

ANAY MEHROTRA

anaymehrotra.com ◊ anaymehrotra1@gmail.com ◊ Google Scholar

EDUCATION

Ph.D. in Computer Science, Yale University 2020 – 2026
Thesis: *Learning Theory in the Wild: Foundations of Missing Data and Language Generation*
Advisors: Manolis Zampetakis and Amin Karbasi

B.Tech. in Computer Science and Engineering, IIT Kanpur 2016 – 2020
GPA 9.9/10, Major GPA 10/10

Exchange Semester, École Polytechnique Fédérale de Lausanne (EPFL) Fall 2018

RESEARCH EXPERIENCE

Motwani Postdoctoral Fellow, Stanford University January 2026 – Present
Hosted by Amin Saberi; developing foundations of language models

Student Researcher, Google Research, USA October – December 2025
Improved inference-time performance of small models

Student Researcher, Robust Intelligence (*acquired* by Cisco), California, USA November 2023 – May 2024
Designed a prompt-generator for making LLMs output harmful text (published at NeurIPS'24; covered by WIRED)

Research Intern, Microsoft Research Lab, Karnataka, India Summer 2022
Studied and designed constrained ranking algorithms to improve hidden utility (published at ACM AIES'23)

SELECTED PUBLICATIONS

In my field, authors are typically listed alphabetically

DDPM Score Matching and Distribution Learning COLT 2026
Best Short Paper at ICLR'25 Workshop on Generative Models S. Chewi, A. Kalavasis, A. Mehrotra, O. Montasser

What Makes Treatment Effects Identifiable? COLT 2025
Best Paper COLT'25 and **Invited Presentation** IJCAI'26 Y. Cai, A. Kalavasis, K. Mamali, A. Mehrotra, M. Zampetakis

On the Limits of Language Generation: Trade-Offs Between Hallucination and Mode Collapse STOC 2025
Highlights Track at Foundations of Responsible Computing (FORC'26) A. Kalavasis, A. Mehrotra, G. Velegkas

Tree of Attacks: Jailbreaking Black-Box LLMs Automatically NeurIPS 2024
Covered by WIRED and TechCrunch A. Mehrotra, M. Zampetakis, ..., A. Karbasi

Toward Controlling Discrimination in Online Ad Auctions ICML 2019
Best Student Paper at Mechanism Design for Social Good (MD4SG'19) E. Celis, A. Mehrotra, N. Vishnoi

SELECTED HONORS AND AWARDS

Selected Research Awards and Honors

- ▶ *Invited Presentation*, International Joint Conference on Artificial Intelligence (IJCAI'26) for Publication [19](#) 2026
- ▶ *Best Paper Award*, Conference on Learning Theory (COLT 2025) for Publication [19](#) 2025
- ▶ *Best Short Paper*, ICLR'25 Workshop on Deep Generative Models in ML for Publication [26](#) 2025
- ▶ *Sri Binay Kumar Sinha Award*, IIT Kanpur for research that addresses a societal problem 2020
- ▶ *Best Student Paper*, Workshop on Mechanism Design for Social Good for Publication [1](#) 2019

Selected Competitive Programming Awards

- ▶ *Finalist*, ICPC World Finals, Rank 33 2021
- ▶ *Winner*, ICPC Asia West Continent Championship, spanning eight countries 2020
- ▶ *Winner (twice)*, ICPC Gwalior–Pune and Kharagpur Regionals 2019

Selected (Non-Research) Academic Honors

- ▶ *Academic Excellence Awards*, For all semesters at IIT Kanpur 2016 – 2020
- ▶ *KVPY Scholar*, National fellowship by the Department of Science and Technology, India 2015

-
29. *Differentially Private Language Generation and Identification in the Limit*
Anay Mehrotra, Grigoris Velegkas, Xifan Yu, Felix Zhou
Conference on Learning Theory COLT 2026
‣ *Other Presentations: Theory and Practice of Differential Privacy (TPDP), 2026*
‣ *Other Presentations: First Workshop on Formal Languages and Neural Networks (FLaNN), 2026*
28. *Language Generation with Infinite Contamination*
Anay Mehrotra, Grigoris Velegkas, Xifan Yu, Felix Zhou
Conference on Learning Theory COLT 2026
‣ *Other Presentations: First Workshop on Formal Languages and Neural Networks (FLaNN), 2026*
27. *Can SGD Select Good Fishermen? Local Convergence Under Self-Selection Biases and Beyond*
Alkis Kalavasis, Anay Mehrotra, Felix Zhou
Conference on Learning Theory COLT 2026
26. *DDPM Score Matching and Distribution Learning*
Sinho Chewi, Alkis Kalavasis, Anay Mehrotra, Omar Montasser
Conference on Learning Theory COLT 2026
Best Short Paper at ICLR'25 Workshop on Deep Generative Models in ML
25. *Linear Regression with Unknown Truncation Beyond Gaussian Features*
Alexandros Kouridakis, Anay Mehrotra, Alkis Kalavasis, Constantine Caramanis
International Conference on Machine Learning ICML 2026
24. *Language Generation with Feedback: Queries and Mistakes*
Steve Hanneke, Amin Karbasi, Anay Mehrotra, Grigoris Velegkas
International Conference on Machine Learning ICML 2026
23. *Smoothed Analysis of Learning from Positive Samples*
Jane H. Lee, Anay Mehrotra, Manolis Zampetakis
ACM Symposium on Theory of Computing STOC 2026
22. *Mean Estimation from Coarse Data: Characterizations and Efficient Algorithms*
Alkis Kalavasis, Anay Mehrotra, Manolis Zampetakis, Felix Zhou, Ziyu Zhu
International Conference on Learning Representations ICLR 2026
21. *Characterizations of Language Generation with Breadth: The Interplay Between Hallucinations, Coverage, and Stability*
Alkis Kalavasis, Anay Mehrotra, Grigoris Velegkas
International Conference on Algorithmic Learning Theory ALT 2026
20. *On Union-Closedness of Language Generation*
Steve Hanneke, Amin Karbasi, Anay Mehrotra, Grigoris Velegkas
Conference on Advances in Neural Information Processing Systems NeurIPS 2025
19. *What Makes Treatment Effects Identifiable? Characterizations and Estimators Beyond Unconfoundedness*
Yang Cai, Alkis Kalavasis, Katerina Mamali, Anay Mehrotra, Manolis Zampetakis
Conference on Learning Theory COLT 2025
Best Paper Award at COLT (2025) and **Invited Presentation** at IJCAI (2026)
18. *On the Limits of Language Generation: Trade-Offs Between Hallucination and Mode Collapse*
Alkis Kalavasis, Anay Mehrotra, Grigoris Velegkas
ACM Symposium on Theory of Computing STOC 2025
Highlights Track at Foundations of Responsible Computing FORC (2026)

ACM Conference on Fairness, Accountability, and Transparency FAccT 2021

▸ *Other Presentations: Workshop on Co-Development of Computer Science and Law, 2020*

3. *The Effect of the Rooney Rule on Implicit Bias in the Long Term*

L. Elisa Celis, Chris Hays, Anay Mehrotra, Nisheeth K. Vishnoi
ACM Conference on Fairness, Accountability, and Transparency FAccT 2021

2. *Interventions for Ranking in the Presence of Implicit Bias*

L. Elisa Celis, Anay Mehrotra, Nisheeth K. Vishnoi
ACM Conference on Fairness, Accountability, and Transparency FAccT 2020

1. *Toward Controlling Discrimination in Online Ad Auctions*

L. Elisa Celis, Anay Mehrotra, Nisheeth K. Vishnoi
International Conference on Machine Learning ICML 2019
Best Student Paper at Mechanism Design for Social Good MD4SG (2019)

PREPRINTS

In my field, authors are listed alphabetically; in exceptions, equal contributors are marked with "★"

6. *Surprises in Proper Positive-Only Learning*

Shai Ben-David, Farnam Mansouri, Anay Mehrotra, Manolis Zampetakis 2026

5. *Reasoning with Sampling: Cutting at Decision Points*

Felix Zhou (★), Anay Mehrotra, Quanquan C. Liu 2026

4. *Improved Guarantees for Heterogeneous Treatment Effect Estimation via Matrix Completion*

Anay Mehrotra, Phuc Tran, Van H. Vu, Manolis Zampetakis 2026

3. *On Language Generation in the Limit with Bounded Memory*

Jon Kleinberg, Anay Mehrotra, Grigoris Velegkas, Amin Saberi 2026

2. *What is Learnable in Valiant's Theory of the Learnable?*

Steve Hanneke, Anay Mehrotra, Grigoris Velegkas, Manolis Zampetakis 2026

1. *Learning-Enabled Estimation: Tight Characterizations under Sample Selection Biases*

Vikram Kher, Jane H. Lee, Anay Mehrotra, Manolis Zampetakis 2026

PATENTS

Systems and Methods for Jailbreaking Black-Box Large Language Models

Anay Mehrotra, Paul Kassianik US20250181836A1

Filed 2024 (Pending)

TEACHING AND MENTORING

Mentoring

▸ *Alexandros Kouridakis* (Ph.D., UT Austin; Publication [25](#)) 2025 – 2026

▸ *Ziyu Zhu* (Undergrad; Publication [22](#); Yale → IMC Trading) 2024

Teaching Experience

▸ Lecturer and Problem Setter, Yale ICPC Club 2024 – 2025

▸ Teaching Fellow, Yale Computer Science: Algorithms; Convex Optimization 2021, 2023

INDUSTRY EXPERIENCE

Quantitative Researcher Intern, Two Sigma, New York Summer 2024

Algorithm Developer Intern, Hudson River Trading, New York Summer 2023

RECENT TALKS (LAST THREE YEARS)

“Smoothed Analysis of Learning from Positive Examples and Applications”

- Stanford Theory Lunch Stanford, April 2026
- MIT A&C Seminar Cambridge, April 2025
- Yale Theory Student Seminar New Haven, March 2025

“An Overview of Jailbreaking LLMs” in the Graduate-Level Course “Special Topics in AI” (IIT Delhi) March 2026

“Language Generation with Infinite Contamination” at Language Generation Day @ Stanford March 2026

“Learning Theory in the Wild: Foundations of Missing Data and Generation” at Google Research February 2026

“What Makes Treatment Effects Identifiable? Characterizations and Estimators Beyond Unconfoundedness”

- Foundations of Replicability and Verifiability in AI (EnCORE workshop) San Diego, April 2026
- Foundations of Data Science (FDS) Colloquium New Haven, September 2025
- Bangalore Theory Seminars at IISc Bengaluru, August 2025
- Conference on Learning Theory (COLT) Lyon, July 2025

“On the Limits of Language Generation”

- Microsoft Research India Bengaluru, August 2025
- Theory Seminar at EPFL Lausanne, July 2025
- COLT Tutorial on Language Generation in the Limit Lyon, July 2025
- Yale NLP Reading Group New Haven, June 2025
- Yang Reading Group New Haven, May 2025
- Nanjing University Online, April 2025
- Columbia University Theory Student Seminar NYC, March 2025

“Efficient Statistics With Unknown Truncation, Polynomial Time Algorithms, Beyond Gaussians”

- Foundations of Computer Science (FOCS) Chicago, October 2024
- Bangalore Theory Seminars at IISc Bengaluru, August 2024
- Yale Theory Student Seminar New Haven, April 2024

“Tree of Attacks: Jailbreaking Black-Box LLMs” at Lunch Seminar at Robust Intelligence Online, May 2024

“Can SGD Select Good Fishermen?” at Yale Theory Student Seminar New Haven, November 2024

SELECTED ACADEMIC SERVICE

Co-Organizer for the following events:

- STOC Workshop: Understanding Large Language Models via a Theoretical Lens June 2026
- Stanford Workshop: Language Generation Day at Stanford [\[Website\]](#) March 2026
- Stanford Reading Group: Rethinking Foundations of Real-world ML [\[Website\]](#) January 2026 – Present
- NeurIPS Workshop: Reliable ML from Unreliable Data at NeurIPS’25 [\[Website\]](#) December 2025
- COLT Tutorial: Language Generation at COLT’25 [\[Website\]](#) June 2025

Program Committee or Area Chair: NeurIPS (2026), COLT (2025, 2026), EC 2026, ACM FAccT (2022, 2023, 2024, 2025)

Reviewer: FOCS (2026, 2024), NeurIPS (2021, 2023, 2024, 2025), ICML (2022, 2025, 2026), ITCS (2024, 2026), AISTATS 2025, STACS 2025, EAAMO 2024, JAIR 2025, Electronic Journal of Statistics 2026

EXTRACURRICULAR ACHIEVEMENTS

Athletics: 5 km under 20 min (2026); Lausanne Half-Marathon (2018)

Fine Arts: Exhibited at All India Fine Art & Crafts Society (2016); Diploma in Fine Arts (2015)