

Yejin Son

sonyejin@cs.ubc.ca | [Homepage](#) | [Google Scholar](#)

I am a Ph.D. student in Computer Science at the University of British Columbia and a Vector Institute affiliate. My research focuses on **Functional Theory of Mind** and **realistic user simulation**. I study how AI can use social reasoning to guide its actions and develop user simulations that capture the diversity of human behavior, with the goal of building AI that meaningfully supports human interaction and relationships.

EDUCATION

University of British Columbia <i>Ph.D. in Computer Science</i> (Advisor: Prof. Vered Shwartz) Vector Institute Affiliate	Sep 2026 – Present
Yonsei University <i>M.S. in Artificial Intelligence</i> (Advisor: Prof. Youngjae Yu) <i>GPA: 4.20/4.30</i>	Sep 2023 – Aug 2026
Yonsei University <i>B.A. in Economics and Applied Statistics</i> <i>GPA: 4.04/4.30</i>	Mar 2021 – Aug 2023

RESEARCH & TEACHING EXPERIENCE

Research Assistant — Vector Institute & University of British Columbia NLP Group <i>Advisor: Prof. Vered Shwartz</i> Conducting research on functional Theory of Mind and realistic user simulation to develop and reliably evaluate socially aware interactive AI agents across diverse human behaviors.	Sep 2026 – Present
Visiting Graduate Researcher — Vector Institute & University of Toronto <i>Advisor: Prof. Zhijing Jin</i> Investigated latent cultural decision boundaries in LLMs through persona-projection divergence, resulting in a first-authored paper at the ICML 2026 Workshop on Pluralistic Alignment.	Jan 2026 – Jul 2026
Research Assistant — MIR Lab, Yonsei University <i>Advisor: Prof. Youngjae Yu</i> Led and co-led research on multimodal reasoning, LLM safety, and alignment, resulting in first- and co-first-author papers at ACL and EMNLP and a publication at COLM.	Mar 2023 – Aug 2026
Teaching Assistant — Multimodal Deep Learning (graduate), Yonsei University	Spring 2024

PUBLICATIONS

* equal contribution † co-corresponding author (8 peer-reviewed publications and 1 preprint)

- [1] S. Lee, J. Jeong, D. Kim, **Y. Son**, Y. Yu. [Mind the Motions: Benchmarking Theory-of-Mind in Everyday Body Language](#). (EMNLP 2026, Findings)
- [2] J. Park, H. Lee, D. Lee, D. Choi, **Y. Son**, H. Park, Y. Yu. [Multimodal Safety Evaluation Should Measure Controllability Beyond Classification](#). (arXiv preprint, 2026)
- [3] S. Kim*, **Y. Son***, J. Park, K. Ka, S. Lee, J. Lee, H. Jang†, A. Oh†, Y. Yu†. [Investigating Counterfactual Unfairness in LLMs towards Identities through Humor](#). (ACL 2026, Main)
- [4] K. Kim, H. Kim, **Y. Son**, S. Kim, K. M. Lee. [Does Persona Make LLMs K-pop Fans? A Pilot Study of LLM-Based Online Concert Audience Agents](#). (ICML 2026 Culture x AI Workshop, Oral)
- [5] **Y. Son**, Y. Yang, R. Faulkner, M. Ratto, S. Lim, Y. Yu, Z. Jin. [The Wedge Questions: Latent Cultural Boundaries in LLMs via Persona Projection Divergence](#). (ICML 2026 Pluralistic Alignment Workshop)
- [6] **Y. Son***, M. Kim*, S. Kim, S. Han, J. Kim, D. Jang, Y. Yu†, C. Y. Park†. [Subtle Risks, Critical Failures: A Framework for Diagnosing Physical Safety of LLMs for Embodied Decision Making](#). (EMNLP 2025, Main)
- [7] **Y. Son***, S. Kim*, D. Min, Y. Yu. [Multimodal UNcommonsense: From Odd to Ordinary and Ordinary to Odd](#). (EMNLP 2025, Findings)
- [8] Y. Choi, Y. Yeo, **Y. Son**, S. Han, Y. Yu. [G1yphD3c0de: Towards Safer Language Models on Visually Perturbed Texts](#). (COLM 2025)
- [9] S. Han, J. Kim, J. Hessel, L. Jiang, J. Chung, **Y. Son**, Y. Choi, Y. Yu. [Reading Books is Great, But Not if You Are Driving! Visually Grounded Reasoning about Defeasible Commonsense Norms](#). (EMNLP 2023, Oral)

AWARDS & HONORS

Four Year Doctoral Fellowship (4YF) , University of British Columbia (stipend & tuition award)	2026
AI Convergence Education Program Visiting Scholarship, IITP (approx. US\$37K; one of 25 awardees nationwide)	2025–2026
AI Seoul Tech Graduate Scholarship, Seoul Scholarship Foundation (approx. US\$3.4K)	2025
Y2Y Merit Scholarship, Yonsei University (approx. US\$8.7K)	2023–2024

ACADEMIC SERVICE

Program Committee: Trustworthy AI for Good Workshops at ICML 2026 and NeurIPS 2026
Reviewer: NeurIPS 2026 | ACL 2026 Demo Track | ICML 2026 Workshop on Pluralistic Alignment | EACL 2026 Student Research Workshop | NAACL 2025 Student Research Workshop

TECHNICAL SKILLS & LANGUAGES

Programming & Tools: Python, PyTorch, TensorFlow, NumPy, LangChain, LangGraph, R, MySQL, Git
Languages: Korean (Native), English (Fluent)