Seungbeen Lee

Mail | Home | LinkedIn | Google Scholar

Research Interest

I truly believe that robots with sophisticated design can fulfill the **basic social needs** of the human. Based on my psychological background and multimodal interest, I want to do **scalable** HRI research which can contribute to **attractiveness** of social agents.

Education

Yonsei University, BA in Psychology and Economics — Mar 2018 – Feb 2023

- GPA: 4.01/4.5
- Coursework: R and Python Programming, Statistical Methodology, Introduction to Quant, Cognitive Psychology, Personality Psychology, Game Theory, Dynamic Macroeconomics, Psychometrics

Yonsei University, MS in Artificial Intelligence — Sept 2023 – **Feb 2026**

- GPA: 3.5/4.5 (Advisor: Youngjae Yu)

Carnegie Mellon University, Visiting researcher, Robotics Institute — Aug 2025 – **Feb 2026**

- Supported by Korea's IITP scholarship program
- Visiting at BIG Group (Advisor: Jean Oh)

Publications

Representation Bending for Large Language Model Safety — ACL 2025

Ashkan Yousefpour; Taeheon Kim; Ryan S. Kwon; ***Seungbeen Lee***; Wonje Jeung; Seungju Han; Alvin Wan; Harrison Ngan; Youngjae Yu; Jonghyun Choi

https://arxiv.org/abs/2504.01550

Verifying the Verifiers: Unveiling Pitfalls and Potentials in Fact Verifiers — COLM 2025

Wooseok Seo; Seungju Han; Jaehun Jung; Benjamin Newman; Seungwon Lim; ***Seungbeen Lee***; Ximing Lu; Yejin Choi; Youngjae Yu

https://arxiv.org/abs/2506.13342

Persona Dynamics: Unveiling the Impact of Personality Traits on Agents in Text-Based Games — ACL 2025

Seungwon Lim; ***Seungbeen Lee***; Dongjun Min; Youngjae Yu

https://arxiv.org/abs/2504.06868

Do LLMs have distinct and consistent personality? Trait: Personality testset designed for LLMs with psychometrics — NAACL 2024

Seungbeen Lee*; Seungwon Lim*; Seungju Han; Giyeong Oh; Hyungjoo Chae; Jiwan Chung; Minju Kim; Beong-woo Kwak; Yeonsoo Lee; Dongha Lee; Jinyeong Yeo; Youngjae Yu

https://arxiv.org/abs/2406.14703

Cactus: Towards Psychological Counseling Conversations using Cognitive Behavioral Theory — EMNLP 2024

Suyeon Lee; Sunghwan Kim; Minju Kim; Dongjin Kang; Dongil Yang; Harim Kim; Minseok Kang; Dayi Jung; Min Hee Kim; ***Seungbeen Lee***; Kyoung-Mee Chung; Youngjae Yu; Dongha Lee; Jinyoung Yeo

https://arxiv.org/abs/2407.03103

Can visual language models resolve textual ambiguity with visual cues? Let visual puns tell you! — EMNLP 2024

Jiwan Chung; Seungwon Lim; Jaehyun Jeon; ***Seungbeen Lee***; Youngjae Yu

https://arxiv.org/abs/2410.01023

Working Papers

Mind the Motions: Benchmarking Theory-of-Mind in Everyday Body Language 2025
Seungbeen Lee; Jinhong Jeong; Donghyun Kim; Yejin Son; Youngjae Yu

Connecting the Dots from Data: LLM-driven Tree-search Career Cartographies of Career Pathways as Your AI Career Explorer 2025
Seungbeen Lee; Keummin Ka; Jinhong Jeong; Chaewon Kim; Youngjae Yu

Selected Media Coverage

- 2025 — ScienceNews, "Are AI chatbot 'personalities' in the eye of the beholder?"

Teaching

TA

- 2025 — Multimodal Deep Learning (AAI5010) (Prof. Youngjae Yu)
- 2024 — Introduction to AI Research (CCO2105) (Prof. Youngwoon Lee)

Other Experience

- 2020 — HR Internship, Human Metrics, Seoul (Human Intelligence Evaluation)