Siddhant Ray

Chicago, IL, USA
+1-(773)-457-4156
siddhant.r98@gmail.com
siddhant-ray.github.io/
https://github.com/Siddhant-Ray (GitHub)

Education

2023 – 2028 **The University of Chicago**, *PhD in Computer Science*
Advisor - Junchen Jiang and Nick Feamster

2020 – 2022 **ETH Zürich**, *MSc in Electrical Engineering and Information Technology*
Advisor - Laurent Vanbever

2016 – 2020 **VIT Vellore**, *B.Tech in Electronics and Communication Engineering*

Experience

Jun 2025 – Sep 2025 **Research Intern**, *Microsoft Research*
- Joint internship with Outlook for scalable and cost-efficient LM architectures for email importance labeling.
- Built a hybrid architecture with SLM cascades and embedding-based classifiers to reduce COGS for labeling upto 150X, with negligible quality loss.

Sep 2023 – Present **Graduate Research Assistant**, *Computer Science Department, The University of Chicago*
- *Project 2:* Joint quality-latency optimization for Retrieval Augmented Generation(RAG) LLM systems with query-level configuration selection and scheduling.
- *Project 1:* Developing Transformer based models for sharp latency change prediction to enable packet-level queue management for tail-latency reduction.

Sep 2022 – Mar 2023 **Cloud Networks Researcher**, *Advanced Network Architectures Lab, UPC Barcelona*
- Developed an approximation for a Mixed-Integer Optimal Matching Algorithm for edge resource allocation to reduce execution time by 2.5-3X.

Oct 2021 – Sep 2022 **Graduate Research Assistant**, *Law, Economics, and Data Science Group, ETH Zurich*
- *Project 2:* Developed improved semantic labeling pipelines using state-of-the-art NLP models to enable better sentence simplification, paraphrase mining, and topic clustering.
- *Project 1:* Developed RoBERTa-based NLP models to analyse political discourse in meat policy documents to enable actor-narrative clustering.

May 2019 – July 2019 **Software Development Intern**, *Capgemini Engineering*
- Developed a K-Shortest Path Searching algorithm in an ONOS based Software Defined Layer 2 VPN.
- Algorithm applied dynamic constraints of network resources (e.g.required edges) for path calculation.

May 2018 – July 2018 **Software Development Intern**, *BlueStacks*
- Developed an ML algorithm to customize the Bluestack's display screen using past users' experiences.
- Built a automation tool for generating SVG cards for Bluestack's game engine and an address verification tool with the EasyPost API.

Publications

2026 **Siddhant Ray**, Xi Jiang, Jack Luo, Nick Feamster, and Junchen Jiang. *SwiftQueue: Optimizing Low-Latency Applications with Swift Packet Queuing*. In *New Ideas in Networked Systems (NINeS'26)*.

2025 Jiayi Yao, Hanchen Li, Yuhan Liu, **Siddhant Ray**, Yihua Cheng, Qizheng Zhang, Kuntai Du, Shan Lu, and Junchen Jiang. *CacheBlend: Fast Large Language Model Serving for RAG with Cached Knowledge Fusion*. In *ACM European Conference on Computer Systems (EuroSys'25)*.

2025 **Siddhant Ray**, Rui Pan, Zhuohan Gu, Kuntai Du, Shaoting Feng, Ganesh Ananthanarayanan, Ravi Netravali, and Junchen Jiang. *METIS: Fast Quality-Aware RAG Systems with Configuration Adaptation*. In *ACM SIGOPS 31st Symposium on Operating Systems Principles (SOSP'25)*.

2024 Yuhan Liu, Hanchen Li, Yihua Cheng, **Siddhant Ray**, Yuyang Huang, Qizheng Zhang, Kuntai Du, Jiayi Yao, Shan Lu, Ganesh Ananthanarayanan, Michael Maire, Henry Hoffmann, Ari Holtzman, and Junchen Jiang. *CacheGen: KV Cache Compression and Streaming for Fast Large Language Model Serving*. In *ACM Special Interest Group on Data Communication (SIGCOMM'2024)*.

2024 Hanchen Li, Yuhan Liu, Yihua Cheng, **Siddhant Ray**, Kuntai Du, and Junchen Jiang. *Eloquent: A More Robust Transmission Scheme for LLM Token Streaming*. In *SIGCOMM Workshop on Networks for AI Computing (NAIC'24)*.

2022 Alexander Dietmüller, **Siddhant Ray**, Romain Jacob, and Laurent Vanbever. *A New Hope for Network Model Generalization*. In *21st ACM Workshop on Hot Topics in Networks (Hotnets'22)*.

2020 **Siddhant Ray** and Budhaditya Bhattacharyya. *Machine Learning based Cell Association for mMTC 5G Communication Networks*. *International Journal of Mobile Network Design and Innovation*.

Honors and Awards

2025 **Travel Grant**, *LDOS PhD Research School*, UT Austin

2025 **Best Paper Award**, *EuroSys 2025*, Rotterdam

2023 – 2028 **Liew Family Graduate Fellowship**, University of Chicago

2022 **Winner at Datathon**, *Microsoft Challenge*, ETH Zurich

2020 **Best Outgoing Student**, *SENSE department*, VIT Vellore

2016 – 2019 **Merit Scholarship for Academic Excellence**, VIT Vellore

Service

2026 **External Review Committee**, *MLSys*

2026 **Reviewer**, *AAAI, ICLR*

2025 **Reviewer**, *ICML*

2025 **Artifact Evaluation Committee**, *ATC, OSDI, CoNEXT*

2025 **Teaching Assistant**, *CS144: Systems Programming II*, UChicago

2024 **Artifact Evaluation Committee**, *CoNEXT*

Talks

Nov 2025 **Seminar Talk**, *Systems Research Seminar*, UIUC

Nov 2025 **Guest Lecture**, *Systems for LLMs and AI Agents Class*, UCSD

Aug 2025 **PhD Research Talk**, *Networking Research Group*, Microsoft Research

April 2024 **Lightning Talk**, *Research Social*, Conviva

Other Research Projects

2022 Advancing Packet-Level Traffic Predictions with Transformers (Master Thesis) - [code, thesis]

2021 Towards a New Framework for Integration of Network Planes (Research Project) - [code]

2021 Attentive Neural Networks for News Classification (Research Project) - [code]

Technical Skills

Programming Python, C++, C, Bash, Rust, SQL, Java, TEX

Software Linux, Git, Docker, P4, Azure, Google Cloud, AWS, Maven, K8s, Helm

Frameworks vLLM, Langchain, Llama-Index, PyTorch, Sklearn, NLTK, Flask, Asyncio, NS3, Mininet, FRRouting