Nouha Dziri Ph.D.

nouha.dziri@gmail.com
http://nouhadziri.com

Research Interests

- Aligning LLMs with human values and preferences
- Red-teaming LLMs, Evaluating LLMs reasoning capabilities.
- Understanding the limits of Transformers LLMs and their inner workings.

Employment

Allen Institute for AI Seattle, US
Research Scientist 2023 - Now

- Advisor: Yejin Choi

Allen Institute for AI Seattle, US
Postdoctoral Fellow 2022 - 2023

- Advisor: Yejin Choi

Mila – Quebec Artificial Intelligence Institute / McGill University Montreal, CA
Visiting Scholar 2021 - 2022

- Advisor: Siva Reddy

Google Research NYC, US
NLP Student Researcher 2020 - 2022

- Advisors: Tal Linzen, David Reitter, Hannah Rashkin

Microsoft Research NYC, US
NLP Research Intern 2019 - 2020

- Advisors: Alessandro Sordoni, Goeff Gordon

Google Research NYC, US
NLP Research Intern 2019

- Advisors: Diyi Yang, Tom Kwiatkowski

Education

Ph.D. Computing Science, University of Alberta Edmonton, Canada
Thesis: Mitigating Hallucinations in Conversational LLMs. 2018 - 2022

- Advisor: Prof. Osmar Zaiane, GPA: 4.00/4.00

MSc. Computing Science, University of Alberta Edmonton, Canada
Thesis: Response Generation For An Open-Ended Conversational Agent 2016 - 2018

- Advisor: Prof. Osmar Zaiane, GPA: 4.00/4.00

Selected Publications

You can find an exhaustive list of my publications in my Google Scholar profile.

1. **WildGuard: Open One-Stop Moderation Tools For Safety Risks, Jailbreaks, and Refusals of LLMs.** Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri.
   Under submission NeurIPS 2024.
2. **WildTeaming at Scale: From In-the-Wild Jailbreaks to (Adversarially) Safer Language Models.** Liwei Jiang, Kavel Rao, Seungju Han, Allyson Ettinger, Faeze Brahman, Sachin Kumar, Niloofar Mireshghallah, Maarten Sap, Yejin Choi, Nouha Dziri.
   Under submission NeurIPS 2024.
3. **RewardBench: Evaluating Reward Models for Language Modeling.** Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A Smith, Hannaneh Hajishirzi
   Under submission NeurIPS 2024.

Selected Publications

4. **WildBench: Benchmarking LLMs with Challenging Tasks from Real Users in the Wild.** Bill Yuchen Lin, Yuntian Deng, Khyathi Chandu, Faeze Brahman, Abhilasha Ravichander, Valentina Pyatkin,Nouha Dziri, Ronan Le Bras, Yejin Choi
   Under submission NeurIPS 2024.

5. **A Roadmap to Pluralistic Alignment.** Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, Yejin Choi.
   ICML 2023.

6. **Faith and Fate: Limits of Transformers on Compositionality.**
   Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jian, Bill Yuchen Lin, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D Hwang, Soumya Sanyal, Sean Welleck, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, Yejin Choi.
   NeurIPS 2023 (Spotlight).

7. **Fine-Grained Human Feedback Gives Better Rewards for Language Model Training.**
   Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A Smith, Mari Ostendorf, Hannaneh Hajishirzi.
   NeurIPS 2023 (Spotlight) .

8. **Self-Refine: Iterative Refinement with Self-Feedback.** Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Sean Welleck, Bodhisattwa Prasad Majumder, Shashank Gupta, Amir Yazdanbakhsh, Peter Clark
   NeurIPS 2023.

9. **The Generative AI Paradox:" What It Can Create, It May Not Understand".** 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.
   ICLR 2024. (* = equal contribution)

10. **Phenomenal Yet Puzzling: Testing Inductive Reasoning Capabilities of Language Models with Hypothesis Refinement.**
    Linlu Qiu, Liwei Jiang, Ximing Lu, Melanie Sclar, Valentina Pyatkin, Chandra Bhagavatula, Bailin Wang, Yoon Kim, Yejin Choi, Nouha Dziri*, Xiang Ren*.
    ICLR 2024 (Oral).

11. **The Unlocking Spell on Base LLMs: Rethinking Alignment via In-Context Learning.**
    Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, Yejin Choi.
    ICLR 2024.

12. **Evaluating Open-Domain Question Answering in the Era of Large Language Models.**
    Ehsan Kamalloo, Nouha Dziri, Charles LA Clarke, Davood Rafiei
    ACL 2023.

13. **On the Origin of Hallucinations in Conversational Models: Is it the Datasets or the Models?** Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, Siva Reddy
    NAACL 2022.

14. **FaithDial: A Faithful Benchmark for Information-Seeking Dialogue**
    Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Osmar Zaiane, Mo Yu, Edoardo M Ponti, Siva Reddy
    TACL 2022.

15. **Decomposed Mutual Information Estimation For Contrastive Representation Learning**
    Alessandro Sordoni*, Nouha Dziri*, Hannes Schulz*, Geoff Gordon, Philip Bachman, Remi Tachet Des Combes
    ICML 2021 (* = equal contribution).

Invited Talks

"What it can create, it may not understand": Studying the Limits of Transformers.
University of Cambridge May 2024

Guest Lecture: Limits of Generative AI Models and their Societal Implications.
Princeton University Dec 2023

Faith and Fate: Limits of Transformers on Compositionality.
The Alan Turing Institute, UK Nov 2023

Faith and Fate: Limits of Transformers on Compositionality.
University of Edinburgh Nov 2023

Faith and Fate: Limits of Transformers on Compositionality.
SAIL workshop on fundamental limits of LLMs, Germany Oct 2023

Faith and Fate: Limits of Transformers on Compositionality.
University of Pittisburgh Oct 2023

Faith and Fate: Limits of Transformers on Compositionality.
Formal Languages and Neural Networks Seminar Sep 2023

Towards Building Hallucination-Free Conversational Models.
Stanford University Aug 2022

FaithDial: A Faithful Benchmark for Information-Seeking Dialogue.
Google Research May 2022

FaithDial: A Faithful Benchmark for Information-Seeking Dialogue.
Amazon Research June 2022

Evaluating Coherence in Dialogue Systems Using Entailment.
Google DeepMind Dec 2019

Awards and Honors

- **Outstanding Reviewer** at ACL 2021 (**top 1%**) 2021
- **Alberta Doctoral Recruitment Scholarship** ($10,000) 2018
- **Mitacs Globalink PhD Graduate Fellowship** ($44,000) 2018
- **Best Poster Award**, ACM Canadian Celebration of Women in Computing ($400) 2017
- **Mitacs Globalink MSc Graduate Fellowship** ($15,000) 2016
- **DAAD Scholarship** for Research Internship, Leipzig, Germany (€10.000) 2015
- **Erasmus Mundus Exchange Scholarship** for BSc studies(€50.000) 2015

Academic Services

Demo Chair: NAACL 2025.

Senior Area Chair: ACL 2025 in the area of Ethics, Bias, and Fairness.

Area Chair: EMNLP 2023, COLM 2024

Reviewer: NeurIPS (2022-2024), ICLR (2022-2024), ACL (2018-2023), EMNLP (2018-2023), NAACL (2018-2023), EACL (2018-2022)

Skills

Languages: English, French.

Programming: `Python`, `Pytorch`, `TensorFlow`.