Haein Yeo

haeinyeo@hanyang.ac.kr |

haein-yeo | haaaein |

521, Fusion Technology Center, Hanyang University, Seoul, South Korea 04763

Research Interests

My research lies at the intersection of Natural Language Processing (NLP) and Human-Computer Interaction (HCI). I focus on LLM-human alignment, human-centered evaluation, and AI for social good, approaching these topics from an AI safety perspective. Ultimately, I aim to design AI systems that align with human values, foster trust, and contribute to positive societal impact.

Keywords: AI Alignment, Human-Centered LLM Evaluation, AI for Policy & Governance, AI for Social Good

Experience

- **Hanyang Human-Centered Computing Laboratory [🌐]** *Sep. 2022 - Present*
  Research Assistant Seoul, Korea

- **NAVER Future AI Center [🌐]** *Jan. 2025 - Feb. 2025*
  AI Safety Research Intern Seoul, Korea

Education

- **Hanyang University** *Sep. 2022 - Present*
  M.S. & Ph.D. Integrated Student, Department of Artificial Intelligence (Advisor: Kyungsik Han) Seoul, Korea
- **Dongduk Women's University** *Aug. 2022*
  Bachelor of Science, Division of Computer Science Seoul, Korea

Projects

- **Responsible Capability Scaling (RCS) of General Purpose AI (GPAI)** *Apr. 2024 - Dec. 2024*
  Collaborated with Telecommunications Technology Association ***(TTA), Center for Trustworthy AI***
  - Developed GPAI Risk Management Framework.
  - Conducted research on methodologies for identifying, classifying, and evaluating risk factors.
- **MindHealth** *Sep. 2023 - Jul. 2024*
  - Investigated correlations between user behavior and depression severity to inform personalized intervention strategies.
  - Analyzed log and text data from mental health applications to identify strategies for enhancing user engagement.
- **MOS** *May. 2022 - Present*
  - Development of an LLM-based explanation generation methodology for explainable recommender systems.
  - Development of a dashboard to support decision-making for fashion experts.

Publications

C=Conference, J=Journal, P=Patent, S=In Submission, T=Thesis

[C.8] **Haein Yeo**, Seungwan Jin, Taehyung Noh, Yejin Shin, Sangyeon Kang, Sangwoo Heo, Jiwon Chung, Hwarim Hyun, Kyungsik Han (2026). "Can LLMs Persuade Humans with Deception?": From a Deceptive Strategy Taxonomy to a Large-Scale Empirical Study. *ACM International Conference on Human Factors in Computing Systems* (**CHI**).

[C.7] Taehyung Noh, Seungwan Jin, **Haein Yeo**, Kyungsik Han (2026). TRIPLE: Theory-Driven Integration of Planned and Habitual Behaviors for LLM-based Personalization. *The AAAI Conference on Artificial Intelligence* (**AAAI**). (Oral)

[C.6] Taehyung Noh, Seungwan Jin, **Haein Yeo**, Kyungsik Han (2025). Externalizing Social-Cognitive Structures for User Modeling: Toward Theory-Driven Profiling with LLMs. *The ACM International Conference on Information and Knowledge Management* (**CIKM**).

[J.2] **Haein Yeo**, Taehyung Noh, Kyungsik Han (2025). LLM-Generated Content-Based Explanations for User Experience in Fashion Recommender Systems. (**Fashion and Textiles [SCI(E) Q1, JCR IF = 3.7]**)

[C.5] **Haein Yeo**, Taehyung Noh, Seungwan Jin, Kyungsik Han (2025). PADO: Personality-induced multiAgents for Detecting OCEAN in human-generated texts. *The International Conference on Computational Linguistics* (**COLING**). (Oral, Top 7.9%)

[J.1] Eunji Kim, **Haein Yeo**, Kyungsik Han (2024). A Study on the Personal Fashion Preference in Social Me-dia using Meta-path based Heterogeneous Graph Modeling. (**Journal of KIISE. (Invited paper from KSC 2023)**)

[C.4] Taehyung Noh, **Haein Yeo**, Myungjin Kim, Kyungsik Han (2023). A Study on User Perception and Experience Differences in Recommendation Results by Domain. *The ACM International Conference on Human Factors in Computing Systems* (**CHI LBW**).

[C.3] Taehyung Noh, **Haein Yeo**, Myungjin Kim, Kyungsik Han (2023). Using Deep Learning-Based Visual Hints to Mitigate Hallucinations in Large Language Model. The Proceedings of the Korea Software Congress (**KSC**).

[C.2] **Haein Yeo**, Taehyung Noh, Kyungsik Han (2023). An Approach to Generating Content-based Recommendation Explanations through a Large Language Model. The Proceedings of the Korea Software Congress (**KSC**).

[C.1] Eunji Kim, **Haein Yeo**, Kyungsik Han (2023). A Study on the Personal Fashion Preference in Social Media using Meta-path based Heterogeneous Graph Modeling. The Proceedings of the Korea Software Congress (**KSC**).

Patents

[P.3] Kyungsik han, **Haein Yeo** (2024). Online Text-Based Personality Prediction System Using Comparative Evaluation of Multi-Agent Framework (PADO).

[P.2] Kyungsik han, Taehyung Noh, **Haein Yeo**, Myungjin Kim (2023). Device and Method for Providing Dashboard Services on Fashion Image Analysis.

[P.1] Kyungsik han, Taehyung Noh, **Haein Yeo**, Myungjin Kim (2023). Device and Method for Similar Image Recommendation Using Fashion Attributes and Image.

Honors and Awards

Best Paper Award, Korea Data Mining Society	Nov. 2024
Best Presentation Award, Korea Software Congress (KSC 2023)	Dec. 2023
Best Inventor Award, Seoul International Invention Fair	Dec. 2022
Participation Award, mySUNI Creative Challenge	Dec. 2022

TEACHING EXPERIENCE

Co-lecturer

- Human-Computer Interaction (In English) Fall 2025
- Human-Computer Interaction (In English) Fall 2024

Teaching Assistant (TA)

- Laboratory practice of Intelligence Computing 2 Fall 2024

ACADEMIC SERVICES

Conference Reviewer

- Conference on Empirical Methods in Natural Language Processing (EMNLP) 2024
- ACM Conference on Human Factors in Computing Systems (CHI) 2024
- ACM International Conference on Information and Knowledge Management (CIKM) 2023

References

1. **Kyungsik Han**
   Associate Professor, Hanyang University
   Department of Artificial Intelligence
   kyungsikhan@hanyang.ac.kr

2. **Yejin Shin**
   Lead Researcher, TTA
   Center for Trustworthy AI
   yepp1252@tta.or.kr

3. **Sangwoo Heo**
   Researcher, NAVER
   AI RM Center
   sangwoo.heo@navercorp.com