Sahil Joshi

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

Education

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

- Research: Large Scale Machine Learning; Randomized Algorithms; Sketching
- Coursework: 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.

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, "Replacing Softmax Similarity with a Sharpened Angular Similarity: Theory and Practice of Scaling to Billion-Context Attention", *arXiv:2510.04008, 2025.*

Projects

Heritage Metadata Extractor
Building an end-to-end application using OCR tools and large language models (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
Developing 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.

Programming Skills

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