Brihi Joshi

in | | |

Email : brihijos@usc.edu
Website : brihijoshi.github.io

Research Interests

Human-AI Interaction and Personalization: I focus on two axes within Human-AI Interaction. First, I evaluate and improve the *personalization* of AI systems for different users. Second, I explore new paradigms of interaction, designing and studying complex multi-turn dialogues and explanation-based systems.
Human-centered Explanations: I use explanations as a medium of communication between AI systems and users, studying both the generation of high-quality explanations by models and their role in enhancing human understanding.

Education

- **University of Southern California** — Los Angeles, CA
  Doctor of Philosophy, Computer Science — *Aug. 2021 – Present*
  - Advised By: Xiang Ren and Swabha Swayamdipta
  - Apple Scholars in AI/ML PhD Fellowship, AY 2024-26
  - Amazon ML PhD Fellow, AY 2023-24
  - USC Annenberg PhD Fellow, AY 2021-22
- **Indraprastha Institute of Information Technology** — New Delhi, India
  Bachelor of Technology, Computer Science and Engineering; CGPA: 9.30/10 — *Aug. 2016 – Dec 2020*
  - Received the Innovative Student Projects Award for **best thesis in Computer Science** from the Indian National Academy of Engineering. One of the highest honors for undergraduates in India.
  - Snap Research Scholarship, 2019

Publications and Preprints

Human-AI Interaction and Personalization

- Sahana Ramnath, Anurag Mudgil, **Brihi Joshi**, Skyler Hallinan, Xiang Ren. Amulet: Putting Complex Multi-Turn Conversations on the Stand with LLM Juries. *EMNLP (Main) 2025*

- **Brihi Joshi**, Xiang Ren, Swabha Swayamdipta, Rik Koncel-Kedziorski, Tim Paek. Improving LLM Personas via Rationalization with Psychological Scaffolds. *Findings of EMNLP 2025*

- Rik Koncel-Kedziorski, **Brihi Joshi**, Tim Paek. PrimeX: A dataset of worldview, opinion, and explanation. *EMNLP (Main) 2025*

- **Brihi Joshi***, Keyu He*, Sahana Ramnath, Sadra Sabouri, Kaitlyn Zhou, Souti Chattopadhyay, Swabha Swayamdipta, Xiang Ren. ELI-Why: Evaluating the Pedagogical Utility of Language Model Explanations. *Findings of ACL 2025*

- Jaspreet Ranjit, **Brihi Joshi**, Rebecca Dorn, Laura Petry, Olga Koumoundouros, Jayne Bottarini, Peichen Liu, Eric Rice, Swabha Swayamdipta. OATH-Frames: Characterizing Online Attitudes Towards Homelessness via LLM Assistants. *EMNLP (Main) 2024*, **Outstanding Paper Award**

Human-centered Explanations

- Keyu He, Tejas Srinivasan*, **Brihi Joshi***, Xiang Ren, Jesse Thomason, Swabha Swayamdipta. Believing without Seeing: Quality Scores for Contextualizing Vision-Language Model Explanations. *Under Submission*

- Isabelle Lee, Emmy Liu, Cathy Jiao, Michael Saxon, **Brihi Joshi**, Dani Yogatama, Fazl Barez, Naomi Saphra. Position: Stop Using Brittle Interpretations. *Under Submission*

- **Brihi Joshi**, Sriram Venkatapathy, Mohit Bansal, Nanyun Peng, Haw-Shiuan Chang. CoKe: A Simple Yet Effective Approach to Story Evaluation via Chain-of-Keyword Rationalization. *Oral at GEM Workshop, ACL 2025*

- Sahana Ramnath, **Brihi Joshi**, Skyler Hallinan, Ximing Lu, Liunian Harold Li, Aaron Chan, Jack Hessel, Yejin Choi, Xiang Ren. Tailoring Self-Rationalizers with Multi-Reward Distillation. *ICLR 2024* and SetLLM@ICLR 2024.

- Aaron Chan*, Zhiyuan Zeng*, Wyatt Lake, **Brihi Joshi**, Hanjie Chen, Xiang Ren. KNIFE: Knowledge Distillation with Free-Text Rationales. *TrustML-(un)Limited@ICLR, 2023*

- **Brihi Joshi***, Ziyi Liu*, Sahana Ramnath, Aaron Chan, Zhewei Tong, Shaoliang Nie, Qifan Wang, Yejin Choi and Xiang Ren. Are Machine Rationales (Not) Useful to Humans? Measuring and Improving Human Utility of Free-text Rationales. *ACL (Main) 2023 and trAIt@CHI 2023*

- Dong-Ho Lee*, Akshen Kadakia*, **Brihi Joshi**, Aaron Chan, Ziyi Liu, Takashi Shibuya, Ryosuke Mitani, Toshiyuki Sekiya, Jay Pujara, Xiang Ren. XMD: An End-to-End Framework for Interactive Explanation-Based Debugging of NLP Models. *ACL Demo Track 2023*

- **Brihi Joshi***, Aaron Chan*, Ziyi Liu*, Shaoliang Nie, Maziar Sanjabi, Hamed Firooz, Xiang Ren. ER-Test: Evaluating Explanation Regularization Methods for NLP Models. *Findings of EMNLP 2022 and TrustNLP@NAACL 2022*

Older Work

- **Brihi Joshi**, Neil Shah, Francesco Barbieri and Leonardo Neves. The Devil is in the Details: Evaluating Limitations of Transformer-based Methods for Granular Tasks. In *The 28th ACM International Conference on Computational Linguistics (COLING 2020).*

- Aditya Chetan*, **Brihi Joshi***, Hridoy Sankar Dutta, Tanmoy Chakraborty. CoReRank: Ranking to Detect Users Involved in Blackmarket-based Collusive Retweeting Activities. In *The 12th ACM International Conference on Web Search and Data Mining (WSDM 2019).* (Acceptance Rate: 16%, CORE2018 A*)

- Udit Arora, Hridoy Sankar Dutta, **Brihi Joshi***,Aditya Chetan*,Tanmoy Chakraborty. Analyzing and Detecting Collusive Users Involved in Blackmarket Retweeting Activities. In *ACM Transactions on Intelligent Systems and Technology (TIST).* (Impact Factor: **3.971**)

- **Brihi Joshi***, Amogh Gulati*, Chirag Jain*, Jainendra Shukla. It's Not What They Play, It's What You Hear: Understanding Perceived vs. Induced Emotions in Hindustani Classical Music. In *22nd ACM International Conference on Multimodal Interaction, Late Breaking Reports (ICMI 2020).*

- **Brihi Joshi***, Shravika Mittal*, Aditya Chetan*. Did You "Read" the Next Episode? Using Textual Cues for Predicting Podcast Popularity. In *First Workshop on NLP for Music and Audio (NLP4MusA) at International Society for Music Information Retrieval Conference (ISMIR 2020).*

- Hridoy Sankar Dutta, **Brihi Joshi***, Aditya Chetan*, Tanmoy Chakraborty. Retweet Us, We Will Retweet You: Spotting Collusive Retweeters Involved in Blackmarket Services. In *The 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2018).* (Acceptance Rate: 15%)

- Nishtha Madaan, Gautam Singh, Sameep Mehta, Aditya Chetan*, **Brihi Joshi***. Generating Clues for Gender based Occupation De-biasing in Text `arXiv:1804.03839 [cs.CL]`

* Equal Contribution

Internships and Work Experience

- **Apple** — Seattle, WA
  ML Research Intern, Apple AIML: Adaptive Devices and Platforms Team — *May 2025 – Aug 2024*
- **Apple** — Seattle, WA
  ML Research Intern, Apple AIML: Adaptive Devices and Platforms Team — *May 2024 – Dec 2024*
- **Amazon** — Cambridge, MA
  Applied Scientist II Intern, Alexa AI Natural Understanding: NLG Team — *May 2023 – Aug. 2023*

- **Goldman Sachs** Bangalore, India
  Analyst (Full Time), Regulatory Engineering Team *Dec. 2020 – July 2021*
- **Snap Inc.** Los Angeles, CA
  Research Intern, Computational Social Science Team *Sep. 2019 – Dec. 2019*
- **Goldman Sachs** Bangalore, India
  Analyst Intern, Regulatory Engineering Team *May 2019 - July 2019*
- **IBM Research** New Delhi, India
  Research Intern, Fairness in ML Team *May 2018 - July 2018*

Awards

- **Apple Scholars in AL/ML PhD Fellowship**: Awarded for PhD research in Human Centred AI and Explainability.
- **Amazon ML PhD Fellowship**: Awarded for work in Explainable NLP, for AY 2023-24.
- **Indian National Academy of Engineering Undergraduate Thesis Award 2021**: Awarded for research done as a part of undergraduate thesis in the Computer and Information Sciences Domain.
- **Annenberg Fellowship**: Awarded for admission to the PhD program at the University of Southern California.
- **Snap Research Scholarship 2019**: Awarded for research done in the field of Machine Learning. Award includes 10,000 USD and an offer to intern at Snap Research, USA. Only scholar from India!
- **AAAI 2020 Undergraduate Consortium**: Accepted to present my thesis at the AAAI 2020 Undergraduate consortium. Includes scholarship to attend to attend AAAI 2020
- **Microsoft Research India Travel Grant**: Awarded travel support of 50000 INR for visiting WSDM 2019
- **ACM-W Scholarship**: Awarded travel support of 1200 USD for visiting WSDM 2019
- **Google Women Techmakers Scholarship, 2018**: Awarded to students who work for diversity and inclusion in the field of Computer Science.
- **Best Technical Poster Runner-up at GHCI 2018**: Received for the project, "Generating Clues for Gender based Occupation De-biasing in Text"
- **Dean's Award for Innovation R&D**: Awarded to students who work on Research projects beyond coursework. Awarded for the academic years 2016-17 and 2017-18.
- **Dean's List of Academic Affairs**: Awarded to students who demonstrate excellence in an academic year. Awarded for the academic year 2016-17 and 2018-19.
- **Grace Hoppers Celebration India (GHCI) Scholarship, 2018**: Awarded travel grant and scholarship to attend the GHCI conference.

Service and Leadership

- **Core Organizer, NLP With Friends**: Core Organizer of NLP with Friends, an online seminar series for PhD students to discuss all things NLP Research.
- **Director, Women Who Code Delhi**: Lead the Delhi Chapter of Women Who Code, a non-profit organisation for upliftment of minorities in techonology.
- **Workshop Organizer, Broadening Research Collaborations in ML, NeurIPS 2022**: PC member and organizer for new workshop at NeurIPS aimed towards making ML research more accessible.
- **Reviewer**: EMNLP 2022, ACL 2023, EMNLP 2023, EMNLP 2024, ACL 2025, EMNLP 2025
- **Grant Writing**: *Utilizing Explanations for Model Refinement* in Alexa: Fairness in AI 2022 award and writing contributions in multiple DARPA and IARPA project proposals under the supervision of Prof. Xiang Ren.
- **Mentoring**: Harshavardhan Alimi (USC MS CS), Keyu He (USC CS Undergrad, CURVE Fellowship Winter, Summer and Fall 2024, Winter 2025, Joined CMU MIIS in Fall 2025), Ziyi Liu (USC MS CS, Joined USC CS PhD in Fall 2023), Zhewei Tong (Viterbi Tsinghua Undergraduate Summer Research Program, Joined CMU MS CS in Fall 2023), Pushpdeep Singh (Indo-US Science and Technology Forum).

Teaching Experience

- **CSCI 544: Graduate Applied Natural Language Processing** USC
  Responsible for mentoring student projects and creating assignments *Aug 2024 - Dec 2024*
- **CSE343: Machine Learning** IIIT, Delhi
  Teaching Assistant for a class of 150 senior undergraduate students *Aug 2020 - Dec 2020*
- **CSE632: Semantic Web** IIIT, Delhi
  Teaching Assistant for a class of 170 senior undergraduate and graduate students *Jan 2020 - May 2020*
- **MTH201: Probability and Statistics** IIIT, Delhi
  Teaching Assistant for a freshman class of 300 students *Jan 2019 - May 2019*