# Karan Tandon

Ph.D. Student in Computer Science, Princeton University
Research interests: computer systems for efficient information retrieval and machine learning.
karan2dec@gmail.com | karantandon.com

## Education

**Doctor of Philosophy (Ph.D.) in Computer Science** Sep 2025 - Present
Princeton University, NJ, United States
Advisors: Jialin Ding, Ravi Netravali

**Bachelor of Technology (B.Tech.) in Computer Science and Engineering** Aug 2017 - Jun 2021
Indraprastha Institute of Information Technology Delhi, India
GPA: 9.67/10.0 (class rank: 2/323)

## Conference Publications

**OPPerTune: Post-Deployment Configuration Tuning of Services Made Easy**
*21st USENIX Symposium on Networked Systems Design and Implementation (NSDI 2024).*
Gagan Somashekar*, **Karan Tandon***, Anush Kini, Chieh-Chun Chang, Petr Husak, Ranjita Bhagwan, Mayukh Das, Anshul Gandhi, Nagarajan Natarajan

## Workshop Papers

**Reward Copilot for RL-driven Systems Optimization**
*NeurIPS Workshop on Machine Learning for Systems, 2024.*
**Karan Tandon**, Manav Mishra, Gagan Somashekar, Mayukh Das, Nagarajan Natarajan

## Preprints

**BeLLMan: Controlling LLM Congestion**
*arXiv:2510.15330, 2025. Selected arXiv talk at the 1st Frontier AI Systems Workshop (FAISys 2025).*
Tella Rajashekhar Reddy*, Atharva Deshmukh*, **Karan Tandon**, Rohan Gandhi, Anjaly Parayil, Debopam Bhattacherjee

**Improving training time and GPU utilization in geo-distributed language model training**
*arXiv:2411.14458, 2024.*
Palak, Tella Rajashekhar Reddy, Bhaskar Kataria, Rohan Gandhi, **Karan Tandon**, Debopam Bhattacherjee, Venkata N. Padmanabhan

## Research Experience

**Princeton University** - Graduate Researcher Sep 2025 - Present
Advisors: Jialin Ding, Ravi Netravali
Developing systems for efficient vector and hybrid retrieval.

**Microsoft Research** - Research Software Engineer 2, India Aug 2022 - Aug 2025
Research mentors: Venkat Padmanabhan, Ranjita Bhagwan, Nagarajan Natarajan

**Reward Copilot: LLM-Assisted Reward Design for Systems Tuning**

- Initiated Reward Copilot to translate operator requirements into reward functions that guide reinforcement-learning-based systems tuning. It uses an LLM to propose candidate functions, tests each against the requirements, and iteratively refines the best one.

**OPPerTune: Automatic Performance Tuning for Deployed Applications**

- Led the design and development of OPPerTune, which tunes applications without knowledge of their internals. Applications expose configurable parameters, request values, and report performance. OPPerTune uses this feedback to choose future settings as workloads and infrastructure change.
- Built a distributed testbed to identify which settings most affect performance and measure how quickly tuning algorithms find good values across workloads.
- Led its open-source release, production deployments across Microsoft services, and integration into Azure Resource Central, Microsoft's prediction-serving platform. The work appeared at NSDI 2024 and resulted in a US patent application.

**CEISER: Cost-Effective Ensemble Inference on a Single GPU**

- Explored cost-efficient inference by replacing one model with an ensemble of smaller models on a single GPU, using model-graph merging and kernel fusion.

**BeLLMan: Response-Length Prediction for Congestion-Aware LLM Serving**

- Helped build BeLLMan's LLM response-length predictor. During congestion, the system instructed the LLM to generate shorter responses, achieving up to 8x lower latency, 25% less energy use, and 19% more completed requests.

**Atlas and BubbleTea: Scheduling Geo-distributed LLM Training**

- Contributed to the design of Atlas and BubbleTea, which schedule wide-area bandwidth and use idle GPU periods for inference during geo-distributed language-model training.

**Networks and Security Lab** - Undergraduate Researcher, IIITD — Jan 2020 - Aug 2021

Bachelor's thesis advisors: Sambuddho Chakravarty, Mukulika Maity

Built a client-independent IP geolocation system that uses Kalman filters to estimate host locations from noisy network-delay measurements.

## Industry Experience

**Microsoft** - Software Engineer, India — Aug 2021 - Jul 2022

Designed and built a Python framework for automated Android device testing, used across several teams. Mentored an undergraduate intern.

**Microsoft** - Software Engineer Intern, India — May 2020 - Jul 2020

Built a Linux kernel driver and Android app to detect device misconfigurations without specialized hardware.

## Teaching

**Introduction to Programming (CSE 101) - Teaching Assistant** — Jan 2021 - Apr 2021

Indraprastha Institute of Information Technology Delhi

Created assignments, programming contests, and labs for 570+ students. Led weekly tutorials for 37 students.

## Technical Skills

C++, Python, C, Go, PyTorch, Nsight Systems, Nsight Compute, wrk2