JISEON KIM

Ph.D. candidate

jiseon_kim@kaist.ac.kr hikoseon12.github.io
jiseon-kim-8ab574136 github.com/hikoseon12

SUMMARY

KAIST Ph.D. candidate specializing in AI Safety and AI for Social Good ●**AI Safety**: Evaluating cultural bias and moral reasoning in LLMs ●**AI for Social Good**: Developing scalable AI frameworks for domain-specific adaptation, policy analytics ●**Skills**: LLM evaluation and benchmarking, dataset construction, domain adaptation of AI models, and data-driven insights

EDUCATION

3/2020-Present **Korea Advanced Institute of Science and Technology (KAIST)** **Daejeon, Korea**
Ph.D. candidate in School of Computing

3/2017-2/2019 **Korea Advanced Institute of Science and Technology (KAIST)** **Daejeon, Korea**
Master in School of Computing

3/2013-2/2017 **Sookmyung Women's University** **Seoul, Korea**
B.S. in Computer Science
Graduated with the highest honor (**1/68**)

EXPERIENCE

8/2025 - 10/2025 **Research Intern @ Max Planck Institute for Security and Privacy (MPI-SP)** **MPI-SP**
- Conducted research on evaluating LLM alignment by analyzing how models understand diverse human personas in moral dilemmas

5/2024 - 5/2024
7/2022 - 8/2022
6/2019 - 8/2019 **Visiting Researcher @ Massachusetts Institute of Technology (MIT)** **MIT**
- Conducted interdisciplinary research with political science and developed a scalable AI framework to understand the U.S. legislative process
- Work published at EMNLP 2021 - "Learning Bill Similarity with Annotated and Augmented Corpora of Bills"
- Preprinted work (2025) on "Measuring Interest Group Positions on Legislation: An AI-Driven Analysis of Lobbying Reports "

3/2023 - 6/2023 **Research Intern @ NAVER AI Lab** **NAVER AI Lab**
- Constructed a Korean bias benchmark dataset to make safer and trustworthy Korean LLM
- Work published at TACL 2024 - "KoBBQ: Korean Bias Benchmark for Question Answering"

PUBLICATION

Preprint 2025 **[P9] Measuring Interest Group Positions on Legislation: An AI-Driven Analysis of Lobbying Reports** **MIT**
Jiseon Kim, Dongkwan Kim, In Song Kim, Alice Oh
- Expanded lobbying position classification beyond binary to include nuanced categories
- Built scalable AI framework with LLMs and GNNs to annotate 279K+ interest group–bill pairs and compute policy preference scores
- Analyzed lobbying strategies influenced by policy area, legislative stage, and group size

Preprint 2024 **[P8] Uncovering Factor Level Preferences to Improve Human-Model Alignment**
Juhyun Oh*, Eunsu Kim*, Jiseon Kim, Wenda Xu, Inha Cha, William Yang Wang, Alice Oh [*]equal contribution
- Developed PROFILE, a framework to explain factors driving LLM-human preference alignment
- Identified key differences in preferences between humans and LLMs across tasks
- Emphasized explainable analysis to enhance human-model alignment and training

EMNLP 2024
Long paper **[P7] Perceptions to Beliefs: Exploring Precursory Inferences for Theory of Mind in Large Language Models** **Allen AI**
Chani Jung, Dongkwan Kim, Jiho Jin, Jiseon Kim, Yeon Seonwoo, Yejin Choi, Alice Oh, Hyunwoo Kim
- Introduced Percept-ToMi and Percept-FANToM datasets to assess ToM precursors in LLMs
- Demonstrated LLMs excel in perception inference but show limitations in perception-to-belief inference
- Developed PercepToM, a method that improves LLM performance on ToM benchmarks

Technical Report
2024 **[P6] HyperCLOVA X Technical Report** **NAVER AI Lab**
Kang Min Yoo et al., Jiseon Kim,...
- Introduced LLM optimized for Korean language and culture, with strong English, math, and coding skills
- Trained on Korean, English, and code data, and evaluated on various benchmarks in both languages
- Contributed to model evaluations, including bias measurement in Korean culture through KoBBQ

TACL 2024, present at ACL 2024

[P5] KoBBQ: Korean Bias Benchmark for Question Answering **NAVER AI Lab**
Jiho Jin*, Jiseon Kim*, Nayeon Lee*, Hanual Yoo*, Alice Oh, Hwaran Lee [*]equal contribution

- Introduced a Korean bias benchmark dataset to address challenges in adapting to non-US cultures
- Proposed a framework for cultural adaptation, categorizing and validating biases via a large-scale survey
- Revealed significant differences in LM biases compared to a machine-translated version, highlighting the need for culturally-sensitive benchmarks

EMNLP 2021 Long paper

[P4] Learning Bill Similarity with Annotated and Augmented Corpora of Bills **MIT**
Jiseon Kim, Elden Griggs, In Song Kim, Alice Oh

- Proposed a 5-class task for bill document semantic similarities to understand bill-to-bill linkage in the legislative process
- Improved model performance by achieving a 5.5% higher F1 score compared to the baseline using data augmentation and multi-stage training
- Quantified the similarities across legal documents at various levels of aggregation

EMNLP 2021 Short paper

[P3] Efficient Contrastive Learning via Novel Data Augmentation and Curriculum Learning
Seonghyeon Ye, Jiseon Kim, Alice Oh

- Proposed a memory-efficient continual pretraining method
- Outperformed baseline models on GLUE benchmark with only 70% computational memory usage

EMNLP 2021 Long paper

[P2] Dimensional emotion detection from categorical emotion
Sungjoon Park, Jiseon Kim, Seonghyeon Ye, Jaeyeol Jeon, Hee Young Park, Alice Oh

- Utilized categorical emotion annotations to train a model predicting fine-grained emotions
- Optmized model with Earth Mover's Distance loss to predict fine-grained and categorical emotions
- Achieved comparable performance to state-of-the-art classifiers in emotion classification

IEEE transactions on intelligent transportation systems 2020

[P1] Denoising recurrent neural networks for classifying crash-related events
Sungjoon Park, Yeon Seonwoo, Jiseon Kim, Jooyeon Kim, Alice Oh

- Developed efficient neural network model with noisy time-series data with missing values for crash event classification
- Outperformed baseline models, improving event classification accuracy in driving scenarios

WORKSHOP

BiAlign @ICLR 2025

[W3] Exploring Persona-dependent LLM Alignment for the Moral Machine Experiment **MPI-SP**
Jiseon Kim*, Jea Kwon*, Luiz Felipe Vecchietti*, Alice Oh, Meeyoung Cha [*]equal contribution

WiML @NeurIPS 2024

[W2] Understanding Lobbying Strategies in Legislative Process: Bill Position Dataset and Lobbying Analysis **MIT**
Jiseon Kim, Dongkwan Kim, Joohye Jeong, In Song Kim, Alice Oh

C3NLP @ACL 2024

[W1] KoBBQ: Korean Bias Benchmark for Question Answering **NAVER AI Lab**
Jiho Jin*, Jiseon Kim*, Nayeon Lee*, Hanual Yoo*, Alice Oh, Hwaran Lee [*]equal contribution

INVITED TALK

ExploreCSR@Google March 21, 2025

Uncovering the Hidden Politics of Lawmaking: How Bills and Lobbying Shape U.S. Policy **KAIST**
Presented on AI for Political Science to understand the legislative process, supported by Google and hosted by KAIST School of Computing.

MPI-SP@Germany Feb 25, 2025

LLMs and the Political-Cultural Lens in Social Science **MPI-SP**
Invited talk at Max Planck Institute for Security and Privacy, hosted by Prof. Meeyoung Cha (Data Science for Humanity).

MLAI@Yonsei Jan 2, 2025

Things I Wish I Had Known Earlier in Grad School **Yonsei University**
Invited talk on networking, self-promotion, and collaboration in academia, hosted by Prof. Kyungwoo Song at the Machine Learning and Artificial Intelligence (MLAI) Lab.

AWARD, SCHOLARSHIP & FUNDING

4/2025 - Present
12/2019 - 8/2024

MISTI Global Seed Funds **MIT**
MIT's Global Seed Funds facilitate international collaborations for addressing global challenges

10/2024

2024 KAIST Graduate Student Outstanding Paper Award **KAIST**
Awarded for KoBBQ: Korean Bias Benchmark for Question Answering

3/2020 - Present

KAIST Support Scholarship (Ph.D.) **KAIST**