Abhilasha Ravichander

aravicha@cs.washington.edu | lasharavichander.github.io

Research Interests

Artificial Intelligence, Large Language Models, Factuality, Transparency, Evaluation, Data-Centric AI

Education

2018–2022	**Carnegie Mellon University**, School of Computer Science, Ph.D. in Language and Information Technologies, **Advisors : Eduard Hovy, Norman Sadeh**
2016–2018	**Carnegie Mellon University**, School of Computer Science M.Sc in Language Technologies. **QPA: 4.0**

Research Experience

2024-Present	**University of Washington**, Paul G. Allen School of Computer Science Postdoctoral Researcher **Advisor : Yejin Choi**
2023-2024	**Allen Institute for Artificial Intelligence**, Seattle, WA Young Investigator
Jun '21–Aug '21	**Allen Institute for Artificial Intelligence**, Seattle, WA Research Intern **Advisors: Ana Marasovic, Matt Gardner**
Jun '19–Aug'19	**Microsoft Research**, Montreal, QC Research Intern **Advisors: Adam Trischler, Kaheer Suleman, Jackie Cheung**
Jun '14–Aug '14	**Institute of Mathematical Sciences**, Chennai, India Visiting Student **Advisor: Venkatesh Raman**

Academic Honors

- TrustNLP 2025 Workshop, **Best Paper Award**
- NAACL 2025 **Best Paper Award Nominee**
- **Rising Star in Generative AI**, 2024
- ACL 2024 **Best Resource Paper Award**
- ACL 2024 **Best Theme Paper Award**
- MASC-SLL 2024 **Best Paper Award**
- **Rising Star in EECS**, 2022
- SoCal NLP Symposium 2022 **Best Paper Award**
- **Rising Star in Data Science**, 2021
- **Outstanding reviewer**, ACL 2020
- **Outstanding reviewer**, EMNLP 2020
- **Outstanding reviewer**, NAACL 2019
- COLING 2018 **Area Chair Favorite Paper Award**
- **$100,000** grant from Amazon for the 2016 Alexa Prize proposal.

Publications

1. **Abhilasha Ravichander***, Shrusti Ghela*, David Wadden, Yejin Choi
   HALoGEN: Fantastic LLM Hallucinations and Where To Find Them
   Under review.
   [TrustNLP Workshop Best Paper Award] *[long paper]*

2. Benjamin Newman, **Abhilasha Ravichander**, Jaehun Jung, Rui Xin, Hamish Ivison, Yegor Kuznetsov, Pang Wei Koh, Yejin Choi
The Curious Case of Factuality Finetuning: Models' Internal Beliefs Can Improve Factuality
Under review. [long paper]

3. Wenting Zhao, Tanya Goyal, Yu Ying Chiu, Liwei Jiang, Benjamin Newman, **Abhilasha Ravichander**, Khyathi Chandu, Ronan Le Bras, Claire Cardie, Yuntian Deng, Yejin Choi
WildHallucinations: Evaluating Long-form Factuality in LLMs with Real-World Entity Queries
Under review. [long paper]

4. Keivan Rezaei, Khyathi Chandu, Soheil Feizi, Yejin Choi, Faeze Brahman, **Abhilasha Ravichander**
RESTOR: Knowledge Recovery through Machine Unlearning
Under review. [long paper]

5. Skyler Hallinan, Jaehun Jung, Melanie Sclar, Ximing Lu, **Abhilasha Ravichander**, Sahana Ramnath, Yejin Choi, Sai Praneeth Karimireddy, Niloofar Mireshghallah, Xiang Ren
The Surprising Effectiveness of Membership Inference with Simple N-Gram Coverage
Under review. [long paper]

6. **Abhilasha Ravichander**, Jillian Fisher, Taylor Sorensen, Ximing Lu, Maria Antoniak, Bill Yuchen Lin, Niloofar Mireshghallah, Chandra Bhagavatula, Yejin Choi
Information-Guided Identification of Training Data Imprint in (Proprietary) Large Language Models
2025 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2025).
[Best Paper Award Nominee] *[long paper]*

7. Nishant Balepur, Feng Gu, **Abhilasha Ravichander**, Shi Feng, Jordan Lee Boyd-Graber, Rachel Rudinger
Reverse Question Answering: Can an LLM Write a Question so Hard (or Bad) that it Can't Answer?
2025 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2025).
[long paper]

8. Bill Yuchen Lin, Yuntian Deng, Khyathi Chandu, Faeze Brahman, **Abhilasha Ravichander**, Valentina Pyatkin, Nouha Dziri, Ronan Le Bras, Yejin Choi
WildBench: Benchmarking LLMs with Challenging Tasks from Real Users in the Wild
International Conference on Learning Representations (ICLR 2025).
[Spotlight] *[long paper]*

9. Faeze Brahman+, Sachin Kumar+, **Abhilasha Ravichander***, Vidhisha Balachandran*, Pradeep Dasigi*, Valentina Pyatkin*, Sarah Wiegreffe*, Nouha Dziri, Khyathi Chandu, Jack Hessel, Yulia Tsvetkov, Noah A. Smith, Yejin Choi, Hannaneh Hajishirzi
The Art of Saying No: Contextual Noncompliance in Language Models
NeurIPS 2024, *Datasets and Benchmarks [long paper]*

10. Nishant Balepur, **Abhilasha Ravichander**, Rachel Rudinger
Artifacts or Abduction: How Do LLMs Answer Multiple-Choice Questions Without the Question?
62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024).
[MASC-SLL 2024 Best Paper Award] *[long paper]*

11. Groeneveld et al.,
OLMo: Accelerating the Science of Language Models
62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024).
[ACL Best Theme Paper Award] [GeekWire Innovation of the Year Award] *[long paper]*

12. Soldaini et al.,
Dolma: an Open Corpus of Three Trillion Tokens for Language Model Pretraining Research
62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024).
[ACL Best Resource Paper Award] *[long paper]*

13. Da Yin, Faeze Brahman, **Abhilasha Ravichander**, Khyathi Chandu, Kai-Wei Chang, Yejin Choi, Bill Yuchen Lin
Lumos: Learning Agents with Unified Data, Modular Design, and Open-Source LLMs
62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024). *[long paper]*

14. Yufei Tian, **Abhilasha Ravichander**, Lianhui Qin, Ronan Le Bras, Raja Marjieh, Nanyun Peng, Yejin Choi, Thomas L Griffiths, Faeze Brahman.
MacGyver: Are Large Language Models Creative Problem Solvers?
2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2024).
[Best Paper Award Nominee] *[long paper]*

15. Yanai Elazar, Akshita Bhagia, Ian Magnusson, **Abhilasha Ravichander**, Dustin Schwenk, Alane Suhr, Evan Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, Hannaneh Hajishirzi, Noah A. Smith, Jesse Dodge
What's In My Big Data?
International Conference on Learning Representations (ICLR 2024).
[Spotlight] *[long paper]*

16. Peter West, Ximing Lu, Nouha Dziri, Faeze Brahman, Linjie Li, Jena D. Hwang, Liwei Jiang, Jillian Fisher, **Abhilasha Ravichander**, Khyathi Chandu, Benjamin Newman, Pang Wei Koh, Allyson Ettinger, Yejin Choi
The Generative AI Paradox: 'What It Can Create, It May Not Understand'
International Conference on Learning Representations (ICLR 2024). *[long paper]*

17. Bill Yuchen Lin, **Abhilasha Ravichander**, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, Yejin Choi
The Unlocking Spell on Base LLMs: Rethinking Alignment via In-Context Learning
International Conference on Learning Representations (ICLR 2024). *[long paper]*

18. Yuanyuan Feng, **Abhilasha Ravichander**, Yaxing Yao, Shikun Zhang, Rex Chen, Shomir Wilson, Norman Sadeh
Understanding How to Inform Blind and Low-Vision Users about Data Privacy through Privacy Question Answering Assistants
USENIX Security 2024. *[long paper]*

19. Ximing Lu, Faeze Brahman, Peter West, Jaehun Jang, Khyathi Chandu, **Abhilasha Ravichander**, Lianhui Qin, Prithviraj Ammanabrolu, Liwei Jiang, Sahana Ramnath, Nouha Dziri, Jillian Fisher, Bill Yuchen Lin, Skyler Hallinan, Xiang Ren, Sean Welleck, Yejin Choi
Inference-Time Policy Adapters (IPA): Tailoring Extreme-Scale LMs without Fine-tuning
Empirical Methods in Natural Language Processing (EMNLP 2023). *[long paper]*

20. **Abhilasha Ravichander***, Joe Stacey*, Marek Rei
When and Why Does Bias Mitigation Work?
Findings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP Findings 2023). *[long paper]*

21. **Abhilasha Ravichander**, Matt Gardner, and Ana Marasović
CondaQA*: A Contrastive Reading Comprehension Dataset for Reasoning about Negation*
2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022).
[SoCal NLP Symposium Best Paper Award] *[long paper]*

22. Yuanyuan Feng, **Abhilasha Ravichander**, Shikun Zhang, Yaxing Yao, and Norman Sadeh
Exploring and Improving the Accessibility of Data Privacy-related Information for People Who Are Blind and Low-vision
7th Workshop on Inclusive Privacy and Security (WIPS 2022). *[long paper]*

23. Yanai Elazar, Nora Kassner, Shauli Ravfogel, Amir Feder, **Abhilasha Ravichander**, Marius Mosbach, Yonatan Belinkov, Hinrich Schütze, Yoav Goldberg
Measuring Causal Effects of Data Statistics on Language Model's Factual Predictions
arXiv, 2022. *[long paper]*

24. Siddhant Arora, Henry Hosseini, Christine Utz, Vinayshekhar Bannihatti Kumar, Tristan O. Dhellemmes, **Abhilasha Ravichander**, Peter Story, Jasmine Mangat, Rex Chen, Martin Degeling, Thomas Norton, Thomas Hupperich, Shomir Wilson and Norman Sadeh
A Tale of Two Regulatory Regimes: Creation and Analysis of a Bilingual Privacy Policy Corpus
13th Language Resources and Evaluation Conference, (LREC 2022). *[long paper]*

25. Dheeraj Rajagopal, Aman Madaan, Niket Tandon, Yiming Yang, Shrimai Prabhumoye, **Abhilasha Ravichander**, Peter Clark, Eduard Hovy. *CURIE: An Iterative Querying Approach for Reasoning About Situations*
First Workshop on Commonsense Representation and Reasoning, (CSRR@ACL 2022) *[long paper]*

26. **Abhilasha Ravichander**, Yonatan Belinkov, Eduard Hovy.
Probing the Probing Paradigm: Does Probing Accuracy Entail Task Relevance?
16th Conference of the European Chapter of the Association for Computational Linguistics, (EACL 2021). *[long paper]*

27. **Abhilasha Ravichander**, Siddharth Dalmia, Maria Ryskina, Florian Metze, Eduard Hovy and Alan Black
NoiseQA: Challenge Sets for User-Centric Question Answering
16th Conference of the European Chapter of the Association for Computational Linguistics, (EACL 2021). *[long paper]*

28. Yanai Elazar, Nora Kassner, Shauli Ravfogel, **Abhilasha Ravichander**, Eduard Hovy, Hinrich Schütze, Yoav Goldberg.
Measuring and Improving Consistency in Pretrained Language Models
Transactions of the Association for Computational Linguistics, (TACL 2021). *[long paper]*

29. **Abhilasha Ravichander**, Alan W Black, Shomir Wilson, Thomas Norton and Norman Sadeh.
Breaking Down Walls of Text: How Can NLP Benefit Consumer Privacy?
59th Annual Meeting of the Association for Computational Linguistics, (ACL 2021). *[long paper]*

30. **Abhilasha Ravichander**, Eduard Hovy, Kaheer Suleman, Adam Trischler, Jackie Chi Kit Cheung .
On the Systematicity of Probing Contextualized Word Representations: The Case of Hypernymy in BERT
2020 Joint Conference on Lexical and Computational Semantics, (*SEM 2020).*[long paper]*

31. **Abhilasha Ravichander***, Aaakanksha Naik*, Carolyn Rose, Eduard Hovy.
EQUATE: A Benchmark Evaluation Framework for Quantitative Reasoning in Natural Language Inference
2019 Conference on Computational Natural Language Learning, (CoNLL 2019).*[long paper]*

32. **Abhilasha Ravichander**, Alan W Black, Shomir Wilson, Thomas Norton and Norman Sadeh
Question Answering for Privacy Policies: Combining Computational and Legal Perspectives
2019 Conference on Empirical Methods in Natural Language Processing (EMNLP 2019).*[long paper]*

33. **Abhilasha Ravichander***, Aaakanksha Naik*, Carolyn Rose, Eduard Hovy
Exploring Numeracy in Word Embeddings
57th Annual Meeting of the Association for Computational Linguistics (ACL 2019).*[short paper]*

34. Peter Story, Sebastian Zimmeck, Daniel Smullen, **Abhilasha Ravichander**, Ziqi Wang, Joel Reidenberg, N. Cameron Russell and Norman Sadeh
MAPS: Scaling Privacy Compliance Analysis to a Million Apps
Proceedings on Privacy Enhancing Technologies (PETS 2019). *[long paper]*

35. Tom Norton, Joel Reidenberg, Norman Sadeh, **Abhilasha Ravichander**
Evaluating How Global Privacy Principles Answer Consumers' Questions About Mobile App Privacy,
4th European Privacy Law Scholars Conference (PLSC 2019).

36. **Abhilasha Ravichander***, Aaakanksha Naik*, Norman Sadeh, Carolyn Rose, Graham Neubig
Stress Test Evaluation for Natural Language Inference,
27th International Conference on Computational Linguistics (COLING 2018)
[COLING Area Chair Favorite Paper Award] *[long paper]*

37. **Abhilasha Ravichander**, Alan Black.
An Empirical Study of Self-Disclosure in Spoken Dialogue Systems
19th Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL 2018). *[long paper]*

38. **Abhilasha Ravichander***, Thomas Manzini*, Matthias Grabmair, Graham Neubig, Eric Nyberg.
How Would You Say It? Eliciting Lexically Diverse Data for Supervised Semantic Parsing
18th Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL 2017). *[long paper]*

39. Paul Michel*, **Abhilasha Ravichander***, Shruti Rijhwani*.
Does the Geometry of Word Embeddings Help Document Classification? A Case Study on Persistent Homology-Based Representations,
Workshop on Representation Learning For NLP, Annual Meeting of the Association for Computational Linguistics (ACL 2017). *[short paper]*

40. Shrimai Prabhumoye*, Fadi Botros*, Khyathi Chandu*, Samridhi Choudhary*, Esha Keni*, Chaitanya Malaviya*, Thomas Manzini*, Rama Pasumarthi*, Shivani Poddar*, **Abhilasha Ravichander***, Zhou Yu, Alan Black.
Building CMU Magnus from User Feedback, Alexa Prize Proceedings, 2017.

Academic Service

- **Action Editor**, ACL Rolling Review, 2024
- **Organizer**, Workshop on Privacy in Natural Language Processing at ACL 2024 (PrivateNLP 2024)
- **Organizer**, Workshop on Representation Learning for NLP at ACL 2023 (Repl4NLP 2023)
- **Session Chair**, EMNLP 2023
- **Area Chair**, EMNLP 2023
- **Area Chair**, ACL 2023
- **Area Chair**, EMNLP 2022
- NAACL DEI **Socio-Cultural Inclusion Chair**, NAACL 2022
- Co-founder, **NLP with Friends** (https://nlpwithfriends.com/)
- **Student Volunteer** for DEI, CMU LTI Faculty Hiring Committee
- **Student Volunteer**, CMU LTI Ph.D. Admissions Committee
- **Reviewer**

  Conferences: NAACL-HLT 2019, ACL 2020, EMNLP 2020, EACL 2021, ACL 2021, ACL Rolling Review 2021, ACL Rolling Review 2022, COLM 2024, Neurips Datasets and Benchmarks 2024, ICLR 2025

  Workshops: ACL SRW 2020, AACL-IJCNLP SRW 2020, EMNLP-SDP 2020, Neurips HAMLETS workshop 2020, RepL4NLP 2021, AmericasNLP 2021, BlackboxNLP 2021, ACL SRW 2022, Blackbox NLP 2022
- **Student Volunteer**, NAACL-HLT, 2019, EMNLP 2019
- **Session Chair**, AAAI Spring Symposium Series, 2019
- CMU AI Research mentor (initiative to mentor under-represented minorities in Computer Science)
- **Research Team Lead**, OurCS 2019 (workshop to introduce undergraduate women to computer science research)
- **Program Committee**, CMU LTI Student Research Symposium 2018.
- **Student Volunteer**, Widening NLP Workshop at NAACL-HLT 2018.

Teaching

- Teaching Assistant, 11-727 Computational Semantics, Carnegie Mellon University, 2020
- Teaching Assistant, 10-606 Mathematical Foundations for Machine Learning, Carnegie Mellon University, 2018
- Teaching Assistant, 10-607 Computational Foundations for Machine Learning, Carnegie Mellon University, 2018
- Stanford Crowd Course Initiative (MOOC), 2015: Taught modules on recursion and computational complexity.

Invited Talks

- "Illuminating Generative AI: Mapping Knowledge in Large Language Models"
  University of California, Santa Cruz, 2025
  Max Planck Institute, 2025
  University of Minnesota, 2025
  University of Utah, 2025
- Invited Panelist, EMNLP 2024 Panel (WiNLP workshop) on '*Navigating Research in the Age of LLMs*', with Isabelle Augenstein, Mrinmaya Sachan, Sunayana Sitaram, and Lu Wang
- "Understanding Factuality in Large Language Models"
  University of Massachusetts, Amherst (rising stars workshop), 2024
- Minds Matter Podcast, 2024
- "How Do We Get to Transparent Large Language Models?"
  University of Massachusetts, Amherst, 2023
  National University of Singapore, 2023
- "Interpreting Neural Model Performance for Robust, Trustworthy NLP"
  University of Texas at Austin, 2022
  Johns Hopkins University, 2022
  Microsoft Research, 2022

University of Washington, 2022
Allen Institute for Artificial Intelligence, 2022
University of Rochester, 2022
George Mason CS, 2022
Emory CS, 2022
Dair.AI 'Women in NLP' seminar, 2022
University of Illinois at Chicago, 2021
University of Bocconi, 2021

- "User-Centric Question Answering"
  Workshop on Search-Oriented Conversational AI, 2021
  University of Chicago (rising stars workshop), 2021
- "How Can NLP Benefit Consumer Privacy?"
  Compass Tech Summit, 2023
  University of St. Gallen, 2022
  NLLP Talk Series, 2021
  TU Munich, 2021
- Invited Panelist, '*Can Large Language Models Solve NLP?*', 2021 Language Technologies Institute Seminar at Carnegie Mellon University, with Yonatan Bisk, Sam Bowman, and Colin Raffel
- Invited Panelist, NAACL 2021 Panel on '*Getting Into NLP Research*', with William Agnew, Pan Xu, Phu Mon Htut, and Elizabeth Salesky
- Invited Panelist, 13th International Conference on Data Protection and Artificial Intelligence (CPDP 2020), with Cameron Russell, Lokke Moerel, and Antoine Bon

Software Development Experience

Sep '15–Jun '16	**Platform Engineer, Sensara Technologies**, Bangalore, India Designed a novel algorithm to finely segment advertisement boundaries i.e detect at a frame-level granularity when advertisements begin and end. This work is currently in production at adbreaks.in ([link]) and used at scale to segment advertisements every day. Relevant talk about this work [link].
Jan '15–Jul '15	**Software Development Engineer Intern, Amazon**, Bangalore, India Worked on integrated module to fetch delivery charges such that they are pincode and quantity aware for Amazon China, India and UK.