Kanishk Goel

Website | Scholar | LinkedIn | k.goel2601@gmail.com

Summary

Research Fellow at Microsoft Research with expertise in systems for AI/ML, large-scale LLM inference, CUDA, parallel programming, and reinforcement learning. Recent EECS Undergrad from IIT Delhi, India.

Education

Indian Institute of Technology Delhi Nov 2020 – May 2024
B.Tech. in Electrical Engineering (CGPA: 8.75/10, Top 7%); Minor in Computer Science (CGPA: 9.59/10)

Experience

Microsoft Research India July 2024 – Present
Research Fellow, AI Infrastructure Team

- Working on **LLM inference efficiency** through scheduling, novel attention mechanisms, and GPU kernel optimizations
- Investigating novel reinforcement learning formulations for improving **training efficiency and system throughput**

IIT Delhi Aug 2023 – Jan 2024
Teaching Assistant, Data Structures & Algorithms (COL106)

- Designed and graded programming assignments in C++, conducted lab sessions and tutorials for 400+ undergraduates

Uber India Systems May 2023 – July 2023
Software Engineering Intern, Platforms Team

- Built an e2e automated retrieval system for database metadata, streamlining debugging for distributed datastore M3DB
- Implemented shard placement queries, bootstrap inspection, and overlapping-shard detection for prod. servers in Go

Selected Projects

System Optimizations for Agentic Inference Oct 2025 – Present

– Optimizing multi-turn agents by overlapping system prompt processing with tool execution to minimize end-to-end latency
– Implementing scheduling strategies to maximize intra- and inter-request KV cache hit rates for higher serving throughput
– Accepted as a poster at **EuroSys'26**. Accepted as a paper at **ASPLOS'27**
Supervisors: Jayashree Mohan, Ramachandran Ramjee (MSR India)

Reinforcement Learning for LLMs Training Efficiency June 2025 – November 2025

– Designing new RL objectives that improve data efficiency and enable higher PPO steps per unit compute, resulting in better system utilization and higher training throughput without sacrificing accuracy
Supervisors: Nipun Kwatra, Ramachandran Ramjee (MSR India)

Niyama: QoS-aware LLM Inference Scheduler Aug 2024 – May 2025

– Achieved 32% higher throughput and 10× fewer SLO violations by dynamically adapting request scheduling
– **[arXiv]**. Accepted as a paper at ASPLOS 2026 (14.5% acceptance rate, 152/1048); accepted as a poster at SOSP 2025
Supervisors: Jayashree Mohan, Nipun Kwatra, Ramachandran Ramjee (MSR India)

Radar-based Imaging for Autonomous Driving Aug 2023 – May 2024

– Proposed a **multi-modal fusion model** integrating radar, lidar, and camera; introduced transformer-based BEV module
– Published as **MMA-Net, ICASSP 2025**
Supervisor: Seshan Srirangarajan (IIT Delhi)

Positions of Responsibility

Core Team Member, Algorithms and Coding Club (ANCC), IIT Delhi May 2021 – May 2024

- Rose from Executive (1 of 25) to Coordinator (1 of 10) to Core Team Member (1 of 5) leading the club over 3 years
- Organized competitive programming contests, lectures, and events to promote campus programming culture

Scholastic Achievements

- **IIT Delhi Dean's List:** Awarded for top 7% academic performance in multiple semesters.
- **National Fellowships & Exams:** KVPY Fellowship (Govt. of India); JEE Advanced Rank 433/250,000+ (Top 0.1%).
- **Science Olympiads:** Ranked in the top 1% nationwide in National Science Olympiads (Physics, Chemistry).

Key Courses

Operating Systems; Computer Architecture; Parallel Programming (OpenMP/MPI/CUDA); Embedded Computing; Multiprocessor Systems; Machine Learning; Deep Learning; Data Structures & Algorithms; Probability & Stochastic Processes; Linear Algebra; Digital Electronics; Digital Signal Processing; Control Engineering; Communication Engineering; Systems and Design Lab