

Faeze Brahman

POSTDOCTORAL RESEARCHER · COMPUTER SCIENCE · AI2 | UNIVERSITY OF WASHINGTON

✉ faezeb@allenai.org | 🏠 <https://fabrahman.github.io> | 🐦 @faeze_brh

Research Interests

Generative AI, Safety & Social AI Alignment, Interpretability

Professional Experience

Postdoctoral Researcher

04/2022-present

ALLEN ALLEN INSTITUTE FOR ARTIFICIAL INTELLIGENCE (AI2) & UNIVERSITY OF WASHINGTON

- Mentor: Yejin Choi

Research Intern

06/2021-09/2021

MICROSOFT RESEARCH - DEEP LEARNING GROUP

- Hosts: Michel Galley, Jianfeng Gao
- Project: Controllable Grounded Long-Form Text Generation

Research Intern

06/2020-09/2020

ALLEN ALLEN INSTITUTE FOR ARTIFICIAL INTELLIGENCE (AI2) - MOSAIC GROUP

- Hosts: Vered Shwartz, Yejin Choi
- Project: Distant Supervision Methods for Explainable AI

Research Intern

06/2018-09/2018

XEROX PARC - INTERACTION & ANALYTIC LAB (IAL)

- Host: Kyle Dent
- Project: RFP Response Assistant System using Semantic Contextualized vectors for retrieval.

Education

University of California, Santa Cruz

Santa Cruz, CA

PHD IN COMPUTER SCIENCE (GPA: 3.87)

2022

- Thesis: Modeling Key Narrative Elements for Automatic Story Generation
- Advisor: Snigdha Chaturvedi

MS IN COMPUTER SCIENCE (GPA: 3.87)

2018

- Advisor: Marilyn Walker

Iran University of Science and Technology

Tehran, Iran

MS IN ELECTRICAL ENGINEERING (GPA: 3.91)

2014

- Thesis: Electrical and Thermal Energy Management of Residential Energy Hubs
- Advisor: Shahram Jadid

BS IN ELECTRICAL ENGINEERING (GPA: 3.54)

2012

- Graduated as Outstanding Student

Publications

PREPRINT AND IN SUBMISSION

The Art of Saying No: Contextual Noncompliance in Language Models

Faeze Brahman^{*}, Sachin Kumar^{*}, Vidhisha Balachandran, Pradeep Dasigi, Valentina Pyatkin, Abhilasha Ravichander, Sarah Wiegrefe, Nouha Dziri, Khyathi Chandu, Jack Hessel, Yulia Tsvetkov, Noah A. Smith, Yejin Choi, and Hannaneh Hajishirzi. @ *Under submission*

Trust or Escalate: LLM Judges with Provable Guarantees for Human Agreement

Jaehun Jung, *Faeze Brahman*, and Yejin Choi. @ *arXiv*

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, Nilofar Mireshghallah, Ximing Lu, Maarten Sap, Nouha Dziri, and Yejin Choi. @ Under submission

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, Ronan Le Bras, and Yejin Choi. @ Under submission

An Ecosystem for Sandboxing Safety Risks in Human-AI Interactions

@ Under prepration

JOURNAL AND PEER-REVIEWED CONFERENCE PAPERS

Agent Lumos: Unified and Modular Training for Open-Source Language Agents

Da Yin, [Faeze Brahman](#), Abhilasha Ravichander, Khyathi Chandu, Kai-Wei Chang, Yejin Choi, and Bill Yuchen Lin. @ ACL 2024

MacGyver: Are Large Language Models Creative Problem Solvers?

Yufei Tian, Abhilasha Ravichander, Lianhui Qin, Ronan Le Bras, Raja Marjieh, Nanyun Peng, Yejin Choi, Thomas L. Griffiths, and [Faeze Brahman](#). @ NAACL 2024

Information-Theoretic Distillation for Reference-less Summarization

Jaehun Jung, Ximing Lu, Liwei Jiang, [Faeze Brahman](#), Peter West, Pang Wei Koh, and Yejin Choi. @ COLM 2024

PlaSma: Making Small Language Models Better Procedural Knowledge Models for (Counterfactual) Planning

[Faeze Brahman](#), Chandra Bhagavatula, Valentina Pyatkin*, Jena D. Hwang*, Xiang Lorraine Li, Hirona J. Arai, Soumya Sanyal, Keisuke Sakaguchi, Xiang Ren, Yejin Choi. @ ICLR 2024

The Generative AI Paradox: “What It Can Create, It May Not Understand”

[Faeze Brahman](#)*, Peter West*, Ximing Lu*, Nouha Dziri*, Linjie Li*, Jena D. Hwang, Liwei Jiang, Jillian Fisher, Abhilasha Ravichander, Khyathi Chandu, Benjamin Newman, Pang Wei Koh, Allyson Ettinger, and Yejin Choi. @ ICLR 2024

Improving Language Models with Advantage-based Offline Policy Gradients

Ashutosh Baheti, Ximing Lu, [Faeze Brahman](#), Ronan Le Bras, Maarten Sap, Mark Riedl. @ ICLR 2024

Impossible Distillation: from Low-Quality Model to High-Quality Dataset & Model for Summarization and Paraphrasing

Jaehun Jung, Peter West, Liwei Jiang, [Faeze Brahman](#), Ximing Lu, Jillian Fisher, Taylor Sorensen, Yejin Choi. @ NAACL 2024

Creativity Support in the Age of Large Language Models: An Empirical Study Involving Emerging Writers

Tuhin Chakrabarty*, Vishakh Padmakumar*, [Faeze Brahman](#), and Smaranda Muresan. @ ACM Creativity & Cognition 2024

SwiftSage: A Generative Agent with Fast and Slow Thinking for Complex Interactive Tasks

Bill Yuchen Lin, Yicheng Fu, Karina Yang, [Faeze Brahman](#), Shiyu Huang, Chandra Bhagavatula, Prithviraj Ammanabrolu, Yejin Choi, Xiang Ren. @ NeurIPS 2023

Inference-Time Policy Adapters (IPA): Tailoring Extreme-Scale LMs without Fine-tuning

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. @ EMNLP 2023

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](#), and Yejin Choi. @ Findings of EMNLP 2023

STEER: Unified Style Transfer with Expert Reinforcement

Skyler Hallinan, [Faeze Brahman](#), Ximing Lu, Jaehun Jung, Sean Welleck, and Yejin Choi. @ Findings of EMNLP 2023

Affective and Dynamic Beam Search for Story Generation

Tenghao Huang, Ehsan Qasemi, Bangzheng Li, He Wang, [Faeze Brahman](#), Muhao Chen, and Snigdha Chaturvedi. @ Findings of EMNLP 2023

REV: Information-Theoretic Evaluation of Free-Text Rationales

Hanjie Chen, [Faeze Brahman](#), Xiang Ren, Yangfeng Ji, Yejin Choi, Swabha Swayamdipta. @ ACL 2023

Generating Sequences by Learning to [Self-]Correct

Sean Welleck*, Ximing Lu*, [Faeze Brahman](#)+, Peter West+, Tianxiao Shen, Daniel Khashabi, Yejin Choi. @ ICLR 2023

* : co-first authors, + co-second authors

Maieutic Prompting: Logically Consistent Reasoning with Recursive Explanations

Jaehun Jung, Lianhui Qin, Sean Welleck, [Faeze Brahman](#), Chandra Bhagavatula, Ronan Le Bras, Yejin Choi. @ EMNLP 2022

Grounded Keys-to-Text Generation: Towards Factual Open-Ended Generation

[Faeze Brahman](#), Baolin Peng, Michel Galley, Sudha Rao, Bill Dolan, Snigdha Chaturvedi, Jianfeng Gao. @ Findings of EMNLP 2022

Towards Inter-character Relationship-driven Story Generation

Anvesh Rao Vijjini, [Faeze Brahman](#), Snigdha Chaturvedi. @ EMNLP 2022

NarraSum: A Large-Scale Dataset for Abstractive Narrative Summarization

Chao Zhao, [Faeze Brahman](#), Kaiqiang Song, Wenlin Yao, Dian Yu, Snigdha Chaturvedi. @ Findings of EMNLP 2022

Revisiting Generative Commonsense Reasoning: A Pre-Ordering Approach

Chao Zhao, [Faeze Brahman](#), Tenghao Huang, Snigdha Chaturvedi. @ Findings of NAACL 2022

Reformulating Sentence Ordering as Conditional Text Generation

Somnath Basu Roy Chowdhury*, [Faeze Brahman](#)* and Snigdha Chaturvedi. @ EMNLP 2021

* : equal contribution

“Let Your Characters Tell Their Story”: A Dataset for Character-Centric Narrative Understanding

[Faeze Brahman](#), Meng Huang, Oyvind Tafjord, Chao Zhao, Mrinmaya Sachan, and Snigdha Chaturvedi. @ Findings of EMNLP 2021

ParsiNLU: A Suite of Language Understanding Challenges for Persian

Daniel Khashabi et al. @ TACL 2021

Uncovering Implicit Gender Bias in Narratives through Commonsense Inference

Tenghao Huang, [Faeze Brahman](#), Vered Shwartz, and Snigdha Chaturvedi. @ Findings of EMNLP 2021

Learning to Rationalize for Nonmonotonic Reasoning with Distant Supervision

[Faeze Brahman](#), Vered Shwartz, Rachel Rudinger, and Yejin Choi. @ AAAI 2021

Modeling Protagonist Emotions for Emotion-Aware Storytelling

[Faeze Brahman](#) and Snigdha Chaturvedi. EMNLP 2020

Cue Me In: Content-Inducing Approaches to Interactive Story Generation

[Faeze Brahman](#), Alexandru Petrusca, and Snigdha Chaturvedi. @ AACL 2020

Effective Forum Curation Via Multi-task Learning

[Faeze Brahman](#), Nikhil Varghese, Suma Bhat, and Snigdha Chaturvedi. @ EDM 2020

Electrical and Thermal Energy Management of a Residential Energy Hub, Integrating Demand Response and Energy Storage System

[Faeze Brahman](#), Masoud Honarmand, and Shahram Jadid. *Journal of Energy and Building*, 2015

Development of a Thermal and Electrical Energy Management System in Residential Building Micro-Grid

Saeed Esmaeili, Behrooz Vahidi, Mehdi Parvizimosaed, [Faeze Brahman](#). *Journal of Renewable and Sustainable Energy*, 2013

WORKSHOP

Towards Emotion-Aware Storytelling Using Reinforcement Learning

[Faeze Brahman](#), and Snigdha Chaturvedi. *Wordplay: When Language Meets Games Workshop @ NeurIPS 2020*

Automatic Story Generation with Human-in-the-Loop

[Faeze Brahman](#), Alexandru Petrusca, and Snigdha Chaturvedi. *Workshop on Narrative Understanding (WNU) @ NAACL 2019*

US PATENT

System and method for artificial intelligence story generation allowing content introduction

Snigdha Chaturvedi, [Faeze Brahman](#), and Alexandru Petrusca. *PUB NO: 20, 200, 311, 341, 2020*

Awards

2021 **Sabbatical Dissertation Fellowship**, Jack Baskin School of Engineering, UC Santa Cruz

2020 **GHC Scholarship**, AnitaB.org
Tapia Celebration of Diversity Scholarship, CMDiT

2019 **GHC Student Scholarship**, AnitaB.org
Women’s Club Scholarship, UC Santa Cruz
Marilyn C. Davis Scholarship, UC Santa Cruz
Second Best Poster Award, UCSC Data Science’s Day
CRA-W Travel Grant, Computing Research Association (CRA)

2016 **Regents Fellowship**, UC Santa Cruz

2012 **Full Scholarship for graduate study**, Iran University of Science and Technology

2008 **Full Scholarship for undergraduate study**, Iran University of Science and Technology

Presentations

TALKS

May 2024. *Creativity, Constrained Reasoning and Problem Solving.*, MilaNLP, Bocconi University.

April 2024. *Creativity, Constrained Reasoning and Problem Solving.*, UBC NLP Seminar.

Mar 2023. *Inference and Learning Frameworks for Consistent Reasoning and Controlled Language Generation.*, UMass NLP Seminars.

Jan 2022. *Modeling Key Narrative Elements for Automatic Story Generation*, University of Southern California.

Jan 2022. *Modeling Key Narrative Elements for Automatic Story Generation*, University of British Columbia.

Nov 2020. *Modeling Protagonist Emotions for Emotion-Aware Storytelling*, EMNLP conference.

Nov 2020. *Cue Me In: Content-Inducing Approaches to Interactive Story Generation*, AACL conference.

Sep 2020. *Weakly-Supervised Rationale Generation for Nonmonotonic Reasoning*, Allen Institute for AI.

Sep 2020. *Automatic Story Generation via Modeling Key Narrative Elements*, UC Santa Cruz, proposal.

Jul 2020. *Effective Forum Curation Via Multi-task Learning*, EDM conference.

GUEST LECTURES

Winter 2023. *Exploration on Language, Knowledge, and Reasoning*. Guest Lecture, University of Washington.

Fall 2018. *Neural Networks*. Guest Lecture, Machine Learning Course, UC Santa Cruz.

Spring 2018. *Neural Networks*. Guest Lecture, Machine Learning Course, UC Santa Cruz.

Spring 2018. *Generative Adversarial Networks*. Guest Lecture, Machine Learning Course, UC Santa Cruz.

Teaching Experience

2020	CSE 115: Software Design Project , Teaching Assistant	UCSC
	CSE 16: Discrete Mathematics , Teaching Assistant	
2019	Natural Language Processing [High school students] , Teaching Assistant	UCSC's COSMOS Program
2018	CSE 142: Machine Learning and Data Mining , Teaching Assistant	UCSC

Mentoring

2024	Xuhui Zhou , AI2 Intern (co-mentor: Ronan Le Bras)
2024	Keivan Rezaei , AI2 Intern (co-mentor: Abhilasha Ravichander)
2023	Da Yin , AI2 Intern (co-mentor: Yuchen Lin & Abhilasha Ravichander)
2023	Yufei Tian , AI2 Intern (co-mentor: Ronan Le Bras & Lianhui Qin)
2023	Skyler Hallinan , Master Student, University of Washington
2022	Kavel Rao , Undergraduate, University of Washington (co-mentor: Liwei Jiang)
2022	Hanji Chen , AI2 Intern (co-mentor: Swabha Swayamdipta)
2021	Tengaho Huang , Undergraduate, UNC Chapel Hill (co-mentor: Snigdha Chaturvedi)
2020	Meng Huang , Master Student, University of Chicago (co-mentor: Mrinmaya Sachan)

Service & Professional Activities

PROGRAM COMMITTEE & REVIEWING

2024 **Senior Area Chair**, NAACL, COLING, ARR
2024 **Program Committee**, Neurips, ICLR
2023 **Program Committee**, NeurIPS, ACL, EMNLP, ARR, NLRSE
2022 **Program Committee**, ARR, ACL, EMNLP
2021 **Program Committee**, ARR, AAAI, ACL, EMNLP, IJCAI
2020 **Program Committee**, AAAI, CoNLL, WNU
2019 **Program Committee**, CoNLL, WNU
2019 **Student Volunteer**, NAACL

WORKSHOP ORGANIZATION

The 6th workshop on Narrative Understanding (WNU) @ EMNLP2024
Upcoming...

The fifth workshop on Narrative Understanding (WNU) @ ACL2023
<https://sites.google.com/view/wnu2022>

The fourth workshop on Narrative Understanding (WNU) @ NAACL2022
<https://sites.google.com/view/wnu2022>
Above 75 attendees

The third workshop on Narrative Understanding, Storylines, and Events (WNU) @ NAACL2021
<https://sites.google.com/view/wnu2021>

DIVERSITY PROMOTING SERVICE

Co-conceptualized and co-organized a Women-in-AI (WiAI) Group at UCSC, 2019.
Mentoring at WiNLP workshop, NAACL 2022. (part of a mentorship and affinity support workshop)

Skills

Python C++ R MATLAB / Pytorch TensorFlow Keras Texar
HTML/CSS JavaScript D3 Bokeh / git \LaTeX Bash script MySQL