JISEON KIM

Ph.D. candidate

✉ jiseon_kim@kaist.ac.kr 🌐 hikoseon12.github.io
in jiseon-kim-8ab574136 G jiseon-google.scholar

SUMMARY

I'm a Ph.D. candidate advised by Alice Oh at KAIST. My research interests lie in natural language processing (NLP) and computational social science (CSS), with a focus on **1) AI alignment with human and societal values** and **2) AI for social good**. In particular, I work on the following topics:

- **LLM-Human Alignment & Evaluation:** I explore LLM alignment with human values and society, examining their behaviors and limitations (e.g., moral decision [W3, W4], cultural bias [P5], AI Alignment [P8], and social reasoning [P7]).
- **AI for Science & Social Impact:** I develop AI frameworks to process large-scale, expertise-driven data, particularly in political science (e.g., legislative processes [P4] and lobbying [P2]), to uncover hidden dynamics, enhance transparency, and assess societal impact.

Keywords: AI Safety, AI for Social Good, AI Alignment, LLM Evaluation, AI for Policy & Governance, Natural Language Processing (NLP), Computational Social Science (CSS)

EXPERIENCE

8/2025 - 10/2025 **Research Intern @ Max Planck Institute for Security and Privacy (MPI-SP)** **Max Planck Institute**
- Conduct research on aligning large language models with human moral decision-making
- Work published at BiAlign workshop at ICLR 2025 - "Exploring Persona-dependent LLM Alignment for the Moral Machine Experiment "
- Collaborate with Jea Kwon, Luiz Felipe Vecchietti, and Meeyoung Cha

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 to understand the US legislative process
- Work published at EMNLP 2021 - "Learning Bill Similarity with Annotated and Augmented Corpora of Bills"
- Collaborated with Elden Griggs and In Song Kim

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"
- Advised by Hwaran Lee

3/2019 – 2/2020 **Researcher @ KAIST** **KAIST**
- Researched multimodal NLP utilizing text and color
- Advised by Alice Oh

6/2015 – 8/2015 **Visiting Student @ UC Berkeley** **UC Berkeley**
- Completed Computer Science 61A, the structure and Interpretation of Computer Programs
- Received support for the UC Berkeley summer session program from the Sookmyung Women's University

EDUCATION

3/2020-Present **Korea Advanced Institute of Science and Technology** **Daejeon, Korea**
Ph.D. candidate in School of Computing
Thesis: Modeling Legislative Politics with Language Models
Advised by Alice Oh

3/2017-2/2019 **Korea Advanced Institute of Science and Technology** **Daejeon, Korea**
M.S. in School of Computing
Thesis: Color Generation for Paragraph Level of Text
Advised by Alice Oh

3/2013-2/2017 **Sookmyung Women's University** **Seoul, Korea**
Bachelor of Science (B.S.) in Computer Science
Graduated with the highest honor (1/68)
GPA: 4.35/4.5

PUBLICATION

Submitted to
ARR Jan 2026 **[P10] Exploring LLM Behavior in Relational Moral Dilemmas: Moral Rightness, Predicted Human Behavior, and Model Decisions** **MPI-SP**
Jiseon Kim, Jea Kwon, Luiz Felipe Vecchietti, Wenchao Dong, Jaehong Kim, Meeyoung Cha
- Developed a context-varying framework for LLM moral decision evaluation
- Revealed cross-perspective inconsistencies in judgments, predictions, and decisions
- Analyzed fairness–loyalty dynamics across perspectives via moral value analysis

Preparing Submissions to Science, 2025

[P9] Measuring Interest Group Positions on Legislation: An AI-Driven Analysis of Lobbying Reports **MIT**

Jiseon Kim, Dongkwan Kim, Joohye Jung, 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

EMNLP-Findings 2025

[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

WiML @ NeurIPS 2025

[W4] Mind the Gap: LLM Actions vs. Human Social Understanding in Moral Dilemmas **Max Planck Institute**

Jiseon Kim, Jea Kwon, Luiz Felipe Vecchietti, Wenchao Dong, Alice Oh, Meeyoung Cha

BiAlign @ ICLR 2025

[W3] Exploring Persona-dependent LLM Alignment for the Moral Machine Experiment **Max Planck Institute**

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

Yonsei AI Innovation Research Institute
November 25, 2025

Coalitions and Conflicts: How U.S. Interest Groups Align and Compete Through Legislative Bill Positions **Yonsei**
Invited talk at Yonsei Institute for AI and Social Innovation for AI for computational social science

MPI-SP@Germany
November 19, 2025

Do As I Believe, Not As You'd Do: How LLMs Favor Their Moral Ideals over Human Behavior in Relational Moral Dilemmas **Max Planck Institute**
Invited talk at Max Planck Institute for Security and Privacy, hosted by Prof. Meeyoung Cha (Data Science for Humanity).

Yonsei x KAIST Conference
November 7, 2025

Coalitions and Conflicts: How U.S. Interest Groups Align and Compete Through Legislative Bill Positions **KAIST**
Presented at "New Frontiers of Humanities and Social Sciences in the Age of Digital Transformation and AI".

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 **Max Planck Institute**
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**

3/2017 - 2/2019

KAIST Support Scholarship (M.S.) **KAIST**

2/2016

Naver Open API Awards in Hackathon **Unithon**
IT community United Hackathon

3/2015 - 3/2017

Korea National Science & Technology Scholarship (B.S.) **Sookmyung Women's University**

TEACHING EXPERIENCE

Fall 2021
Spring 2021

Machine Learning for NLP **KAIST**
Teaching Assistant

Fall 2021

Advanced Data Mining **KAIST**
Teaching Assistant

Spring 2020

Artificial Intelligence and Machine Learning **KAIST**
Head Teaching Assistant

Fall 2018
Spring 2018
Fall 2017

Data Structure **KAIST**
Teaching Assistant, Developed assignments

ACADEMIC SERVICE

Reviwer	Feb/May ACL Rolling Review (ARR) 2025 BiALign Workshop @ ICLR 2025 Feb/Apr/June ACL Rolling Review (ARR) 2024
Volunteer	BiALign Workshop @ ICLR 2025 FAccT 2022 COLING 2022
Undergraduate Research Program @ KAIST	Spring 2024 (Received an Encouragement Award)
Individual Research Mentoring @ KAIST	Spring 2024, Fall 2024 Spring 2022, Fall 2023 Spring 2021, Fall 2021 Spring 2020, Fall 2020

SKILL

Language	Python, Latex, PostgreSQL
Framework	Pytorch, Docker, Git

REFERENCE

Alice Oh	Professor in School of Computing, KAIST (alice.oh@kaist.edu)
Meeyoung Cha	Scientific Director at MPI-SP, Professor at KAIST, (mia.cha@mpi-sp.org)
In Song Kim	Associate Professor of Political Science, MIT (insong@mit.edu)