Seungju Han

Ph.D. Student @ Stanford Computer Science
Research Intern @ NVIDIA

Email: seungju@stanford.edu
Home: https://seungjuhan.me

Education

Stanford University Sep, 2025—
Ph.D. Student in Computer Science Stanford, CA

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.

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

Experiences

NVIDIA Nov, 2024—Aug, 2025
Language and Cognition Research (LACR) Team Santa Clara, CA (Remote)
Research Intern

Yonsei University Mar, 2023—Present
MIR Lab Seoul, Korea
Visiting Researcher

Allen Institute for AI Sep, 2022—Aug, 2024
Mosaic Team Seattle, WA (Remote)
Visiting Researcher

Hyperconnect Apr, 2019—Aug, 2022
AI Lab Seoul, Korea
Machine Learning Engineer

Publications

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

Preprints & Tech Reports

5. Prismatic Synthesis: Gradient-based Data Diversification Boosts Generalization in LLM Reasoning
Jaehun Jung, **Seungju Han***, Ximing Lu*, Skyler Hallinan*, David Acuna, Shrimai Prabhumoye, Mostofa Patwary, Mohammad Shoeybi, Bryan Catanzaro, Yejin Choi
2025, pdf, blog

4. Subtle Risks, Critical Failures: A Framework for Diagnosing Physical Safety of LLMs for Embodied Decision Making
Yejin Son*, Minseo Kim*, Sungwoong Kim, **Seungju Han**, Jian Kim, Dongju Jang, Youngjae Yu, Chanyoung Park
2025, pdf

3. Nemotron-CrossThink: Scaling Self-Learning beyond Math Reasoning
Syeda Nahida Akter, Shrimai Prabhumoye, Matvei Novikov, **Seungju Han**, Ying Lin, Evelina Bakhturi, Eric Nyberg, Yejin Choi, Mostofa Patwary, Mohammad Shoeybi, Bryan Catanzaro
2025, pdf

2. Retro-Search: Exploring Untaken Paths for Deeper and Efficient Reasoning
Ximing Lu*, **Seungju Han***, David Acuna*, Hyunwoo Kim*, Jaehun Jung*, Shrimai Prabhumoye, Niklas Muennighoff, Mostofa Patwary, Mohammad Shoeybi, Bryan Catanzaro, Yejin Choi
2025, pdf

1. Nemotron-H: A Family of Accurate and Efficient Hybrid Mamba-Transformer Models
NVIDIA Team, contributed to build pretraining data
2025, pdf

Conference & Workshop Papers

18. Verifying the Verifiers: Unveiling Pitfalls and Potentials in Fact Verifiers
Wooseok Seo*, **Seungju Han***, Jaehun Jung, Benjamin Newman, Seungwon Lim, Seungbeen Lee, Ximing Lu, Yejin Choi, Youngjae Yu
COLM 2025, pdf

17. G1yphD3c0de: Towards Safer Language Models on Visually Perturbed Texts
Yejin Choi, Yejin Yeo, Yejin Son, Seungju Han, Youngjae Yu
COLM 2025

16. MAPoRL: Multi-Agent Post-Co-Training for Collaborative Large Language Models with Reinforcement Learning
Chanwoo Park, **Seungju Han**, Xingzhi Guo, Asuman Ozdaglar, Kaiqing Zhang, Joo-Kyung Kim
ACL 2025, pdf

15. Representation Bending for Large Language Model Safety
Ashkan Yousefpour*, Taeheon Kim*, Ryan S Kwon, Seungbeen Lee, Wonje Jeung, **Seungju Han**, Alvin Wan, Harrison Ngan, Youngjae Yu, Jonghyun Choi
ACL 2025, pdf

14. 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
ICLR 2025 (**Oral Presentation**), pdf

13. 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
Findings of NAACL 2025, code, pdf

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 (**Oral Presentation**) code, pdf

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

8. 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 (**Oral Presentation**), code, blog, pdf

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

6. Measuring and Improving Semantic Diversity of Dialogue Generation
Seungju Han, Beomsu Kim, Buru Chang
Findings of EMNLP 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 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 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

Ph.D. Scholarship from **Korea Foundation for Advanced Studies (KFAS)** 2023—2024
Doctoral study fellowship, declined since started Ph.D. a year later

Undergraduate Scholarship from **Kwanjeong Educational Foundation** 2019—2023
Full scholarship (awarded to up to 50 people every year)

Undergraduate Scholarship from **Seoul National University** 2017—2019
Full scholarship (merit-based)

Scholarship from **Hanseong Son Jae Han Scholarship Foundation** 2015—2016
Hanseong Nobel scholarship (awarded to up to 200 people every year)

Professional Activities

Volunteer
EMNLP 2022

Reviewer (∗ indicates outstanding reviewer)
EMNLP 2021
ACL ARR 2022–

NeurIPS	2023, 2024*
ICLR	2024
ICML	2025