Shivam Patel

Personal Webpage | Google Scholar

Education

Carnegie Mellon University 2024-28 (Expected)
PhD in Electrical and Computer Engineering
Advisor: Prof. Gauri Joshi

Indian Institute of Technology Bombay 2020-24
Bachelor of Technology in Electrical Engineering
Advisor: Prof. Vivek Borkar; CPI: 9.66/10

Research Interests

Efficient and Low-Latency LLM Inference, Systems-Algorithm Co-Design for Inference Optimization.
Previously: Convex Optimization, Stochastic Systems, Game Theory.

Work Experience

AI Research Intern
Robert Bosch Gmbh, Pittsburgh *May - August '26*
Advisor: Dr. Bingqing Chen. Working on large scale multi-GPU inference optimization.

Graduate Researcher
Carnegie Mellon University, Pittsburgh *August '24 – Present*
Advisor: Prof. Gauri Joshi. Working on compute and memory efficiency of LLMs as a PhD Candidate at CMU.

Visiting Undergraduate Researcher
USC Viterbi School of Engineering, Los Angeles *May – Aug '23*
Advisor: Prof. Meisam Razaviyayn. Worked on statistical fairness and robust model objectives.

Publications and Preprints

[6] Beyond Accuracy and Cost: Latency-Aware LLM Query Routing for Dynamic Workloads
Shivam Patel*, Akaash Parthasarathy*, Ankur Mallick, Gauri Joshi. arXiv Preprint, 2026.

[5] LOCUS: Low-Dimensional Model Embeddings for Efficient Model Exploration, Comparison, and Selection
Shivam Patel, William Cocke, Gauri Joshi. AdaptFM-ICML 2026.

[4] Federate the Router: Learning Language Model Routers with Sparse and Decentralized Evaluations
Baris Askin*, **Shivam Patel***, Anupam Nayak*, Andrea Vigano, Jiin Woo,
Gauri Joshi, Carlee Joe-Wong. arXiv preprint, 2026.

[3] ProxRouter: Proximity-Weighted LLM Query Routing for Improved Robustness to Outliers
Shivam Patel, Neharika Jali, Ankur Mallick, Gauri Joshi.
International Conference on Artificial Intelligence and Statistics, AISTATS 2026.

[2] An Asymptotic CVaR Measure of Risk for Markov Chains
Shivam Patel, Vivek S. Borkar.
IEEE Signal Processing and Communications Conf., SPCOM 2024.

[1] f-FERM: A Scalable Framework for Robust Fair Empirical Risk Minimization
Sina Baharlouei, **Shivam Patel**, Meisam Razaviyayn.
The Twelfth International Conference on Learning Representations, ICLR 2024.

* Equal Contribution.

Relevant Coursework

At CMU: Algorithms for Language Modeling, Generative AI, Statistical Methods in Machine Learning, Large Scale Optimization for Distributed ML

At IITB: Stochastic Optimization, Markov Chains and Queueing Systems, Advanced ML, Information Theory Estimation and Identification, Reinforcement Learning, Game Theory

Scholastic Achievements

- Exemplary PhD Qualifying Examination Performance by ECE, CMU *(Spring 26)*
- Awarded the William Messner Endowed Fellowship at ECE, CMU *(2025-26)*
- Conferred the Carnegie Institute of Technology Dean's Fellowship at CMU *(2024-25)*
- Awarded the IUSSTF-Viterbi Scholarship for pursuing summer research at USC Viterbi School by Department of Science & Technology, Govt. of India. One of 15 selected students across India. *(2023)*
- Accorded the Kishore Vaigyanik Protsahan Yojana (KVPY) fellowship by IISc and DST India *(2020)*
- Achieved an All India Rank 219 in JEE Advanced among 0.225 million candidates *(2020)*