Hwaran Lee

Contact Information	NAVER AI Lab at NAVER Cloud hwaran.lee@gmail.com hwaranlee.github.io

Research Interests	My current primary research interests are Ethics, Safety, and Trustworthiness of Large Language Models. I am also interested in controllable language generation, dialog systems, and machine learning for language models.

Education

KAIST — Daejeon, South Korea
Ph.D., Electrical Engineering — Mar. 2013 – Aug. 2018
Dissertation: *Neural Representations for Speech Recognition and Natural Language Generation*
Advisor: Prof. Soo-Young Lee

KAIST — Daejeon, South Korea
B.S., Mathematical Science — Feb. 2008 – Aug. 2012
Minor: Financial Engineering Program
Magna Cum Laude

Research and Work Experiences

NAVER Cloud — Seongnam, South Korea
- *Leader*, Safety Team, HyperScale AI — Apr. 2023 – Present
- *Leader*, Language Research Team, NAVER AI Lab — Apr. 2023 – Present
- *Tech Leader*, NAVER AI Lab — Jan. 2023 – Mar. 2023
- *Research Scientist*, NAVER AI Lab — Jan. 2023 – Present

NAVER — Seongnam, South Korea
- *Tech Leader*, NAVER AI Lab — Jul. 2022 – Dec. 2022
- *Research Scientist*, NAVER AI Lab — Mar. 2021 – Dec. 2022

SK Telecom — Seoul, South Korea
- *Research Scientist*, T-Brain, AI Center — Nov. 2018 – Feb. 2021

Brain Science Research Center — Daejeon, South Korea
- *Undergraduate Researcher* — Sep. 2012 – Feb. 2013

Publications

International Journal

[J1] **Hwaran Lee**, Seokhwan Jo, HyungJun Kim, Sangkeun Jung, and Tae-Yoon Kim, "SUMBT+LaRL: Effective Multi-domain End-to-end Neural Task-oriented Dialog System", *IEEE Access*, 9 (2021): 116133-116146.

[J2] Geonmin Kim, **Hwaran Lee**, Bo-Kyeong Kim, Sang-Hoon Oh, and Soo-Young Lee, "Unpaired Speech Enhancement by Acoustic and Adversarial Supervision for Speech Recognition", *IEEE Signal Processing Letters*, (2019): 159-163.

[J3] Ho-Gyeong Kim, **Hwaran Lee**, Geonmin Kim, Sang-Hoon Oh, and Soo-Young Lee, "Rescoring of N-best Hypotheses using Top-down Selective Attention for Automatic Speech Recognition", *IEEE Signal Processing Letters*, (2018): 199-203.

[J4] **Hwaran Lee**, Geonmin Kim, Ho-Gyeong Kim, Sang-Hoon Oh, and Soo-Young Lee, "Deep CNNs Along the Time Axis With Intermap Pooling for Robustness to Spectral Variations", *IEEE Signal Processing Letters* 23.10 (2016): 1310-1314.

[J5] **Hwaran Lee**, Nadeem Iqbal, Wonil Chang, and Soo-Young Lee, "A Calibration Method for Eye-Gaze Estimation Systems Based on 3D Geometrical Optics", *IEEE Sensors Journal* 13, no. 9 (2013): 3219-3225.

[J6] Nadeem Iqbal, **Hwaran Lee**, and Soo-Young Lee, "Smart User Interface for Mobile Consumer Devices Using Model-Based Eye-Gaze Estimation", *IEEE Transactions on Consumer Electronics* 59, no. 1 (2013): 161-166.

International Conference

[C1] Siwon Kim, Sangdoo Yun, **Hwaran Lee**, Martin Gubri, Sungroh Yoon, Seong Joon Oh, "ProPILE: Probing Privacy Leakage in Large Language Models", arXiv preprint arXiv:2305.15060 (2023) ***NeurIPS (Spotlight)**, 2023*

[C2] **Hwaran Lee***, Seokhee Hong*, Joonsuk Park, Takyoung Kim, Meeyoung Cha, Yejin Choi, Byoungpil Kim, Gunhee Kim, Eun-Ju Lee, Yong Lim, Alice Oh, Sangchul Park, and Jung-Woo Ha, "SQuARe: A Large-Scale Dataset of Sensitive Questions and Acceptable Responses Created through Human-Machine Collaboration", ***ACL (Oral)***, 2023

[C3] **Hwaran Lee***, Seokhee Hong*, Joonsuk Park, Takyoung Kim, Gunhee Kim, and Jung-Woo Ha, "KoSBI: A Dataset for Mitigating Social Bias Risks Towards Safer Large Language Model Applications", ***ACL***, 2023

[C4] Deokjae Lee, JunYeong Lee, Jung-Woo Ha, Jin-Hwa Kim, Sang-Woo Lee, **Hwaran Lee**, and Hyun Oh Song, "Query-Efficient Black-Box Red Teaming via Bayesian Optimization", ***ACL***, 2023

[C5] Minbeom Kim, Hwanhee Lee, Kang Min Yoo, Joonsuk Park, **Hwaran Lee**†, Kyomin Jung†, "Critic-Guided Decoding for Controlled Text Generation", ***ACL** (Findings)*, 2023

[C6] Miyoung Ko, Ingyu Seong, **Hwaran Lee**, Joonsuk Park, Minsuk Chang, Minjoon Seo, "Beyond Fact Verification: Comparing and Contrasting Claims on Contentious Topics", ***ACL** (Findings)*, 2023

[C7] Hwanhee Lee, Kang Min Yoo, Joonsuk Park, **Hwaran Lee**†, Kyomin Jung†, "Masked Summarization to Generate Factually Inconsistent Summaries for Improved Factual Consistency Checking", *In Findings of the Association for Computational Linguistics:* **NAACL**, 2022.

[C8] Kyungjae Lee, Wookje Han, Seung-won Hwang, **Hwaran Lee**, Joonsuk Park, Sang-Woo Lee, "Plug-and-Play Adaptation for Continuously-updated QA", *In Findings of the Association for Computational Linguistics:* **ACL**, 2022.

[C9] John Yoon Young Chung, Wooseok Kim, Kang Min Yoo, **Hwaran Lee**, Eytan Adar, Minsuk Chang, "TaleBrush: Sketching Stories with Generative Pretrained Language Models", *In Proceedings of the 2022 **CHI** Conference on Human Factors in Computing Systems*, 2022.

[C10] Gi-Cheon Kang, Junseok Park, **Hwaran Lee**, Byoung-Tak Zhang, and Jin-Hwa Kim, "Reasoning Visual Dialog with Sparse Graph Learning and Knowledge Transfer", *In Findings of the Association for Computational Linguistics:* **EMNLP**, 2021.

[C11] **Hwaran Lee***, Jinsik Lee*, and Tae-Yoon Kim, "SUMBT: Slot-Utterance Matching for Universal and Scalable Belief Tracker", *The 57th Annual Meeting of the Association for Computational Linguistics* (**ACL**), 2019.

†corresponding authors
*these authors contributed equally to this work

[C12] Geonmin Kim, **Hwaran Lee**, Bo-Kyeong Kim, and Soo-Young Lee, "Compositional Sentence Representation from Character within Large Context Text", *International Conference on Neural Information Processing* (**ICONIP**), 2017.

[C13] Ho-Gyeong Kim, Jihyeon Roh, **Hwaran Lee**, Geonmin Kim, and Soo-Young Lee, "Active Learning for Large-scale Object Classification: from Exploration to Exploitation" *In Proceedings of the 3rd International Conference on Human-Agent Interaction*, (**HAI**), 2015.

[C14] Bo-Kyeong Kim, **Hwaran Lee**, Jihyeon Roh, and Soo-Young Lee, "Hierarchical committee of deep CNNs with exponentially-weighted decision fusion for static facial expression recognition", *In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction* (**ICMI**), 2015.

International Workshop

[W1] Jaimeen Ahn, **Hwaran Lee**, Jin-Hwa Kim, Alice Oh, "Why Knowledge Distillation Amplifies Gender Bias and How to Mitigate from the Perspective of DistilBERT", *In Proceedings of the 4rd Workshop on Gender Bias in Natural Language Processing*, 2022.

[W2] John Yoon Young Chung, Wooseok Kim, Kang Min Yoo, **Hwaran Lee**, Eytan Adar, Minsuk Chang, "TaleBrush: Visual Sketching of Story Generation with Pretrained Language Models ", *CHI EA 22: CHI Conference on Human Factors in Computing Systems Extended Abstracts*, 2022.

[W3] Geonmin Kim*, **Hwaran Lee***, CheongAn Lee, Eunmi Hong, Byunggeun Kim, Soo-Young Lee, "A Deep Chatbot for QA and Chitchat.", *The Conversational Intelligence Challenge section on NIPS 2017 Competition Track Workshop,* 2017.

[W4] **Hwaran Lee**, Geonmin Kim, Jihyeon Roh, and Soo-Young Lee, "Learning Tonotopically Organized Auditory Feature-map from Speech by an Intermap Pooling Layer in a Deep CNN", *15th China-Japan-Korea Joint Workshop on Neurobioloby and Neuroinformatics* (**NBNI**), 2015. (only abstract)

[W5] Geonmin Kim, **Hwaran Lee**, Jaemyung Yu, and Soo-Young Lee, "Spoken Sentence Embedding from Character by Jointly Learning Character-level Compositional Word Model and RNN Sentence Encoder", *15th China-Japan-Korea Joint Workshop on Neurobioloby and Neuroinformatics* (**NBNI**), 2015. (only abstract)

Pre-prints

[A1] Jiho Jin, Jiseon Kim, Nayeon Lee, Haneul Yoo, Alice Oh, Hwaran Lee, "KoBBQ: Korean Bias Benchmark for Question Answering", arXiv preprint arXiv:2307.16778

[A2] Taehyun Lee, Seokhee Hong, Jaewoo Ahn, Ilgee Hong, **Hwaran Lee**, Sangdoo Yun, Jamin Shin, Gunhee Kim, "Who Wrote this Code? Watermarking for Code Generation", arXiv preprint arXiv:2305.15060 (2023)

[A3] Seungone Kim *, Jamin Shin *, Yejin Cho *, Joel Jang, Shayne Longpre, **Hwaran Lee**, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, Minjoon Seo, "Prometheus: Inducing Fine-grained Evaluation Capability in Language Models", Arxiv, 2023

Patents

[P1] Tae-Yoon Kim, Jin Kim, Hyungjoon Kim, Jinsik Lee, **Hwaran Lee**, Heewon Jeon, Seokhwan Jo,"Method and Apparatus for Providing Hybrid Intelligent Customer Consultation", Korea Patent Application 10-2019-0136035, filed October 2019, Patent Pending.

[P2] **Hwaran Lee**, Jinsik Lee, Tae-Yoon Kim, "Method and Apparatus for Dialogue State Tracking for Use in Goal-oriented Dialog System",

- Korea Patent Application 10-2019-0086380, filed July 17, 2019, Patent Pending.
- PCT/KR2020/008832, filed July 7, 2020, Patent Pending.
- China Patent 202080051265.8, filed July 7, 2020, Patent Pending.
- US Patent 17/619,568, filed July 7, 2020, Patent Pending.

Honors and Awards

- Annual Roll Award, KAIST EE — Apr. 2018
- Ranked 3rd, ConvAI challenge, NIPS 2017 Competition Track Workshop — 2017
- Challenge Winner, ICMI EmotiW2015 — 2015
- Best Paper Award, HAI — 2015
- Qualcomm Innovation Award — 2015
- BK21 Plus Financial Support for Graduates Long Term Training — May. 2014
- KAIST Graduate Scholarship — Mar. 2013 - Aug. 2018
- Australian Endeavour Student Exchange Grant (AUD$ 5000), The University of Queensland — Apr. 2011
- National Excellence Scholarship, KOSAF — Feb.2008 - Feb. 2012

Academic Services

- Organizing Committee
  - ACM FAccT 2022 CRAFT HyperscaleFAccT
- Area Chair
  - NeurIPS 2023 Datasets & Benchmarks 2023
- Reviewer
  - ARR 2021-2022, ACL 2021-2023, EMNLP 2021-2023, COLING 2020, 2022
  - NeurIPS 2021-2023, ICLR 2021-2023
  - WWW 2022
  - ACL'22 In2Writing Workshop
- Samsung Humantech Paper Awards Committee — 2020
- Qualcomm Innovation Awards Committee — 2019
- Speech Communication — 2019
- IEEE Transactions on Neural Networks and Learning Systems — 2017 - 2018
- Neural Processing Letters — 2015

Outside Activities

- Committee member of the 2nd Forum on Artificial Intelligence Ethics and Policy, organized by the Ministry of Science and ICT, South Korea. — 2023
- Organizing committee of AI Ethics Forum for Human at NAVER — 2022

Invited Talks

- Ethical Problems in Language Models
  - Intellectual Property High Court, Daejeon, Apr. 2023
  - GIST, Gwang-ju, Feb. 2023
  - Hanyang Univ., Seoul, Sep. 2022
  - KAIST, Daejeon, Jun. 2022
  - SNU, Seoul, Jun. 2022
- Introduction to deep learning for dialogue systems
  - Inha Univ., Seoul, Oct. 2020
  - Yeonsei Univ., Seoul, Jun. 2020
  - Sookmyung Women's Univ., Seoul, Oct. 2019

- Toward end-to-end neural dialog systems for multi-domain task completion
  KAIST, Daejeon, Dec. 2019