# Dongjun Kim

LinkedIn | GitHub | Website | Email

LLM Researcher specializing in AI Safety through Mechanistic Interpretability and Model Evaluation.

## Education

**Korea University** — Seoul, South Korea
*Master of Science in Computer Science* — *Expected Feb 2026*
- Advisor: Prof. Heuiseok Lim
- Research Focus: LLM Evaluation, Mechanistic Interpretability, and AI Safety

**University of South Florida** — Tampa, FL
*Bachelor of Science in Computer Science (Cum Laude)* — *May 2023*

## Publications

**Benchmark Profiling: Mechanistic Diagnosis of LLM Benchmarks**
**Kim, D.**, Shim, G., Chun, Y. C., Kim, M., Park, C., & Lim, H.
**Proceedings of EMNLP 2025** (Oral Presentation)

**KoLEG: On-the-Fly Korean Legal Knowledge Editing with Continuous Retrieval**
Seo, J., Jung, D., Lee, J., Chun, Y. C., **Kim, D.**, Ryu, H., Shin, D., & Lim, H.
**Findings of EMNLP 2025**

**MMA-ASIA: A Multilingual and Multimodal Alignment Framework for Culturally-Grounded Evaluation**
Zheng W., ..., **Kim, D.**, ..., & Chen, N. F.
**arXiv Preprint, 2025,** ICLR 2026 Under Review

**A Self-Improving Leaderboard for Robust and Evolving Natural Language Processing (NLP) Evaluation**
Park, C., Moon, H., **Kim, D.**, Lee, S., Seo, J., Eo, S., Lim, H.
**arXiv Preprint, 2025**

**Exploring Coding Spot: Understanding Parametric Contributions to LLM Coding Performance**
**Kim, D.**, Kim, M., Chun, Y. C., Park, C., & Lim, H.
**arXiv Preprint, 2024,** EACL 2026 Under Review

**Computational Comprehension**
**Kim, D.**, Shim, G., Kim, M., Park, C., & Lim, H.
ICT Express 2026 Under Review

**Exploring Inherent Biases in LLMs within Korean Social Context**
Lee, S., **Kim, D.**, Jung, D., Park, C., & Lim, H.
**NAACL 2024 Student Research Workshop**

**CitySEIRCast: An Agent-Based City Digital Twin for Pandemic Analysis**
Bilal, S., ... **Kim, D.**, ... & Michael, E.
**Complex & Intelligent Systems, 2024**

## Projects

**Independent AI Foundation Model Project (WBL)** — *NC AI Consortium*
- Implemented an **evaluation framework** (50+ benchmarks) covering reasoning, safety, and robustness with reproducible pipelines and CI triggers.
- Designed a unified metric layer and *Weights & Biases* dashboards for time series tracking, regression checks, and cross model slice analysis.
- Conducted pre and post training analyses on successive checkpoints, identified failure modes, and provided recommendations used for data and recipe updates.
- Introduced contamination checks (de duplication and overlap scans) and standardized runbooks to support fair and comparable evaluations across systems.

**KoLEG: Korean Legal Knowledge Editing** *Collaboration with KT Corporation*

- Co developed an **on the fly legal knowledge editing framework** with continuous retrieval and timestamp aware sequential updates (team of 8).
- Led crawling and construction of the Korean Legislative Amendment dataset, aligning implementation periods and applying high precision filtering.
- Contributed mechanistic interpretability analyses on locality and generalization in edited knowledge.
- Built expert evaluation demos and protocols for human assessment, enabling attorney review and iterative error analysis.

**KULLM 3 & KULLM R & Ko Gemma Model Training** *NLP&AI Lab*

- Contributed on **post training** for a team of 10, including instruction tuning, training framework, and multilingual and code switch datasets.
- Curated code and math corpora, established quality gates, and ran capability evaluations for coding and math.
- **KULLM Reasoning:** implemented reinforcement learning with custom reward functions and **GRPO** and applied **adaptive response length** (concise on easier tasks, deeper chains on harder ones).
- Observed improved Korean reasoning and math accuracy versus a Qwen3 reasoning baseline while reducing unnecessary verbosity in internal evaluations.

**Self Improving Leaderboard** *NLP&AI Lab*

- Implemented daily crawlers across multiple news categories, **real time QA generation, and automated multi LLM evaluation** on daily refreshed data.
- Launched a live leaderboard with time aware ranking and quarterly stability and volatility metrics to track consistency over time.
- Maintained scheduling, monitoring, and data hygiene to support regular refreshes and clear longitudinal comparisons.

## Experience

**NLP & AI Researcher** Jan 2024 – Feb 2026

*NLP&AI Lab, Korea University* *Seoul, South Korea*

- **Mechanistic Interpretability:** Designed Benchmark Profiling to probe layer and module behavior using gradient-based importance and targeted ablations, with config-driven, seed-controlled experiment suites and reproducible analysis toolchains.
- **SAE And Feature Steering:** Trained sparse autoencoders on residual and MLP activation streams to learn disentangled feature dictionaries, mapped features to behaviors with intervention tests, and applied feature steering during decoding using calibrated coefficients and sparsity constraints, with closed loop evaluation of causal effects using shims and ablations.
- **Evaluation Engineering & Automation:** Codified evaluation-as-code with deterministic runs, regression gates, and time series tracking, introduced unified metric registries, slice-based reporting, and single click reruns bound to code and dataset snapshots.
- **Data Systems:** Stood up pipelines for multilingual and code switch corpora and code and math curation, implemented contamination and overlap scans, dataset manifests with checksums, and audit-ready metadata for fair, replayable training and evaluation.
- **Demo & Evaluation UX:** Developed side-by-side streaming comparison harnesses on vLLM backends with think and answer separation, generation-time logging, and per-turn metrics to accelerate triage and stakeholder reviews.
- **MLOps & HPC Administration:** Operated lab infrastructure with 40+ NVIDIA A100 and H100 GPUs for 30+ researchers, including Slurm job templates and queues, Docker based reproducible environments, artifact and version hygiene, and uptime monitoring for multi node training and high throughput evaluation.
- **Safety:** Built contamination checks and evaluation presets for fair comparisons, added refusal and jailbreak test suites, toxicity and PII filters, prompt injection probes, and policy aligned refusal and redirection auditing, with dashboards for risk exposure, traceability, and regression alerts.

**RLHF Data Trainer** Mar 2023 – Jan 2024

*Scale AI* *San Francisco, CA (Remote)*

- Contributed to the alignment and safety tuning for flagship foundation models from leading AI labs, including projects for Google (Flamingo, Bard), Meta (Llama series), and OpenAI.
- Crafted expert-level preference data for RLHF and instruction tuning, utilizing advanced platforms including OpenAI's proprietary Feather tool to improve model safety, factuality, and reasoning.

**AR/VR Software Engineering Intern** May 2023 – Nov 2023

*Simacro* *Cambridge, MA*

- Developed a Unity-based VR/AR application to transform static P&IDs into interactive digital twins, streamlining operations for industrial clients including Hyundai Oil Bank.
- Built a computer vision API (Python/C#) for symbol detection that automated 80% of manual labeling with 95% accuracy, accelerating the P&ID digitization pipeline.
- Integrated Virnect's image tracking SDK into the Unity app to create robust, precisely anchored AR overlays of diagrams on physical machinery in industrial environments.

**High Performance Computing Intern** Aug 2022 – May 2023

*Dr. Edwin Michael's Lab, USF* *Tampa, FL*

- Contributed to a city-scale digital twin (CitySEIRCast) for pandemic forecasting, handling large datasets via parallel and distributed computing (MPI, OpenMP, CUDA).
- Implemented data pipelines in Python, integrating `NumPy`, `Pandas`, and SQL to manage simulation I/O.
- Enhanced simulation performance by optimizing C++ and Python code for HPC clusters, enabling faster data processing.

**Mixed Reality Research Assistant** Jan 2022 – May 2023

*USF Mixed Reality Lab* *Tampa, FL*

- Created an automatic room mapping method for Mixed Reality, reducing manual mapping time by 40%.
- Published research on novel data collection and modeling techniques in MR at IPMV 2023.