Hyunwoo Kim

Incoming Assistant Professor at KAIST AI | Postdoctoral Researcher at NVIDIA

hyunwookim.com

google scholar | linkedIn | github | email

Research Interest

Building machines capable of reasoning and cognition. I'm interested in enhancing machines' cognitive capabilities through computational methods, focusing on reasoning (e.g., social, cognitive, multimodal) and interactive agents guided by these abilities. I also like to take an interdisciplinary approach to my research, connecting AI and cognitive science. In particular, I like to distill insights from human social cognition into models for reasoning. Currently, my research centers on general reasoning capabilities of AI.

Experience

NVIDIA — Santa Clara, United States
Postdoctoral Researcher — *Feb 2025 - Current*
Working on reasoning, advised by Yejin Choi

Allen Institute for AI — Seattle, United States
Young Investigator — *Jun 2023 - Nov 2024*
Worked 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

Coupang — Seoul, Korea
Software Engineering Intern — *Winter 2016*
Worked on immutable infrastructure and continuous integration

Education

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

Publication (* Denotes equal contribution)

Preprints

Privasis: Synthesizing the Largest "Public" Private Dataset from Scratch — *arXiv 2026*
Hyunwoo Kim*, Niloofar Mireshghallah*, Michael Duan, Rui Xin, Shuyue Stella Li, Jaehun Jung, David Acuna, Qi Pang, Hanshen Xiao, G. Edward Suh, Sewoong Oh, Yulia Tsvetkov, Pang Wei Koh, Yejin Choi

Golden Goose: A Simple Trick to Synthesize Unlimited RLVR Tasks from Unverifiable Internet Text — *arXiv 2026*
Ximing Lu, David Acuna, Jaehun Jung, Jian Hu, Di Zhang, Shizhe Diao, Yunheng Zou, Shaokun Zhang, Brandon Cui, Mingjie Liu, Hyunwoo Kim, Prithviraj Ammanabrolu, Jan Kautz, Yi Dong, Yejin Choi

Long Grounded Thoughts: Distilling Compositional Visual Reasoning Chains at Scale *arXiv 2025*
{David Acuna*, Chao-Han Huck Yang*, Yuntian Deng*, Jaehun Jung*, Ximing Lu*}, Prithviraj Ammanabrolu, Hyunwoo Kim, Yuan-Hong Liao, Yejin Choi

Social World Models: Universal Structured Representation for Social Reasoning *arXiv 2025*
Xuhui Zhou, Jiarui Liu, Akhila Yerukola, Hyunwoo Kim, Maarten Sap

Retro-Search: Exploring Untaken Paths for Deeper and Efficient Reasoning *arXiv 2025*
{Ximing Lu*, Seungju Han*, David Acuna*, Hyunwoo Kim*, Jaehun Jung*},
Shrimai Prabhumoye, Niklas Muennighoff, Mostofa Patwary, Mohammad Shoeybi, Bryan Catanzaro, Yejin Choi

Conferences

SimpleToM: Exposing the Gap between Explicit ToM Inference and Implicit ToM Application in LLMs *ICLR 2026*
Yuling Gu, Oyvind Tafjord, Hyunwoo Kim, Jared Moore, Ronan Le Bras, Peter Clark, Yejin Choi

A False Sense of Privacy: Evaluating Textual Data Sanitization Beyond Surface-level Privacy Leakage *SaTML 2026*
Rui Xin*, Niloofar Mireshghallah*, Shuyue Stella Li, Michael Duan, Hyunwoo Kim, Yejin Choi,
Yulia Tsvetkov, Sewoong Oh, Pang Wei Koh

Socratic-MCTS: Test-Time Visual Reasoning by Asking the Right Questions *EMNLP 2025*
{David Acuna*, Ximing Lu*, Jaehun Jung*, Hyunwoo Kim*}, Amlan Kar, Sanja Fidler, Yejin Choi

Hypothesis-Driven Theory-of-Mind Reasoning for Large Language Models *COLM 2025*
Hyunwoo Kim, Melanie Sclar, Tan Zhi-Xuan, Lance Ying, Sydney Levine, Yang Liu, Joshua B. Tenenbaum, Yejin Choi

HAICOSYSTEM: An Ecosystem for Sandboxing Safety Risks in Human-AI Interactions *COLM 2025*
Xuhui Zhou, Hyunwoo Kim, Faeze Brahman, Liwei Jiang, Hao Zhu, Ximing Lu, Frank Xu, Bill Yuchen Lin,
Yejin Choi, Niloofar Mireshghallah, Ronan Le Bras, Maarten Sap

Alpaca against Vicuna: Using LLMs to Uncover Memorization of LLMs *NAACL 2025*
Aly M. Kassem*, Omar Mahmoud*, Niloofar Mireshghallah*, Hyunwoo Kim, Yulia Tsvetkov, Yejin Choi,
Sherif Saad, Santu Rana

Perceptions to Beliefs: Exploring Precursory Inferences for Theory of Mind in Large Language Models *EMNLP 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 *EMNLP 2024*
Xuhui Zhou, Zhe Su, Tiwalayo Eisape, Hyunwoo Kim, Maarten Sap

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

Commendation
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
- **Area Chair**: ACL Rolling Review (ACL, EMNLP, NAACL)
- **Reviewer**: ACL, EMNLP, NAACL, EACL, ACL Rolling Review, NeurIPS, ICLR, ICML, COLM