Last updated in October 2025

Wonje Jeung

✉ Email 🎓 Scholar 🌐 Website in LinkedIn

Education

Yonsei University *Sept 2024 – Aug 2026* (Expected)

MS in Artificial Intelligence

- GPA: 4.24/4.3
- **Coursework:** Information Theory, Statistical Pattern Recognition, Multimodal Deep Learning, Text Understanding, Large Scale Learning and Inference, Medical Imaging, Responsible AI.
- Advisor: Albert No

BS in Computer Science *Mar 2020 – Aug 2024*

- GPA: 4.09/4.3
- **Coursework:** Machine Learning, Computer Vision, Deep Learning Theory and Practice, Signal Processing, Probability and Statistics, Computer Systems, Computer Architectures, Operating Systems, Big Data.
- Advisor: Jonghyun Choi

Publications

Conference Proceedings

C1. An Information Theoretic Evaluation Metric For Strong Unlearning
Dongjae Jeon*, **Wonje Jeung***, Taeheon Kim, Albert No, Jonghyun Choi
The Association for the Advancement of Artificial Intelligence (**AAAI 2026**) [PDF] ↗.

C2. SAFEPATH: Preventing Harmfulness Reasoning in Chain-of-Thought via Early Alignment.
Wonje Jeung, Sangyeon Yoon, Minsuk Kahng, Albert No.
Conference on Neural Information Processing Systems (**NeurIPS 2025**) [PDF] ↗.

C3. SEPS: A Separability Measure for Robust Unlearning in LLMs
Wonje Jeung*, Sangyeon Yoon*, Albert No.
Conference on Empirical Methods in Natural Language Processing (**EMNLP 2025**) [PDF] ↗.

C4. R-TOFU: Unlearning in Large Reasoning Models
Sangyeon Yoon, **Wonje Jeung**, Albert No.
Conference on Empirical Methods in Natural Language Processing (**EMNLP 2025**) [PDF] ↗.

C5. Large Language Models Still Exhibit Bias in Long Text
Wonje Jeung, Dongjae Jeon, Ashkan Yousefpour, Jonghyun Choi.
Annual Meeting of the Association for Computational Linguistics (**ACL 2025**) [PDF] ↗.

C6. 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.
Annual Meeting of the Association for Computational Linguistics (**ACL 2025**) [PDF] ↗.

C7. ReALFRED: An Embodied Instruction Following Benchmark in Photo-Realistic Environments
Taewoong Kim*, Cheolhong Min*, Byeonghwi Kim, Jinyeon Kim, **Wonje Jeung**, Jonghyun Choi
European Conference on Computer Vision (**ECCV 2024**) [PDF] ↗.

C8. Learning Equi-angular Representations for Online Continual Learning
Minhyuk Seo, Hyunseo Koh, **Wonje Jeung**, Minjae Lee, San Kim, Hankook Lee, Sungjun Cho, Sungik Choi, Hyunwoo Kim, Jonghyun Choi
Conference on Computer Vision and Pattern Recognition (**CVPR 2024**) [PDF] ↗.

Workshops

W1. Adversarial Sample-Based Approach for Tighter Privacy Auditing in Final Model-Only Scenarios
Sangyeon Yoon*, **Wonje Jeung***, Albert No
Workshop on Statistical Foundations of LLMs and Foundation Models (**NeurIPS SFLLM 2024**) [PDF] ↗.

W2. Multi-Level Knowledge Distillation and Dynamic Self-Supervised Learning for Continual Learning
SNUMPR TEAM
Workshop on CLVISION (**CVPR CLVISION 2024**), 2nd place [PDF] ↗.

Preprints

P1. A2D: Any-Order, Any-Step Safety Alignment for Diffusion Language Models
Wonje Jeung*, Sangyeon Yoon*, Dongjae Jeon, Sangwoo Shin, Hyesoo Hong, Yoonjun Cho, Albert No
International Conference on Learning Representations (**ICLR 2026**), Under Review [PDF] ↗.

P2. DUSK: Do not unlearn shared knowledge
Wonje Jeung*, Sangyeon Yoon*, Hyesoo Hong, Soeun Kim, Seungju Han, Youngjae Yu, Albert No
International Conference on Learning Representations (**ICLR 2026**), Under Review [PDF] ↗.

P3. Rethinking Benign Relearning: Syntax as the Hidden Driver of Unlearning Failures
Sangyeon Yoon, Hyesoo Hong, **Wonje Jeung**, Albert No
International Conference on Learning Representations (**ICLR 2026**), Under Review.

P4. Rainbow Padding: Mitigating Early Termination in Instruction-Tuned Diffusion LLMs
BumJun Kim*, Dueun Kim*, Dongjae Jeon*, **Wonje Jeung**, Albert No.
International Conference on Learning Representations (**ICLR 2026**), Under Review [PDF] ↗..

Research Experience

AI-ISL (P.I. Albert No) *Yonsei University*
Graduate Research Assistant *Aug 2024 – Present*

- Developed safety alignment methods across diverse language models, including LLMs, LRMs, and dLLMs.
- Designed a privacy auditing algorithm providing tight differential privacy bounds, tailored for realistic scenarios where only the final model's outputs are accessible, using canary samples.
- Exposed vulnerabilities in existing unlearning evaluations and introduced an information-theoretic metric (IDI), a separability measure (SEPS), and realistic unlearning benchmarks (R-TOFU, DUSK).

Vnl Lab (P.I. Jonghyun Choi) *Yonsei University*
Undergraduate Research Intern *Mar 2023 – July 2024*

- Developed strategies to improve vision classification performance in online continual learning settings.
- Hired and managed hundreds of annotators via Amazon MTurk to collect labeled data and build an embodied instruction-following benchmark in real-world scenarios.
- Simulated and analyzed diverse embodied benchmarks, including RoboThor, Gibson, and Habitat.
- Revealed that recent LLMs exhibit bias in long-text generation and proposed a mitigation strategy.

Professional Activities

Conference Reviewer

○ International Conference on Learning Representations (**ICLR**)	*2026*
○ Empirical Methods in Natural Language Processing (**EMNLP**)	*2025*
○ IEEE Transactions on Information Forensics and Security (**T-IFS**)	*2025*
○ **NeurIPS Workshop** on Socially Responsible Language Modelling Research	*2024*

Invited Talks

- Gave a talk at the MLSYS group on deep safety alignment specialized for different language model architectures (LLMs, LRMs, dLLMs). *Oct 2025*
- Presented at the MLSYS group on leveraging reasoning to achieve safety. *June 2025*

Honors and Scholarship

Academic Excellence Award *Yonsei University*
1st semester 2020, 2nd semester 2020, 1st semester 2023, 1st semester 2024
Academic Excellence Scholarship *Yonsei University*
Covering tuition fees of $1,397 (1st semester 2020), $1,665 (2nd semester 2020).
Software Maestro Fellowship IITP & ICT
Full Stipend ($6,154), elite software training program.
Excellence Award in RC Creativity Platform *Yonsei University*
Awarded $769 prize money.

Work Experience

SW Maestro *Seoul, Korea*
Research Trainee, funded by Ministry of Science and ICT of Korean Government *Apr 2022 – Dec 2022*

- Built a family app leveraging ML methods to enhance emotional connections, serving as team leader.
- Designed and implemented scalable AWS architecture and backend framework.

R2C Company *Seoul, Korea*
Software Developer / Intern *Oct 2021 – Apr 2022*

- Built a system to collect in-app user behavior data, enabling adaptive surveys based on user actions.
- Streamlined cross-platform development by unifying Android and iOS codebases.

Skills

Research Pytorch, Huggingface, DeepSpeed, vLLM, MTurk Amazon Survice.
Development Spring, Node, Express, React, Flutter, Cloud Service (Lambda, S3, CloudFront, EC2).
Languages English: professional proficiency, Korean: native.

Reference

Prof. Albert No
Associate Professor, Department of Artificial Intelligence, Yonsei University.
Email: albertno@yonsei.ac.kr — Phone: +82 (2) 2123-7808
Prof. Minsuk Kahng
Asssistant Professor, Department of Computer Science and Engineering, Yonsei University.
Email: minsuk@yonsei.ac.kr
Prof. Jean Oh
Associate Professor, School of Computer Science (Robot Institute), Carnegie Mellon University.
Email: hyaejino@andrew.cmu.edu