Sangyeon Yoon

Seoul, Republic of Korea

2025324135@yonsei.ac.kr Scholar Website LinkedIn

Research Interest

I am an M.S. student in Artificial Intelligence at Yonsei University, advised by Prof. Albert No. **My research interests are LLM safety, machine unlearning, natural language processing and LLM reasoning.**

Education

Yonsei University *Sep 2025 – Present*
M.S. in Artificial Intelligence

Hongik University *Mar 2019 – Aug 2025*
B.S. in Computer Engineering

- Includes a 2-year mandatory military service in South Korea.

Research Experience

EXAONE Lab, LG AI Research *Sep 2025 – Feb 2026*
Research Intern

- Contributed to foundation model development, including EXAONE-4.5 and K-EXAONE-236B-A23B.
- Contributed to the post-training team by designing synthetic data for reasoning.

AI-ISL Lab, Yonsei University *Mar 2024 – Present*
Graduate Reseacher (under **Prof. Albert No**)

Publications

* indicates equal contribution.

Tech Report

1. EXAONE 4.5 Technical Report: LG's First Open-Weight Vision-Language Model for Industrial Intelligence
   LG AI Research; contributed to build synthetic post-training data
   2026 [Link] [Model]

2. K-EXAONE Technical Report: Journey to Frontier-Level Performance of Foundation Models
   LG AI Research; contributed to build synthetic post-training data
   2026 [Link] [Model] *Hugging Face* ***#1 Paper of the Day***

Preprints

1. Few-Shot Truly Benign DPO Attack for Jailbreaking LLMs
   Sangyeon Yoon*, Wonje Jeung*, Yoonjun Cho, Dongjae Jeon, Albert No
   Under Review

2. VLMs Trace Without Tracking: Diagnosing Failures in Visual Path Following
   Hyesoo Hong, Min Soo Kim, Wonje Jeung, **Sangyeon Yoon**, Dongjae Jeon, Albert No
   Under Review

3. BenchPreS: A Benchmark for Context-Aware Personalized Preference Selectivity of Persistent-Memory LLMs
   Sangyeon Yoon, Sunkyoung Kim, Hyesoo Hong, Wonje Jeung, Yongil Kim, Wooseok Seo, Heuiyeen Yeen, Albert No
   Under Review [Link]

Conference & Workshop Papers

1. Position: The Term "Machine Unlearning" Is Overused in LLMs
   Sangyeon Yoon*, Yeachan Jun*, Albert No
   International Conference on Machine Learning (**ICML 2026**)

2. DUSK: Do Not Unlearn Shared Knowledge
   Wonje Jeung*, **Sangyeon Yoon***, Hyesoo Hong*, Soeun Kim, Seungju Han, Youngjae Yu, Albert No
   Findings of the Association for Computational Linguistics (**ACL Findings 2026**)

3. Rethinking Benign Relearning: Syntax as the Hidden Driver of Unlearning Failures
   Sangyeon Yoon, Hyesoo Hong, Wonje Jeung, Albert No
   Selected as a BK21 Academy Research Fellow
   International Conference on Learning Representations (**ICLR 2026**) [Link]

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

5. SAFEPATH: Preventing Harmful Reasoning in Chain-of-Thought via Early Alignment
   Wonje Jeung, **Sangyeon Yoon**, Minsuk Kang, Albert No
   Conference on Neural Information Processing Systems (**NeurIPS 2025**) [Link]

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

7. 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 Main**) [Link]

8. Adversarial Sample-Based Privacy Auditing
   Sangyeon Yoon*, Wonje Jeung*, Albert No
   Workshop on Statistical Foundations of LLMs and Foundation Models (**NeurIPS 2024@SFLLM**) [Link]

Reviewer

NeurIPS *2026*

Honors and Scholarship

Academy Research Fellowship, BK21 *2026*
Awarded for the research paper ($1.5K).

Undergraduate Scholarship, Hongik University *2019 – 2025*
Full scholarship ($58K over 4 years).