DASOL CHOI

dasolchoi@yonsei.ac.kr | +82 10-2431-0366 | Github |LinkedIn | Google Scholar

RESEARCH SUMMARY

Machine learning researcher specializing in privacy, responsible AI, and safety of foundation models, with publications in AAAI, WACV, JMIR, and top workshops.

EDUCATION

Yonsei University, Seoul, South Korea — 03.2024 -- Present
Master's in Digital Analytics, College of Computing
Coursework: Intermediate Programming, Machine Learning, Practical Applications of Big Data, Database, Deep Learning, Topics in Responsible AI, Natural Language Processing, etc — GPA: 4.24/4.50

Kyung Hee University, Seoul, South Korea — 03.2013 -- 02.2019
Bachelor of Arts in Japanese Language (Major)
Bachelor of Physical Education (Double Major) — GPA: 4.07/4.50

EXPERIENCE

AIM Intelligence #Research Scientist — 07.2025 -- Present

- Conducted collaborative research with BMW Group on organization-specific policy alignment for LLMs; conceptualized and submitted COMPASS framework to ARR, Oct 2025.
- Developing AIM Multimodal Guard for proactive policy-violation filtering across image, video, audio, and text modalities; provisional patent filed.
- In collaboration with LG Electronics, investigated benign-sounding audio jailbreaks in Audio–Language Models; co-authored paper under review at AAAI 2026.

SIONIC AI #ML/DL Research Intern — 12.2024 -- 02.2025

- Led development of inclusive multilingual reasoning benchmarks for Korean and Japanese language models, focusing on cultural and linguistic diversity [Dataset].
- Designed and implemented an evaluation leaderboard to systematically compare model performance on the benchmark [Leaderboard].

Samsung Medical Center #Data Analysis Researcher — 07.2023 -- 03.2024

- Developed an evaluation framework for LLM-generated patient records, specifying error taxonomies and quality standards for medical AI systems.
- Led federated learning model development for breast cancer prognosis prediction in a government-funded project (Ministry of Trade, Industry, and Energy).

SELECTED PUBLICATIONS

Distribution-Level Feature Distancing for Machine Unlearning: Towards a Better Trade-off Between Model Utility and Forgetting (AAAI 2025)

- **Dasol Choi**, Dongbin Na (First author)

LLM-Based Medical Document Evaluation: Integrating Human Expert Insights (MedInfo 2025 - Best Student Paper Award)

- Junhyuk Seo*, **Dasol Choi***, Wonchul Cha, Taerim Kim (Co-first author)

Evaluation Framework of Large Language Models in Medical Documentation: Development and Usability Study (JMIR 2024)

- Junkyuk Seo*, **Dasol Choi***, Wonchul Cha, Haanju Yoo, Namkee Oh, Yongjin Yi, Gyehwa Lee, Edward Choi, Taerim Kim (Co-first author)

Better Safe Than Sorry? Overreaction Problem of Vision Language Models in Visual Emergency Recognition (under review at WACV 2026)

- **Dasol Choi**, Seunghyun Lee, Youngsook Song (First author)

When Cars Have Stereotypes: Auditing Demographic Bias in Objects from Text-to-Image Models (under review at AAAI 2026)

- **Dasol Choi**, Jihwan Lee, Minjae Lee, Minsuk Kahng (First author)

COMPASS: A Framework for Evaluating Organization-Specific Policy Alignment in LLMs (submitted to ACL Rolling Review, October 2025)

- **Dasol Choi***, DongGeon Lee*, Brigitta Jesica Kartono*, Helena Berndt, Haon Park, Hwanjo Yu, Minsuk Kahng (Co-first author)

When Good Sounds Go Adversarial: Jailbreaking Audio-Language Models with Benign Inputs (under review at AAAI 2026)

- Bodam Kim*, Hiskias Dingeto*, Taeyoun Kwon*, **Dasol Choi**, DongGeon Lee, Haon Park, JaeHoon Lee, Jongho Shin (Co-author)

Towards Efficient Machine Unlearning with Data Augmentation: Guided Loss-Increasing (GLI) to Prevent the Degradation of Model Utility (CVPR 2024 Workshop)

- **Dasol Choi**, Soora Choi, Eunsun Lee, Jinwoo Seo, Dongbin Na (First author)

Towards Machine Unlearning Benchmarks: Forgetting the Personal Identities in Facial Recognition Systems (AAAI 2024 Workshop)

- **Dasol Choi***, Dongbin Na* (Co-first author)

KoMultiText: Large-Scale Korean Text Dataset for Classifying Biased Speech in Real-World Online Services (NeurIPS 2023 Workshop)

- **Dasol Choi**, Jooyoung Song, Eunsun Lee, Jinwoo Seo, Heejune Park, Dongbin Na (First author)

COMMUNITY INVOLVEMENT

HAE-RAE Open-source Research Community 05.2024 -- Present
Researcher, Contributor to Korean LLM Research

- Led development of HAERAE-Vision, a comprehensive Korean vision-language model benchmark for culturally-grounded multimodal evaluation.
- Co-authored research papers focusing on making Korean LLMs more accessible and culturally inclusive.
- Developed unified evaluation framework for Korean LLM benchmarking with standardized APIs and extensible architecture [GitHub].

AAAI 2026, Program Committee Member 08.2025 -- 10.2025

- Served as a reviewer for the **Main Track** and **Alignment Track**, evaluating submissions on responsible AI, alignment, and safety.

AWARDS & ACHIEVEMENTS

- **MedInfo2025 Best Student Paper Award** 08.2025
  Awarded for "LLM-Based Medical Document Evaluation: Integrating Human Expert Insights," recognizing innovative approaches to medical AI evaluation.
- **Google East Asia Student Travel Grants**, AAAI 2025 03.2025
  Awarded travel grant for the accepted paper "Distribution-Level Feature Distancing for Machine Unlearning: Towards a Better Trade-off Between Model Utility and Forgetting."

LANGUAGES

Korean: native **English:** TOEFL 95/120 **Japanese:** JLPT N1