

Pardis Sadat Zahraei

✉ zahraei2@illinois.edu 📧 Pardis Sadat Zahraei 🌐 pardis-zahraei 🐦 @PardisZahraei 😊 PardisSzah

Education

PhD in Computer Science

Aug 2025 – May 2030

University of Illinois at Urbana-Champaign (UIUC)

Illinois, USA

- Advised by Prof. Gökhan Tür and Prof. Dilek Hakkani-Tür

B.Sc in Computer Engineering

Nov 2020 - Jul. 2025

Sharif University of Technology

Final Grade: 18.27/20

Selected Publications

Emergent Unfaithfulness: How Alignment Training Causes Language Models to Silently Override Task Faithfulness

*Authors: **Pardis Sadat Zahraei**, Janvijay Singh, Gokhan Tur, Dilek Hakkani-Tür*

COLM 2026 🏆 Outstanding Paper Award @ KnowFM Workshop (ACL 2026)

Asymmetrically Unfair Decision Making and Reasoning in LLMs

*Authors: **Pardis Sadat Zahraei**, Faeze Brahman, Dilek Hakkani-Tür, Gokhan Tur, Maarten Sap*

Under review ICLR 2027

The Alignment Veto: How Safety Training Suppresses Cultural Knowledge in LLMs

*Authors: **Pardis Sadat Zahraei**, Gokhan Tur, Dilek Hakkani-Tür, Ehsaneddin Asgari*

EMNLP 2026 - Oral 🎤

AgentAbstain: Do LLM Agents Know When Not to Act?

Authors: Xun Liu[†], Yi Evie Zhang[†], Vira Kasprova, Parisa Rabbani*, **Pardis Sadat Zahraei***, et al.*

NeurIPS 2026 Evaluations and Datasets - Oral 🎤

Prior Beliefs Prejudice LLM-as-Judge: Evidence from Persuasion Evaluation

*Authors: **Pardis Sadat Zahraei**, Xiaoning Wang, Beyza Bozdog, Gokhan Tur, Dilek Hakkani-Tur*

Findings of ACL 2026

WSC+: Enhancing The Winograd Schema Challenge Using Tree-of-Experts

*Authors: **Pardis Sadat Zahraei**, Ali Emami*

EACL 2024 - Oral 🎤

TuringQ: Benchmarking AI Comprehension in Theory of Computation

*Authors: **Pardis Sadat Zahraei**, Ehsaneddin Asgari*

Findings of EMNLP 2024

Translate With Care: Addressing Gender Bias, Neutrality, and Reasoning in LLM Translations

*Authors: **Pardis Sadat Zahraei**, Ali Emami*

Findings of ACL 2025 ⚡ Lightning Talk at GeBNLP Workshop

Research Experience

Student Researcher at AI2

Jul 2026 – Aug 2026

Open Ecosystem Team, Allen Institute for AI (Supervised by Prof. Maarten Sap)

Research Interests

- LLM Personas & Pragmatic Interpretability
- LLM Evaluation and Benchmarking
- AI Safety and Alignment
- Data-Centric Machine Learning

Academic Service

- **Reviewer:** NeurIPS 2026 (**Top Reviewer** 🏆) ICLR 2027, COLM 2026, ACL, EACL, NAACL and EMNLP (ARR Cycles, starting 2024)