

Nikhil Prakash

Khoury College of Computer Sciences
Northeastern University
Boston, MA, USA
✉ prakash.nik@northeastern.edu
🌐 nix07.github.io
🗝 kUfq-fEAAAAAJ

Research Interests

I'm interested in understanding the cognitive abilities of deep neural networks, such as reasoning and theory of mind, as well as exploring their downstream applications.

Education

- Sept 2022 – **Ph.D. & Masters**, Northeastern University, Boston, MA.
May 2027* Computer Science, Bau Lab (advisor: Prof. David Bau )
GPA: 4/4
- Aug 2016 – **Bachelor of Eng.**, RV College of Eng., Bangalore, India
Aug 2020 Department of Telecommunication
Cumulative GPA: 8.31/10 (First Class with Distinction)

Peer-Reviewed Publications

Conference publications

- ICML 2026 **The Dual Mechanisms of Spatial Reasoning in Vision–Language Models**
(submitted) Kelly Cui, Nikhil Prakash, Ayush Raina, David Bau, Antonio Torralba, Tamar Rott Shaham
- ACL 2026 **CircuitTuning: Improving Math Reasoning in LLMs via Targeted Sub-Network Updates**
(submitted) Nikhil Prakash, Donghao Ren, Dominik Moritz, Yannick Assogba.
- ICLR 2026 **Language Models use Lookbacks to Track Beliefs**
Nikhil Prakash, Natalie Shapira, Arnab Sen Sharma, Christoph Riedl, Yonatan Belinkov, Tamar Rott Shaham, David Bau, Atticus Geiger.

- ICML 2025 **MIB: A Mechanistic Interpretability Benchmark**
Aaron Mueller, Atticus Geiger, Sarah Wiegreffe, Dana Arad, Iván Aruschin, Adam Belfki, Yik Siu Chan, Jaden Fried Fiotto-Kaufman, Tal Haklay, Michael Hanna, Jing Huang, Rohan Gupta, Yaniv Nikankin, Hadas Orgad, Nikhil Prakash, Anja Reusch, Aruna Sankaranarayanan, Shun Shao, Alessandro Stolfo, Martin Tutek, Amir Zur, David Bau, Yonatan Belinkov.
- ICLR 2025 **NNsight and NDIF: Democratizing Access to Foundation Model Internals**
Jaden Fiotto-Kaufman, Alexander R Loftus, Eric Todd, Jannik Brinkmann, Caden Juang, Koyena Pal, Can Rager, Aaron Mueller, Samuel Marks, Arnab Sen Sharma, Francesca Lucchetti, Michael Ripa, Adam Belfki, Nikhil Prakash, Sumeet Multani, Carla Brodley, Arjun Guha, Jonathan Bell, Byron Wallace, David Bau.
- ICLR 2024 **Fine-Tuning Enhances Existing Mechanisms: A Case Study on Entity Tracking**
Nikhil Prakash, Tamar Rott Shaham, Tal Haklay, Yonatan Belinkov, David Bau
- IUI 2023 **Supporting Requesters in Writing Clear Crowdsourcing Task Descriptions Through Computational Flaw Assessment**
Zahra Nouri, Nikhil Prakash, Ujwal Gadiraju, Henning Wachsmuth
[Journal publications](#)
- Computational Linguistics **The Quest for the Right Mediator: A History, Survey, and Theoretical Grounding of Causal Interpretability**
Aaron Mueller, Jannik Brinkmann, Millicent Li, Samuel Marks, Koyena Pal, Nikhil Prakash, Can Rager, Aruna Sankaranarayanan, Arnab Sen Sharma, Jiuding Sun, Eric Todd, David Bau, Yonatan Belinkov
[Workshop and demonstration publications](#)
- NeurIPS 2025 **Language Models use Lookbacks to Track Beliefs**
Nikhil Prakash, Natalie Shapira, Arnab Sen Sharma, Christoph Riedl, Yonatan Belinkov, Tamar Rott Shaham, David Bau, Atticus Geiger
First Workshop on CogInterp: Interpreting Cognition in Deep Learning Models
Mechanistic Interpretability Workshop at NeurIPS 2025

ICML 2023	Discovering Variable Binding Circuitry with Desiderata Xander Davies*, Max Nadeau*, <u>Nikhil Prakash*</u> , Tamar Rott Sham, David Bau Workshop on Challenges in Deployable Generative AI.
HCOMP 2021	iClarify – A Tool to Help Requesters Iteratively Improve Task Descriptions in Crowdsourcing Zahra Nouri, <u>Nikhil Prakash</u> , Ujwal Gadiraju, Henning Wachsmuth Work-in-Progress & Demonstration, Human Computation and Crowdsourcing 2021.
NeurIPS 2020	Conceptualization and Framework of Hybrid Intelligence Systems <u>Nikhil Prakash</u> and Kory W. Mathewson Human And Machine in-the-Loop Evaluation and Learning Strategies Workshop
COMSNETS 2019	A Grid-based Model for Generating Scale-Free Networks Amit Kumar Verma and <u>Nikhil Prakash</u> Social Networking Workshop

Work Experience

May–August 2025	Apple Inc., Machine Learning Intern Mechanistically investigating reasoning capabilities of language models, mentored by Yannick Assogba  .
July–August 2024	Practical AI Alignment and Interpretability Research (Pr(Ai)²R) Group, Research Intern Mechanistically investigating <i>theory of mind</i> capabilities in language models, mentored by Dr. Atticus Geiger  .
June–July 2023	Stanford Existential Risks Initiative ML Alignment Theory Scholars (SERI-MATS) Investigated superposition in attention heads of GPT2-XL that retrieve factual information from subject token residual stream, mentored by Neel Nanda  .
Jan–Aug 2022	Max Planck Institute for Security and Privacy (MPI-SP), Visiting Scholar Worked under the supervision of Prof. Asia J. Biega  to conduct exploratory user studies for investigating well-being of microtask crowdworkers.
May 2021–Sept 2022	Delft University of Technology (TU Delft), Research Assistance (Voluntary) Worked under the supervision of Prof. Ujwal Gadiraju  and Prof. Henning Wachsmuth  , to develop an assistive tool that would help crowdsourcing requesters in creating clear task descriptions quickly.

- July 2020–Jan 2022 **Accolite Software India**, *Senior Software Engineer*
Worked as full-stack engineer to design, develop, and maintain complex production-ready web applications. Primary technology stack was Angular, Java, and IBM DB2.
- Jan–May 2020 & June–Aug 2019 **Korea Advanced Institute of Science and Technology (KAIST)**, *Visiting Student Researcher*
Worked under the supervision of Prof. Juho Kim  to develop an intelligent interactive system for personalizing food recipes based on users' preferences and dietary constraints.
- May–June 2019 **Samsung R&D Institute India - Bangalore**, *Student Trainee*
Worked as a student trainee at Samsung Research, Bangalore (SRIB) in the Service PF team to design and develop the offline newsfeed functionality in the Samsung Internet.
- June–July 2018 **Indian Institute of Technology, Ropar (IIT Ropar)**, *Research Intern*
Worked under the supervision of Prof. Sudarshan Iyengar  to investigate the dynamics of collaboration process on English Wikipedia.

Invited Talks and Presentations

- 2025 Boston University NLP Group.
2025 Algorithms and Behavioral Science Coffee, MIT Economics.
2025 Apple AIML Visualization Team Meetings.
2025 Ploutos.dev.
2025 New England NLP 2025, Yale University.
2024 Practical AI Alignment and Interpretability Research Group.
2024 Computational Linguistics and Complex Social Networks Group at Indian Institute of Technology Gandhinagar.
2024 New England NLP 2024, Brown University.

Awards and Recognitions

- 2024 Top Reviewer at NeurIPS; Received complimentary registration, worth \$450.
2024 Google Gemma Academic Program GCP Credit Award, worth \$5k.
2022 Khoury College of Computer Sciences start-up fund, worth \$5k.
2020 NeurIPS complimentary registration (Travel Award), worth \$100.
2020 ACM CoDS-COMAD Student Travel Grant, worth \$80.

Voluntary Services

- 2026 Reviewer for ICML 2026, COLM 2026.

- 2025 Student Representative for Tenure Track Hiring Committee at Northeastern University.
- 2025 Volunteer for Code to PhD. Helping prospective PhD students with their applications.
- 2025 Reviewer for ICLR 2026.
- 2025 Reviewer for Mech Interp & CogInterp workshops at NeurIPS 2025.
- 2025 Reviewer for Visualization for AI Explainability workshop at IEEE VIS.
- 2025 Reviewer for BlackboxNLP at EMNLP 2025.
- 2025 Reviewer for XLLM-Reason-Plan workshop at COLM 2025.
- 2025 Reviewer for ICML 2025, COLM 2025, TMLR 2025, NeurIPS 2025 (main conference & workshop proposals).
- 2025 Reviewer for Actionable Interpretability Workshop at ICML 2025.
- 2024 Reviewer for ICLR 2025.
- 2024 Reviewer for Interpretable AI: Past, Present and Future, ATTRIB, and MINT workshops at NeurIPS 2024.
- 2024 Reviewer for NeurIPS 2024 (main conference & workshop proposals).
- 2024 Co-organizing Mechanistic Interpretability Social at ICLR 2024.
- 2023 Program committee member of ATTRIB 2023 workshop at NeurIPS.
- 2022 Student volunteer at Ph.D. admissions committee to review prospective Ph.D. student applications.
- 2019 Student volunteer at India HCI.

Engineering Skills

- Languages Python, C++, Java, JavaScript, HTML, CSS, SQL
- Technologies & Frameworks PyTorch, Transformers, NNsight, Transformer-lens, Scikit-learn, NLTK, Pandas, Numpy, Angular, Bootstrap, jQuery, Docker, Linux