Faeze Brahman

Postdoctoral Researcher · Computer Science · Allen Institute for AI | University of Washington

fae.brahman@gmail.com | https://fabrahman.github.io | @faeze_brh

Research Interests

NLP, Human AI Alignment, Trustworthy AI, Human Computer Interaction

Professional Experience

Postdoctoral Researcher *04/2022-present*

Allen Allen Institute for Artificial Intelligence (Courtesy appointment at the 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

Google Scholar: https://scholar.google.com/citations?user=viCG2ikAAAAJhl=en
Semantic Scholar: https://www.semanticscholar.org/author/Faeze-Brahman/9252833

Preprints and In Submissions

An Ecosystem for Sandboxing Safety Risks in Human-AI Interactions
Xuhui Zhou, *Faeze Brahman**, Hyunwoo Kim*, Liwei Jiang, Hao Zhu, Ximing Lu, Frank F. Xu, Bill Yuchen Lin, Yejin Choi, Niloofar Mireshghallah, Ronan Le Bras, Maarten Sap @ *Under submission*

Hybrid Preferences: Learning to Route Instances for Human vs. AI Feedback
Lester James Validad Miranda, Yizhong Wang, Yanai Elazar, Sachin Kumar, Valentina Pyatkin, *Faeze Brahman*, Noah A. Smith, Hannaneh Hajishirzi, Pradeep Dasigi. *@ Under submission*

RESTOR: Knowledge Recovery through Machine Unlearning
Keivan Rezaei, Khyathi Chandu, Soheil Feizi, Yejin Choi, *Faeze Brahman*, Abhilasha Ravichander. *@ Under submission*

TÜLU 3: Pushing Frontiers in Open Language Model Post-Training
*Faeze Brahman**, Nathan Lambert*, Jacob Morrison*, Valentina Pyatkin*, Shengyi Huang*, Hamish Ivison*, and et al. *@ arXiv*

Journal and Peer-reviewed Conference papers

Trust or Escalate: LLM Judges with Provable Guarantees for Human Agreement
Jaehun Jung, *Faeze Brahman*, and Yejin Choi. *ICLR 2025*

AI-LieDar : Examine the Trade-off Between Utility and Truthfulness in LLM Agents
Zhe Su, Xuhui Zhou, Sanketh Rangreji, Anubha Kabra, Julia Mendelsohn, *Faeze Brahman*, Maarten Sap . *NAACL 2025*

The Art of Saying No: Contextual Noncompliance in Language Models
*Faeze Brahman**, Sachin Kumar*, Vidhisha Balachandran, Pradeep Dasigi, Valentina Pyatkin, Abhilasha Ravichander, Sarah Wiegreffe, Nouha Dziri, Khyathi Chandu, Jack Hessel, Yulia Tsvetkov, Noah A. Smith, Yejin Choi, and Hannaneh Hajishirzi. *@ Neurips D&B Track*

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. *ICLR 2025*

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, Ximing Lu, Maarten Sap, Nouha Dzir, and Yejin Choii. *@ Neurips 2024*

In Search of the Long-Tail: Systematic Generation of Long-Tail Inferential Knowledge via Logical Rule Guided Search
Huihan Li, Yuting Ning, Zeyi Liao, Siyuan Wang, Xiang Lorraine Li, Ximing Lu, Wenting Zhao, *Faeze Brahman*, Yejin Choi, Xiang Ren. *@ EMNLP 2024*

How to Train Your Fact Verifier: Knowledge Transfer with Multimodal Open Models
Jaeyoung Lee, Ximing Lu, Jack Hessel, *Faeze Brahman*, Youngjae Yu, Yonatan Bisk, Yejin Choi, Saadia Gabriel. *@ Findings of EMNLP 2024*

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*

Tailoring with Targeted Precision: Edit-Based Agents for Open-Domain Procedure Customization
Yash Kumar Lal, Li Zhang, *Faeze Brahman*, Bodhisattwa Prasad Majumder, Peter Clark, Niket Tandon. *@ Findings of 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*

"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 Varghase, 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*

Selected Honors & Awards

2024	**Rising Star in Generative AI**, University of Massachusetts, Amherst (Awarded to 9 people) — <unable to attend>
2021	**Sabbatical Dissertation Fellowship**, Jack Baskin School of Engineering, UC Santa Cruz
2020	**Grace Hopper Scholarship**, Anita Borg Institute **Tapia Celebration of Diversity Scholarship**, CMDiT
2019	**Grace Hopper Scholar Scholarship**, Anita Borg Institute **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**, ACM Computing Research Association
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

Teaching Experience

Assistant

2020	**CSE 115: Software Design Project**, Teaching Assistant **CSE 16: Discrete Mathematics**, Teaching Assistant	*UCSC*
2019	**ML and NLP [Summer Program for High School Scholars]**, Teaching Assistant	*UCSC's COSMOS Program*
2018	**CSE 142: Machine Learning and Data Mining**, Teaching Assistant	*UCSC*
2017	**TIM 50: Business Information Systems**, Teaching Assistant	*UCSC*

Guest Lectures

2023	**Language, Knowledge, and Reasoning**, Graduate Advanced NLP Seminar	*UW*
2018	**Introduction to Neural Networks**, Undergraduate ML course (Fall)	*UCSC*
2018	**Introduction to Neural Networks**, Undergraduate ML course (Spring)	*UCSC*
2018	**Generative Adversarial Networks**, graduate ML course (Fall)	*UCSC*

Invited 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 Language Reasoning and 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.

Mentoring

(Met at least weekly during the course of project)

2024	**Xuhui Zhou**, AI2 Intern (co-mentor: Maarten Sap) → *submission to ICLR*
2024	**Keivan Rezaei**, AI2 Intern (co-mentor: Abhilasha Ravichander) → *submission to NAACL*
2023	**Da Yin**, AI2 Intern (co-mentor: Yuchen Lin) → *publication @ACL 2024*
2023	**Yufei Tian**, AI2 Intern (co-mentor: Ronan Le Bras) → *publication @NAACL 2024*
2023	**Skyler Hallinan**, Master Student, University of Washington → *publication @ EMNLP Findings 2023*
2022	**Kavel Rao**, Undergraduate, University of Washington (co-mentor: Liwei Jiang) → *publication @EMNLP Findings 2023*
2022	**Hanji Chen**, AI2 Intern (co-mentor: Swabha Swayamdipta) → *publication @ACL 2023*
2021	**Tengaho Huang**, Undergraduate, UNC Chapel Hill (co-mentor: Snigdha Chaturvedi) → *publication @EMNLP Findings 2021, 2023*
2020	**Meng Huang**, Master Student, University of Chicago (co-mentor: Mrinmaya Sachan) → *publication at EMNLP Findings 2021*

Service & Professional Activities

Program Committee & Reviewing

2024	**(Senior) Area Chair**, NAACL, COLING, ACL Rolling Review (ARR)
2024	**Program Committee**, Neurips, ICLR, ACL Rolling Review (ARR)
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
`https://bit.ly/wnu2024`

- The first workshop on Creative AI across modalities @ AAAI2023
`https://creativeai-ws.github.io`

- The fifth workshop on Narrative Understanding (WNU) @ ACL2023
`https://sites.google.com/umass.edu/wnu2023`

- 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)

References

Yejin Choi (yejin@cs.washington.edu)
Brett Helsel Professor
Paul G. Allen School of Computer Science and Engineering, University of Washington

Snigdha Chaturvedi (snigdha@cs.unc.edu)
Associate Professor
Computer Science, UNC-Chapel Hill

Maarten Sap (maartensap@cmu.edu)
Assistant Professor
LTI, Carnegie Mellon University

Xiang Ren (xiangren@usc.edu)
Associate Professor
Computer Science, University of Southern California

Vered Shwartz (vered.shwartz@ubc.ca)
Assistant Professor
Computer Science, University of British Columbia