Hyderabad, India*

- Led the creation of an in-house React Component Library and utilized it in a client's product using micro-front end architecture resulting in $0.1M savings.
- Implemented React virtualization and lazy-loading to serve data efficiently on the front end, enhancing user experience and reducing load times

PROJECTS

Food-weight estimation iOS app | https://weight-estimation.streamlit.app/ **Jan 2025 - May 2025**
Applied Research Project *Atlanta, USA*

- Built iOS app that can estimate food weight with just one picture.
- Created own custom dataset using Apple's Object Capture API to generate 3D objects.
- Fine-tuned a custom ResNet model that maps multiple 2D images to 3D quantity.

Skills

- **Technical Skills**: Python, React, JavaScript, HTML/CSS, PyTorch, AWS, GCP, SQL, Generative AI, LLMs, Computer Vision
- **Core Cs Fundamentals**: Algorithms, Data Structures
- **Product & Leadership**: Product Ownership