Sai Surya Duvvuri

UT Austin CS department
2317 Speedway, Austin, TX 78712

subramanyamdvss@gmail.com
Google Scholar

Education

University of Texas at Austin *August 2021 - present*
Advisor: Prof. Inderjit Dhillon
PhD Student in Computer Science, GPA: 3.94
Expected Graduation: Dec 2026

Indian Institute of Technology, Kharagpur *July 2015 - May 2019*
Bachelor of Technology
Department of Computer Science and Engineering, GPA: 8.98

Publications

- **The Art of Scaling Reinforcement Learning Compute for LLMs**
  Devvrit Khatri, Lovish Madaan, Rishabh Tiwari, Rachit Bansal, **Sai Surya Duvvuri**, Manzil Zaheer, Inderjit S. Dhillon, David Brandfonbrener, Rishabh Agarwal
  International Conference on Learning Representations (ICLR), 2026 (Oral)
- **Let's (not) just put things in Context: Test-Time Training for Long-Context LLMs**
  Rachit Bansal, Aston Zhang, Rishabh Tiwari, Lovish Madaan, **Sai Surya Duvvuri**, Devvrit Khatri, David Brandfonbrener, David Alvarez-Melis, Prajjwal Bhargava, Mihir Sanjay Kale, Samy Jelassi
  International Conference on Learning Representations (ICLR), 2026
- **LASER: Attention with Exponential Transformation**
  Sai Surya Duvvuri, Inderjit S. Dhillon
  International Conference on Machine Learning (ICML), 2025
- **LoRA Done RITE: Robust Invariant Transformation Equilibration for LoRA Optimization**
  Jui-Nan Yen, Si Si, Zhao Meng, Felix Yu, **Sai Surya Duvvuri**, Inderjit S. Dhillon, Cho-Jui Hsieh, Sanjiv Kumar
  International Conference on Learning Representations (ICLR), 2025 (Oral)
- **Combining Axes Preconditioners through Kronecker Approximation for Deep Learning**
  Sai Surya Duvvuri, Fnu Devvrit, Rohan Anil, Cho-Jui Hsieh, Inderjit S. Dhillon
  International Conference on Learning Representations (ICLR), 2024
- **A Computationally Efficient Sparsified Online Newton Method**
  Fnu Devvrit*, **Sai Surya Duvvuri***[1], Rohan Anil, Vineet Gupta, Cho-Jui Hsieh, Inderjit S. Dhillon
  Neural Information Processing Systems (NeurIPS), 2023
- **Block Low-Rank Preconditioner with Shared Basis for Stochastic Optimization**
  Jui-Nan Yen, **Sai Surya Duvvuri**, Inderjit S. Dhillon, Cho-Jui Hsieh
  Neural Information Processing Systems (NeurIPS), 2023
- **Unsupervised Neural Text Simplification**
  Sai Surya Duvvuri, Abhijit Mishra, Anirban Laha, Parag Jain, Karthik Sankaranarayanan
  in main conference of **Association for Computational Linguistics (ACL) 2019**
- **iBox: Internet in a Box**
  Sachin Ashok, **Sai Surya Duvvuri**, Nagarajan Natarajan, Venkata N. Padmanabhan, Sundararajan Sellamanickam, Johannes Gehrke
  in **ACM Hotnets 2020 workshop**

Preprints

- **LUCID: Attention with Preconditioned Representations**
  Sai Surya Duvvuri*, Nirmal Patel*, Nilesh Gupta, Inderjit Dhillon
  Under review in ICML 2026
- **Interleaved Head Attention**
  Sai Surya Duvvuri*, Chanakya Ekbote*, Rachit Bansal, Rishabh Tiwari, Devvrit Khatri, David Brandfonbrener, Paul Liang, Inderjit Dhillon, Manzil Zaheer
  Under review in ICML 2026

[1] *equal contribution

- **Adaptive Regularization through Coupled Kronecker Factoring**
  Sai Surya Duvvuri, Cho-Jui Hsieh, Inderjit S. Dhillon
  Under review in ICML 2026

- **Fast and Simplex: 2-Simplicial Attention in Triton**
  Aurko Roy, Timothy Chou, **Sai Surya Duvvuri**, Sijia Chen, Jiecao Yu, Xiaodong Wang, Manzil Zaheer, Rohan Anil

Work Experience

Google
Student Researcher

- **Diffusion and Recursion** ***Jan 2026 - Present***
  Host: Inderjit S. Dhillon
  - Working on applications of recursive transformers for Diffusion and reasoning.

Meta, Menlo Park, CA
Visiting Researcher

- **Novel Attention Mechanisms** ***May 2025 - January 2026***
  Host: Manzil Zaheer
  - Working on novel attention mechanisms aiding in thinking LLMs.

Google, Mountain View, CA
Student Researcher

- **Novel Attention Mechanisms for Decoder LLMs** *May 2024 - May 2025*
  Host: Inderjit S. Dhillon
  - Developed LASER attention, which conducts attention in exponential transformation of values.
  - Working towards using techniques from state space models in softmax-attention.

Google DeepMind, Mountain View, CA
Summer and Fall 2023 Student Researcher

- **Second-order optimization for Large Language Models** *May - August and Oct - January 2023*
  Host: Rohan Anil
  - Developed CASPR optimizer and showed performance improvements in Large Language models over Adam and Shampoo.
  - CASPR also demonstrates a tighter convergence bound over Shampoo.

Microsoft Research India, Bengaluru
Research Fellow in machine learning and optimization group

- **Robust Unsupervised Learning Algorithms** *July 2020 - present*
  Advisors: Dr. Ankit Garg, Dr. Neeraj Kayal
  - Worked on algorithms for unsupervised learning problems using algebraic techniques.
  - The focus of this project was on tensor decomposition, subspace clustering and general gaussian mixtures.

- **Data-Driven Network-Simulation** *July 2019 - July 2020*
  Advisors: Dr. Nagarajan Natarajan, Dr. Venkat Padmanabhan, Dr. Sundararajan Sellamanickam
  - Automatic simulation of network paths using machine learning algorithms to make congestion control protocol testing more reliable.
  - Worked on ML models trained on real-world network traces to simulate delays and reordering.

IBM Research India, Bengaluru *May 2018 - July 2018*
Research Intern in natural language generation group
Advisors: Dr. Abhijit Mishra & Dr. Karthik Sankaranarayanan

- **Unsupervised Neural Text Simplification**
  - Proposed an unsupervised training method for the automatic text simplification task.
  - Created unlabeled dataset from en-wikipedia dump and trained sequence-to-sequence models in conjunction with GAN-based loss functions to simplify input sentences.

Selected Awards and Honors

- Best B.Tech Thesis Award 2019, CSE Department IIT Kharagpur *May 2019*
- Secured an AIR of **272** in JEE-Advanced among **0.17 million** candidates *May 2015*
- Selected for JBNSTS scholarship provided by West Bengal gvt. *Jan 2016*

Relevant Courses

Graduate:

Theoretical Statistics,Numerical Analysis: Linear Algebra, Optimization II, Distributed Systems, Randomized Algorithms, Datacenters, Supervised Teaching in Computer Science

Undergraduate:

Algorithms – I, Algorithms - II, Compilers, Probability and Statistics, Computer Org & Arch, Cryptography, Software Engineering, Discrete structures, Formal Language and Automata Theory, Machine Learning, Database Management Systems, Operating Systems, Computer Networks, Speech and Natural Language Processing, Image Processing, Parallel and Distributed Algorithms, Theory of Computing, Cognitive Information Processing. Deep Learning, Advanced Machine Learning, High Performance Parallel Programming