

Faeze Brahman

Senior Research Scientist, Allen Institute for AI (Ai2)

faezeb@allenai.org | fabrahman.github.io | [Google Scholar](#)

Work Experience

2024 – present	Senior Research Scientist, Allen Institute for AI (Ai2) OLMo Post-training team
Apr 2022 – 2024	Postdoctoral Researcher, Allen Institute for AI (Ai2) and University of Washington Mentor: Yejin Choi
Jun – Sep 2021	Research Intern, Microsoft Research Host: Michel Galley, Jianfeng Gao
Jun – Sep 2020	Research Intern, Allen Institute for AI (Mosaic) Host: Vered Shwartz, Yejin Choi
Jun – Sep 2018	Research Intern, Xerox PARC Host: Kyle Dent

Education

2022	Ph.D. in Computer Science, University of California, Santa Cruz Advisor: Snigdha Chaturvedi Thesis: <i>Modeling Key Narrative Elements for Automatic Story Generation</i>
2018	M.S. in Computer Science, University of California, Santa Cruz Advisor: Marilyn Walker
2014	M.S. in Electrical Engineering, Iran University of Science and Technology
2012	B.S. in Electrical Engineering, Iran University of Science and Technology

Awards & Honors

2025	ICLR Oral (top 1.8%) for <i>Trust or Escalate</i>
2024	Rising Star, GenAI Workshop, UMass Amherst
2024	Best Paper Nomination at NAACL for <i>MacGyver</i>
2021	Sabbatical Dissertation Fellowship, UC Santa Cruz
2019–20	Grace Hopper Celebration and Tapia Diversity Scholarships; CRA-W Travel Grant
2016	Regents Fellowship, UC Santa Cruz

Selected Publications

* denotes equal contribution. For the full publication list, please check [Google Scholar](#).

Language Model Post-training & Development

F Brahman^{*}, N Lambert^{*}, J Morrison^{*}, V Pyatkin^{*}, S Huang^{*}, H Ivison^{*}, L Miranda^{*}, ..., H Hajishirzi. [TÜLU 3: Pushing Frontiers in Open Language Model Post-Training](#). *COLM* (2025).

Olmo Team (core contributors, including **F Brahman**). [Olmo 3](#). 2025.

OLMo Team, ..., **F Brahman**, ..., N Smith, H Hajishirzi. [2 OLMo 2 Furious](#). *COLM* (2025).

R Shao, A Asai, S Shen, H Ivison, ..., **F Brahman**, ..., P Koh. [DR Tulu: Reinforcement Learning with Evolving Rubrics](#)

for Deep Research. *Preprint* (2025).

T Chen, A Asai, L Zettlemoyer, H Hajishirzi, **F Brahman**. [Train for Truth, Keep the Skills: Binary Retrieval-Augmented Reward Mitigates Hallucinations](#). *Preprint* (2025).

L Miranda*, Y Wang*, Y Elazar, S Kumar, V Pyatkin, **F Brahman**, N Smith, H Hajishirzi, P Dasigi. [Hybrid Preferences: Learning to Route Instances for Human vs. AI Feedback](#). *ACL* (2025).

Y Wang, **F Brahman**, S Feng, T Xiao, H Hajishirzi, Y Tsvetkov. [Small Reward Models via Backward Inference](#). *Preprint* (2026).

T Xiao, Y Yuan, H Ivison, H Zhu, **F Brahman**, N Lambert, P Dasigi, N Smith, H Hajishirzi. [Meta-Reinforcement Learning with Self-Reflection for Agentic Search](#). *COLM* (2026).

H Ivison, M Zhang, **F Brahman**, P Koh, P Dasigi. [Large-Scale Data Selection for Instruction Tuning](#). *Preprint* (2025).

Safety, Evaluation & Agents

L Jiang, K Rao, S Han, A Ettinger, **F Brahman**, S Kumar, N Mireshghallah, X Lu, M Sap, Y Choi, N Dziri. [WildTeaming at Scale: From In-the-Wild Jailbreaks to \(Adversarially\) Safer Language Models](#). *NeurIPS* (2024).

F Brahman*, S Kumar*, V Balachandran, P Dasigi, V Pyatkin, A Ravichander, S Wiegrefe, N Dziri, K Chandu, J Hessel, Y Tsvetkov, N Smith, Y Choi, H Hajishirzi. [The Art of Saying No: Contextual Noncompliance in Language Models](#). *NeurIPS D&B* (2024).

J Jung, **F Brahman**, Y Choi. [Trust or Escalate: LLM Judges with Provable Guarantees for Human Agreement](#). *ICLR (Oral)* (2025).

B Lin, Y Deng, K Chandu, **F Brahman**, A Ravichander, V Pyatkin, N Dziri, R Le Bras, Y Choi. [WildBench: Benchmarking LLMs with Challenging Tasks from Real Users in the Wild](#). *ICLR* (2025).

X Zhou, **F Brahman***, H Kim*, L Jiang, H Zhu, X Lu, F Xu, B Lin, Y Choi, N Mireshghallah, R Le Bras, M Sap. [HAICOSYSTEM: An Ecosystem for Sandboxing Safety Risks in Human-AI Interactions](#). *COLM* (2025).

Z Su, X Zhou, S Rangrej, A Kabra, J Mendelsohn, **F Brahman**, M Sap. [AI-LieDar: Examine the Trade-off Between Utility and Truthfulness in LLM Agents](#). *NAACL* (2025).

P Röttger, M Hinck, V Hofmann, K Hackenburg, V Pyatkin, **F Brahman**, D Hovy. [IssueBench: Millions of Realistic Prompts for Measuring Issue Bias in LLM Writing Assistance](#). *TACL* (2026).

T Chen, **F Brahman**, J Liu, N Mireshghallah, W Shi, P Koh, L Zettlemoyer, H Hajishirzi. [ParaPO: Aligning Language Models to Reduce Verbatim Reproduction of Pre-training Data](#). *COLM* (2025).

K Rezaei, K Chandu, S Feizi, Y Choi, **F Brahman**, A Ravichander. [RESTOR: Knowledge Recovery through Machine Unlearning](#). *TMLR* (2025).

Reasoning, Personalization & Inference-time Algorithm

S Li*, A Bose*, **F Brahman**, S Du, P Koh, M Fazel, Y Tsvetkov. [PrefDisco: Benchmarking Proactive Personalized Reasoning](#). *ICLR* (2026).

Z Zheng, V Balachandran, C Park, **F Brahman**, S Kumar. [Reasoning Up the Instruction Ladder for Controllable Language Models](#). *Findings of ACL* (2026).

F Brahman*, P West*, X Lu*, N Dziri*, L Li*, J Hwang, L Jiang, J Fisher, A Ravichander, K Chandu, B Newman, P Koh, A Ettinger, Y Choi. [The Generative AI Paradox: “What It Can Create, It May Not Understand”](#). *ICLR* (2024).

Y Tian, A Ravichander, L Qin, R Le Bras, R Marjeh, N Peng, Y Choi, T Griffiths, **F Brahman**. [MacGyver: Are Large Language Models Creative Problem Solvers?](#) *NAACL (Best Paper Nominee)* (2024).

F Brahman, C Bhagavatula, V Pyatkin*, J Hwang*, X Li, H Arai, S Sanyal, K Sakaguchi, X Ren, Y Choi. [PlaSma: Making Small Language Models Better Procedural Knowledge Models for \(Counterfactual\) Planning](#). *ICLR* (2024).

D Yin, **F Brahman**, A Ravichander, K Chandu, K Chang, Y Choi, B Lin. [Agent Lumos: Unified and Modular Training for Open-Source Language Agents](#). *ACL* (2024).

S Welleck*, X Lu*, **F Brahman**⁺, P West⁺, T Shen, D Khashabi, Y Choi. [Generating Sequences by Learning to Self-Correct](#).

ICLR (2023).

X Lu, **F Brahman**, P West, J Jung, K Chandu, A Ravichander, L Qin, P Ammanabrolu, L Jiang, S Ramnath, ..., S Welleck, Y Choi. [Inference-Time Policy Adapters \(IPA\): Tailoring Extreme-Scale LMs without Fine-tuning](#). *EMNLP* (2023).

J Jung, L Qin, S Welleck, **F Brahman**, C Bhagavatula, R Le Bras, Y Choi. [Maieutic Prompting: Logically Consistent Reasoning with Recursive Explanations](#). *EMNLP* (2022).

Narrative Understanding & Story Generation

F Brahman, B Peng, M Galley, S Rao, B Dolan, S Chaturvedi, J Gao. [Grounded Keys-to-Text Generation: Towards Factual Open-Ended Generation](#). *Findings of EMNLP* (2022).

F Brahman, M Huang, O Tafjord, C Zhao, M Sachan, S Chaturvedi. [“Let Your Characters Tell Their Story”: A Dataset for Character-Centric Narrative Understanding](#). *Findings of EMNLP* (2021).

F Brahman, S Chaturvedi. [Modeling Protagonist Emotions for Emotion-Aware Storytelling](#). *EMNLP* (2020).

Invited Talks

2025	Keynote, Singapore Symposium on NLP (SSNLP) UW RAISE Seminar Series UCLA NLP Seminar Series
2024	Bocconi University (MilaNLP) UBC NLP Group
2023	UMass Amherst NLP Seminar

Service, Outreach & Mentorship

Chairing **Ethics Chair**, NAACL 2027; Area Chair, COLM 2025 and COLING 2025; Senior Area Chair, NAACL 2024

Reviewing Program Committee for NeurIPS, ICLR, ACL, EMNLP, AAAI, IJCAI, ARR

Organization [Second Workshop on Customizable NLP \(CustomNLP4U\)](#) at ACL 2026; tutorial [Guardrails and Security for LLMs: Safe, Secure and Controllable Steering of LLM Applications](#) at ACL 2025; Workshop on Narrative Understanding (WNU) at NAACL 2021, NAACL 2022, ACL 2023, and EMNLP 2024

Teaching Guest lecturer, UW CSE 599 (2023); TA for Machine Learning and Data Mining (2018) and Software Design (2020) at UCSC

Diversity Co-founded the Women-in-AI group at UCSC (2019); mentor at WiNLP, NAACL 2022

Mentorship Mentored 13+ interns, PhD/MS students, and pre-docs (UW, CMU, UCLA, ETH, UMD, UNC), with student-led work published at ACL, NAACL, EMNLP, COLM, and TMLR. Mentees include Yufei Tian (UCLA, now OpenAI), Hanjie Chen (UVA, now Rice), Afra Amini (ETH), Da Yin (UCLA), Skyler Hallinan (UW, now USC PhD), and Tong Chen and Yike Wang (UW).