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, O2, O4], (2) teaching LMs to express their thought process [C3, O1], (3) ensuring LMs to share commonsense knowledge as humans do [C2], and (4) enforcing LMs to generate helpful & factual responses [O3].

Education

- **Yonsei University** — **Seoul, Korea**
  B.S. in Computer Science — *March 2018 – February 2023*
  Advisor : Jinyoung Yeo
- **Korea Advanced Institute of Science and Technology (KAIST)** — **Seoul, Korea**
  M.S. in Artificial Intelligence — *March 2023 –*
  Advisor : Minjoon Seo

Publications

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]** *Rationale Tuned Introspective Language Models are Strong Zero-Shot Learners*
  Planning for Submission at EMNLP 2023
  Seungone Kim*, Sejune Joo*, Doyoung Kim, Joel Jang, Seonghyeon Ye, Jamin Shin, Minjoon Seo

- **[O2]** *BranchLM: Branching and Merging Generalist Language Models into Specialized Language Models*
  Planning for Submission at EMNLP 2023
  Joel Jang*, Yujin Kim*, Kyungjae Lee, **Seungone Kim**, Colin Raffel, Luke Zettlemoyer, Minjoon Seo, Moontae Lee, Lajanugen Logeswaran
- **[O3]** *Rethinking the Role of Human Feedback: What Makes Reinforcement Learning with Human Feedback Work?*
  Planning for Submission at NeurIPS 2023
  Hoyeon Chang, **Seungone Kim**, Jinho Park, Joel Jang, Minjoon Seo
- **[O4]** *Towards Responsible Language Generation: Can Language Models Truly Unfollow Toxic Instructions*
  Planning for Submission at AAAI 2024
  Seungone Kim, Doyoung Kim, Minjoon Seo

Research Experience

- **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 a distributed expert language model that tackles the weakness of a single Instruction-tuned LM such as T0 and FLAN [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.

Technical and Personal skills

- **Programming Languages:** Python(Proficient), C++(Proficient), C, Java
- **Frameworks / Software Skills:** PyTorch(Proficient), PyTorch Lightning(Proficient), Transformers(Proficient), FairSeq, Torch Geometric, NLTK, SpaCy
- **Software Skills:** Linux, Google Cloud Platform, Azure, Docker, LaTeX, Git, Wandb

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**
  Planning to present paper accepted at COLING 2022[C2] and EACL 2023[C3] — *May 2023 (Expected)*
  Host: Prof. Wonseok Lee (Yonsei) and Yonsei BigData Conference (YBIGTA)
- **Innovation Session Series** — **Frankfurt, Germany**
  Planning to present paper accepted at EACL 2023[C3] — *June 2023 (Expected)*
  Host: Laux Michael (SAP)

Services

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

Honors and Awards

- **Certificate of Achievement** — **US Department of the Army**
  KATUSA Sergeant (Korean Army Augmented to United States Army) — *August 2020*
  I have served as a KATUSA between January 2019 - August 2020. I worked with Chief Warrant Officers who would fly helicopters to move Generals across the peninsula.
- **Excellence Award (Top 8 papers)** — **Korea Information Processing Society**
  Annual Conference of KIPS, 2021 — *November 2021*
  Awarded with my paper at ACK 2021[C1].
- **Grand Prize in Graduation Capstone** — **Yonsei University**
  Top 3 projects — *July 2022*
  Awarded with my paper at COLING 2022[C2].
- **NLP Best Paper Award** — **Samsung Research, Yonsei University**
  Yonsei AI Workshop — *October 2022*
  Awarded with my paper at COLING 2022[C2].
- **Semester Academic Excellence Scholarship** — **Yonsei University**
  Yonsei University — *Fall 2020, Spring 2021, Fall 2021, Spring 2022, Fall 2022*
  Undergraduate Studies
- **High / Highest Academic Honors** — **Yonsei University**
  Yonsei University — *Fall 2020, Spring 2021, Spring 2022*
  Undergraduate Studies

Teaching Experience

- **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.