Siddharth Sriraman

Email: sidsr@gatech.edu

Santa Clara, CA / LinkedIn / Website / GitHub

Looking for full-time SDE/MLE roles starting May 2025

Education

- **Georgia Institute of Technology** — Atlanta, GA
  Master of Science in Computer Science; GPA: 4.0/4.0 — Aug 2023 - May 2025
  Coursework: Graduate Algorithms, Machine Learning, Computer Vision, Data Analytics; Graduate Researcher under Prof. Munmun De Choudhury
- **Anna University** — Chennai, India
  Bachelor of Engineering in Computer Science; GPA: 3.95/4; Department Rank 5/221 — Aug 2018 - May 2022

Work Experience

- **NVIDIA** — Santa Clara, CA
  Software Engineer Intern — May 2024 - Present
  - Developing tools to streamline org-wide functional safety compliance in autonomous vehicles as part of NVIDIA DRIVE.
  - Designed and deployed LLM-powered productivity tooling to assist engineers in analysing safety and security risks in stakeholder requirements and architecture.
  - **Technologies**: Docker, NVIDIA NIM, LLM, Elasticsearch, Kibana
- **Amazon** — Chennai, India
  Software Development Engineer — July 2022 - June 2023
  - Owned the end-to-end delivery of a VP goal with Amazon Science to build an AI-assisted manual data labeling pipeline, which is now the single source of labeled data for all ML-based content moderation across Alexa.
  - Launched 5+ NLP models that block offensive content in Alexa interactions worldwide in real-time across 9 languages with ∼95% precision, enhancing the safety of 10M+ daily interactions.
  - Implemented systems for scientists to analyse customer impact of NLP models on 1M+ interactions worldwide in a few clicks, with serverless orchestration that integrates with the labeling pipeline to automate metrics measurement.
  - Designed the API schema for sensitive content detection in long-form chat conversations for the LLMs powering Alexa.
  - **Technologies**: Python, Java, REST APIs, PyTorch, AWS S3, Lambda, EC2, SageMaker, DynamoDB
- **Amazon** — Chennai, India
  Software Development Engineer Intern — Feb 2022 - June 2022
  - Reduced the inference latency of production NLP models by up to 40%, and hosting costs by 8.8x compared to GPUs, by designing a software layer to integrate AWS Inferentia chips into the Alexa Sensitive Content team's ML infrastructure.
  - Sped up model development cycles by 25% (∼1 week) by developing an automated ML benchmarking tool for researchers to seamlessly analyse inference latency of models at scale with different AWS SageMaker instances from a local CLI.
  - **Technologies**: Docker, CI/CD, IaC, AWS Step Functions, SageMaker, Python
- **Indian Institute of Technology Madras** — Chennai, India
  Machine Learning Research Intern — May 2021 - Oct 2021
  - Trained ML models to predict the location of loudspeaker sources from sound pressure data in Prof. K Srinivasan's lab.
  - Processed and visualised 3+ GB of pressure signals from microphone arrays in a semi-anechoic chamber.
  - Built a complex-valued regression model with inference time 5 orders of magnitude lesser than conventional methods, while matching their localisation performance. Co-authored a paper in the Journal of the Acoustical Society of America.
  - **Technologies**: Python, TensorFlow, Pandas

Skills

- **Programming Languages**: Java, Python, JavaScript, TypeScript, C++, SQL, HTML, CSS, Linux scripting
- **Tools/Technologies**: Git, Docker, Jupyter, Pandas, Kubernetes, CI/CD, MongoDB, React, REST

Projects

- **Optimising AutoML Pipelines for MLOps** (paper)
  - Researching transfer learning to improve the runtime of automatic data cleaning systems for ML by orders of magnitude, guided by Prof. Kexin Rong, as part of a research seminar course on human-in-the-loop data analytics at Georgia Tech.
- **Q-Snake, Interactive Reinforcement Learning** (website/code)
  - Developed a web app to visualise how RL agents learn to play the game Snake with tabular Q-learning coded from scratch.
  - Utilised by PolyHx, the CS society at Université de Montréal, in 2021 to teach beginners about core RL concepts.