

DASOL CHOI

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

RESEARCH SUMMARY

Machine learning researcher specializing in AI safety, alignment, and evaluation across multimodal and language models, focusing on building reliable and trustworthy AI systems.

EDUCATION

Yonsei University , Seoul, South Korea M.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	03.2024 – Present GPA: 4.24/4.50
Kyung Hee University , Seoul, South Korea Bachelor of Arts in Japanese Language (Major) Bachelor of Physical Education (Double Major)	03.2013 – 02.2019 GPA: 4.07/4.50

EXPERIENCE

AIM Intelligence <i>#Research Scientist</i> • Led BMW Group collaboration on organization-specific LLM policy alignment, developing <i>COMPASS</i> evaluation framework. • Developed Multimodal Guard, a proactive safety system for filtering policy violations across image, video, audio, and text modalities. • Investigated benign-sounding audio jailbreak attacks in Audio-Language Models with LG Electronics collaboration.	07.2025 – Present
SIONIC AI <i>#ML/DL Research Intern</i> • Developed multilingual reasoning benchmarks for Korean and Japanese language models, emphasizing cultural and linguistic diversity. • Built systematic evaluation leaderboard for cross-model performance comparison.	12.2024 – 02.2025
Samsung Medical Center <i>#Data Analysis Researcher</i> • Created LLM evaluation framework for medical documentation with comprehensive error taxonomies and quality standards. • Led federated learning model development for breast cancer prognosis prediction in Ministry of Trade, Industry, and Energy-funded project.	07.2023 – 03.2024

SELECTED PUBLICATIONS

** indicates equal contribution. Full publication list available at Google Scholar.*

Peer-Reviewed Publications

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

- Dasol Choi, Dongbin Na
- *AAAI Conference on Artificial Intelligence (AAAI), 2025*

Better Safe Than Sorry? Overreaction Problem of Vision Language Models in Visual Emergency Recognition

- Dasol Choi, Seunghyun Lee, Youngsook Song
- *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2026*

LLM-Based Medical Document Evaluation: Integrating Human Expert Insights

- Junhyuk Seo*, Dasol Choi*, Wonchul Cha, Taerim Kim
- *MedInfo, 2025 Best Student Paper Award*

Evaluation Framework of Large Language Models in Medical Documentation: Development and Usability Study

- Junhyuk Seo*, **Dasol Choi***, Wonchul Cha, Haanju Yoo, Namkee Oh, Yongjin Yi, Gyehwa Lee, Edward Choi, Taerim Kim
- *Journal of Medical Internet Research (JMIR)*, 2024

Under Review

When Cars Have Stereotypes: Auditing Demographic Bias in Objects from Text-to-Image Models

- **Dasol Choi**, Jihwan Lee, Minjae Lee, Minsuk Kahng

COMPASS: A Framework for Evaluating Organization-Specific Policy Alignment in LLMs

- **Dasol Choi***, DongGeon Lee*, Brigitta Jesica Kartono*, Helena Berndt, Haon Park, Hwanjo Yu, Minsuk Kahng

When Good Sounds Go Adversarial: Jailbreaking Audio-Language Models with Benign Inputs

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

Redefining Evaluation Standards: A Unified Framework for Evaluating the Korean Capabilities of Language Models

- Hanwool Lee*, **Dasol Choi***, Sooyong Kim, Ilgyun Jung, Sangwon Baek, Guijin Son, Inseon Hwang, Naeun Lee, Seunghyeok Hong

What Users Leave Unsaid: Under-Specified Queries Limit Vision-Language Models

- **Dasol Choi***, Guijin Son*, Hanwool Lee*, Minhyuk Kim, Hyunwoo Ko, Teabin Lim, Eungyeol Ahn, Jungwhan Kim, Seunghyeok Hong, Youngsook Song

PATENTS

- **Method and System for Automated Evaluation of a Conversational Language-Model Assistant Against Organization-Specific Policies**
EP Patent Application 25216794.5 Nov 18, 2025
- **Method and Apparatus for Evaluating Safety of an Artificial Intelligence Model**
KR Patent Application 10-2025-0148814 Oct 15, 2025

COMMUNITY INVOLVEMENT

HAE-RAE Open-source Research Community

05.2024 – Present

Researcher, Contributor to Korean LLM Research

- Led development of HAERA-E-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."