Siddharth Sriraman

Email : ssriraman8@gatech.edu
LinkedIn / Website / GitHub

Education

- **Georgia Institute of Technology** — Atlanta, GA
  Master of Science in Computer Science; ML Specialisation — *Aug 2023*
- **Anna University (SSN College of Engineering)** — Chennai, India
  Bachelor of Engineering in Computer Science; GPA: 9.54/10 - Rank #5/221 — *Aug 2018 - May 2022*

Skills

- **Languages**: Java, Python, JavaScript, C/C++, SQL, Unix scripting
- **Tools (production experience)**: PyTorch, transformers, Docker, MySQL, React, AWS Lambda, DynamoDB, S3

Experience

- **Amazon** — Chennai, India
  Software Development Engineer (Alexa Sensitive Content Intelligence, Devices Privacy) — *July 2022 - July 2023*
  - **Data Labeling Pipeline**: Led a high-visibility project with Amazon Science and Alexa Content Policy to improve the precision/objectivity of the manual data annotation systems that generate ML training data, through computational social science. Serves as the single source of ground truth data for all Alexa content moderation use cases across nine languages.
  - **Security and Privacy Compliance**: Delivered high-level and low-level engineering design for the data labeling pipeline, covering threat models, encryption setup, data retention and auditing to ensure private data is handled safely by working with the security certification team.
  - **ML Model Launches**: Built systems that serve NLP models that detect and block offensive Alexa responses in millions of customer interactions in all supported languages, making Alexa trustworthy and safe.
  - **Model Performance Monitoring**: Built a serverless orchestration layer for workflows that automate the calculation of key business metrics of NLP models used in production, allowing scientists to evaluate the customer impact of model launches in across millions of Alexa interactions worldwide in a few clicks.
- **Amazon** — Chennai, India
  Software Development Engineer Intern (Alexa Sensitive Content Intelligence, Devices Privacy) — *Feb 2022 - June 2022*
  - **Inference Latency Optimisation**: Integrated hardware accelerators into our ML deployment infrastructure which used ahead-of-time compilation to reduce the inference latency of production NLP models by 50%.
  - **Model Performance Benchmarking**: Developed an automated ML benchmarking tool to allow researchers and engineers to seamlessly load-test models at scale with different hardware and model configurations and analyse latency reports.
- **Indian Institute of Technology Madras** — Chennai, India
  ML Research Intern (Dept. of Mechanical Engineering) — *May 2021 - October 2021*
  - Researched the performance of ML in acoustic source localisation (predicting the location of loudspeaker sources with sound pressure data from microphone arrays) in an interdisciplinary research group under Prof. K Srinivasan.
  - Trained a complex-valued regression model by collecting data from semi-anechoic chambers and studied its performance over various distances. Co-authored a paper at the Journal of the Acoustical Society of America.
- **Fidelity Investments** — Chennai, India
  Fullstack Engineer Intern (Fund and Investment Operations) — *June 2021 - July 2021*
  - Built an administrative monitoring tool for a batch-processing system that manages ingestion/distribution of critical fund data to generate financial statements.
  - Automated 10+ previously manual and intricate database tasks and built a central UI to configure distributed job scheduling, reducing the team's operations load to minutes.

Research Publications

- S.K. Chaitanya, **Siddharth Sriraman**, Srinath Srinivasan, K. Srinivasan; "Machine learning aided near-field acoustic holography based on equivalent source method". Journal of the Acoustical Society of America (2023) (link)
- **Siddharth Sriraman**, Sneha Kannan, Sonali Ravishankar, B. Bharathi; "An On-device Federated Learning System for SMS Spam Classification". 8th IEEE MIT Undergraduate Research Technology Conference (2022) (link)

Academic Projects

- **English Assessment Platform**: Built a full-stack app for customised listening and reading tests in my university's humanities lab, with 50+ computers connected via REST APIs to a Node.js backend. Designed systems to automate overall/student-wise test report generation, reducing hours of manual faculty work to minutes. (Feb 2020)
- **Q-Snake, Interactive Reinforcement Learning**: Developed a web app to visualise how RL agents learn to play the game Snake through tabular Q-learning coded from scratch. Used by PolyHx, the CS society at Université de Montréal, to teach beginners about core RL concepts. (Oct 2020)