

DASOL CHOI

dasolchoi@yonsei.ac.kr | +82 10-2431-0366 | [Homepage](#) | [GitHub](#) | [LinkedIn](#) | [Google Scholar](#)

RESEARCH SUMMARY

AI safety and evaluation researcher for language and multimodal models. I design and run evaluations end-to-end, covering benchmark and red-teaming study design, execution, analysis, and public release. My work includes first-author publications at ACL, ECCV, AAAI, WACV, and LREC, and evaluation collaborations with Microsoft, BMW Group, and Korea AISI, among others.

EXPERIENCE

AIM Intelligence — Principal Researcher, Evaluation Team Lead

07.2025 – Present

- Led *Judgement Day: AI Red Team Arena*, an online multimodal red-teaming competition co-hosted with Korea AISI — ran the project end-to-end, including adversarial scenario design (2 tracks, 8 high-stakes scenarios), multimodal attack channels, and dynamic bounty scoring.
- Led *XL-SafetyBench*, a country-grounded cross-cultural LLM safety benchmark with 5,500 test cases across 10 country-language pairs, in collaboration with Microsoft, Korea AISI, and KT.
- Led the BMW Group collaboration on organization-specific LLM policy alignment, developing the *COMPASS* evaluation framework with 5,920 validated queries across 8 industry scenarios (ACL 2026); core method filed as an EP patent.
- Investigated benign-sounding audio jailbreaks (*WhisperInject*) in audio-language models in collaboration with LG Electronics.

SIONIC AI — ML/DL Research Intern

12.2024 – 02.2025

- Developed multilingual reasoning benchmarks for Korean and Japanese language models, emphasizing cultural and linguistic diversity.
- Built a systematic evaluation leaderboard for cross-model performance comparison.

Samsung Medical Center — Data Analysis Researcher

07.2023 – 03.2024

- Created an LLM evaluation framework for medical documentation with comprehensive error taxonomies and quality standards (JMIR 2024; Best Student Paper at MedInfo 2025).
- Led federated learning model development for breast cancer prognosis prediction in a Ministry of Trade, Industry, and Energy-funded project.

EDUCATION

Yonsei University, Seoul, South Korea

03.2024 – 02.2026

M.S. in Digital Analytics, College of Computing

GPA: 4.26/4.50

Kyung Hee University, Seoul, South Korea

03.2013 – 02.2019

B.A. in Japanese Language and Physical Education (Double Major)

GPA: 4.07/4.50

SELECTED PUBLICATIONS

Selected publications; full list on [Google Scholar](#).

(* equal contribution; † corresponding author)

Peer-Reviewed Publications

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
- *ACL 2026*

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

- Dasol Choi, Jihwan Lee, Minjae Lee, Minsuk Kahng
- *ECCV 2026*

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
- *ACL 2026 Findings*

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

- Dasol Choi, Dongbin Na
- *AAAI 2025*

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

- **Dasol Choi**, Seunghyun Lee, Youngsook Song
- WACV 2026

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
- LREC 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, Gye-hwa Lee, Edward Choi, Taerim Kim
- Journal of Medical Internet Research (JMIR), 2024

Pushing on Multilingual Reasoning Models with Language-Mixed Chain-of-Thought (co-author)

- Guijin Son, Donghun Yang, Hitesh Laxmichand Patel, Amit Agarwal, Hyunwoo Ko, Chanuk Lim, Srikant Panda, Minhyuk Kim, Nikunj Drolia, **Dasol Choi**, Kyong-Ha Lee, Youngjae Yu
- ICLR 2026

Under Review

XL-SafetyBench: A Country-Grounded Cross-Cultural Benchmark for LLM Safety and Cultural Sensitivity

- **Dasol Choi**[†], Eugenia Kim, Jaewon Noh, Sang Seo, Eunmi Kim, Myunggyo Oh, Yunjin Park, Brigitta Jesica Kartono, Josef Pichlmeier, Helena Berndt, Sai Krishna Mendu, Glenn Johannes Tungka, Özlem Gökçe, Suresh Gehlot, Katherine Pratt, Amanda Minnich, Haon Park[†]

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

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

When Context Flips, Safety Breaks: Diagnosing Brittle Safety in Aligned Language Models

- **Dasol Choi**^{*†}, Alex Kwon^{*}

EgoSafetyBench: A Diagnostic Egocentric Video Benchmark for Evaluating Embodied VLMs as Runtime Safety Guards

- Siddhant Panpatil, Arth Singh, Mijin Koo, Chaeyun Kim, Haon Park, **Dasol Choi**[†]

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

SERVICE & COMMUNITY INVOLVEMENT

HAE-RAE Open-source Research Community

05.2024 – Present

Researcher, Contributor to Korean LLM Research

- Led development of **HAERAE-Vision**, a Korean vision-language benchmark for culturally grounded multimodal evaluation.
- Developed the **Haerae-Evaluation-Toolkit (HRET)**, a unified evaluation framework for Korean LLM benchmarking with standardized APIs and an extensible architecture.

Seoul Forum on AI Safety & Security (SFASS) 2026, Organizing Member & Speaker

07.2026

- Co-organized the forum program, including the Red-Teaming Workshop and Live Red-Teaming Challenge.
- Talk: “Judgement Day: Insights from Multimodal Red-Teaming Challenge.”

AAAI 2026, 2027, Program Committee Member

- **Main Track** and **Alignment Track** (responsible AI, alignment, and safety)

AWARDS & ACHIEVEMENTS

- **Best Student Paper Award**, MedInfo 2025 08.2025
- **Google East Asia Student Travel Grant**, AAAI 2025 03.2025