Noam Razin

Postdoctoral Fellow
Princeton Language and Intelligence, Princeton University

Email: noamrazin@princeton.edu
Homepage: noamrazin.github.io

Research Interests

Fundamentals of Artificial Intelligence, Alignment, Deep Learning Theory, Tensor Analysis

Academic Positions

2024–present **Postdoctoral Fellow**, *Princeton Language and Intelligence, Princeton University*, NJ, USA
Host: Prof. Sanjeev Arora

Education

2019–2024 **PhD in Computer Science**, *Tel Aviv University*, Israel
Advisor: Prof. Nadav Cohen
Thesis: Understanding Deep Learning via Notions of Rank

2015–2018 **BSc in Computer Science**, *The Hebrew University of Jerusalem*, Israel
Graduated summa cum laude
GPA 99.2/100 (rank 1 of 283 in class of 2018)

Industry Positions

2023–2023 **Research Intern**, *Apple Machine Learning Research*, Cupertino, CA, USA
Optimization pitfalls in reinforcement finetuning of language models

2019–2020 **Research Intern**, *Microsoft*, Israel
Deep learning for visual and textual content-based product recommendations

2017–2019 **Software Engineer**, *Microsoft*, Israel
Distributed systems for servicing product recommendations in Microsoft stores

2013–2015 **Software Engineer**, *Jive*, Israel
Server-side development of a scalable gamification module for a work collaboration tool

Awards and Honors

2025 Best paper runner-up award, Reliable ML from Unreliable Data Workshop, NeurIPS 2025
2024 Israeli Council for Higher Education (VATAT) Fellowship for Oustanding Postdoctoral Scholars
2024 Zuckerman Postdoctoral Scholarship
2023 Deutsch Prize in Computer Science for PhD candidates
2022 Apple Scholars in AI/ML PhD Fellowship (one of 15 recipients worldwide)
2021 Tel Aviv University Center for AI and Data Science Excellence Fellowship
2018 Summa cum laude graduate of The Hebrew University of Jerusalem BSc Computer Science
2018 The Hebrew University of Jerusalem dean's honor list
2016–2017 The Hebrew University of Jerusalem rector's honor list
2015–2018 Amirim honors program for outstanding undergraduate students

Publications

[*] denotes equal contribution

Conference Proceedings

1. **Why is Your Language Model a Poor Implicit Reward Model?**
   Noam Razin, Yong Lin, Jiarui Yao, Sanjeev Arora
   International Conference on Learning Representations (ICLR), 2026
   Best paper runner-up, Reliable ML from Unreliable Data Workshop, NeurIPS 2025

2. **What Makes a Reward Model a Good Teacher? An Optimization Perspective**
   Noam Razin, Zixuan Wang, Hubert Strauss, Stanley Wei, Jason D. Lee, Sanjeev Arora
   Conference on Neural Information Processing Systems (NeurIPS), 2025

3. **The Implicit Bias of Structured State Space Models Can Be Poisoned With Clean Labels**
   Yonatan Slutzky*, Yotam Alexander*, **Noam Razin**, Nadav Cohen
   Conference on Neural Information Processing Systems (NeurIPS), 2025

4. **Unintentional Unalignment: Likelihood Displacement in Direct Preference Optimization**
   Noam Razin, Sadhika Malladi, Adithya Bhaskar, Danqi Chen, Sanjeev Arora, Boris Hanin
   International Conference on Learning Representations (ICLR), 2025

5. **Implicit Bias of Policy Gradient in Linear Quadratic Control: Extrapolation to Unseen Initial States**
   Noam Razin*, Yotam Alexander*, Edo Cohen-Karlik, Raja Giryes, Amir Globerson, Nadav Cohen
   International Conference on Machine Learning (ICML), 2024

6. **Vanishing Gradients in Reinforcement Finetuning of Language Models**
   Noam Razin, Hattie Zhou, Omid Saremi, Vimal Thilak, Arwen Bradley, Preetum Nakkiran, Joshua M. Susskind, Etai Littwin
   International Conference on Learning Representations (ICLR), 2024

7. **What Algorithms Can Transformers Learn? A Study in Length Generalization**
   Hattie Zhou, Arwen Bradley, Etai Littwin, **Noam Razin**, Omid Saremi, Joshua M. Susskind, Samy Bengio, Preetum Nakkiran
   International Conference on Learning Representations (ICLR), 2024

8. **What Makes Data Suitable for a Locally Connected Neural Network? A Necessary and Sufficient Condition Based on Quantum Entanglement**
   Yotam Alexander*, Nimrod De La Vega*, **Noam Razin**, Nadav Cohen
   Conference on Neural Information Processing Systems (NeurIPS), 2023

9. **On the Ability of Graph Neural Networks to Model Interactions Between Vertices**
   Noam Razin, Tom Verbin, Nadav Cohen
   Conference on Neural Information Processing Systems (NeurIPS), 2023

10. **Implicit Regularization in Hierarchical Tensor Factorization and Deep Convolutional Neural Networks**
    Noam Razin, Asaf Maman, Nadav Cohen
    International Conference on Machine Learning (ICML), 2022

11. **Implicit Regularization in Tensor Factorization**
    Noam Razin*, Asaf Maman*, Nadav Cohen
    International Conference on Machine Learning (ICML), 2021

12. **Implicit Regularization in Deep Learning May Not Be Explainable by Norms**
    Noam Razin, Nadav Cohen
    Conference on Neural Information Processing Systems (NeurIPS), 2020

13. **RecoBERT: A Catalog Language Model for Text-Based Recommendations**

Itzik Malkiel, Oren Barkan, Avi Caciularu, **Noam Razin**, Ori Katz, Noam Koenigstein
Findings of the Association for Computational Linguistics: EMNLP, 2020

14. **Scalable Attentive Sentence-Pair Modeling via Distilled Sentence Embedding**
Oren Barkan*, **Noam Razin***, Itzik Malkiel, Ori Katz, Avi Caciularu, Noam Koenigstein
AAAI Conference on Artificial Intelligence (AAAI), 2020

Preprints

1. **Retaining by Doing: The Role of On-Policy Data in Mitigating Forgetting**
Howard Chen, **Noam Razin**, Karthik Narasimhan, Danqi Chen
Preprint, 2025

Miscellaneous

1. **Lecture Notes on Linear Neural Networks: A Tale of Optimization and Generalization in Deep Learning**
Nadav Cohen, **Noam Razin**
Lecture Notes, 2024

2. **Understanding Deep Learning via Notions of Rank**
Noam Razin
PhD Thesis, 2024

Patents

1. Machine Learning Multiple Features of Depicted Item
Oren Barkan, Noam Razin, Noam Koenigstein, Roy Hirsch, Nir Nice
US Patent 16725652, 2022

2. Searching Using Changed Feature of Viewed Item
Oren Barkan, Noam Razin, Roy Hirsch, Noam Koenigstein, Nir Nice
US Patent 16725461, 2021

3. Sentence Similarity Scoring Using Neural Network Distillation
Oren Barkan, Noam Razin, Noam Koenigstein
US Patent 16789385, 2021

Invited Talks

2026

One World MINDS Seminar (virtual)

2025

Microsoft Research New York City Lab Seminar (virtual)

Conference on Neural Information Processing Systems (NeurIPS), Reliable Machine Learning from Unreliable Data Workshop, San Diego, CA, USA

Conference on Neural Information Processing Systems (NeurIPS), Aligning Reinforcement Learning Experimentalists and Theorists Workshop, San Diego, CA, USA

University of Pennsylvania Machine Learning Seminar, Philadelphia, PA, USA

EPFL AI Fundamentals Seminar, Lausanne, Switzerland

Ludwig Maximilian University of Munich AI Seminar, Munich, Germany

MPI MiS + UCLA Math Machine Learning Seminar, Los Angeles, CA, USA (virtual)

Intel Labs Seminar (virtual)

Princeton Language and Intelligence Seminar, Princeton, NJ, USA

Deep Learning: Classics and Trends Seminar (virtual)

2024

Apple Machine Learning Seminar (virtual)

Fundamental AI Research at Meta Seminar (virtual)

Flatiron Institute Machine Learning Seminar, New York, NY, USA

Princeton Language and Intelligence Seminar, Princeton, NJ, USA

EUROPT Conference on Advances in Continuous Optimization, Lund, Sweden

Oberwolfach Applied Harmonic Analysis and Data Science Workshop, Oberwolfach, Germany

Bosch Research, Haifa, Israel

Technion Machine Learning Seminar, Haifa, Israel

MPI MiS + UCLA Math Machine Learning Seminar, Los Angeles, CA, USA (virtual)

Tel Aviv University Natural Language Processing Seminar, Tel Aviv, Israel

EPFL Foundations of Learning and AI Research Seminar, Lausanne, Switzerland

Deep Learning: Classics and Trends Seminar (virtual)

2023

New York University Machine Learning Group Meeting, New York, NY, USA

Princeton ALG-ML Seminar, Princeton, NJ, USA

Apple Machine Learning Seminar (virtual)

The Hebrew University of Jerusalem Machine Learning Seminar, Jerusalem, Israel

Complex Network Analysis group at NCSR Demokritos, Athens, Greece (virtual)

Learning on Graphs and Geometry Reading Group (virtual)

Technion Machine Learning Seminar, Haifa, Israel (virtual)

2022

Apple Siri Natural Language Processing Reading Group (virtual)

ICTP Youth in High-Dimensions Conference, Trieste, Italy (virtual)

MPI MiS + UCLA Math Machine Learning Seminar, Los Angeles, CA, USA (virtual)

Ludwig Maximilian University of Munich Mathematical Foundations of Artificial Intelligence Seminar, Munich, Germany (virtual)

Caltech Machine Learning Group Meeting, Pasadena, CA, USA

New York University Machine Learning Group Meeting, New York, NY, USA

Harvard Machine Learning Group Meeting, Cambridge, MA, USA

Princeton ALG-ML Seminar, Princeton, NJ, USA

Apple Machine Learning Seminar (virtual)

2021

Oberwolfach Applied Harmonic Analysis and Data Science Workshop, Germany (virtual)

RWTH Aachen Mathematics of Information Processing Seminar, Aachen, Germany (virtual)

The Hebrew University of Jerusalem Machine Learning Seminar, Jerusalem, Israel

Caltech Machine Learning Group Meeting, Pasadena, CA, USA (virtual)

Princeton ALG-ML Seminar, Princeton, NJ, USA (virtual)

Technion Machine Learning Seminar, Haifa, Israel (virtual)

2020

Technion Machine Learning Seminar, Haifa, Israel (virtual)

Princeton Theoretical Machine Learning Reading Group, Princeton, NJ, USA (virtual)

Tel Aviv University Machine Learning Seminar, Tel Aviv, Israel (virtual)

Teaching Experience

2025–2026 Guest Lecturer, Fundamentals of Deep Learning (COS 514), Princeton University, NJ, USA

2024–2025 Guest Lecturer, Introduction to Reinforcement Learning (COS 435), Princeton University, NJ, USA

2021–2024 Guest Lecturer, First Steps in Research for Excellent Students, Tel Aviv University, Israel

2021–2023 Teaching Assistant, Foundations of Deep Learning, Tel Aviv University, Israel

Academic Service

Reviewer

International Conference on Machine Learning (ICML)

Conference on Neural Information Processing Systems (NeurIPS)

International Conference on Learning Representations (ICLR)

Journal of Machine Learning Research (JMLR)

Foundations of Computational Mathematics (FoCM)