ANAY MEHROTRA

anaymehrotra.com ◇ anaymehrotra1@gmail.com

EDUCATION

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

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

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

RESEARCH EXPERIENCE

Motwani Postdoctoral Fellow, Stanford University January 2026 – Present
Hosted by Amin Saberi; co-organizing reading group on foundations of machine learning

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)

HONORS AND AWARDS

- *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 Paper 1 2025
- *Finalist,* ICPC World Finals (Rank 33) 2021
- *Winner,* ICPC Asia West Continent Championship, spanning eight countries 2020
- *Winner,* At both at ICPC Gwalior–Pune and Kharagpur Regionals (twice in) 2019
- *Sri Binay Kumar Sinha Award,* IIT Kanpur for research that addresses a societal problem 2020
- *Academic Excellence Awards,* For all semesters at IIT Kanpur 2016 – 2020
- *Best Student Paper,* Workshop on Mechanism Design for Social Good for Publication 1 2019
- *KVPY Scholar,* A national fellowship by the Department of Science and Technology, India 2015

SELECTED PUBLICATIONS

In my field, authors are typically listed alphabetically

Smoothed Analysis of Learning from Positive Samples STOC 2026
Jane H. Lee, Anay Mehrotra, Manolis Zampetakis

What Makes Treatment Effects Identifiable? Characterizations and Estimators Beyond Unconfoundedness COLT 2025
Yang Cai, Alkis Kalavasis, Katerina Mamali, Anay Mehrotra, Manolis Zampetakis
Best Paper Award at COLT (2025) and **Invited Presentation** at IJCAI (2026)

DDPM Score Matching and Distribution Learning Preprint 2025
Sinho Chewi, Alkis Kalavasis, Anay Mehrotra, Omar Montasser
Best Short Paper at ICLR'25 Workshop on Deep Generative Models in ML

On the Limits of Language Generation: Trade-Offs Between Hallucination and Mode Collapse STOC 2025
Alkis Kalavasis, Anay Mehrotra, Grigoris Velegkas
Highlights Track at Foundations of Responsible Computing FORC (2026)

Tree of Attacks: Jailbreaking Black-Box LLMs Automatically NeurIPS 2024
Anay Mehrotra, Manolis Zampetakis, Paul K., B. Nelson, Hyrum Anderson, Yaron Singer, Amin Karbasi
Covered by WIRED and TechCrunch

Toward Controlling Discrimination in Online Ad Auctions ICML 2019
L. Elisa Celis, Anay Mehrotra, Nisheeth K. Vishnoi
Best Student Paper at Mechanism Design for Social Good MD4SG (2019)

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)

17. *Tree of Attacks: Jailbreaking Black-Box LLMs Automatically*
Anay Mehrotra (⋆), Manolis Zampetakis, Paul K., B. Nelson, Hyrum Anderson, Yaron Singer, Amin Karbasi
Conference on Advances in Neural Information Processing Systems NeurIPS 2024
Covered by WIRED and TechCrunch

16. *Efficient Statistics With Unknown Truncation, Polynomial Time Algorithms, Beyond Gaussians*
Jane H. Lee, Anay Mehrotra, Manolis Zampetakis
IEEE Symposium on Foundations of Computer Science FOCS 2024

15. *Smaller Confidence Intervals From IPW Estimators via Data-Dependent Coarsening*
Alkis Kalavasis, Anay Mehrotra, Manolis Zampetakis
Conference on Learning Theory COLT 2024

14. *Fair Classification with Partial Feedback: An Exploration-Based Data-Collection Approach*
Vijay Keswani (⋆), Anay Mehrotra (⋆), L. Elisa Celis
International Conference on Machine Learning ICML 2024
▹ *Other Presentations: ACM Conference on Equity and Access in Algorithms, Mechanisms & Optimization'24*

13. *Bias in Evaluation Processes: An Optimization-Based Model*
L. Elisa Celis, Amit Kumar, Anay Mehrotra, Nisheeth K. Vishnoi
Conference on Advances in Neural Information Processing Systems NeurIPS 2023

12. *Sampling Individually-Fair Rankings that are Always Group Fair*
Sruthi Gorantla (⋆), Anay Mehrotra (⋆), Amit Deshpande, Anand Louis
ACM Conference on Artificial Intelligence, Ethics, and Society AIES 2023

11. *Subset Selection Based On Multiple Rankings in the Presence of Bias: Effectiveness of Fairness Constraints for Multiwinner Voting Score Functions*

Niclas Boehmer, L. Elisa Celis, Lingxiao Huang, Anay Mehrotra, Nisheeth K. Vishnoi
International Conference on Machine Learning ICML 2023

10. *Maximizing Submodular Functions for Recommendation in the Presence of Biases*
Anay Mehrotra and Nisheeth K. Vishnoi
ACM Web Conference WWW 2023

9. *Fair Ranking with Noisy Protected Attributes*
Anay Mehrotra and Nisheeth K. Vishnoi
Conference on Advances in Neural Information Processing Systems NeurIPS 2022

8. *Revisiting Group Fairness Metrics: The Effect of Networks*
Anay Mehrotra (⋆), Jeffrey Sachs (⋆), L. Elisa Celis
ACM Conference on Computer-Supported Cooperative Work & Social Computing CSCW 2022

7. *Selection in the Presence of Implicit Bias: The Advantage of Intersectional Constraints*
Anay Mehrotra, Bary S. R. Pradelski, Nisheeth K. Vishnoi
ACM Conference on Fairness, Accountability, and Transparency FAccT 2022

6. *Fairness for AUC via Feature Augmentation*
Hortense Fong, Vineet Kumar, Anay Mehrotra, Nisheeth K. Vishnoi
ACM Conference on Fairness, Accountability, and Transparency FAccT 2022

5. *Fair Classification with Adversarial Perturbations*
L. Elisa Celis, Anay Mehrotra, Nisheeth K. Vishnoi
Conference on Advances in Neural Information Processing Systems NeurIPS 2021

4. *Mitigating Bias in Set Selection with Noisy Protected Attributes*
Anay Mehrotra (⋆) and L. Elisa Celis
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. *Differentially Private Language Generation in the Limit*
Anay Mehrotra, Grigoris Velegkas, Xifan Yu, Felix Zhou 2026
   - ▹ *Other Presentations: Theory and Practice of Differential Privacy (TPDP), 2026*
   - ▹ *Other Presentations: First Workshop on Formal Languages and Neural Networks (FLaNN), 2026*

5. *Language Generation with Feedback: Queries and Mistakes*
Steve Hanneke, Amin Karbasi, Anay Mehrotra, Grigoris Velegkas 2026

4. *Linear Regression with Unknown Truncation Beyond Gaussian Features*
Alexandros Kouridakis, Anay Mehrotra, Alkis Kalavasis, Constantine Caramanis 2026

3. *Language Generation with Infinite Contamination*
   Anay Mehrotra, Grigoris Velegkas, Xifan Yu, Felix Zhou 2025
   - ▹ *Other Presentations: First Workshop on Formal Languages and Neural Networks (FLaNN), 2026*
2. *Can SGD Select Good Fishermen? Local Convergence Under Self-Selection Biases*
   Alkis Kalavasis, Anay Mehrotra, Felix Zhou 2025
1. *DDPM Score Matching and Distribution Learning*
   Sinho Chewi, Alkis Kalavasis, Anay Mehrotra, Omar Montasser 2025
   Best Short Paper at ICLR'25 Workshop on Deep Generative Models in ML

PATENTS

Systems and Methods for Jailbreaking Black-Box Large Language Models US20250181836A1
Anay Mehrotra, Paul Kassianik Filed 2024 (Pending)

INDUSTRY EXPERIENCE

Quantitative Researcher Intern, Two Sigma, New York Summer 2024
Algorithm Developer Intern, Hudson River Trading, New York Summer 2023

TEACHING AND MENTORING

Mentoring

- ▹ Alexandros Kouridakis (Ph.D., UT Austin) November 2024 – January 2026
  Preprint 4, first algorithm for linear regression with unknown truncation with non-gaussian features
- ▹ Ziyu Zhu (Undergraduate, Yale'24 → IMC Trading) Spring 2024
  Senior thesis on "Learning with Friction" (Thesis's extension published in ICLR 2026; Publication 22)

Teaching Experience

Lecturer, Yale ICPC Club 2025
Problem Setter, Yale ICPC Club 2024 – 2025
Teaching Fellow, Yale Computer Science:

- ▹ Algorithms (CPSC 365/ECON 365) – 120+ undergraduates Fall 2023
- ▹ Algorithms via Continuous Optimization (CPSC 368/516) – Graduate course Spring 2023
- ▹ Algorithms via Continuous Optimization (CPSC 463/563) – Graduate course Fall 2021

Academic Mentor, IIT Kanpur – Tutored 50+ students in Electrodynamics 2017 – 2018

RECENT TALKS (LAST THREE YEARS)

"Smoothed Analysis of Learning from Positive Examples and Applications"

- ▹ Stanford Theory Lunch Stanford, April 2026

"An Overview of Jailbreaking LLMs"

- ▹ Guest Lecture in the Graduate-Level Course "Special Topics in AI" IIT Delhi, March 2026

"Language Generation in Presence of Contamination"

- ▹ Language Generation Day @ Stanford Stanford, March 2026

"Learning Theory in the Wild: Foundations of Missing Data and Generation"

- ▹ Google Research Mountain View, 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

"Learning with Positive and Imperfect Unlabeled Data"

- MIT A&C Seminar — Cambridge, April 2025
- Yale Theory Student Seminar — New Haven, 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"

- Lunch Seminar at Robust Intelligence — Online, May 2024

"Can SGD Select Good Fishermen?"

- 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

Leadership: President, Robotics Club, Delhi Public School (hosted national robotics event; 200+ participants) 2015

EXTRACURRICULAR ACHIEVEMENTS

Athletics: 5 km under 20 min (2026); Lausanne Half-Marathon (2018)
Fine Arts: Exhibited at All India Fine Art Society (2016); Diploma in Fine Arts (2015)