Hoyun Song

hysong@kaist.ac.kr • Google Scholar • linkedin/hoyun-song-

RESEARCH INTERESTS

I am a dedicated NLP/AI researcher focused on developing socially responsible AI, driven by the belief that high-quality, well-refined data is essential for robust and human-aligned model performance. I am currently a postdoctoral researcher at KAIST, collaborating with Professor KyungTae Lim. My core research interests include: (1) Domain-Specific Modeling and Knowledge Integration for critical sectors such as mental healthcare and online safety, emphasizing the incorporation of domain expertise into modeling and data generation processes; (2) Personalization and Human-Alignment to ensure language models mirror nuanced value preferences and maintain consistent personas; and (3) Social Simulation and Behavior Analysis utilizing computational social science to assess how LLMs handle systemic roles, obligations, and social complexities. These interests focus on creating evaluative benchmarks, knowledge-integrated modeling, and simulating complex institutional behaviors to steer the development of AI systems that aim to serve and benefit humanity.

KEY AREAS

1) Domain-Specific Modeling & Knowledge Integration

- Mental Healthcare: Integrating domain-specific expertise and clinical frameworks (e.g., DSM-5) into LLM modeling and high-fidelity data generation.
- Safety & Moderation: Detecting hate content, online trolling, and generating strategic counter-speech.
- Technical Methods: Knowledge Distillation, Data Synthesis, Supervised Fine-tuning, Reinforcement Learning

2) Personalization & Human-Aligned Behavior

- Persona Fidelity: Evaluating and diversifying linguistic patterns through personality-based persona generation.
- Human-aligned Modeling: Modeling value preferences and human-aligned reasoning to ensure models mirror nuanced human judgment.

3) Social Simulation & LLM Behavior Analysis

- Social Bias Assessment: Quantifying social biases from diverse demographic perspectives.
- LLM Behavior Analysis: Understanding and interpreting model behaviors through the lens of computational social science and cognitive patterns.
- Applying Social Theory: Simulating multi-agent social scenarios to assess how LLMs internalize and execute human-centric social norms and expectations.

EDUCATION

Korea Advanced Institute of Science and Technology (KAIST) — Daejeon, Republic of Korea
Ph.D. degree, School of Computing — Sep 2019 – Aug 2025
Dissertation: Integration of Domain-Specific Knowledge for Mental Health Diagnostics

Korea Advanced Institute of Science and Technology (KAIST) — Daejeon, Republic of Korea
M.S. degree, School of Computing — Mar 2017 – Feb 2019
Thesis: Interpretable Depression Detection from Social Media using Hierarchical Attention Network with Depression Indicators

Dongguk University — Seoul, Republic of Korea
B.S. degree, Computer Engineering — Mar 2011 – Feb 2017
Minor in Business School

Seoul Foreign Language High School — Mar 2007 – Feb 2010

EMPLOYMENT

M.S. Researcher | *KAIST* — Mar 2019 - Sep 2019

Collaborator: Prof. Jong C. Park

Postdoctoral Researcher | *KAIST* Aug 2025 - Current

InnoCORE Postdoctoral Fellow

Collaborator: Prof. KyungTae Lim

PUBLICATIONS

Technical Report{#}

TR1 KORMo: Korean Open Reasoning Model for Everyone
Minjun Kim, Hyeonseok Lim, Hangyeol Yoo, Inho Won, Seungwoo Song, Minkyung Cho, Junhun Yuk, Changsu Choi, Dongjae Shin, Huije Lee, **Hoyun Song**, Alice Oh, and KyungTae Lim
https://arxiv.org/abs/2510.09426

International{Conference/Journal}{#}

IC19 MENTALBENCH: A Benchmark for Evaluating Psychiatric Diagnostic Capability of Large Language Models
Hoyun Song*, Migyeong Kang*, Jisu Shin, Jihyun Kim, Chanbi Park, Hangyeol Yoo, Jihyun An, Alice Oh, Jinyoung Han, KyungTae Lim
Under review, https://arxiv.org/abs/2602.12871.

IC18 MENTOR: A Reinforcement Learning Framework for Enabling Tool Use in Small Models via Teacher-Optimized Rewards
ChangSu Choi*, **Hoyun Song***, Dongyeon Kim, WooHyeon Jung, Minkyung Cho, Sunjin Park, NohHyeob Bae, Seona Yu, KyungTae Lim
Under review, https://arxiv.org/abs/2510.18383.

IC17 RoleConflictBench: A Benchmark of Role Conflict Scenarios for Evaluating LLMs' Contextual Sensitivity
Jisu Shin, **Hoyun Song**, Juhyun Oh, Changgeon Ko, Eunsu Kim, Chani Jung, and Alice Oh
Under review, https://arxiv.org/abs/2509.25897.

IC16 TReX: Tokenizer Regression for Optimal Data Mixture
Inho Won, Hangyeol Yoo, Minkyung Cho, Jungyeul Park, **Hoyun Song**†, and KyungTae Lim†
In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2026.

IC15 A Multi-Task Benchmark for Abusive Language Detection in Low-Resource Settings
Fitsum Gaim, **Hoyun Song**, Huije Lee, Changgeon Ko, Eui Jun Hwang, and Jong C. Park
In *NeurIPS 2025 Datasets and Benchmarks Track*, 2025.

IC14 Does Rationale Quality Matter? Enhancing Mental Disorder Detection via Selective Reasoning Distillation
Hoyun Song, Huije Lee, Jisu Shin, Sukmin Cho, Changgeon Ko, and Jong C. Park
In *Findings of the Association for Computational Linguistics: ACL 2025 (Findings of ACL)*, 2025.

IC13 Spotting Out-of-Character Behavior: Atomic-Level Evaluation of Persona Fidelity in Open-Ended Generation
Jisu Shin, Juhyun Oh, Eunsu Kim, **Hoyun Song**, and Alice Oh
In *Findings of the Association for Computational Linguistics: ACL 2025 (Findings of ACL)*, 2025.

IC12 EXIT: Context-Aware Extractive Compression for Enhancing Retrieval-Augmented Generation
Taeho Hwang, Sukmin Cho, Soyeong Jeong, **Hoyun Song**, SeungYoon Han, and Jong C. Park
In *Findings of the Association for Computational Linguistics: ACL 2025 (Findings of ACL)*, 2025.

IC11 Temporal Information Retrieval via Time-Specifier Model Merging
SeungYoon Han, Taeho Hwang, Sukmin Cho, Soyeong Jeong, **Hoyun Song**, Huije Lee, and Jong C. Park
In *The workshop on Towards Knowledgeable Foundation Models at ACL 2025 (KnowFM@ACL 2025)*, 2025.

IC10 Lossless Acceleration of Large Language Models with Hierarchical Drafting based on Temporal Locality in Speculative Decoding
Sukmin Cho, Sangjin Choi, Taeho Hwang, Jeongyeon Seo, Soyeong Jeong, Huije Lee, **Hoyun Song**, Jong C. Park, and Youngjin Kwon
In *Findings of the Association for Computational Linguistics: NAACL 2025 (Findings of NAACL)*, 2025.

IC09 Different Bias Under Different Criteria: Assessing Bias in LLMs with a Fact-Based Approach
Changgeon Ko, Jisu Shin, **Hoyun Song**, Jeongyeon Seo, and Jong C. Park
In *The workshop on Socially Responsible Language Modelling Research at NeurIPS 2024 (SoLaR@NeurIPS 2024)*, 2024.

IC08 Towards Effective Counter-Responses: Aligning Human Preferences with Strategies to Combat Online Trolling
Huije Lee, **Hoyun Song**, Jisu Shin, Sukmin Cho, Seungyoon Han, and Jong C. Park
In *Findings of the Association for Computational Linguistics: EMNLP 2024 (Findings of EMNLP)*, 2024.

IC07 Ask LLMs Directly, "What shapes your bias?": Measuring Social Bias in Large Language Models
Jisu Shin, **Hoyun Song**, Huije Lee, Soyeong Jeong, and Jong C. Park
In *Findings of the Association for Computational Linguistics: ACL 2024 (Findings of ACL)*, 2024.

IC06 Generation of Korean Offensive Language by Leveraging Large Language Models via Prompt Design
Jisu Shin, **Hoyun Song**, Huije Lee, Fitsum Gaim, and Jong C. Park
In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers) (IJCNLP-AACL)*, 2023.

IC05 A Simple and Flexible Modeling for Mental Disorder Detection by Learning from Clinical Questionnaires
Hoyun Song, Jisu Shin, Huije Lee, and Jong C. Park
In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL)*, 2023.

IC04 ELF22: A Context-based Counter-Trolling Dataset to Combat Internet Trolls
Huije Lee, Young Ju Na, **Hoyun Song**, Jisu Shin, and Jong C. Park
In *Proceedings of the 13th Language Resources and Evaluation Conference (LREC)*, 2022.

IC03 Optimizing Domain Specificity of Transformer-based Language Models for Extractive Summarization of Financial News Articles in Korean
Huije Lee, Wonsuk Yang, ChaeHun Park, **Hoyun Song**, Eugene Jang, and Jong C. Park
In *35th Pacific Asia Conference on Language on Language, Information and Computation (PACLIC 35)*, 2021.

IC02 A Large-scale Comprehensive Abusiveness Detection Dataset with Multifaceted Labels from Reddit
Hoyun Song*, Soo Hyun Ryu*, Huije Lee, and Jong C. Park
In *Proceedings of the 25th Conference on Computational Natural Language Learning (CoNLL)*, 2021.

IC01 Feature Attention Network: Interpretable Depression Detection from Social Media
Hoyun Song*, Jinseon You*, Jin-Woo Chung, and Jong C. Park
In *32nd Pacific Asia Conference on Language, Information and Computation (PACLIC 32)*, 2018.

Domestic{Conference/Journal}{#}

DC09 Korean Needle in a Haystack: A Benchmark for Evaluating Korean Long-Context Understanding
Hangyeol Yoo, KyungTae Lim, and **Hoyun Song**
In *Proceedings of the 2025 Joint Conference on Human and Cognitive Language Technology (HCLT)*, 2025.

DC08 Beyond word Form: Semantic Reconstruction of Korean LLMs under Jamo Scrambling
Dongyeon Kim, **Hoyun Song**, Changsu Choi, and KyungTae Lim
In *Proceedings of the 2025 Joint Conference on Human and Cognitive Language Technology (HCLT)*, 2025.
(Selected as Best Paper)

DC07 Iterative Feedback-based Personality Persona Generation for Diversifying Linguistic Patterns in Large Language Models
Taeho Hwang, **Hoyun Song**, Jisu Shin, Sukmin Cho, and Jong C. Park
In *Proceedings of the 35th Annual Conference on Human and Cognitive Language Technology (HCLT)*, 2023.

DJ06 Detecting Implicitly Abusive Language by Applying Out-of-Distribution Problem
Jisu Shin, **Hoyun Song**, and Jong C. Park
In *Journal of KIISE (JOK)*, 49(11), 2022.

DC05 Stopwords Mask Pooling for Dense Retrieval in Medical Domain
Dongho Choi, **Hoyun Song**, Soyeong Jeong, Sukmin Cho, and Jong C. Park
In *Proceedings of the Korea Computer Congress (KCC 2022)*, 2022.

DC04 Constructing Korean Abusive Language Dataset using Machine Translation
Jisu Shin, **Hoyun Song**, Huije Lee, and Jong C. Park
In *Proceedings of Korea Computer Congress (KCC)*, 2022.

DC03 Detecting Implicitly Abusive Language by Applying Out-of-Distribution Problem
Jisu Shin, **Hoyun Song**, and Jong C. Park
In *Proceedings of Korea Software Congress (KSC)*, 2021.
(Selected as Best Paper)

DC02 BERT-based Personality Disorder Detection Model with Abusive Language Marking from Social Media
Jisu Shin, **Hoyun Song**, and Jong C. Park
In *Proceedings of Korea Computer Congress (KCC)*, 2021.

DC01 Predicting Symptoms of Depression for Social Media Users via Linguistic Patterns
Hoyun Song, Hancheol Park, Wonsuk Yang, and Jong C. Park
In *Korea Software Congress (KSC)*, 2017.

PATENTS

- System for Correcting Verbal Violence and Enhancementing Text Reliability Using Abusive Language and Credibility Dependencies, and the Method Thereof
  KR 10-2021-0166872, Registration Date: Nov 29, 2021 [50% Contribution]
- System and Its Method for Credibility Prediction from Dialogues Considering Personalities of System User and Interlocutor
  KR 10-2020-0162062, Registration Date: Nov 2, 2022 [40% Contribution]
- Method and System for Predicting and Improving Persuasiveness of Documents Based on Concreteness and the Order of Persuasion Strategy
  KR 10-2020-0127837, Registration Date: Jul 6, 2022 [10% Contribution]
- System for Generating Customized Conversation for Enhancing Credibility by Transferring Compatible Personality Traits of Conversation Partner and the Method Thereof
  KR 10-2021-0166873, Filing Date: Nov 29, 2021 [50% Contribution]
- Method and System for Personality Recognition from Dialogues
  KR 10-2020-0011541, Registration Date: Oct 25, 2021 [15% Contribution]
- System and Method for Analyzing and Utilizing How Authors Refer to an Event in a Web Document to Predict the Distribution of How Much Readers Consider Such Document Credible
  KR 10-2018-0128806, Registration Date: Feb 27, 2020 [10% Contribution]

PROJECTS

Abusive language detection and automatic corrective feedback for large language models Mar 2023 – Aug 2025
Project Manager | Funded by National Research Foundation of Korea (NRF)

- Generated fluent and culturally aligned synthetic data in Korean for detecting abusive language using LLMs.
- Developed a framework to measure social bias in LLMs by quantifying social perceptions from diverse demographic perspectives.
- Developed a human preference-aligned approach that leverages LLMs to generate effective counter-speech against online trolling.

A system for offensive language detection and automatic feedback with correction Mar 2020 – Feb 2023
Project Manager | Funded by National Research Foundation of Korea (NRF)

- Constructed a dataset of abusive language by collecting a diverse set of social media texts that include various forms of offensive language.
- Designed the annotation tasks on Amazon Mechanical Turk (AMT) with clear labeling instructions and a structured scheme, enabling crowd workers to perform accurate annotations efficiently.
- Developed an automated model to detect abusive language, fine-tuned on our curated dataset of such language.

- Constructed a dataset for counter-speech generation by in-house data annotation, including detailed troll-type annotation and strategy-labeled counter-responses.
- Developed a model for context-aware counter-speech generation, utilizing troll-type and response-strategy annotations.

An automatic feedback system for the prevention and early treatment of depressive symptoms through language use analysis Mar 2017 – Feb 2020

Project Manager | Funded by National Research Foundation of Korea (NRF)

- Developed a knowledge-based, interpretable model that detects multiple mental disorders by leveraging specialized domain knowledge such as DSM-5.
- Constructed a large-scale dataset that captures four mental disorders—depression, anxiety disorder, bipolar disorder, and borderline personality disorder—by gathering social media posts.
- Developed an interpretable deep learning-based model that detects depression in social media posts and explains its detection results.
- Constructed a Korean depression dataset from social media and developed an automatic detection model.

Emotional Intelligence Technology to Infer Human Emotion and Carry on Dialogue Accordingly Mar 2017 – Feb 2020

Research Assistant | Funded by Institute of Information & communications Technology Planning & Evaluation (IITP)

- Constructed an annotated dataset to train an emotion prediction model leveraging Ekman's six basic emotions theory.
- Constructed an annotated dataset for training a personality prediction model using the Big Five personality traits framework.
- Developed CNN- and RNN-based models for emotion and personality prediction.

SKILLS

Programming Languages - Python, C

Frameworks - Pytorch, Tensorflow

Tools - Hugging Face Transformers, PEFT, vLLM, GitHub

Libraries - NumPy, Pandas, Scikit-learn, SpaCy, NLTK

Language - Korean (native), English (fluent)