Sahil Joshi

LinkedIn : `linkedin.com/sahiljoshi515`
Email : `sahiljoshi759@gmail.com`
Google Scholar: `scholar.google.com/sahil`

Education

Rice University Houston, TX
Doctor of Philosophy in Computer Science *Jan 2025 – Present*

- Research: Randomized Algorithms; Large Scale Machine Learning; Sketching
- Coursework: Deep Learning Systems, Probabilistic Data Structures and Algorithms, Prallel Programming with CUDA (Udemy), LLM Engineering (Udemy), IoT Programming and Data Analysis

University of Massachusetts, Amherst Amherst, MA
Bachelor of Science in Computer Science and Mathematics *Sept 2019 – May 2023*

- GPA: 4.0; Chancellor's Scholarship; Baystate Fellowship
- Outstanding Undergraduate Award; Dean's List; Outstanding Undergraduate Course Assistant Award (2021)
- Coursework: Machine Learning, Artificial Intelligence, Introduction to Algorithms, Information Systems
- Honors Thesis: Detection of Distributed Denial Of Service Attacks Using Binary Signals

Professional Experience

Rush Lab Houston, TX
Researcher (Advisor: Dr. Anshumali Shrivastava) *June 2025 – Present*

- Developed a novel linear-time, linear-memory attention mechanism using Repeated Arrays-of-Count Estimators (RACE), enabling efficient scaling to extremely large context windows while maintaining competitive performance with baselines on publicly available benchmarks.
- Proposed a novel Soft LSH formulation that empirically outperforms traditional LSH for sparse attention, and engineered high-performance CUDA and Triton kernels for practical deployment.

Dell Technologies Round Rock, TX
Software Engineer 1 (DevOps Engineer) *July 2023 – Jan 2025*

- Enhanced the user interface and optimized backend services of the Pipeline Dashboard application, used for managing Object Scale components similar to AWS S3.
- Implemented CI/CD automation using Jenkins and GitLab for deployment pipelines, improving release efficiency and reducing manual intervention across Object Scale services.
- Integrated observability tools (Prometheus, Grafana) to monitor service health, enabling proactive issue detection and reducing downtime.

Lutron Electronics Philadelphia, PA
Software Engineer Co-Op (iOS App Developer) *May 2022 – Dec 2022*

- Improved the design and workflow for the Lutron app, which is used for controlling lights, fans and shades. Fixed bugs on a bi-weekly basis to improve the iOS app's ratings.
- Designed the feature for Widget Zone Control. This feature lets the users control individual devices from the widget by providing them more flexibility from outside the app.
- Partnered with other students to think about test strategies for the new feature and wrote unit tests to minimize the errors in the app.

Publications

S. Joshi, A. Chowdhury, A. Kanakamedala, E. Singh, E. Tu, A. Shrivastava. *RACE Attention: An Inherently Efficient Linear-Time Attention Mechanism for Long Sequences*. ICLR 2026.

S. Joshi, A. Chowdhury, W. Bellinger, A. Kanakamedala, E. Singh, H. Le, A. Desai, A. Shrivastava. *SOCKET: SOft Collison Kernel EsTimator for Sparse Attention.* arXiv preprint, 2026.

H. Le, **S. Joshi**, Z. Yang, Z. Xu, A. Shrivastava. *Scout Before You Attend: Sketch-and-Walk Sparse Attention for Efficient LLM Inference.* arXiv preprint, 2026.

Projects

Sparse Attention Hub
Core contributor to an end-to-end sparse attention infrastructure for off-the-shelf large languagde models (LLMs), enabling rapid experimentation and fair, apples-to-apples evaluation of decode-time sparse attention algorithms.

Heritage Metadata Extractor
Built an end-to-end application using OCR tools and (LLMs) to digitize handwritten heritage documents, extract key metadata, store it in a structured database, and support semantic search through natural language queries.

OMI: Metagenomics Assistant
Developed an LLM-powered assistant that uses Retrieval-Augmented Generation (RAG) to translate natural language prompts into YAML configurations for a custom Nextflow pipeline, enabling non-experts to run long-read metagenomic analyses with automated tool selection, parameter tuning, and workflow documentation.

Teaching

Teaching Assistant — Rice University
COMP 518: IoT Programming and Data Analysis — *Spring 2026*

- Held weekly office hours to support students with course concepts, programming assignments, and project debugging; assisted with grading and feedback.

Activities

Rice University Computer Science Graduate Student Association — Houston, TX
President — *June 2025 – Present*

- Organized seminar series by inviting computer science faculty from Texas and neighboring states to present research to graduate students.
- Led initiatives to strengthen graduate student support, including launching Teaching Assistant Awards, organizing community events, and enabling travel funding for master’s students.

Programming Skills

Languages: Python, CUDA, Java, C++, SQL, LaTeX
Frameworks: PyTorch, NumPy, Pandas, Scikit-learn