Nikhil Prakash

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

Research Interests

I'm interested in understanding the internal mechanisms of deep neural networks to enhance human-AI collaboration and prevent misalignment.

Education

Sept 2022 –
May 2027*
Ph.D., *Northeastern University*, Boston, Massachusetts.
Computer Science, Bau Lab (advisor: Prof. David Bau 🌐)
GPA: 4/4

Aug 2016 –
Aug 2020
Bachelor of Eng., *RV College of Eng.*, Bangalore, India
Department of Telecommunication
Cumulative GPA: 8.31/10 (First Class with Distinction)

Peer-Reviewed Publications

Conference publications

ICLR 2024 **Fine-Tuning Enhances Existing Mechanisms: A Case Study on Entity Tracking**
Nikhil Prakash, Tamar Rott Shaham, Tal Haklay, Yonatan Belinkov, David Bau
International Conference on Learning Representations 2024.

IUI 2023 **Supporting Requesters in Writing Clear Crowdsourcing Task Descriptions Through Computational Flaw Assessment**
Zahra Nouri, Nikhil Prakash, Ujwal Gadiraju, Henning Wachsmuth
Intelligent User Interfaces 2023.

Journal publications

TMLR **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

ICML 2023 **Discovering Variable Binding Circuitry with Desiderata**
Xander Davies*, Max Nadeau*, Nikhil Prakash*, Tamar Rott Shaham, David Bau
Workshop on Challenges in Deployable Generative AI, International Conference on Machine Learning 2023.

HCOMP 2021	**iClarify – A Tool to Help Requesters Iteratively Improve Task Descriptions in Crowdsourcing** Zahra Nouri, Nikhil Prakash, Ujwal Gadiraju, Henning Wachsmuth Work-in-Progress & Demonstration, Human Computation and Crowdsourcing 2021.
NeurIPS 2020	**Conceptualization and Framework of Hybrid Intelligence Systems** Nikhil Prakash and Kory W. Mathewson Human And Machine in-the-Loop Evaluation and Learning Strategies Workshop, Neural Information Processing Systems 2020.
COMSNETS 2019	**A Grid-based Model for Generating Scale-Free Networks** Amit Kumar Verma and Nikhil Prakash Social Networking Workshop, Conference on Communication Systems & Networks 2019.

Other Publications

ArXiv 2024	**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.

Work Experience

July–August 2024	**Practical AI Alignment and Interpretability Research (Pr(Ai)²R) Group**, *Research Intern* Mechanistically investigating *theory of mind* capabilities in language models, mentored by 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 🌐.
June 2023	**Alignment Research Engineer Accelerator (ARENA)** Participated in this fully-funded upskilling initiative focused on technical AI alignment research.
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.
Sept 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

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.

Fellowships and Awards

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

2024	Reviewer for ICLR 2025.
2024	Reviewer for Interpretable AI: Past, Present and Future and ATTRIB 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