nouha.dziri@gmail.com
http://nouhadziri.com
+1 206 617 2801

Nouha Dziri Ph.D.

Research Interests

- Post-training LLMs
- Advancing LLMs reasoning capabilities, Evaluating LLMs capabilities.
- Agent, Safety alignment, Red-teaming LLMs

Employment

Allen Institute for AI Seattle, US
Research Scientist 2023 - Now
- Advisor: Yejin Choi

Allen Institute for AI Seattle, US
Postdoctoral Fellow 2023
- Advisor: Yejin Choi

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

Google DeepMind NYC, US
Student Researcher 2020 - 2022
- Advisors: Tal Linzen, David Reitter, Hannah Rashkin

Microsoft Research Montreal, CA
Research Intern 2019 - 2020
- Advisors: Alessandro Sordoni, Goeff Gordon

Google DeepMind NYC, US
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

Preprints

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

1. **OMEGA: Can LLMs Reason Outside the Box in Math? Evaluating Exploratory, Compositional, and Transformative Generalization.** Yiyou Sun, Georgia Zhou, Hao Wang, Dacheng Li, Nouha Dziri*, Dawn Song*
   Arxiv 2025 (* = equal advising role)
2. **Climbing the Ladder of Reasoning: What LLMs Can-and Still Can't-Solve after SFT?.** Yiyou Sun, Shawn Hu, Georgia Zhou, Ken Zheng, Hannaneh Hajishirzi, Nouha Dziri, Dawn Song
   Arxiv 2025.
3. **OpenAgentSafety: A Comprehensive Framework for Evaluating Real-World AI Agent Safety.** Sanidhya Vijayvargiya, Aditya Bharat Soni, Xuhui Zhou, Zora Zhiruo Wang, Nouha Dziri, Graham Neubig, Maarten Sap
   Arxiv 2025.

Journal & Conference Publications

4. **On the Trustworthiness of Generative Foundation Models: Guideline, Assessment, and Perspective** Yue Huang, Chujie Gao, Siyuan Wu, Haoran Wang, Xiangqi Wang, Yujun Zhou, ..., **Nouha Dziri**, Yu Su, ..., Mohit Bansal, Nitesh V Chawla, Jian Pei, Jianfeng Gao, Michael Backes, Philip S Yu, Neil Zhenqiang Gong, Pin-Yu Chen, Bo Li, Xiangliang Zhang
   Arxiv 2025

5. **2 OLMo 2 Furious.** Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, ..., **Nouha Dziri**, Noah A. Smith, Hannaneh Hajishirzi
   COLM 2025

6. **TÜLU 3: Pushing Frontiers in Open Language Model Post-Training.** Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, **Nouha Dziri**, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi and Hannaneh Hajishirzi.
   COLM 2025.

7. **SafetyAnalyst: Interpretable, Transparent, and Steerable LLM Safety Moderation** Jing-Jing Li, Valentina Pyatkin, Max Kleiman-Weiner, Liwei Jiang, **Nouha Dziri**, Anne GE Collins, Jana Schaich Borg, Maarten Sap, Yejin Choi, Sydney Levine
   ICML 2025

8. **AI as Humanity's Salieri: Quantifying Linguistic Creativity of Language Models via Systematic Attribution of Machine Text against Web Text.** Ximing Lu, Melanie Sclar, Skyler Hallinan, Niloofar Mireshghallah, Jiacheng Liu, Seungju Han, Allyson Ettinger, Liwei Jiang, Khyathi Chandu, **Nouha Dziri**, Yejin Choi.
   ICLR 2025 **(Oral, 1.8% acceptance rate)**.

9. **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
   ICLR 2025.

10. **Steering Masked Discrete Diffusion Models Via Discrete Denoising Posterior Prediction** Jarrid Rector-Brooks, Mohsin Hasan, Zhangzhi Peng, Zachary Quinn, Chenghao Liu, Sarthak Mittal, **Nouha Dziri**, Michael Bronstein, Yoshua Bengio, Pranam Chatterjee, Alexander Tong, Avishek Joey Bose
    ICLR 2025.

11. **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
    NAACL 2025.

12. **Rel-AI: An Interaction-Centered Approach To Measuring Human-LM Reliance.** Kaitlyn Zhou, Jena D Hwang, Xiang Ren, **Nouha Dziri**, Dan Jurafsky, Maarten Sap
    NAACL 2025 **(Best Paper Runner-Up 2025)** .

13. **To Err is AI: A Case Study Informing LLM Flaw Reporting Practices** Sean McGregor, Allyson Ettinger, Nick Judd, Paul Albee, Liwei Jiang, Kavel Rao, Will Smith, Shayne Longpre, Avijit Ghosh, Christopher Fiorelli, Michelle Hoang, Sven Cattell, **Nouha Dziri**.
    AAAI 2025 .

14. **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**.
    NeurIPS 2024.

15. **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**.
    NeurIPS 2024.

Journal & Conference Publications

16. **Multi-Attribute Constraint Satisfaction via Language Model Rewriting** Ashutosh Baheti, Debanjana Chakraborty, Faeze Brahman, Ronan Le Bras, Ximing Lu, Nouha Dziri, Yejin Choi, Mark Riedl, Maarten Sap
    TMLR 2024.

17. **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 2024.

18. **Elastic Weight Removal For Faithful and Abstractive Dialogue generation** Nico Daheim, Nouha Dziri, Mrinmaya Sachan, Iryna Gurevych, Edoardo Ponti
    NAACL 2024.

19. **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, 2% acceptance rate).

20. **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, 2% acceptance rate).

21. **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.

22. **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)

23. **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, 1.8% acceptance rate). (* = equal advising role)

24. **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.

25. **Value Kaleidoscope: Engaging AI with pluralistic human values, rights, and duties**
    Taylor Sorensen, Liwei Jiang, Jena Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, Maarten Sap, John Tasioulas, Yejin Choi
    AAAI 2024.

26. **Culture-gen: Revealing global cultural perception in language models through natural language prompting**
    Huihan Li, Liwei Jiang, Jena D Hwang, Hyunwoo Kim, Sebastin Santy, Taylor Sorensen, Bill Yuchen Lin, Nouha Dziri, Xiang Ren, Yejin Choi
    COLM 2024.

27. **What Makes it Ok to Set a Fire? Iterative Self-distillation of Contexts and Rationales for Disambiguating Defeasible Social and Moral Situations** Kavel Rao, Liwei Jiang, Valentina Pyatkin, Yuling Gu, Niket Tandon, Nouha Dziri, Faeze Brahman, Yejin Choi
    EMNLP 2023.

Journal & Conference Publications

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

29. **CHAMPAGNE: Learning Real-world Conversation from Large-Scale Web Videos**
    Seungju Han, Jack Hessel, Nouha Dziri, Yejin Choi, Youngjae Yu
    ICCV 2023.

30. **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.

31. **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.

32. **Neural Path Hunter: Reducing Hallucination in Dialogue Systems via Path Grounding**
    Nouha Dziri, Andrea Madotto, Osmar Zaiane, Avishek Joey Bose
    EMNLP 2021.

33. **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).

34. **Evaluating Coherence in Dialogue Systems Using entailment**
    Nouha Dziri*, Ehsan Kamalloo*, Kory W Mathewson, Osmar Zaiane
    NAACL 2019 (* = equal contribution).

Invited Panels

Computer Use Agents
ICML workshop, Vancouver July 2025

Data in Generative Models
ICML workshop, Vancouver July 2025

AI Reasoning
Cross Future AI & Technology Summit, Vancouver July 2025

Trustworthy Foundation Models
International Symposium on Trustworthy Foundation Models, UAE May 2025

Meta-Generation Algorithms for Large Language Models
NeurIPS Tutorial, Vancouver Dec 2024

Analyzing the Plasticity of LLMs During Language Interactions.
Language Gamification Workshop NeurIPS, Vancouver Dec 2024

Towards A Global Standard for AI Product Safety
MLCommons, San Francisco Nov 2024

Invited Talks

LLM Reasoning: Advanced Inference-time Strategies
Armenia LLM Summer School, Armenia July 2025

Can LLMs Reason Outside the Box in Math?
Apple Workshop on Reasoning and Planning, Cupertino July 2025

Faith and Fate: Limits of Transformers in Reasoning
Cross Future AI & Technology Summit, Vancouver July 2025

Invited Talks

Trustworthy LLMs: How Data Quality Shapes Performance and Where It Falls Short?
ICML 2025 Workshop on Data in Generative Models, Vancouver July 2025

OpenAgentSafety: A Comprehensive Framework for Evaluating AI Agent Safety
ICML 2025 Workshop on Computer-Use Agents, Vancouver July 2025

Does Scaling Guarantee Trustworthy LLMs
International Symposium on Trustworthy Foundation Models, UAE May 2025

In-Context Learning in LLMs: Potential and Limits
Causality in the Era of Foundation Models Workshop, Barbados Feb 2025

Red-teaming and Safeguarding LLMs
International Association for Safe and Ethical AI, Paris Feb 2025

In-Context Learning in LLMs: Potential and Limits
Language Gamification Workshop NeurIPS, Vancouver Dec 2024

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

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 Pittsburgh Oct 2023

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

Towards Building Hallucination-Free Conversational Models
Stanford University Aug 2022

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

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

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

Guest Lecturer

11-430/830 Ethics, Safety, and Social Impact in NLP and LLMs
***Instructor**: Maarten Sap*
Carnegie Mellon University *Winter 2025*
***Lecture**: Redteaming and Safegarding in LLMs*

Generative AI Seminar Course
***Instructor**: Adji Bousso Dieng*
Princeton University *Fall 2023*
***Lecture**: Limits of Generative AI Models and their Societal Implications.*

Guest Lecturer

CSE 599 D1: (Grad) Exploration on Language, Knowledge, and Reasoning
***Instructor**: Yejin Choi*
University of Washington *Winter 2023*
***Lecture**: Can LLMs reason?*

Students Mentoring

- **Ximing Lu**, Predoctoral student at Ai2 –> PhD student at Stanford 2022-2024
- **Liwei Jiang**, PhD student at UW 2023-2025
- **Seungju Han**, BSc at Seoul National University -> PhD at Stanford 2023-2025
- **Kavel Rao**, MSc student at UW -> Researcher at Jane Street 2023-2025
- **Linlu Qiu**, PhD student at MIT 2023
- **Huihan Li**, PhD student at University of Southern California 2023-2024
- **Yiyou Sun**, Postdoc at University of California, Berkeley 2025-current
- **Shrey Jain**, MSc student at Carnegie Mellon University 2025-current

Workshop Co-organizer

- The first workshop for Research on Agent Language Models (REALM) ACL 2025
- The Reasoning and Planning for Large Language Models ICLR 2025
- System 2 Reasoning at Scale NeurIPS 2024
- The third workshop on Document-Grounded Dialogue Systems (DialDoc) ACL 2023

Academic Services

Demo Chair: NAACL 2025.

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

Area Chair: EMNLP 2023, COLM (2024-2025)

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

PRESS

- **Quanta Magazine**: *Chatbot Software Begins to Face Fundamental Limitations*, 2025.
- **LeMonde**: *Should we be concerned about the 'hallucinations' of AI like ChatGPT or Gemini?*, 2024.
- **ScienceNews**: *AI's reasoning skills can't be assessed by current tests*, 2024.
- **TechCrunch**: *Treating a chatbot nicely might boost its performance — here's why*, 2024.
- **Podcast**: *Women in AI Research*, 2025.

Awards and Honors

- **Runner-up Best Paper** at NAACL 2025 **(top 0.1%)** 2025
- **Outstanding Reviewer** at ACL 2021 **(top 1%)** 2021
- **Alberta Doctoral Recruitment Scholarship** ($10,000) 2018
- **Mitacs Globalink PhD Graduate Fellowship** ($44,000) 2018
- **National Scholarship for MSc and PhD studies** ($70,000) 2016
- **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