





JISEON KIM

Ph.D candidate

 jiseon_kim@kaist.ac.kr

 hikoseon12.github.io

 jiseon-kim-8ab574136

 github.com/hikoseon12

SUMMARY

I'm a PhD 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 [C9], cultural bias [C5], social reasoning [C7]).
- **AI for Science & Social Impact:** I develop AI frameworks to process large-scale, expertise-driven data, particularly in political science (e.g., legislative processes[C4], lobbying[W2]), to uncover hidden dynamics, enhance transparency, and assess societal impact.

Keywords: AI Alignment, LLM Evaluation, AI for Policy & Governance, AI for Social Good, NLP, Computational Social Science

EXPERIENCE

3/2023 - 6/2023	Research Intern @ NAVER AI Lab	NAVER AI Lab
	<ul style="list-style-type: none">• 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	
5/2024 - 5/2024	Visiting Researcher @ MIT	MIT
7/2022 - 8/2022	<ul style="list-style-type: none">• Conducted interdisciplinary research with political science to understand the US legislative process	
6/2019 - 8/2019	<ul style="list-style-type: none">• Work published at EMNLP 2021 - "Learning Bill Similarity with Annotated and Augmented Corpora of Bills"• Collaborated with Elden Griggs and In Song Kim	
3/2019 - 2/2020	Research Fellow @ KAIST	KAIST
	<ul style="list-style-type: none">• Researched multimodal NLP utilizing text and color• Advised by Alice Oh	
6/2015 - 8/2015	Visiting Student @ UC Berkeley	UC Berkeley
	<ul style="list-style-type: none">• 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 Ph.D. in School of Computing Advised by Alice Oh	Daejeon, Korea
3/2017-2/2019	Korea Advanced Institute of Science and Technology Master in School of Computing Thesis: Color Generation for Paragraph Level of Text Advised by Alice Oh	Daejeon, Korea
3/2013-2/2017	Sookmyung Women's University B.S. Student in Computer Science Graduated with the highest honor (1/68)	Seoul, Korea

PUBLICATION

under review 2025	[C9] Exploring Persona-dependent LLM Alignment for the Moral Machine Experiment <u>Jiseon Kim</u> [*] , Jea Kwon [*] , Luiz Felipe Vecchietti [*] , Alice Oh, Meeyoung Cha ^{[*]equal contribution}
	<ul style="list-style-type: none">• Analyzed how sociodemographic personas impact LLM moral decisions, revealing higher variability than humans• Developed a metric to measure alignment between LLM and human moral judgments• Identified political bias as a key factor, raising concerns about bias amplification in deployment

under review 2025	<p>[C8] 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}</p> <ul style="list-style-type: none"> • Introduced a framework to explain and quantify factors influencing alignment between human and LLM preferences • Uncovered discrepancies between human and LLM preferences, especially in generation tasks • Offered explainable insights into misaligned factors, helping improve LLM alignment through targeted adjustments
EMNLP 2024	<p>[C7] Perceptions to Beliefs: Exploring Precursory Inferences for Theory of Mind in Large Language Models Chani Jung, Dongkwan Kim, Jiho Jin, Jiseon Kim, Yeon Seonwoo, Yejin Choi, Alice Oh, Hyunwoo Kim</p> <ul style="list-style-type: none"> • 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	<p>[C6] HyperCLOVA X Technical Report NAVER AI Lab Kang Min Yoo, Jaegeun Han, Sookyo In, Heewon Jeon, ... Jiseon Kim*...(additional authors)</p> <ul style="list-style-type: none"> • 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	<p>[C5] KoBBQ: Korean Bias Benchmark for Question Answering NAVER AI Lab Jiho Jin*, Jiseon Kim*, Nayeon Lee*, Hanual Yoo*, Alice Oh, Hwaran Lee ^{[*]equal contribution}</p> <ul style="list-style-type: none"> • 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	<p>[C4] Learning Bill Similarity with Annotated and Augmented Corpora of Bills MIT Jiseon Kim, Elden Griggs, In Song Kim, Alice Oh</p> <ul style="list-style-type: none"> • 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	<p>[C3] Efficient Contrastive Learning via Novel Data Augmentation and Curriculum Learning Seonghyeon Ye, Jiseon Kim, Alice Oh</p> <ul style="list-style-type: none"> • Proposed a memory-efficient continual pretraining method • Outperformed baseline models on GLUE benchmark with only 70% computational memory usage
EMNLP 2021 Long paper	<p>[C2] Dimensional emotion detection from categorical emotion Sungjoon Park, Jiseon Kim, Seonghyeon Ye, Jaeyeol Jeon, Hee Young Park, Alice Oh</p> <ul style="list-style-type: none"> • Utilized categorical emotion annotations to train a model predicting fine-grained emotions • Optimized 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	<p>[C1] Denoising recurrent neural networks for classifying crash-related events Sungjoon Park, Yeon Seonwoo, Jiseon Kim, Jooyeon Kim, Alice Oh</p> <ul style="list-style-type: none"> • 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 2024	<p>[W2] Understanding Lobbying Strategies in Legislative Process: Bill Position Dataset and Lobbying Analysis Jiseon Kim, Dongkwan Kim, Joohye Jeong, In Song Kim, Alice Oh</p> <ul style="list-style-type: none"> • Will be presented at the 19th Women in Machine Learning Workshop at NeurIPS (WiML 2024)
C3NLP, Co-located with ACL 2024	<p>[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}</p> <ul style="list-style-type: none"> • Presented at the 2nd Workshop on Cross-Cultural Considerations in NLP (C3NLP)

ONGOING PROJECT

Ongoing	Competing Interests in U.S. Politics: Who Supports and Who Opposes Congressional Bills? • Developed a dataset capturing real-world legislative activities and diverse lobbying positions • Developed a scalable AI framework to measure interest groups' bill positions • Analyzed lobbying strategies, providing insights into legislative interactions	MIT
---------	--	-----

AWARDS, SCHOLARSHIPS & FUNDINGS

10/2024	2024 KAIST Graduate Student Outstanding Paper Award Awarded for KoBBQ: Korean Bias Benchmark for Question Answering	KAIST
12/2019 - 8/2024	MISTI Global Seed Funds MIT's Global Seed Funds facilitate international collaborations for addressing global challenges	MIT
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 IT community United Hackathon	Unithon
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 Teaching Assistant	KAIST
Fall 2021	Advanced Data Mining Teaching Assistant	KAIST
Spring 2020	Artificial Intelligence and Machine Learning Head Teaching Assistant	KAIST
Fall 2018 Spring 2018 Fall 2017	Data Structure Teaching Assistant, Developed assignments	KAIST

ACADEMIC SERVICE

Reviewer	Feb ARR 2025, ICLR 2025 Bi-Align Workshop, Feb/Apr/June ARR 2024
Volunteer	FACCT 2022, COLING 2022
Undergraduate Research Program @ KAIST	Spring 2024 (Received an Encouragement Award)
Individual Research Mentoring @ KAIST	Summer/Fall 2020, Spring/Fall 2021, Spring 2022, Fall 2023 Spring/Fall 2024

SKILL

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

LANGUAGE

English	Professional
Korean	Native

REFERENCE

Alice Oh	Professor in School of Computing, KAIST (alice.oh@kaist.edu)
-----------------	--

