

# HWARAN LEE

@ hwaran.lee@gmail.com | [LinkedIn](#) | [GitHub](#) | [Homepage](#) | [NAVER AI Lab, South Korea](#)

Updated 2024-04-18

## RESEARCH INTERESTS

---

My research is committed to develop human-like and trustworthy Artificial Intelligence that can positively contribute to society. Recent primary interests has been building **trustworthy and safe Large Language Models (LLMs)**, with a focus on: (1) construction of safety datasets, benchmarks, and evaluation metrics<sup>[1, 8, 31, 14, 15]</sup>; (2) controllable language generation<sup>[10, 16]</sup>; (3) LLM security, including adversarial attack and red-teaming<sup>[11, 13]</sup>; (4) safety alignment and learning methods<sup>[30, 18]</sup>.

Before the advent of LLMs, my work was focused on deep neural network-based Dialog Systems<sup>[36, 2, 20, 27]</sup> and Speech Recognition Systems<sup>[3, 4, 5]</sup>. This involved devising effective learning methodologies and network architectures, with a focus on neural representations of language and speech. Additionally, my earlier work was enhancing the accuracy of Eye-gaze Systems<sup>[6, 7]</sup> for human-computer interface.

## EMPLOYMENT HISTORY

---

### NAVER Cloud

- Leader, Safety Team, HyperCLOVA X and HyperScale AI
- Leader, Language Research Team, NAVER AI Lab
- Tech Leader, Language Research Team, NAVER AI Lab

Seongnam, South Korea  
Apr 2023 – **Present**  
Apr 2023 – **Present**  
Jan 2023 – Mar 2023

### NAVER

- Leader, AI Safety Lab, Future AI Center
- Tech Leader, Language Research Team, NAVER AI Lab
- Research Scientist, Language Research Team, NAVER AI Lab

Seongnam, South Korea  
Jan 2024 – **Present**  
Jul 2022 – Dec 2022  
Mar 2021 – Jun 2022

### SK Telecom

- Research Scientist, T-Brain, AI Center

Seoul, South Korea  
Nov 2018 – Feb 2021

### Brain Science Research Center, KAIST

- Undergraduate Researcher,

Daejeon, South Korea  
Sep 2012 – Feb 2013

## EDUCATION

---

### Ph.D., KAIST

Electrical Engineering

Dissertation: Neural Representations for Speech Recognition and Natural Language Generation

Deajeon, South Korea  
Mar 2013 – Aug 2018

Advisor: [Soo-Young Lee](#)

### B.S., KAIST

Mathematical Science & Financial Engineering Program (minor)

Magna Cum Laude

Deajeon, South Korea  
Mar 2008 – Aug 2012

## PUBLICATIONS

---

\* indicates equal contributions, † indicates corresponding author(s)

### JOURNAL ARTICLES

---

- [1] J. Jin, J. Kim, N. Lee, H. Yoo, A. Oh, and **H. Lee**<sup>†</sup>. “[KoBBQ: Korean Bias Benchmark for Question Answering](#)”. In: *Transactions of the Association for Computational Linguistics* (2024).
- [2] **H. Lee**, S. Jo, H. Kim, S. Jung, and T.-Y. Kim. “[SUMBT+ LaRL: Effective Multi-domain End-to-end Neural Task-oriented Dialog System](#)”. In: *IEEE Access* (2021).
- [3] G. Kim, **H. Lee**, B.-K. Kim, S.-H. Oh, and S.-Y. Lee. “[Unpaired Speech Enhancement by Acoustic and Adversarial Supervision for Speech Recognition](#)”. In: *IEEE Signal Processing Letters* 26.1 (2019), pp. 159–163.

- [4] H.-G. Kim, **H. Lee**, G. Kim, S.-H. Oh, and S.-Y. Lee. “Rescoring of N-Best Hypotheses Using Top-Down Selective Attention for Automatic Speech Recognition”. In: *IEEE Signal Processing Letters* 25.2 (2018), pp. 199–203.
- [5] **H. Lee**, G. Kim, H.-G. Kim, S.-H. Oh, and S.-Y. Lee. “Deep CNNs Along the Time Axis With Intermap Pooling for Robustness to Spectral Variations”. In: *IEEE Signal Processing Letters* 23.10 (2016), pp. 1310–1314.
- [6] N. Iqbal, **H. Lee**, and S.-Y. Lee. “Smart user interface for mobile consumer devices using model-based eye-gaze estimation”. In: *IEEE Transactions on Consumer Electronics* 59.1 (2013), pp. 161–166.
- [7] **H. Lee**, N. Iqbal, W. Chang, and S.-Y. Lee. “A calibration method for eye-gaze estimation systems based on 3D geometrical optics”. In: *IEEE Sensors Journal* 13.9 (2013), pp. 3219–3225.

## CONFERENCE PAPERS

---

- [8] M. Kim, J. Koo, H. Lee, J. Park<sup>†</sup>, **H. Lee**, and K. Jung<sup>†</sup>. “LifeTox: Unveiling Implicit Toxicity in Life Advice”. In: *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL)* (2024).
- [9] S. Kim<sup>\*</sup>, J. Shin<sup>\*</sup>, Y. Cho<sup>\*</sup>, J. Jang, S. Longpre, **H. Lee**, S. Yun, S. Shin, S. Kim, J. Thorne, et al. “Prometheus: Inducing Fine-grained Evaluation Capability in Language Models”. In: *ICLR*. 2024.
- [10] M. Kim, H. Lee, K. M. Yoo, J. Park, **H. Lee**<sup>†</sup>, and K. Jung<sup>†</sup>. “Critic-Guided Decoding for Controlled Text Generation”. In: *Findings of the Association for Computational Linguistics: ACL 2023* (2023).
- [11] S. Kim, S. Yun, **H. Lee**, M. Gubri, S. Yoon, and S. J. Oh. “ProPILE: Probing Privacy Leakage in Large Language Models”. In: *NeurIPS 2023* (2023).
- [12] M. Ko, I. Seong, **H. Lee**, J. Park, M. Chang, and M. Seo. “ClaimDiff: Comparing and Contrasting Claims on Contentious Issues”. In: *Findings of the Association for Computational Linguistics: ACL 2023*. 2023, pp. 4711–4731.
- [13] D. Lee, J. Lee, J.-W. Ha, J.-H. Kim, S.-W. Lee, **H. Lee**, and H. O. Song. “Query-Efficient Black-Box Red Teaming via Bayesian Optimization”. In: *The 61st Annual Meeting Of The Association For Computational Linguistics (ACL)*. 2023.
- [14] **H. Lee**<sup>\*</sup>, S. Hong<sup>\*</sup>, J. Park, T. Kim, M. Cha, Y. Choi, B. P. Kim, G. Kim, E.-J. Lee, Y. Lim, A. Oh, S. Park, and J.-W. Ha. “SQuARe: A Large-Scale Dataset of Sensitive Questions and Acceptable Responses Created Through Human-Machine Collaboration”. In: *The 61st Annual Meeting Of The Association For Computational Linguistics (ACL)* (2023).
- [15] **H. Lee**<sup>\*</sup>, S. Hong<sup>\*</sup>, J. Park, T. Kim, G. Kim, and J.-W. Ha. “KoSBI: A Dataset for Mitigating Social Bias Risks Towards Safer Large Language Model Application”. In: *The 61st Annual Meeting Of The Association For Computational Linguistics (ACL)*. 2023.
- [16] J. J. Y. Chung, W. Kim, K. M. Yoo, **H. Lee**, E. Adar, and M. Chang. “TaleBrush: sketching stories with generative pretrained language models”. In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 2022, pp. 1–19.
- [17] H. Lee, K. M. Yoo, J. Park, **H. Lee**<sup>†</sup>, and K. Jung<sup>†</sup>. “Masked Summarization to Generate Factually Inconsistent Summaries for Improved Factual Consistency Checking”. In: *Findings of the Association for Computational Linguistics: NAACL 2022* (2022), pp. 1019–1030.
- [18] K. Lee, W. Han, S.-w. Hwang, **H. Lee**, J. Park, and S.-W. Lee. “Plug-and-Play Adaptation for Continuously-updated QA”. In: *Findings of the Association for Computational Linguistics: ACL 2022* (2022), pp. 438–447.
- [19] G.-C. Kang, J. Park, **H. Lee**, B.-T. Zhang, and J.-H. Kim. “Reasoning Visual Dialog with Sparse Graph Learning and Knowledge Transfer”. In: *Findings of Empirical Methods in Natural Language Processing (Findings of EMNLP)* (2021).
- [20] **H. Lee**<sup>\*</sup>, J. Lee<sup>\*</sup>, and T.-Y. Kim. “SUMBT: Slot-Utterance Matching for Universal and Scalable Belief Tracking”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics. 2019, pp. 5478–5483.

- [21] G. Kim, **H. Lee**, B. Kim, and S.-y. Lee. “[Compositional Sentence Representation from Character within Large Context Text](#)”. In: *International Conference on Neural Information Processing*. Springer, Cham. 2017, pp. 674–685.
- [22] H.-G. Kim, J. Roh, **H. Lee**, G. Kim, and S.-Y. Lee. “[Active Learning for Large-scale Object Classification: from Exploration to Exploitation](#)”. In: *Proceedings of the 3rd International Conference on Human-Agent Interaction*. ACM. 2015, pp. 251–254.
- [23] B.-K. Kim, **H. Lee**, J. Roh, and S.-Y. 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*. 2015, pp. 427–434.

## WORKSHOP PAPERS

---

- [24] S. Kim<sup>\*</sup>, J. Shin<sup>\*</sup>, Y. Cho<sup>\*</sup>, J. Jang, S. Longpre, **H. Lee**, S. Yun, S. Shin, S. Kim, J. Thorne, et al. “[Prometheus: Inducing Evaluation Capability in Language Models](#)”. In: *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*. 2023.
- [25] J. Ahn, **H. Lee**, J. Kim, and A. Oh. “[Why Knowledge Distillation Amplifies Gender Bias and How to Mitigate from the Perspective of DistilBERT](#)”. In: *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*. 2022, pp. 266–272.
- [26] J. J. Y. Chung, W. Kim, K. M. Yoo, **H. Lee**, E. Adar, and M. Chang. “[TaleBrush: Visual Sketching of Story Generation with Pretrained Language Models](#)”. In: *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 2022, pp. 1–4.
- [27] G. Kim<sup>\*</sup>, **H. Lee**<sup>\*</sup>, C. Lee, E. Hong, B. Kim, and S.-Y. Lee. “[A Deep Chatbot for QA and Chitchat](#)”. In: *The Conversational Intelligence Challenge section on NIPS 2017 Competition Track Workshop*. Neural Information Processing Systems Foundation. 2017.

## PREPRINTS

---

- [28] J. Ahn, T. Lee, J. Lim, J.-H. Kim, S. Yun, **H. Lee**, and G. K. Kim. “[TimeChara: Evaluating Point-in-Time Character Hallucination of Role-Playing Large Language Models](#)”. In: *Under Review* (2024).
- [29] M. Gubri, D. Ulmer, **H. Lee**, S. Yun, and S. J. Oh. “[TRAP: Targeted Random Adversarial Prompt Honey-pot for Black-Box Identification](#)”. In: *arXiv preprint arXiv:2402.12991* (2024).
- [30] M. Kim, H. Lee, J. P. Park, **H. Lee**<sup>†</sup>, and K. Jung<sup>†</sup>. “[AdvisorQA: Towards Helpful and Harmless Advice-seeking Question Answering with Collective Intelligence](#)”. In: *arXiv preprint arXiv:2404.11826* (2024).
- [31] J. Lee<sup>\*</sup>, M. Kim<sup>\*</sup>, S. Kim, J. Kim, S. Won, **H. Lee**, and E. Choi. “[KorNAT: LLM Alignment Benchmark for Korean Social Values and Common Knowledge](#)”. In: *arXiv preprint arXiv:2402.13605* (2024).
- [32] **H. X. Team**. “[HyperCLOVA X Technical Report](#)”. In: *arXiv preprint arXiv:2404.01954* (2024).
- [33] D. Ulmer, M. Gubri, **H. Lee**, S. Yun, and S. J. Oh. “[Calibrating Large Language Models Using Their Generations Only](#)”. In: *arXiv preprint arXiv:2403.05973* (2024).
- [34] T. Lee<sup>\*</sup>, S. Hong<sup>\*</sup>, J. Ahn, I. Hong, **H. Lee**, S. Yun, J. Shin<sup>†</sup>, and G. Kim<sup>†</sup>. “[Who Wrote this Code? Watermarking for Code Generation](#)”. In: *arXiv preprint arXiv:2305.15060* (2023).

## PATENTS

---

- [35] T.-Y. Kim, J. Kim, H. Kim, J. Lee, **H. Lee**, H. Jeon, and S. Jo. [Method and Apparatus for Providing Hybrid Intelligent Customer Consultation](#). KR Patent KR102488886B1, filed March 2022 and issued January 2023.
- [36] **H. Lee**, T.-Y. Kim, and J. Lee. [Method and device for tracking dialogue state in goal-oriented dialogue system](#). KR Patent KR102281581B1, filed July 2019 and issued July 2021; CN Patent CN114127711A; WO Patent WO2021010636A1; US Patent US11687731B2, filed July 2020 and issued Jan 2021.

## ACADEMIC SERVICES

---

### Organizing Committee

- ACL 2025 Diversity and Inclusion Co-Chair
- FAccT'22 Craft HyperScaleFAccT

### Area Chair

- CoLM 2024
- NeurIPS Data & Benchmark 2023-2024

### Conference Reviewer

- ARR 2021-2024, ACL 2021-2023, EMNLP 2021-2023, COLING 2020-2022
- NeurIPS 2021-2023, ICLR 2021-2024
- WWW 2022

### Journal Reviewer

- TMLR 2024
- Speech Communication 2019
- IEEE Transactions on Neural Networks and Learning Systems 2017-2018
- Neural Processing Letters 2015

## EXTERNAL ACTIVITIES

---

- |  |             |
|--|-------------|
| Committee member of the Forum on Artificial Intelligence Ethics and Policy | 2023 - 2024 |
| • Organized by the Ministry of Science and ICT, South Korea                |             |

## INVITED TALKS (SELECTED)

---

- |  |      |
|--|------|
| A Tutorial on Large Language Models Safety   | 2024 |
| • IEIE AI Signal Processing Society Winter School  |      |
| Towards Safer Large Language Models  | 2023 |
| • KAIST, Seoul National University, Sogang University, GIST, Korea University, Institute of Basic Science, HKUST, University of Tübingen |      |
| AI Ethics and Policies for Everyone  | 2023 |
| • NAVER DAN Conference 2023  |      |
| Ethical Problems in Language Models  | 2022 |
| • KAIST, Seoul National University, Hanyang University   |      |
| Introduction to deep learning for dialogue systems   | 2020 |
| • KAIST, Sookmyung Women's Univ. Yeonsei Univ. Inha Univ.  |      |