SAI NARAYAN SUNDARESAN

+91-9083521442 saisundaresan01@gmail.com LinkedIn Website Google Scholar

RESEARCH INTERESTS

My research focuses engineering Systems for Machine Learning, aiming to optimize inference efficiency in generative models by uncovering redundancies in the generation process. I have developed and published caching-based strategies that reduce inference cost and latency for LLMs, and I am currently working on improving the efficiency of video generation models.

EDUCATION

- **Dual Degree (B.Tech, M.Tech), Industrial and Systems Engineering** July 2019 – April 2024
  Indian Institute of Technology Kharagpur, India GPA: 9.15/10.00
  Micro-specialization: Artificial Intelligence and Applications

RESEARCH EXPERIENCE

- **Research Associate – Adobe Inc. (Systems and Insights Group)** Jun 2024 – Present
  Mentors: Dr. Subrata Mitra, Dr. Atanu Sinha, Dr. Shiv Saini
  - Worked on engineering efficient systems for LLM serving and Video Generation by incorporating caching techniques
  - Published 2 papers and filed 3 patents within one year and successfully integrated research innovations in 2 products
  - Selected for a session on world models at Adobe Tech Summit 2025, a company-wide internal technical conference

- **Research Intern – Sarvam.AI** Jan 2024 – May 2024
  under Prof. Pratyush Kumar, Dr. Vivek Raghavan
  - Built an end-to-end teaching tool that creates a guided audio lesson from text using custom speech-to-text pipelines
  - Deployed an efficient speech recognition service for production using Nvidia's RIVA toolkit for low resource languages
  - Trained expressive TTS models with a StyleTTS2-based architecture spanning diverse emotions and speech patterns

- **Research Intern – Adobe Inc. (BigData Intelligence Lab)** May 2023 – Aug 2023
  Brand-guided Poster Generation
  - Built a design creation tool for making posters given a product image and brand reference based on Custom Diffusion
  - Devised an algorithm to determine the optimal latents for multiple reference concepts based on latent interpolation
  - Implemented the algorithm in Custom Diffusion to reduce the dependence on initial latent in the diffusion process

- **Data Engineering Inern – AI4Bharat, IIT Madras** Jan 2023 – Apr 2023
  under Prof. Mitesh M. Khapra
  - Collated benchmarks for ASR across 12 Indian languages accounting for dialect, domain and, sound quality variations
  - Performed alignment to create datasets from long form audio content like NPTEL using Needleman-Wunsch algorithm
  - Built a large-scale training dataset for Indian language ASR with 10,736 hours of diverse data across 12 languages
  - Finetuned the Whisper multilingual model on the dataset achieving an average 4.1 WER reduction over existing models

- **Machine Learning Intern – Honeywell Technology Solutions Lab** May 2022 – Jul 2022
  under Mr. Srikanth Viswaraju
  - Designed a contextual search engine for flight manuals based on natural language processing using a BERT based model
  - Achieved an accuracy of 79.38% for textual queries and 54.69% for tabular queries on a set of 224 crowdsourced queries
  - Showcased the search method at Honeywell AI/ML Roadshow as an efficient alternative to traditional keyword search

PUBLICATIONS

[1] (EMNLP' 25) **Sai Sundaresan**, Harshita Chopra, Atanu R. Sinha, Koustava Goswami, Nagasai Saketh Naidu, Raghav Karan, N Anushka. **Subjective Behaviors and Preferences in LLM: Language of Browsing.** In *The 30th Conference on Empirical Methods in Natural Language Processing*, 2025.

[2] (SIGMOD '25) Shubham Agarwal*, **Sai Sundaresan***, Subrata Mitra, Debabrata Mahapatra, Archit Gupta, Rounak Sharma, Nirmal Joshua Kapu, Tong Yu, Shiv Saini. **Cache-Craft: Managing Chunk-Caches for Efficient Retrieval-Augmented Generation.** In *The 44th International Conference on Management of Data*, 2025.

[3] (ICLST '24) **Sai Sundaresan**, Anand Abraham. **Single Resource Capacity Control Model for Hidden City Ticketing.** In *The International Conference on Logistics, Supply Chain and Transportation*, 2024.

[4] (ORSI '24) **Sai Sundaresan**, Anand Abraham. **Clearance Sale Models under Competition.** In *The International Conference on Trends in Business Analytics & Management*, 2024.

[5] (INTERSPEECH '23) Kaushal Santosh Bhogale*, **Sai Sundaresan***, Abhigyan Raman, Tahir Javed, Mitesh M Khapra, Pratyush Kumar. **Vistaar: Diverse Benchmarks and Training Sets for Indian Language ASR.** In *The 24th INTERSPEECH Conference*, 2023.

[6] (INTERSPEECH '23) Tahir Javed, Sakshi Joshi, Vignesh Nagarajan, **Sai Sundaresan**, Janki Nawale, Abhigyan Raman, Kaushal Bhogale, Pratyush Kumar, Mitesh M. Khapra. **Svarah: Evaluating English ASR Systems on Indian Accents.** In *The 24th INTERSPEECH Conference*, 2023.

SELECTED PROJECTS

- **Efficient Video Generation through Patch Level Caching** ongoing
  - Developing a context-aware system to reuse intermediate states across similar patches via rectified flow based interpolation
  - The algorithm exploits the linear trajectories followed to estimate complete path updates from a subset of computations
  - Initial results show 1.2× improvement in latency over Teacache for certain domains while maintaining similar FVD scores

- **Efficient LLM Serving for RAG** *paper published in* [SIGMOD '25]
  - Built a KV-cache reuse system for RAG, cutting redundant attention computations by allowing prefix-independent reuse.
  - Developed algorithms to determine cache reusability and fixing chunk-caches via recomputation of high attention tokens
  - Delivered 51% computation reduction, 1.6× throughput, 2× latency gains over prefix-caching in production workloads

- **Heterogenity Aware User Behaviour Prediction** *paper published in* [EMNLP '25]
  - Proposed a heterogeneity-aware, clusterwise language model training approach for subjective user browsing behaviors.
  - Demonstrated that clusterwise page-level tokenized small LMs outperform larger LMs on user browsing sequencing tasks.
  - Achieved higher mean and lower variance in page generation and outcome prediction metrics, improving user-level alignment.

PATENTS

[1] **Sai Narayan Sundaresan**, Atanu R Sinha, Harshita Chopra, Koustava Goswami, Raghav Karan, Nagasai Saketh Naidu, Anushka N. **Heterogenous LLMs for Subjective Behaviors.** [Filed] (US Patent App. 19/215,758)

[2] Harshita Chopra, Nagasai Saketh Naidu, Raghav Karan, Anushka N, Atanu R Sinha, Koustava Goswami, **Sai Narayan Sundaresan**. **Utilizing Digital Page Sequence Tokens with Large Language Models to Generate Digital User Activity Predictions.** [Filed] (US Patent App. 19/050,836)

[3] Shubham Agarwal, **Sai Sundaresan**, Subrata Mitra, Debabrata Mahapatra, Archit Gupta, Rounak Sharma, Nirmal Joshua Kapu, Tong Yu, Shiv Saini. **Managing Chunk Caches for Efficient Retrieval-Augmented Generation.** [Filed] (US Patent App. 19/074,061)

TECHNICAL SKILLS

- **Programming Languages** Python, C++, C, Triton
- **ML/Systems Frameworks** vLLM, Transformers, Diffusers, PyTorch, Git

RELEVANT COURSEWORK

Graphical and Generative Models, Machine Learning Foundations, Artificial Intelligence Foundations, Algorithms, Programming and Data Structures, Probability and Statistics, AI for Cyber-Physical Systems, AI for Economics

ACTIVITIES AND ACHIEVEMENTS

- Received best paper award and merit paper award respectively for work presented at ICLST '24 and ORSI '24
- Mentored over 6 undergraduate interns and collaborated with 1 PhD intern during summer internships at Adobe
- Contributed to multiple paper reviews for technical conferences including Usenix ATC, SIGMOD, EMNLP, etc.
- Recieved Certificate of Merit for achieving the top rank in Computer Science, AISSCE 2019 (D.A.V. School, Chennai)
- Served as Head of the Computer Graphics Society, organizing events and promoting engagement in game development.