Seungone Kim

louisdebroglie@kaist.ac.kr • https://github.com/SeungoneKim

My main research goal is to make techniques that ensure neural models 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) enabling LMs to truly follow human instructions [C4, O3], (2) teaching LMs to express their thought process [C3, P1], (3) ensuring LMs share commonsense knowledge as humans do [C2], and (4) reinforcing LMs to generate helpful & harmless responses and evaluating them [O1, O2, O4].

Education

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

- **Yonsei University** — **Seoul, Korea**
  B.S. in Computer Science — *March 2018 – February 2023*
  Advisor : Jinyoung Yeo

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]

Peer-Reviewed Conference Papers

- **[C4]** *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]

- **[C3]** *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]

- **[C2]** *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]

- **[C1]** *SG-MLP : Switch Gated Multi-Layer Perceptron Model for Natural Language Understanding*
  Annual Conference of Korea Information Processing Society, 2021 (ACK 2021)

Guijin Son, **Seungone Kim**, Se June Joo, Woojin Cho, JeongEun Nah
[Paper] [Code]

Ongoing

- **[O1]** *Controllable Alignment Tuning: Balancing the Degree of Helpfulness and Harmlessness with Weight Merging*

  Planning for Submission at ICLR 2024
  Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Hwaran Lee, Sangdoo Yun, James Thorne, Minjoon Seo

- **[O2]** Is Reinforcement Learning With Human Feedback Generally Preferred Over Supervised Learning?

  Planning for Submission at ICLR 2024
  Hoyeon Chang, **Seungone Kim**, Jinho Park, Joel Jang, Kee-Eung Kim, Minjoon Seo

- **[O3]** Advancing Real-World Planning with Tool Augmented Expert Language Models

  Planning for Submission at ICLR 2024
  Sejune Joo, **Seungone Kim**, Juyoung Suk, Joel Jang, Seonghyeon Ye, Jamin Shin, Minjoon Seo

- **[O4]** FLASK: Fine-grained Language Skillset Evaluation of Large Language Models

  Planning for Submission at ICLR 2024
  Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, **Seungone Kim**, Yongrae Jo, James Thorne, Juho Kim, Minjoon Seo

Research Experience

- **NAVER AI Lab** — **Seongnam, Korea**
  Language Research Team — *Jul 2022 – Jan 2023*

  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 [P2].

- **Yonsei University** — **Seoul, Korea**
  Undergraduate Thesis advised at Conversational Intelligence Laboratory — *Sep 2021 – Present*

  I have published my undergraduate thesis at COLING 2022 (long) as a first author [C2]. I proposed a method that injects commonsense knowledge inferences to solve abstractive dialogue summarization.

- **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.

- **Yonsei University** — **Seoul, Korea**
  Internship at Soft Computing Laboratory — *Apr 2021 – Aug 2021*

  I have participated in 'Industry - University Assignments' between Yonsei University and Samsung Electronics Research. I have built a neural model that transforms documents into tabular data. I built the overall pipeline along with a demo system to test it out.

Invited Talks

- **Commonsense Knowledge Consortium 1st Workshop** — **Seoul, Korea**
  Presented paper accepted at COLING 2022[C2] — *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[C2] — *October 2022*
  Host: Prof. Jonghyeon Choi (Yonsei)

- **Conversation with Senior Event** — **Seoul, Korea**
  Presented paper accepted at COLING 2022[C2] and EACL 2023[C3] — *May 2023*
  Host: Prof. Wonseok Lee (Yonsei) and Yonsei BigData Conference (YBIGTA)

- **Innovation Session Series** — **Frankfurt, Germany**
  Presented paper accepted at EACL 2023[C3] — *June 2023*
  Host: Laux Michael (SAP)

Services

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

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.