CHAEYUN KIM

+82(10) 3691-3754 ◦ Gwanak-gu, Seoul
golddohyun@snu.ac.kr ◦ Google Scholar ◦ Github Page

RESEARCH INTERESTS

Vision-Language Compositional Understanding, Robustness and Reasoning of (M)LLMs, AI Safety & Alignment, Red Teaming & Adversarial Attacks, Evaluation and Benchmarks

EDUCATION

Seoul National University (SNU) **Seoul, Korea**

M.S. in Data Science Mar. 2023 – Feb. 2025

- Relevant Coursework: Machine Learning & Deep Learning I, II, Big Data & Knowledge Management Systems I, II, Computing for Data Science I, Scalable High-Performance Computing, Data Science Project, Machine Learning for Visual Understanding, Conversational AI & Dialogue Systems

B.S. in International Relations Mar. 2017 – Feb. 2022

- CGPA: 4.15/4.3, Graduated *Summa Cum Laude*
- Dean's List, Academic Scholarship for 5 semesters (2018-2020)

PUBLICATIONS

[Journal/Conference (including in-review)]

- Seongsu Ha*, Chaeyun Kim*, Donghwa Kim*, Junho Lee, Sangho Lee, Joonseok Lee**, "Finding NeMo: Negative-mined Mosaic Augmentation for Referring Image Segmentation," European Conference on Computer Vision **(ECCV 24)**. Springer, 2025. Available: https://arxiv.org/pdf/2411.01494
- Chaeyun Kim*, Seunghoon Yi*, Yejin Kim, Yohan Jo, Joonseok Lee, "Towards Motion-aware Referring Image Segmentation". **(AISTATS 26, accepted)**
- Chaeyun Kim, Yongtaek Lim, Kihyun Kim, Junghwan Kim, Minwoo Kim, "CAGE: A Framework for Culturally Adaptive Red-Teaming Benchmark Generation". **(ICLR 26, accepted)**
- Joonki Min*, Chaeyun Kim*, Hyungwook Choi, Yejin Kim, Kihyun Kim, Yohan Jo, Joonseok Lee, "Fine-Grained Multi-Image Object Hallucination Benchmark". **(CVPR 26, accepted)**
- Min-jae Jung, Yongtaek Lim, Chaeyun Kim, Junghwan Kim, Kihyun Kim, Minwoo Kim, "A Strategy-Response Multiplex Network Approach to Automated LLM Red Teaming" (ARR 26 Jan. under review)
- Chaeyun Kim*, Daeyoung Park*, Junghwan Kim, Jinyoung Jeong, Eunji Song, Minwoo Kim, "FinRED: An Expert-Guided Red-Teaming Benchmark for Financial LLM Safety" (KDD 26, under review)

[Books]

- Kim, S.B., **Kim, C.Y.,** et al, "The 4th Industrial Revolution and U.S.-China Power Politics: A Perspective from Information Politics," Social Criticism Academy, May 2020 (authored subsection of Chp 8), Link
- Kim, U.Y., **Kim, C.Y.,** et al, "2020 Civic Politics Yearbook: Theories and Practices of Community-based Education," Pureungil, Jan. 2021 (authored Chapter 3), Link

RESEARCH EXPERIENCE

AIM Intelligence **Seoul, Korea**

Physical/Multimodal AI Researcher **Jan 2026 – Present**

- Developing a universal **cross-embodiment action latent** framework that aligns human motion semantics with robot proprioceptive values to enforce robust Physical AI safety

SelectStar (DATUMO Inc.) **Seoul, Korea**

LLM Safety Researcher **May 2025 – Dec 2025**

[1] "**FinRED: Expert-Guided Red-Teaming Benchmark for Financial LLM Safety**" *(WWW 26, In review)*

- Developed the first comprehensive **financial domain red-teaming benchmark** in collaboration with **FSI (Financial Security Institute)**, featuring a 2-level taxonomy of financial risks
- Created an expert-validated finance-specific safety rubric demonstrating substantially higher alignment with human judgments compared to generic rubrics.

[2] "**STAR-Teaming: Strategy-Response Multiplex Network for Automated LLM Red Teaming**" *(ARR Oct. Submission)*

- Participated in automated LLM red teaming research (STAR-Teaming) by contributing to main experiments and paper writing.
- Responsible for refactoring and optimizing the research-level code for commercial product integration, bridging the gap between research and deployment.

[3] "CAGE : A Culturally Adaptive Red-Teaming Benchmark for LLMs" *(ICLR 2026, **accepted**)*

- Pioneered the **CAGE framework**, which uses flexible **'Semantic Molds'** as content guidelines to generate diverse and realistic adversarial prompts tailored to any cultural and legal context.
- Built **KorSET**, the first large-scale Korean safety benchmark, and demonstrated the framework's high **efficiency and scalability** (e.g., generating over 1,250 prompts in a single run).
- Validated the framework's generalizability by extending it to low-resource languages (e.g. Khmer), quantitatively proving its superior efficacy over translation-based methods.

SNU Visual Information Processing Lab **Seoul, Korea**

Graduate Researcher **Sep. 2024 – Feb. 2025**

[1] "Fine-Grained Multi-Image Object Hallucination Benchmark" *(CVPR 2026, **accepted**)*

- Developed a systematic evaluation framework for multi-image object hallucination across four tasks (existence, counting, attribute, position) and three reasoning patterns (comprehensive, comparative, selective).
- Formulated four controllable adversarial pressures; **revealed that** hallucination stems from **integration failure** (error to bind attributes to specific image location), rather than simple perceptual errors.
- **Conducted Large-Scale Empirical Analysis of 30+ Models**

[2] "Towards Motion-aware Referring Image Segmentation" *(AISTATS 2026, **accepted**)*

- **Exposed a critical failure in state-of-the-art RIS models**: their significant underperformance on motion-centric queries due to an over-reliance on static visual cues.
- Proposed a unified framework that combines **text**(hard positive) **augmentation** strategy to supplement under-represented motion-centric phrases with novel **Multimodal Radial Contrastive Loss (MRaCL)** for context-aware semantic alignment.
- Introduced **M-Bench**, a novel benchmark for evaluating motion comprehension.

Graduate Research Assistant (Advisor: Prof. Joonseok Lee) Jan. 2023 – Aug. 2024

- Developed negative-mined mosaic augmentation for referring image segmentation, improving model comprehension of visual-linguistic cues in ambiguous scenarios; **co-first author (ECCV 2024)**.

PROJECT EXPERIENCE

Selectstar (DATUMO Inc.) **Seoul, Korea**

"Reasoning-based LLM Safety Guardrail Development & Commercialization"

Lead AI Safety Researcher Aug. 2025 – Present

- Led development of Korean reasoning-based safety guardrail, overcoming limitations of rule-based judges (poor explainability, performance degradation on policy changes)
- Designed multi-task GRPO training pipeline using 4,500 high-quality examples, achieving **91% accuracy** on internal benchmarks, outperforming SFT baselines; Commercialized solution with Woori Bank

"Collaboration with Samsung SDS for AI Guardrail Safety Enhancement"

Lead AI Safety Researcher & Engineer Sep. 2025 – Nov. 2025

- Led construction of 10,000 adversarial query-response pairs to strengthen Korean safety judgment capabilities of SDS guardrail system, applying CAGE framework for base harmful query generation.
- **Developed 5,000 adversarial attack scenarios using diverse jailbreaking** techniques, including:
  - Instruction Indirection (**enhanced STAR-Teaming** with advanced scorer/judge prompts and interpretable sampling)
  - Prompt Injection, Suffix/Prefix attacks, and Generation Glide methods (cipher, encoding obfuscation, etc).

"WBL AI Foundation Model Safety Consortium (SKT)"

Tech Lead & Engineer Sep. 2025 – Dec. 2025

- Led construction of safety/red-teaming dataset in SKT's AI foundation model development project.
- Built multi-turn focused harmless datasets and RL-ready triplet format (harmful query, safe response, unsafe response) for robust post-training alignment.

Seoul National University - Graduate School of Data Science **Seoul, Korea**

"ECOLENS: Real-time Object Recognition for Recycling Guidance"

Machine Learning Engineer & Back-end Developer Sep. 2023 – Dec. 2023

- Developed a system providing user-customized text and speech recycling instructions for objects based on photos and information about available bins.
- Constructed the application backend using the Flask framework; implemented a MongoDB database for efficient data management.

"Data-Effective Semantic Segmentation on 4D Point Cloud Videos"

Team Member Mar. 2023 – Jun. 2023

- Designed a data-loading pipeline optimized for processing large-scale datasets.
- Proposed a framework combining active learning and contrastive learning to efficiently utilize feature representations in irregular and unordered point clouds.

"GSDS PleaseGraduate: Graduation Requirement Checker and Course Recommender"

Team Leader Oct. 2022 – Jan. 2023

- Spearheaded the development of a customized graduation requirement checking system for the Graduate School of Data Science.
- Designed user interface and led front-end development using HTML, CSS, jQuery, and Django; managed the data preprocessing pipeline.

EXCHANGE PROGRAMS

University of Washington, Seattle **Seattle, WA**

Henry M. Jackson School of International Studies Mar. 2020 – Jun. 2020

Vrije Universiteit Brussel **Brussels, Belgium**

Institute for European Studies Jul. 2019 – Aug. 2019

- Participated in SNU in the EU program; took a lecture on the European Union's international mission.

TEACHING ASSISTANTSHIPS

"Core Computing: Thinking with Computers – Data Structure and Algorithms," SNU (TA)	Fall. 2024
"Principles and Applications of Data Science," SNU (TA)	Fall. 2023
"Machine Learning and Deep Learning I," SNU (TA)	Fall. 2023
"ExploreCSR Program (with Google)," SNU (tutor, mentor)	Summer. 2023
"Basic Computing: First Adventures in Computing," SNU (tutor)	Fall. 2022

SCHOLARSHIPS & AWARDS

Graduate

Academic Scholarship : Full Tuition Exemption, Seoul National University — Sep.2023
Data Science Graduate School Scholarship, Seoul National University — May.2023
4th Stage BK21 Scholarship, Seoul National University — Mar.2023

Undergraduate

International Scholarship, SNU Office of International Affairs — Sep.2020
Dean's List, Academic Scholarship, Seoul National University — Sep.2018 – Dec.2020
Global Leadership Award, Seoul National University — Feb.2022
1st Place, Division of Policy Strategy, Future Aerospace Conference — Oct.2021
1st Place, Undergraduate Social Science Research Grant, SNU — Feb.2020
3rd Place, KOICA 16th International Development Cooperation Paper Competition — May.2019

LEADERSHIP

Vice President, SNU Graduate School of Data Science Student Association — Aug. 2023 – Feb. 2024

- Served as the cohort's deputy leader and an active member, representing the Data Science department.
- Organized various on-campus events, bi-weekly student seminars, and community-building activities in collaboration with school administrators.

ACADEMIC ACTIVITIES

SNU-HIT International Joint Seminar — Sep. 2017 – Oct. 2017, Sep. 2019 – Oct. 2019

- Participated in the international exchange seminar between Seoul National University Department of Political Science & International Relations and Hitotsubashi University Faculty of Law
  - 2017: Presented and led discussions on the "Korea-Japan Security Cooperation Plan" and "General Security of Military Information Agreement (GSOMIA) and Japan's Constitutional Amendment."
  - 2019: Hosted seminars in the History Department; facilitated discussions on the Supreme Court's ruling on forced labor, individual claim issues, and strategies to improve the Korea-Japan relations."

12th Asian Future Political Leaders' Association (AFPLA) — May 2018 – Aug. 2018

- Served as Politics Division Session Leader; presented and led discussions on "Definition, Importance, and the Evolution of 'Citizen' and 'Citizen Culture' in Past, Present, and Future Contexts."

PROFICIENCY IN SKILLS

Programming: Python, R, SQL, C/C++
Frameworks & Libraries: PyTorch, TensorFlow, OpenCV, NumPy, Pandas, Sckit-learn, Seaborn, Flask
Languages: Korean (native), English (full-professional fluency), Mandarin Chinese, Spanish (intermediate)