

Ranajoy Sadhukhan

Carnegie Mellon University

📞 412-430-1472 📩 rsadhukh@andrew.cmu.edu 💬 [ranajoy-sadhukhan](#) 💬 [ranonrkm](#)

RESEARCH INTEREST

Efficient Machine Learning, Information Retrieval, Natural Language Processing.

EDUCATION

- **Carnegie Mellon University** 2023-present
PhD Candidate, Electrical and Computer Engineering GPA: 3.81/4.0
Advisor: [Dr. Beidi Chen](#)
- **Indian Institute of Technology Kharagpur** 2016-2021
Dual-degree(B.Tech+M.Tech) in Electrical Engineering, Specialization in Signal Processing GPA : 9.39/10
Minor in Computer Science & Engineering GPA : 9.92/10

PUBLICATIONS

- **STEM: Scaling Transformers with Embedding Modules**
by **Ranajoy Sadhukhan**, Sheng Cao, Harry Dong, Changsheng Zhao, Attiano Purpura-Pontoniere, Yuandong Tian, Zechun Liu, Beidi Chen
Under submission in International Conference on Learning Representations 2026
- **Kinetics: Rethinking Test-Time Scaling Laws**
by **Ranajoy Sadhukhan***, Zhuoming Chen*, Haizhong Zheng, Yang Zhou, Emma Strubell, Beidi Chen
In Neural Information Processing Systems 2025 [\[Paper\]](#) [\[Code\]](#)
- **MagicDec: Breaking the Latency-Throughput Tradeoff for Long Context Generation with Speculative Decoding**
by **Ranajoy Sadhukhan***, Jian Chen*, Vashisth Tiwari, Zhuoming Chen, Ruihang Lai, Jinyuan Shi, Ian En-Hsu Yen, Avner May, Tianqi Chen, Beidi Chen
In International Conference on Learning Representations 2025 [\[Paper\]](#) [\[Code\]](#)
- **MagicPIG: LSH Sampling for Efficient LLM Generation**
by Zhuoming Chen, **Ranajoy Sadhukhan**, Zihao Ye, Yang Zhou, Jianyu Zhang, Niklas Nolte, Yuandong Tian, Matthijs Douze, Léon Bottou, Zhihao Jia, Beidi Chen
In International Conference on Learning Representations 2025 [\[Paper\]](#) [\[Code\]](#)
- **Memory Mosaics**
by Jianyu Zhang, Niklas Nolte, **Ranajoy Sadhukhan**, Beidi Chen, Léon Bottou
In International Conference on Learning Representations 2025 [\[Paper\]](#) [\[Code\]](#)
- **Taxonomy Driven Learning Of Semantic Hierarchy Of Classes**
by **Ranajoy Sadhukhan**, Ankita Chatterjee, Jayanta Mukhopadhyay, Amit Patra
In IEEE International Conference on Image Processing 2022 [\[Paper\]](#) [\[Code\]](#)
- **Knowledge Distillation Inspired Fine-Tuning of Tucker Decomposed CNNs and Adversarial Robustness Analysis**
by **Ranajoy Sadhukhan**, Abhinav Saha, Jayanta Mukhopadhyay, Amit Patra
In IEEE International Conference on Image Processing 2020 [\[Paper\]](#) [\[Code\]](#)

INDUSTRY EXPERIENCE

- **Meta - Monetization Generative AI** Aug '21–Jan '23
Advisors – [Rick Cao](#) & [Dr. Zechun Liu](#) May '25-Present
 - Designing architecture-level modifications of Large Language Models that reduce end-to-end LLM serving latency while improving response quality at scale.

- Microsoft Research India Aug '21–Jan '23
Advisors – Dr. Harsha Vardhan Simhadri & Dr. Manik Varma
 - Developed retrieval-metric-aware learnable Product Quantization (PQ) for memory-efficient, high-recall dense retrieval, achieving $64\times$ compression.
 - Improved DiskANN index construction for out-of-distribution queries, yielding up to 45% lower latency at comparable recall on 100 million-scale databases.
- Samsung R&D Institute Bangalore May '20–Jul '20
Advisor – Pankaj Kumar Bajpai
 - Built a 6 MB lightweight DNN for joint monocular depth estimation and panoptic segmentation; introduced adaptive batch sampling and task-specific loss fusion to handle asymmetrically annotated datasets.

ACADEMIC RESEARCH EXPERIENCE

- Efficient LLM serving for long-context applications May '24–Apr '25
Advisor – Dr. Beidi Chen Carnegie Mellon University
 - Formulated the *Kinetics Scaling Law* for a more practical and effective test-time scaling of Large Language Models.
 - Designed a hardware-efficient LLM decoding stack combining speculative decoding and KV-cache compression, achieving up to $2.51\times$ speedup in long-context, large-batch inference.
- Hierarchically Self-Decomposing CNN Feb '20–May '21
Advisor – Dr. Jayanta Mukhopadhyay Indian Institute of Technology Kharagpur
 - Devised an explainable hierarchical decomposition of pretrained CNNs via a semantic loss, producing task-specific sub-networks with up to $4\times$ speedup and 75% fewer parameters for limited-class tasks, without fine-tuning.

SERVICES

- Peer Review
Reviewer: ICLR'25, NeurIPS'25, ES-FoMo@ICML'24
- Teaching Assistant – 18-789: Deep Generative Modeling Jan '24–May '24
Instructors: Dr. Beidi Chen, Dr. Giulia Fanti
- Teaching Assistant – EE19001: Electrical Technology Lab Dec '20–Mar '21
Instructors: Dr. Souvik Chattopadhyay, Dr. Dipankar Debnath

ACHIEVEMENTS

- Selected for the highly decorated **Mitacs Globalink Research Internship Program** and **DAAD Research Internship Program** (2019).
- Awarded **Best Project Award** for **Masters Thesis Project** in Electrical Engineering at IIT Kharagpur.
- Awarded **Merit-cum-Means scholarship, IIT Kharagpur**, for academic excellence among 1400 students (2016).