

# WONJE JEUNG

43, Yeonhui-ro, Mapo-gu, Seoul, Republic of Korea

 Email  Google Scholar  Website  Linked In

## EDUCATION

**Yonsei University** B.S. in Computer Science  
GPA: 4.09 / 4.3 (142 credits), Magna Cum Laude.

2020.03-2024.08

**Yonsei University** M.S. in Artificial Intelligence Department

2024.08-

## PUBLICATIONS

### Safety, Privacy and Alignment

(8) SAFEPATH: Preventing Harmful Reasoning in Chain-of-Thought via Early Alignment.

**Wonje Jeung**, Sangyeon Yoon, Minsuk Kang, Albert No. **NeurIPS**, UNDER REVIEW, (2025). [\[pdf\]](#) [\[website\]](#)

(7) DUSK: Do not unlearn shared knowledge.

**Wonje Jeung\***, Sangyeon Yoon\*, Haesoo Hong, Soeun Kim, Seungju Han, Youngjae Yu, Albert No. **NeurIPS**, UNDER REVIEW, (2025). [\[pdf\]](#) [\[website\]](#)

(6) R-TOFU: Unlearning in Large Reasoning Models.,

Sangyeon Yoon, **Wonje Jeung**, Albert No. **EMNLP**, Oral Presentation (2025). [\[pdf\]](#) [\[website\]](#)

(5) SEPS: A Separability Measure for Robust Unlearning in LLMs.

**Wonje Jeung\***, Sangyeon Yoon\*, Albert No. **EMNLP**, Main (2025). [\[pdf\]](#)

(4) Representation Bending for Large Language Model Safety.

Ashkan Yousefpour\*, Taeheon Kim\*, Ryan Sungmo Kwon, Seungbeen Lee, **Wonje Jeung**, Seungju Han, Alvin Wan, Harrison Ngan, Youngjae Yu, Jonghyun Choi. **ACL**, Main, (2025). [\[pdf\]](#)

(3) Adversarial Sample-Based Approach for Tighter Privacy Auditing in Final Model-Only Scenarios.

Sangyeon Yoon\*, **Wonje Jeung\***, Albert No. **NeurIPS Statistical Foundations of LLMs and Foundation Models**, (2024). [\[pdf\]](#)

(2) An Information Theoretic Metric for Evaluating Unlearning Models.

Dongjae Jeon\*, **Wonje Jeung\***, Taeheon Kim, Albert No, Jonghyun Choi. (2025). [\[pdf\]](#)

(1) Large Language Models Still Exhibit Bias in Long Text.

**Wonje Jeung**, Dongjae Jeon, Ashkan Yousefpour, Jonghyun Choi. **ACL**, Findings (2025). [\[pdf\]](#)

### Vision & Robotics

(3) Multi-Level Knowledge Distillation and Dynamic Self-Supervised Learning for Continual Learning.

**Technical Report**, 5th CLVISION Challenge (Second Place), **CVPR** (2024) [\[pdf\]](#)

(2) ReALFRED: An Embodied Instruction Following Benchmark in Photo-Realistic Environments.

Taewoong Kim\*, Cheolhong Min\*, Byeonghwi Kim, Jinyeon Kim, **Wonje Jeung**, Jonghyun Choi. **ECCV** (2024). [\[pdf\]](#)

(1) 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. **CVPR** (2024). [\[pdf\]](#)

## RESEARCH EXPERIENCE

### Vnl Lab, SNUMPR Lab

Jonghyun Choi

Research Student

Mar 2023 to July 2024

- Demonstrated vulnerabilities in existing machine unlearning metrics by altering only the model's head, achieving significant performance on current metrics. Developed a new metric based on mutual information of samples to more accurately measure the degree of unlearning.
- Revealed that even recent Large Language Models (LLMs) exhibit explicit unfair bias in long text generation scenarios. Created a specialized dataset for testing and proposed a mitigation strategy.
- Developed an effective method for various continual learning scenarios, including unrelated data, and data without labels.
- Investigated strategies for online continual learning (preparatory data training and residual connections). Conducted extensive experiments to validate and maximize the performance of proposed methods.
- Designed and managed a website for data annotation, utilizing MTurk to obtain labeled data from diverse people. Simulated and analyzed benchmarks such as RoboTHOR, Gibson, and Habitat.

- Conducted research on detecting and explaining memorization behaviors in diffusion models.
- Worked on a privacy auditing algorithm to achieve tight bounds for differential privacy, specifically tailored for practical scenarios with only final models.
- Worked on safety alignment inspired by machine unlearning for LLMs and LRMs.
- Worked on unlearning metric and benchmarks.
- Studying / Researching agent and robot safety with reinforcement learning.

AWARDS & SCHOLARSHIPS

Academic Award	Yonsei University
high honors (1st semester 2020), honors (2nd semester 2020, 1st semester 2023, 1st semester 2024).	
Academic Excellence Scholarship	Yonsei University
covering tuition fees of \$1,397 (1st semster 2020), \$1,665 (2nd semester 2020).	
RC Creativity Platform	Yonsei University
Excellence Award, awarded \$769 prize money.	2020.12
Startup Express Competition	Korea University
3rd place for making data-driven survey application.	2021.06
Software Maestro Fellowship	National IT Promotion Agency (IITP) & Ministry of Science and ICT, South Korea
Full stipend (\$6,154), elite software training program.	2022.06-2022.11
CVPR Workshop Competition	CVPR (CLVISION)
2nd Place in Class-Incremental with Repetition (CIR) using Unlabeled Data.	2024.03
Capstone Project	Yonsei University
1st place for developing real-world continual learning setup.	2024.08
Outstanding Project Award	Ministry of Science and ICT, South Korea
Recognized as a nominee for SW festival, with a potential prize of \$385 to \$2,308. Final decision forthcoming.	2024.10

TEACHING & SERVICES

- Reviewer | NeurIPS Workshop, T-IFS, EMNLP.
- Teaching Assistant | Introduction to mathematics for deep learning (AAI2230.01-00), Calculus (2021), Geometry (2021).

WORK EXPERIENCE

Pumasi	
Software Developer / Intern	June 2020 to August 2021
• Developed prototypes swiftly to validate product concepts, collaborating closely with cross-functional teams to align with user needs and preferences through weekly feedback cycles.	
• Led iterative testing and refinements to enhance product-market fit, focusing on adaptability and efficiency in early-stage development.	
R2C Company	
Software Developer / Intern	Oct 2021 to April 2022
• Built a system for collecting in-app user behavior data, enabling targeted surveys based on user demographics, with optimized efficiency by locally storing data until API-triggered batch transmission.	
• Streamlined cross-platform deployment by unifying separate Android and iOS codebases into a React Native framework, re-designing interface layouts and architecture for a cohesive experience.	

ACTIVITY & PROJECT

- **SW Maestro** (2022.04-2022.11): Participated as a team leader in a software education program organized by the Ministry of Science and ICT. Developed an album app for families, focusing on improving emotional connections with ML methods.
- **Industry-Academia Collaboration Project** (2023.09-2023.12): Responsible for researching a method for effectively separating bone segments in noisy and blurry images. I served as the team leader.
- **Others:** Automatic Scheduling Application (SPARCS Hackerton), Sentiment Analysis Diary Application, MODULAB RG lead.

SKILLS

**Research** Python, Pytorch, Jax, Huggingface, Deepspeed, MTurk Amazon Service.  
**Development** Backend (Spring, Node, Express), Frontend (Vanila, React, Flutter), Cloud Service (Lambda, S3, CloudFront, EC2)  
**Languages** English: professional proficiency. Korean: native.

REFERENCE

**Prof. Jonghyun Choi**  
Associate Professor, Department of ECE, Seoul National University University  
Email: jonghyunchoi@snu.ac.kr | Phone: +82 (2) 880-1766

**Prof. Albert No**  
Associate Professor, Department of Artificial Intelligence, Yonsei University  
Email: albertno@yonsei.ac.kr | Phone: +82 (2) 2123-7808

**Prof. Jongduk Baek**

Professor, Department of Artificial Intelligence, Yonsei University

Email: [jongdukbaek@yonsei.ac.kr](mailto:jongdukbaek@yonsei.ac.kr) | Phone: +82 (2) 2123-5737

**Dr. Ashkan Yousefpour**

Post-Doctoral Yonsei University (Prof. Youngjae Yu) and Seoul National University (Prof. Jonghyun Choi)

Email: [yousefpour.ashkan@gmail.com](mailto:yousefpour.ashkan@gmail.com) | Phone: +1 (914) 803-6303