

# SAI NARAYAN SUNDARESAN

☎ +91-9083521442 ✉ [saisundaresan01@gmail.com](mailto:saisundaresan01@gmail.com) in [LinkedIn](#) 🏠 [Website](#) 🎓 [Google Scholar](#)

## RESEARCH INTERESTS

My research focuses on the intersection of Machine Learning and Systems, 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 Intern – 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 *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] ([ORSI '24](#)) **Sai Sundaresan**, Anand Abraham. **Proof of Federated Training: Accountable Cross-Network Model Training and Inference**. In *International Conference on Trends in Business Analytics & Management Sciences*, 2024.
- [4] ([ICLST '24](#)) **Sai Sundaresan**, Anand Abraham. **Single Resource Capacity Control Model for Hidden City Ticketing**. In *Recent Advances in Logistics, Supply Chain and Transportation*, Volume 1, 2025.
- [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
  - Worked on a system that enables reuse of intermediate states across similar patches through rectified flow based interpolation
  - Developed a Key-Value reuse mechanism that decomposes prefill states into reusable chunks, achieving a 2.7× reduction in prefill compute costs and latency while maintaining output quality.
- **Efficient LLM Serving for RAG** paper published in [\[SIGMOD '25\]](#)
  - Built a KV-cache reuse system for RAG-based LLMs, cutting redundant attention computations.
  - Devised efficient chunk-cache reuse and minimal recomputation strategies, plus hardware-aware cache management.
  - Delivered 51–75% computation reduction and 1.6× throughput, 2× latency gains over prefix-caching in production.
- **Heterogeneity Aware User Behaviour Prediction** paper published in [\[EMNLP '25\]](#)
  - Proposed a heterogeneity-aware, clusterwise LM training approach for subjective user behaviors.
  - Demonstrated that small LMs with page-level tokenization outperform larger pretrained or finetuned LMs on browsing data.
  - Achieved higher mean performance with lower variance, improving user-level alignment and consistency.

## 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)

## SKILLS

---

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

## ACTIVITIES AND ACHIEVEMENTS

---

- Mentored 6 undergraduate interns and collaborated with 1 PhD intern during summer internships at Adobe
- Received best paper award and merit paper award respectively for work presented at ORSI '24 and ICLST '24
- Branch Rank 1 in Quality Engineering Design and Manufacturing (highest GPA)
- Certificate of Merit for top rank in Computer Science, AISSCE 2019 (D.A.V. School, Chennai)
- Headed the Computer Graphics society with an active interest in Game Development

## 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