Hyunwoo Kim

hyunwookim.com

Young Investigator, Allen Institute for AI
google scholar | semantic scholar | github | mail

Research Interest

Natural language processing inspired by human cognition. I'm interested in taking an interdisciplinary approach to my research, connecting natural language processing and psychology. In particular, I like to distill insights from social cognition into models for reasoning. Currently, my research centers on the social reasoning capabilities of large language models, such as theory of mind.

Education

Seoul National University — Seoul, Korea
Ph.D. in Computer Science and Engineering; GPA: 4.24 / 4.3 (4.0 / 4.0) — *Mar 2019 – May 2023*

Advisor: Gunhee Kim
Thesis: Towards Conversational Agents with Social Cognition and Commonsense
Committee: Seung-won Hwang, Gunhee Kim, Byung-gon Chun, Minjoon Seo, and Yejin Choi

Yonsei University — Seoul, Korea
B.A. in Psychology & B.S. in Computer Science; GPA: 4.20 / 4.3 (3.98 / 4.0) — *Mar 2012 – Feb 2018*
Compulsory military service in the Republic of Korea Air Force (2013-2015)

Experience

Allen Institute for AI — Seattle, United States
Young Investigator (Postdoctoral Researcher) — *Jun 2023 - Current*
Working on social reasoning capabilities of large language models, advised by Yejin Choi

Research Intern — *Oct 2021 - May 2023*
Worked on social commonsense and dialogues, advised by Yejin Choi

NAVER — Seongnam, Korea
Research Intern — *Winter 2019*
Worked on exploration in reinforcement learning, advised by Ji-hoon Kim

SNU Institute for Industrial Systems Innovation — Seoul, Korea
Assistant Researcher — *Mar 2018 - Nov 2018*
Worked on summarization for online forum texts, advised by Gunhee Kim

Coupang — Seoul, Korea
Software Engineering Intern — *Winter 2016*
Worked on Immutable Infrastructure and Continuous Integration

Publication (* Denotes equal contribution)

Preprints

Perceptions to Beliefs: Exploring Precursory Inferences for Theory of Mind in Large Language Models — *arXiv 2024*
Chani Jung, Dongkwan Kim, Jiho Jin, Jiseon Kim, Yeon Seonwoo, Yejin Choi, Alice Oh, Hyunwoo Kim

Is this the real life? Is this just fantasy? The Misleading Success of Simulating Social Interactions With LLMs — *arXiv 2024*
Xuhui Zhou, Zhe Su, Tiwalayo Eisape, Hyunwoo Kim, Maarten Sap

Alpaca against Vicuna: Using LLMs to Uncover Memorization of LLMs — *arXiv 2024*
Aly M. Kassem*, Omar Mahmoud*, Niloofar Mireshghallah*, Hyunwoo Kim, Yulia Tsvetkov, Yejin Choi, Sherif Saad, Santu Rana

Conferences

Deal or no deal (or who knows)? Forecasting Uncertainty in Conversations using Large Language Models *ACL 2024*
Anthony Sicilia, Hyunwoo Kim, Khyathi Raghavi Chandu, Malihe Alikhani, Jack Hessel *Findings*

Can LLMs Keep a Secret? Testing Privacy Implications of Language Models via Contextual Integrity Theory *ICLR 2024*
Hyunwoo Kim*, Niloofar Mireshghallah*, Xuhui Zhou, Yulia Tsvetkov, Maarten Sap, Reza Shokri, Yejin Choi *Spotlight*

SODA: Million-scale Dialogue Distillation with Social Commonsense Contextualization *EMNLP 2023*
Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, *Outstanding Paper Award*
Ronan Le Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi

FANToM: A Benchmark for Stress-testing Machine Theory of Mind in Interactions *EMNLP 2023*
Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Le Bras, Gunhee Kim, Yejin Choi, and Maarten Sap *Oral presentation*

ProsocialDialog: A Prosocial Backbone for Conversational Agents *EMNLP 2022*
Hyunwoo Kim*, Youngjae Yu*, Liwei Jiang, Ximing Lu, Daniel Khashabi, Gunhee Kim, Yejin Choi, and Maarten Sap

Perspective-Taking and Pragmatics for Generating Empathetic Responses Focused on Emotion Causes *EMNLP 2021*
Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim

KLUE: Korean Language Understanding Evaluation *NeurIPS Datasets and Benchmarks 2021*
Sungjoon Park*, Jihyung Moon*, Sungdong Kim*, Won Ik Cho*, ..., Hyunwoo Kim, ...,
Alice Oh**, Jung-Woo Ha**, Kyunghyun Cho** (31 authors)

How Robust are Fact Checking Systems on Colloquial Claims? *NAACL 2021*
Byeongchang Kim*, Hyunwoo Kim*, and Gunhee Kim

Will I Sound Like Me? Improving Persona Consistency in Dialogues through Pragmatic Self-Consciousness *EMNLP 2020*
Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim *Oral presentation*

Curiosity Bottleneck: Exploration by Distilling Task-Specific Novelty *ICML 2019*
Youngjin Kim, Hyunwoo Kim*, Wontae Nam*, Ji-hoon Kim, and Gunhee Kim

Abstractive Summarization of Reddit Posts with Multi-level Memory Networks *NAACL 2019*
Byeongchang Kim, Hyunwoo Kim, and Gunhee Kim *Oral presentation*

Workshops

Perspective-Taking and Pragmatics for Generating Empathetic Responses Focused on Emotion Causes *NeurIPS MiC 2021*
Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim *Contributed talk*

Public Self-Consciousness for Endowing Dialogue Agents with Consistent Persona *ICLR BAICS 2020*
Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim *Contributed talk*

Awards

Outstanding Paper Award *Dec 2023*
SODA: Million-Scale Dialogue Distillation with Social Commonsense Contextualization, EMNLP 2023

Distinguished Doctoral Dissertation Award *Aug 2023*
Department of Computer Science and Engineering, Seoul National University

AI Star Scholarship *Jul 2022*
Yulchon foundation

Outstanding Researcher Fellowship *Mar 2022*
Brain Korea (BK21) FOUR Intelligence Computing, Seoul National University

Star Student Researcher Award *Feb 2022*
Brain Korea (BK21) FOUR Intelligence Computing, Seoul National University

NAVER Ph.D. Fellowship *Dec 2021*
Scholarship award, NAVER

Qualcomm Innovation Fellowship Korea *Nov 2021*
Scholarship award, Qualcomm Korea

NRF Ph.D. Student Research Funding *Jun 2021*
Next Generation Researcher support program, National Research Foundation (Korea)

Kwanjeong Scholarship *2019 – 2020*
Full tuition and fees for 2 years of graduate studies, Kwanjeong Educational Foundation

Best Presentation Award *Nov 2020*
Korean Society for Brain and Neural Sciences, 23rd KSBNS conference

National Excellence Scholarship *2012 – 2018*
Full tuition and fees for 4 years of undergraduate studies, Korean Student Aid Foundation (KOSAF)

Commendation *Feb 2015*
For excellence in mission, the 3rd Air Defense Artillery Brigade Commander, Republic of Korea Air Force

Academic Service

- **2023 ICML Theory of Mind Workshop Organizing Committee**: Organizer
- **2022 NAACL Organizing Committee**: Coordinator of Volunteers
- **Reviewer**: ACL, EMNLP, NAACL, EACL, ACL Rolling Review, NeurIPS, ICLR, ICML
- **Teaching Assistant at Seoul National University**: Discrete Mathematics, Probablistic Graphical Models, Introduction to Deep Learning, Introduction to Artificial Intelligence, Computer Vision