HWARAN LEE

@ hwaran.lee@gmail.com | 🔗 LinkedIn | 🐙 GitHub | 🌐 Homepage | 📍 Sogang University, South Korea
Updated 2025-04-15

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[1, 13, 15, 23, 24]; (2) controllable language generation[19, 25]; (3) LLM security, including adversarial attack and red-teaming[40, 20, 22]; (4) safety alignment and learning methods[8, 38, 27].

Before the advent of LLMs, my work was focused on deep neural network-based Dialog Systems[42, 2, 29, 36] and Speech Recognition Systems[3, 4, 5]. 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[6, 7] for human-computer interface.

EMPLOYMENT HISTORY

Sogang University — *Seoul, South Korea*
- *Assistant Professor, Department of Artificial Intelligence, Computer Science & Engineering* — *Mar 2025 – **Present***

NAVER Cloud — *Seongnam, South Korea*
- *Leader, Language Research Team, NAVER AI Lab* — *Apr 2023 – Feb 2025*
- *Leader, Safety Team, HyperCLOVA X and HyperScale AI* — *Apr 2023 – Jun 2024*
- *Tech Leader, Language Research Team, NAVER AI Lab* — *Jan 2023 – Mar 2023*

NAVER — *Seongnam, South Korea*
- *Leader, AI Safety Lab, Future AI Center* — *Jan 2024 – Feb 2025*
- *Tech Leader, Language Research Team, NAVER AI Lab* — *Jul 2022 – Dec 2022*
- *Research Scientist, Language Research Team, NAVER AI Lab* — *Mar 2021 – Jun 2022*

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

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

EDUCATION

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

B.S., KAIST — Deajeon, South Korea
Mathematical Science & Financial Engineering Program (minor) — *Mar 2008 – Aug 2012*
Magna Cum Laude

PUBLICATIONS

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

Journal Articles

[1] J. Jin, J. Kim, N. Lee, H. Yoo, A. Oh, and **H. Lee**[†]. "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, H. Lee, J. P. Park, **H. Lee**[†], and K. Jung[†]. "AdvisorQA: Towards Helpful and Harmless Advice-seeking Question Answering with Collective Intelligence". In: *NAACL* (2025).

[9] M. Kim, T. Thonet, J. Rozen, **H. Lee**, K. Jung, and M. Dymetman. "Guaranteed Generation from Large Language Models". In: *ICLR* (2025).

[10] Y. Yang, H. Yoo, and **H. Lee**[†]. "MAQA: Evaluating Uncertainty Quantification in LLMs Regarding Data Uncertainty". In: *NAACL Findings* (2025).

[11] 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: *Findings of the Association for Computational Linguistics: ACL* (2024).

[12] M. Gubri, D. Ulmer, **H. Lee**, S. Yun, and S. J. Oh. "TRAP: Targeted Random Adversarial Prompt Honeypot for Black-Box Identification". In: *Findings of the Association for Computational Linguistics: ACL* (2024).

[13] M. Kim, J. Koo, H. Lee, J. Park[†], **H. Lee**, and K. Jung[†]. "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).

[14] S. Kim[*], J. Shin[*], Y. Cho[*], 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.

[15] J. Lee[*], M. Kim[*], S. Kim, J. Kim, S. Won, **H. Lee**, and E. Choi. "KorNAT: LLM Alignment Benchmark for Korean Social Values and Common Knowledge". In: *Findings of the Association for Computational Linguistics: ACL* (2024).

[16] T. Lee[*], S. Hong[*], J. Ahn, I. Hong, **H. Lee**, S. Yun, J. Shin[†], and G. Kim[†]. "Who Wrote this Code? Watermarking for Code Generation". In: *The 62nd Annual Meeting Of The Association For Computational Linguistics (ACL)* (2024).

[17] J. Myung, N. Lee, Y. Zhou, J. Jin, R. A. Putri, D. Antypas, H. Borkakoty, E. Kim, C. Perez-Almendros, A. A. Ayele, V. Gutierrez-Basulto, Y. Ibanez-Garcia, **H. Lee**, S. H. Muhammad, K. Park, A. S. Rzayev, N. White, S. M. Yimam, M. T. Pilehvar, N. Ousidhoum, J. Camacho-Collados, and A. Oh. "BLEnD: A Benchmark for LLMs on Everyday Knowledge in Diverse Cultures and Languages". In: *NeurIPS Dataset & Benchmark* (2024).

[18] D. Ulmer, M. Gubri, **H. Lee**, S. Yun, and S. J. Oh. "Calibrating Large Language Models Using Their Generations Only". In: *The 62nd Annual Meeting Of The Association For Computational Linguistics (ACL)* (2024).

[19] M. Kim, H. Lee, K. M. Yoo, J. Park, **H. Lee**[†], and K. Jung[†]. "Critic-Guided Decoding for Controlled Text Generation". In: *Findings of the Association for Computational Linguistics: ACL 2023* (2023).

[20] 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).

[21] 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.

[22] 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.

[23] **H. Lee**[*], S. Hong[*], 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).

[24] **H. Lee**[*], S. Hong[*], 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.

[25] 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.

[26] H. Lee, K. M. Yoo, J. Park, **H. Lee**[†], and K. Jung[†]. "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.

[27] 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.

[28] 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).

[29] **H. Lee**[*], J. Lee[*], 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.

[30] 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.

[31] 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.

[32] 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

[33] S. Kim[*], J. Shin[*], Y. Cho[*], 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.

[34] 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.

[35] 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.

[36] G. Kim*, **H. Lee***, 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

[37] M. Kim, K.-i. Lee, S. Joo, **H. Lee**, and K. Jung. "Drift: Decoding-time Personalized Alignments with Implicit User Preferences". In: *arXiv preprint arXiv:2502.14289* (2025).

[38] **H. X. Team**. "HyperCLOVA X Technical Report". In: *arXiv preprint arXiv:2404.01954* (2024).

[39] H. Yoo, C. Park, S. Yun, A. Oh, and **H. Lee**†. "Code-Switching Curriculum Learning for Multilingual Transfer in LLMs". In: *arXiv preprint arXiv:2411.02460* (2024).

[40] H. Yoo, Y. Yang, and **H. Lee**†. "CSRT: Evaluation and Analysis of LLMs using Code-Switching Red-Teaming Dataset". In: *arXiv preprint arXiv:2406.15481* (2024).

Patents

[41] 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.

[42] **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 Chair
- FAccT'22 Craft HyperScaleFAccT

Area Chair
- ARR (Action Editor) 2024-2025
- NeurIPS Data & Benchmark 2023-2024
- CoLM 2024-2025

Conference Reviewer
- ARR, ACL, EMNLP 2021-2024, COLING 2020-2022
- NeurIPS, ICLR 2021-2025
- 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)

Safe and Trustworthy AI   2025
- Google Research Australia, Singapore University of Technology and Design

Conference on Frontier AI Safety Frameworks @ Berkeley, U.S.   2024

Multilinguality and LLMs, Panelist @ CoLM   2024

Safety and Trustworthiness of Large Language Models   2024
- IEIE AI Signal Processing Society Winter School, Hanyang Univ., KAIST, SKKU

Towards Safer Large Language Models   2023
- KAIST, Seoul National University, Sogang University, GIST, Korea University, Institute of Basic Science, HKUST, University of Tubingen

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.