Danish Pruthi

Assistant Professor
Department of Computational Data Sciences (CDS)
Indian Institute of Science (IISc), Bangalore, 560012, India
Email: danishp@iisc.ac.in
https://danishpruthi.com/

Research Interests

I started and currently lead a research group working broadly in the area of Responsible AI. Specifically, we are interested in (a) detecting AI-generated content, and broadly, curbing unhealthy reliance on AI, (b) measuring and improving geo-cultural representation in AI, and (c) evaluating large language models, with an emphasis to enable responsible use.

Education

2016–2021	PhD in Language & Information Technologies Carnegie Mellon University Advisors: Graham Neubig & Zachary C. Lipton Thesis Committee: William W. Cohen & Michael Collins (+ advisors)
2016–2018	Masters in Language Technologies Carnegie Mellon University
2011–2015	B.E. (Hons.) in Computer Science Birla Institute of Technology and Science, Pilani (BITS Pilani)

Professional Experience

2022–2023	Applied Scientist, Amazon Web Services, Santa Clara
2020 Fall	Student Researcher, Google Research, Pittsburgh (Host: William W. Cohen)
2020 Summer	Research Intern, Google Research, Pittsburgh (Host: William W. Cohen)
2019 Summer	Research Intern, Facebook AI Research (FAIR), New York (Host: Brenden Lake)
2015–2016	Project Assistant, Indian Institute of Science (IISc), Bangalore (Host: Partha Talukdar)
2015 Spring	Research Intern, Microsoft Research, Bangalore
2014 Summer	Software Engineering Intern, Google, Hyderabad

Selected Awards

2025	BITS Pilani 30 Under 30 Award (to recognize exceptional young alumni globally)
2025	Outstanding Paper Award at ACL 2025
2023–2026	Schmidt Science AI2050 Early Career Fellowship
2023–2025	Pratiksha Trust Young Investigator Fellowship
2018–2019	CMU Presidential Fellowship
2017–2018	Siebel Scholarship
2019	Best Demo Runner-up Award at NAACL 2019
2011	KVPY Scholarship (declined to pursue engineering)

Publications

Total citations: 1800+, h-index: 17, i10-index: 23 (Google Scholar: `http://bit.ly/danish037`)

Refereed Journal Papers

[1] Assisting Human Decisions in Document Matching
Joon Sik Kim, Valerie Chen, Danish Pruthi, Nihar B. Shah, Ameet Talwalkar
Transactions on Machine Learning Research (TMLR, 2023)

[2] Evaluating Explanations: How much do explanations from the teacher aid students?
Danish Pruthi, Rachit Bansal, Bhuwan Dhingra, Livio Baldini Soares, Michael Collins,
Zachary C. Lipton, Graham Neubig, William W. Cohen
Transactions of the Association for Computational Linguistics (TACL, 2021)

Refereed Conference Papers

[1] TALES: A Taxonomy and Analysis of Cultural Representations in LLM-generated Stories
Kirti Bhagat, Shaily Bhatt, Athul Velagapudi, Aditya Vashistha, Shachi Dave, Danish Pruthi
ACM Conference on Human Factors in Computing Systems (CHI 2026)

[2] Beyond World Models: Rethinking Understanding in AI Models
Tarun Gupta, Danish Pruthi
AAAI Conference on Artificial Intelligence (AAAI 2026)

[3] All That Glitters is Not Novel: Plagiarism in AI Generated Research
Tarun Gupta, Danish Pruthi
Association for Computational Linguistics (ACL 2025)
Recipient of the **Outstanding Paper Award**

[4] FairI Tales: Evaluation of Fairness in Indian Contexts with a Focus on Bias and Stereotypes
Janki Atul Nawale*, Mohammed Safi Ur Rahman Khan*, Janani D, Mansi Gupta, Danish Pruthi, Mitesh M Khapra
Association for Computational Linguistics (ACL 2025)

[5] Silencing Empowerment, Allowing Bigotry: Auditing the Moderation of Hate Speech on Twitch
Prarabdh Shukla*, Wei Yin Chong*, Yash Patel*, Brennan Schaffner, Danish Pruthi, Arjun Bhagoji
Association for Computational Linguistics (ACL 2025)
Senior Area Chairs' Highlights (Top 3% of accepted papers)

[6] STAMP Your Content: Proving Dataset Membership via Watermarked Rephrasings
Saksham Rastogi, Pratyush Maini, Danish Pruthi
International Conference on Machine Learning (ICML 2025)

[7] Knowledge Graph Guided Evaluation of Abstention Techniques
Kinshuk Vasisht, Navreet Kaur, Danish Pruthi
Annual Conference of the Nations of the Americas Chapter of the ACL (NAACL, 2025)

[8] Richer Output for Richer Countries: Uncovering Geographical Disparities in Generated Stories and Travel Recommendations
Kirti Bhagat, Kinshuk Vasisht, Danish Pruthi
Findings of the Annual Conference of the Nations of the Americas Chapter of the ACL (NAACL, 2025)

[9] Revisiting the Robustness of Watermarking to Paraphrasing Attacks
Saksham Rastogi, Danish Pruthi
Conference on Empirical Methods in Natural Language Processing (EMNLP 2024)

[10] Performance Trade-offs of a Family of Text Watermarks
Anirudh Ajith, Sameer Singh, Danish Pruthi
Findings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2024)

[11] Evaluating Large Language Models for Health-related Queries with Presuppositions
Navreet Kaur, Monojit Choudhury, Danish Pruthi
Findings of the Association for Computational Linguistics (ACL, 2024)

[12] Goodhart's Law Applies to NLP's Explanation Benchmarks
Jennifer Hsia, Danish Pruthi, Aarti Singh, Zachary C. Lipton
Findings of European Chapter of the Association for Computational Linguistics (EACL, 2024)

[13] Model-tuning Via Prompts Makes NLP Models Adversarially Robust
Mrigank Raman, Pratyush Maini, Zico Kolter, Zachary C. Lipton, Danish Pruthi
Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)

[14] Geographical Erasure in Language Generation
Pola Schwöbel, Jacek Golebiowski, Michele Donini, Cedric Archambeau, Danish Pruthi
Findings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)

[15] Inspecting Geographical Representativeness of Images from Text-to-Image Models
Abhipsa Basu, R. Venkatesh Babu, Danish Pruthi
International Conference on Computer Vision (ICCV, 2023)

[16] Learning the Legibility of Visual Text Perturbations
Dev Seth, Rickard Stureborg, Danish Pruthi, Bhuwan Dhingra
European Chapter of the Association for Computational Linguistics (EACL, 2023)

[17] Learning to Scaffold: Optimizing Model Explanations for Teaching
Patrick Fernandes, Marcos Treviso, Danish Pruthi, André F. T. Martins, Graham Neubig
Conference on Neural Information Processing Systems (NeurIPS, 2022)

[18] Measures of Information Reflect Memorization Patterns
Rachit Bansal, Danish Pruthi, Yonatan Belinkov
Conference on Neural Information Processing Systems (NeurIPS, 2022)

[19] Explain, Edit, and Understand: Rethinking User Study Design for Evaluating Model Explanations
Siddhant Arora*, Danish Pruthi*, Norman Sadeh, William W. Cohen, Zachary C. Lipton, Graham Neubig
Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI, 2022)

[20] Do Context-Aware Translation Models Pay the Right Attention?
Kayo Yin, Patrick Fernandes, Danish Pruthi, Aditi Chaudhary, André F. T. Martins, Graham Neubig.
The Annual Meeting of the Association for Computational Linguistics (ACL, 2021)

[21] Weakly- and Semi-supervised Evidence Extraction
Danish Pruthi, Bhuwan Dhingra, Graham Neubig, Zachary C. Lipton
Findings of Empirical Methods in Natural Language Processing (EMNLP, 2020)

[22] Why and when should you pool? Analyzing Pooling in Recurrent Architectures
Pratyush Maini, Keshav Kolluru, Danish Pruthi, Mausam
Findings of Empirical Methods in Natural Language Processing (EMNLP, 2020)

[23] Learning to Deceive with Attention-Based Explanations
Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, Zachary C. Lipton
The Annual Meeting of the Association for Computational Linguistics (ACL, 2020)

[24] Combating Adversarial Misspellings with Robust Word Recognition
Danish Pruthi, Bhuwan Dhingra, Zachary C. Lipton
The Annual Meeting of the Association for Computational Linguistics (ACL, 2019)

[25] Simple and Effective Semi-Supervised Question Answering
Bhuwan Dhingra*, Danish Pruthi*, Dheeraj Rajagopal*
Meeting of the North American Chapter of the ACL (NAACL, 2018)

[26] SPINE: SParse Interpretable Neural Embeddings
Danish Pruthi*, Harsh Jhamtani*, Anant Subramanian*, Taylor Berg-Kirkpatrick, Eduard Hovy
AAAI Conference on Artificial Intelligence (AAAI, 2018)

[27] Discovering Response Eliciting Factors in Social Question Answering: A Reddit Inspired Study
Danish Pruthi, Yogesh Dahiya, Partha Talukdar
AAAI Conference on Web and Social Media (ICWSM, 2016)

[28] Maxxyt: An Autonomous Wearable Device for Real-time Tracking of a Wide Range of Exercises
Danish Pruthi, Ayush Jain, KrishnaMurthy Jatavallabhula, Ruppesh Nalwaya, and Puneet Teja
International Conference on Modelling and Simulation. (UKSim, 2015)

* denotes equal contribution

Refereed System Demonstrations

[1] NeuSpell: A Neural Spelling Correction Toolkit
Sai Muralidhar Jayanthi, Danish Pruthi, Graham Neubig
Conference on Empirical Methods in Natural Language Processing (EMNLP, 2020)

[2] `compare-mt`: A Tool for Holistic Comparison of Language Generation Systems
Graham Neubig, Zi-Yi Dou, Junjie Hu, Paul Michel, Danish Pruthi, Xinyi Wang, John Wieting
Meeting of the North American Chapter of the ACL (NAACL, 2019)
Recipient of the Best Demo Runner-up Award

Technical Reports

[1] What Can Natural Language Processing Do for Peer Review?
Ilia Kuznetsov, Osama Mohammed Afzal, Koen Dercksen, Nils Dycke, Alexander Goldberg, Tom Hope, Dirk Hovy, Jonathan K. Kummerfeld, Anne Lauscher, Kevin Leyton-Brown, Sheng Lu, Mausam, Margot Mieskes, Aurélie Névéol, Danish Pruthi, Lizhen Qu, Roy Schwartz, Noah A. Smith, Thamar Solorio, Jingyan Wang, Xiaodan Zhu, Anna Rogers, Nihar B. Shah, Iryna Gurevych
Technical Report. Dagstuhl Seminar 24052

Grants

As an individual Principal Investigator (PI):

[1] Google.org Research Grant I
(Amount: USD 70, 000 or INR 61.6 lakhs)

[2] Google.org Research Grant II
(Amount: USD 30, 000 or INR 27.2 lakhs)

[3] Microsoft Research Unrestricted Grant
(Amount: INR 20 lakhs)

[4] Adobe Research Award x 3
(Amount: USD 50, 000 or INR 43.4 lakhs)

[5] Schmidt Sciences AI2050 Early-career Fellowship
(Amount: USD 196, 350 or INR 1.64 Cr)

[6] Google Unrestricted Gift
(Amount: USD 25, 000 or INR 20.8 lakhs)

Contributed to preparation of the following grants:

[1] Expert-in-the-Loop Neural Summarization for Consequential Domains
National Science Foundation (NSF) Medium Research Award
(Amount: USD 583,746; PI: Zachary Lipton; Timeline: 2022–2026)

[2] Robustifying NLP by Exploiting Invariances Learned via Human Interaction
Facebook Research Award
(Amount: USD 80,000; PI: Zachary Lipton; Timeline: 2019–2020)

Teaching Experience

- Course Instructor for Introduction to NLP (DS 207)
  Course webpage: `https://danishpruthi.com/teaching/ds-207-jan-2025/`

- Course Instructor for Ethics in AI (DS 307)
  Course webpage: `https://danishpruthi.com/teaching/ds-307-aug-2024/`
- Teaching Assistant for Introduction to Machine Learning (PhD) (10-701)
  My responsibilities included conducting recitations, holding office hours, creating and grading asssignments.
- Competitive Programming Special Interest Group (CPSIG)
  Led the special interest group at BITS Pilani. Delivered lectures spanning data structures, algorithms, graph theory and game theory. Conducted similar workshops in sister campuses of BITS Goa and BITS Hyderabad.

Invited Talks & Panels

- Plagiarism in AI-Generated Research
  Keynote, Pre-ACL Workshop in Copenhagen, 2025
  Jio Talk (Keynote) at Reliance Jio, 2025
  IIT Bombay, 2025

- Richer Output for Richer Countries
  Keynote, Workshop on Widening NLP (WiNLP) at EMNLP, 2024
  IIT Madras, 2024
  Keynote, Workshop on Building Geo-Diverse and Culturally Aware Models at CVPR, 2025

- Evaluating Models and their Explanations
  RAISE Seminar, University of Washington, 2024
  MBZUAI, Abu Dhabi, 2024
  Google Research, India 2023
  Microsoft Research, India 2023
  Adobe Research, India 2023
  Samsung Research, India 2023
- Watermarking Language Models
  Invited Tutorial at IndoML, IIT Bombay, 2023
  Winter School at IIT Jodhpur, 2024
  Shell.ai Devcon Generative AI week, India 2023
- Fireside Chat
  On AI Opportunities, Google Research Week, 2024
  On Responsible AI, Capital One ML Summit, 2023
  Flipkart Billion AI Event, 2023
- Evaluating Model Explanations
  Conference on Deployable AI, March 2022
  Allen Institute for AI, January 2022
  Amazon, December 2021
  Google AI, November 2021
- Towards Model Understanding
  IISc Bangalore, April 2022
  IIT Bombay, October 2021

IIT Delhi, October 2021
IIT Madras, October 2021
Unbabel AI Seminar, September 2021

- A Tale of Evidence and Explanations
  Data Science Seminar at University of Utah, December 2020
  Machine Learning Seminar at Twitter Inc., October 2020
  NLP Weekly at Google AI, July 2020
- Attention and its Interpretation
  IIT Delhi, January 2020
  ACL, July 2020
- Model Interpretation
  Alumni Research Talk at BITS Pilani, January 2020
  Guest Lecture for Neural Networks for NLP course at CMU, Spring 2019, 2020, 2021
  Guest Lecture for Computational Semantics Course at CMU, Spring 2019
- Interpreting Word Representations
  Indian Institute of Science Bangalore, January 2018
  Student Research Symposium at CMU, August 2017 **(Honorable mention for the best presentation)**
- Panel Discussion on Higher Education for Undergraduates
  Session for undergraduate student researchers at ACL, July 2020
  IIT Jodhpur, November 2020
- Panel Discussion on Life at LTI & CMU
  Open house for admitted students at CMU, Spring 2020, 2019

Professional Services

- Co-Organizer Deployable AI (DAI) Workshop at AAAI 2025
- Co-Chair Diversity & Inclusion Committee (EMNLP 2024)
- Senior Area Chair
  2023: EMNLP
- Area Chair
  2025: NAACL, ACL, EMNLP
  2024: ACL
  2022: EMNLP
  2021: ICLR Workshop on Responsible AI
- Reviewing
  2025: ICLR
  2024: NeurIPS
  2021: ACL, NAACL, FAccT, JAIR
  2020: EMNLP **(Outstanding Reviewer)**, ACL, ICLR, AAAI
  2019: EMNLP, NAACL
- Volunteering
  2021: Faculty Hiring Committee
  2020: Graduate Application Support Program

Selected Press

[1] What counts as plagiarism? AI-generated papers pose new risks
News Feature in Nature (https://www.nature.com/articles/d41586-025-02616-5)

[2] How AI reduces the world to stereotypes.
Rest of World (https://restofworld.org/2023/ai-image-stereotypes/)

[3] AI image generators like DALL-E and Stable Diffusion have a representation problem
Fast Company (https://tinyurl.com/iisc-fast-company-study)

[4] AI Image Generators Lack Global Diversity
CoCreations AI (https://tinyurl.com/iisc-co-creations-study)

[5] A BTech in AI may soon trump a BTech in CompSci as IITians' top choice
The Ken

[6] Fast-learning AI assistants set for smarter roles
Deccan Herald

[7] It's Too Easy to Hide Bias in Deep-Learning Systems.
IEEE Spectrum

[8] How to Measure The Performance Of Explainability Models.
Analytics India Magazine

[9] 12 Interesting Papers From ACL 2020.
Analytics India Magazine

References

Available upon request.