Seungju Han

Email: wade3han@snu.ac.kr Homepage: https://seungjuhan.me Twitter/X: @SeungjuHan3

Education

Seoul National University Mar, 2017—Aug, 2024*
B.S. in Electrical and Computer Engineering Seoul, Korea
∗ includes a 3-year mandatory military service in South Korea.

University of Washington Sep, 2022—Dec, 2022
Exchange Student Seattle, WA

Seoul Science High School Mar, 2014—Feb, 2017
Specialized high school for students talented in math and science. Seoul, Korea

Experiences

University of Washington, xlab Aug, 2024—Present
Undergraduate Research Intern Seattle, WA (Remote)
Advisor: Prof. Yejin Choi

Yonsei University, MIR Lab Mar, 2023—Present
Visiting Researcher Seoul, Korea
Advisor: Prof. Youngjae Yu

Allen Institute for AI, Mosaic Team Sep, 2022—Aug, 2024
Visiting Predoctoral Researcher Seattle, WA (Remote)
Advisor: Prof. Yejin Choi / Hosts: Dr. Nouha Dziri, Dr. Jack Hessel, Prof. Youngjae Yu

Hyperconnect Apr, 2019—Aug, 2022
Machine Learning Engineer Seoul, Korea
Collaborator: Prof. Buru Chang, Dr. Dongyoung Kim, Dr. Sungjoo Ha

Publications and Preprints

Google Scholar: https://scholar.google.com/citations?hl=en&user=g_anRqAAAAAJ#
∗ indicates equal contribution.

Preprints

2. AI as Humanity's Salieri: Quantifying Linguistic Creativity of Language Models via Systematic Attribution of Machine Text against Web Text
   Ximing Lu, Melanie Sclar, Skyler Hallinan, Niloofar Mireshghallah, Jiacheng Liu, **Seungju Han**, Allyson Ettinger, Liwei Jiang, Khyathi Chandu, Nouha Dziri, Yejin Choi
   Under review

1. Do LLMs Have Distinct and Consistent Personality? TRAIT: Personality Testset designed for LLMs with Psychometrics
   Seungbeen Lee*, Seungwon Lim*, **Seungju Han**, Giyoung Oh, Minju Kim, Beongwoo Kwak, Jiwan Chung, Hyungjoo Chae, Dongha Lee, Jinyoung Yeo, Youngjae Yu
   Preprint, code, pdf

Publications

12. WILDGUARD: Open One-stop Moderation Tools for Safety Risks, Jailbreaks, and Refusals of LLMs
    Seungju Han*, Kavel Rao*, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, Nouha Dziri
    NeurIPS 2024 Datasets & Benchmarks, code, pdf

11. WILDTEAMING at Scale: From In-the-Wild Jailbreaks to (Adversarially) Safer Language Models
Liwei Jiang, Kavel Rao*, **Seungju Han***, Faeze Brahman, Sachin Kumar, Niloofar Mireshghallah, Ximing Lu, Marteen Sap, Yejin Choi, Nouha Dziri
NeurIPS 2024, code, pdf

10. Selective Vision is the Challenge for Visual Reasoning: A Benchmark for Visual Argument Understanding
Jiwan Chung, Sungje Lee, Minseo Kim, **Seungju Han**, Ashkan Yousefpour, Jack Hessel, Youngjae Yu
EMNLP 2024 (long) code, pdf

9. Reading Books is Great, But Not if You Are Driving! Visually Grounded Reasoning about Defeasible Commonsense Norms
Seungju Han, Junhyeok Kim, Jack Hessel, Liwei Jiang, Jiwan Chung, Yejin Son, Yejin Choi, Youngjae Yu
EMNLP 2023 (long) (**Oral Presentation**), code, project page, pdf

8. Multimodal Laughter Reasoning with Textual Audio-Visual Representation
Hyun Lee, Sung Bin Kim, **Seungju Han**, Youngjae Yu, Tae Hyun Oh
Findings of NAACL (long) 2024, code, pdf
ICCV Workshop, What is Next in Multimodal Foundation Models? 2023

7. CHAMPAGNE: Learning Real-world Conversation from Large-Scale Web Videos
Seungju Han, Jack Hessel, Nouha Dziri, Yejin Choi, Youngjae Yu
ICCV 2023, code, project page, pdf

6. Measuring and Improving Semantic Diversity of Dialogue Generation
Seungju Han, Beomsu Kim, Buru Chang
Findings of EMNLP (long) 2022, code, pdf

5. Meet Your Favorite Character: Open-domain Chatbot Mimicking Fictional Characters with only a Few Utterances
Seungju Han*, Beomsu Kim*, Jin Yong Yoo*, Seokjun seo, Sangbum Kim, Enkhbayar Erdenee, Buru Chang
NAACL (short) 2022, code, pdf

4. Understanding and Improving the Exemplar-based Generation for Open-domain Conversation
Seungju Han*, Beomsu Kim*, Seokjun seo*, Enkhbayar Erdenee*, Buru Chang
ACL 4th Workshop on NLP4ConvAI (**Oral Presentation, Outstanding Paper**) 2022, code, pdf

3. Distilling the Knowledge of Large-scale Generative Models into Retrieval Models for Efficient Open-domain Conversation
Beomsu Kim*, Seokjun seo*, **Seungju Han***, Enkhbayar Erdenee*, Buru Chang
Findings of EMNLP (long) 2021, code, pdf

2. Disentangling Label Distribution for Long-tailed Visual Recognition
Youngkyu Hong*, **Seungju Han***, Kwanghee Choi*, Seokjun seo, Beomsu Kim, Buru Chang
CVPR 2021, code, blog, pdf

1. Attentron: Few-Shot Text-to-Speech Utilizing Attention-Based Variable-Length Embedding
Seungwoo Choi*, **Seungju Han***, Dongyoung Kim*, Sungjoo Ha
Interspeech 2020, pdf

Awards and Honors

Research Grant from **Korea Foundation for Advanced Studies (KFAS)** 2023—2024
Doctoral study fellowship (awarded to six people this year in Computer Science)

Undergraduate Scholarship from **Kwanjeong Educational Foundation** 2019—2023
Full scholarship, approx. 20,000 USD in total (awarded to up to 50 people every year)

Undergraduate Scholarship from **Seoul National University** 2017—2019
Full scholarship, approx. 9,000 USD in total (merit-based)

Scholarship from **Hanseong Son Jae Han Scholarship Foundation** 2015—2016
Hanseong Nobel scholarship, approx. 10,000 USD in total (awarded to up to 200 people every year)

Invited Presentations

Learning representations from large-scale videos, Tech seminar @ Dalpha AI 2023 Summer
A path towards AGI, Tech seminar @ Corca AI 2023 Summer
Research in start-up, Tech seminar @ FriendliAI 2022 Fall
Research in start-up, Tech seminar @ Corca AI 2022 Summer
Open-domain conversation agent, AI retreat @ Seoul National University 2021 Spring

Professional Activities

Volunteer

EMNLP (Virtual) 2022

Reviewer

EMNLP 2021
ACL ARR 2022—
NeurIPS 2023—
ICLR 2024—