Seungone Kim

✉ seungone@kaist.ac.kr • 🌐 https://github.com/SeungoneKim

My main research goal is to make techniques that ensure language models (LMs) are aligned with human values in order to reduce the substantial risks they might have (also known as the '*Alignment Problem*'). Specifically, I am interested in (1) ensuring LMs to truly follow human instructions (e.g., Instruction Tuning, RLHF, Fine-grained Evaluation) [C3, P2, O1, O2], (2) teaching LMs to express their thought process with chain-of-thoughts [C2, P1], and (3) grounding LMs to share commonsense knowledge as humans do [C1].

Education

○ **Korea Advanced Institute of Science and Technology (KAIST)** **Seoul, Korea**
M.S. in Artificial Intelligence *March 2023 – Present*
Advisor : Minjoon Seo

○ **Yonsei University** **Seoul, Korea**
B.S. in Computer Science *March 2018 – February 2023*

Publications

Preprints

○ **[P1]** *The CoT Collection: Improving Zero-shot and Few-shot Learning of Language Models via Chain-of-Thought Fine-Tuning*
Submitted at EMNLP 2023
Seungone Kim*, Sejune Joo*, Doyoung Kim, Joel Jang, Seonghyeon Ye, Jamin Shin, Minjoon Seo
[Paper] [Code]

○ **[P2]** *FLASK: Fine-grained Language Model Evaluation based on Alignment Skill Sets*
Planning for Submission at ICLR 2024
Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, **Seungone Kim**, Yongrae Jo, James Thorne, Juho Kim, Minjoon Seo
[Paper] [Code]

Peer-Reviewed Conference Papers

○ **[C3]** *Exploring the Benefits of Training Expert Language Models over Instruction Tuning*
The 40th International Conference on Machine Learning (ICML 2023)
Joel Jang, **Seungone Kim**, Seonghyeon Ye, Doyoung Kim, Lajanugen Logeswaran, Moontae Lee, Kyungjae Lee, Minjoon Seo
[Paper] [Code]

○ **[C2]** *CoTEVer: Chain of Thought Prompting Annotation Toolkit for Explanation Verification*
The 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2023)
Seungone Kim, Se June Joo, Yul Jang, Hyungjoo Chae, Jinyoung Yeo
[Paper] [Code]

○ **[C1]** *Mind the Gap! Injecting Commonsense Knowledge for Abstractive Dialogue Summarization*
Proceedings of the 29th International Conference on Computational Linguistics (COLING 2022)
Seungone Kim*, Se June Joo*, Hyungjoo Chae*, Chaehyeong Kim*, Seungwon Hwang, Jinyoung Yeo

[Paper] [Code]

Ongoing

- **[O1]** *Aligning Mixture of Expert Language Models on Fine-grained Rewards*
  Planning for Submission at ICLR 2024
  Seungone Kim*, Jamin Shin*, Yejin Cho*, Joel Jang, Shayne Longpre, Hwaran Lee, James Thorne, Minjoon Seo
- **[O2]** Teaching Language Models to Browse the Web with Natural Language Feedback
  Planning for Submission at NAACL 2024
  Sejune Joo*, **Seungone Kim***, Juyoung Suk, Joel Jang, Seonghyeon Ye, Jamin Shin, Minjoon Seo

Research Experience

- **NAVER AI Lab** **Seongnam, Korea**
  Research Internship at Language Research Team *Jul 2023 – Present*
  I am leading a research project, focusing on reinforcing language models to be aligned to human values [O1].
- **Korea Advanced Institute of Science and Technology (KAIST)** **Seoul, Korea**
  Research Internship at Language and Knowledge Laboratory *Jul 2022 – Jan 2023*
  I participated in a research project building Expert language models, overcoming the weakness of a single Instruction-tuned LM [C3].
- **Seoul National University** **Seoul, Korea**
  Research Internship at Language and Data Intelligence Laboratory *Sep 2021 – Jun 2022*
  I have participated in a research project building a multilingual language model for low-resource languages via continual cross-lingual pre-training. Also, I have participated in the initial stage of building a Korean Language Model where I pre-trained a 11B sized decoder-only model with korean text corpus from scratch.

Invited Talks

- **Commonsense Knowledge Consortium 1st Workshop** **Seoul, Korea**
  Presented paper accepted at COLING 2022[C1] *July 2022*
  Host: Prof. Il Chul Moon(KAIST), Chanyoung Park(KAIST), Jinyoung Yeo(Yonsei), Seungwon Hwang(SNU), Joseph Lim(KAIST), Jonghyeon Choi (Yonsei)
- **Yonsei University AI Workshop** **Seoul, Korea**
  Presented paper accepted at COLING 2022[C1] *October 2022*
  Host: Prof. Jonghyeon Choi (Yonsei)
- **Innovation Session Series** **Frankfurt, Germany**
  Presented paper accepted at EACL 2023[C2] *June 2023*
  Host: Laux Michael (SAP)

Services

- **Conference Reviewer:** EACL 2023, EMNLP 2023
- **Secondary Reviewer:** ACL 2023

Mentoring

○ **Juyoung Suk:** B.S. Student, KAIST (July 2023 - Present)

Teaching Experience & Extracirricular Experience

○ **Korea Army Augmented to US Army (KATUSA)** **United States Army Garrison Humphreys**
Sergeant *January 2019 – August 2020*
I have worked as a translator and a Senior Katusa for Chief Warrant Officers flying Helicopters across the Peninsula, serving for General Robert B. Abrams, the 25th Commander of U.S. Forces Korea.

○ **Samsung Dream Class** **Samsung Welfare**
Mentor *March 2022 – February 2023*
I have participated in the Dream Class program as a Mentor, managed by Samsung in Korea. I have taught Python programming to middle-school students including object-oriented programming, machine learning with numpy and pytorch, and data visualization with seaborn.