Dongjun Kim

LinkedIn | GitHub | Website | Email

LLM Engineer specializing in AI Safety through Mechanistic Interpretability and Model Evaluation.
Seeking an LLM Engineer position as a 전문연구요원 (병역특례) to advance robust and interpretable AI systems.

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

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

Towards Computational Comprehension:
A Non-Anthropocentric Framework for Evaluating LLM Understanding
Kim, D., Shim, G., Kim, M., Park, C., & Lim, H.
ICT Express 2026 Under Review

From Snapshot to Stram:
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

Enhancing Automatic Term Extraction with Large Language Models via Syntactic Retrieval
Chun, Y., Kim, M., **Kim, D.**, Park, C., & Lim, H.
Findings of ACL 2025

KITE: A Benchmark for Evaluating Korean Instruction-Following Abilities in Large Language Models
Kim, D., Park, C., Park, C., & Lim, H.
arXiv Preprint, 2025 IEEE Access 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) *Collaboration with NC AI, ETRI (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, Korea University*

- 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, Korea University*

- 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.

LLM Lesion Brain Mapping *NLP&AI Lab, Korea University*

- Studied how targeted parameter perturbations relate to language deficits in LLMs across syntax, semantics, and retrieval behaviors.
- Explored correspondences between internal subspaces and neuro linguistic functionality with layer and module level analysis.
- Prototyped a **lesion based causal tracing** method linking internal changes to observable behaviors with reproducible analysis tooling.

Technical Skills

Core ML & Deep Learning: PyTorch, JAX, Hugging Face (Transformers, Accelerate), Polars

LLM Training & Scaling: DeepSpeed, FSDP, Megatron-LM, PEFT (LoRA, QLoRA)

Advanced RL & Evaluation: GRPO, GSPO, RLHF (TRL), LM-Eval-Harness, EvalChemy

LLM Inference Optimization: vLLM, SGLang, TensorRT-LLM, KV Cache Optimization, Quantization

Infrastructure & MLOps: CUDA, Triton, Docker, Kubernetes, Terraform, AWS (SageMaker), W&B

Experience

NLP & AI Researcher Jan 2024 – Feb 2026
NLP&AI Lab, Korea University *Seoul, South Korea*

- **Evaluation Engineering & Automation:** Codified evaluation-as-code, enabling deterministic runs, regression gates, and time series tracking; introduced unified metric registries, slice-based reporting, and single click reruns tied to code and dataset snapshots.
- **Data Systems:** Developed automated pipelines for multilingual and code switch corpora, as well as code and math dataset curation; implemented contamination and overlap scans, generated dataset manifests with checksums, and established audit-ready metadata for fair, replayable training and evaluation.
- **SAE And Feature Steering:** Trained sparse autoencoders on residual and MLP activation streams to learn disentangled feature dictionaries, mapped features to behaviors via intervention tests, and applied feature steering during decoding with calibrated coefficients and sparsity constraints, conducting closed loop causal evaluations with shims and ablations.
- **MLOps & GPU Infrastructure Management:** Managed lab infrastructure with 40+ NVIDIA A100 and H100 GPUs, supporting 30+ researchers; maintained Kubernetes clusters and Docker-based reproducible environments, implemented artifact and version control hygiene, and monitored uptime for multi-node training and high throughput evaluation.
- **Demo & Evaluation UX:** Created side-by-side streaming comparison harnesses on vLLM backends with think and answer separation, generation-time logging, and per-turn metrics, expediting triage and stakeholder reviews.
- **Safety:** Engineered contamination checks and evaluation presets for fair comparisons, developed refusal and jailbreak test suites, implemented toxicity and PII filters, prompt injection probes, and policy-aligned refusal and redirection auditing; designed dashboards for risk exposure, traceability, and regression alerts.

RLHF Data Trainer Mar 2023 – Jan 2024
Scale AI *San Francisco, CA (Remote)*

- Supported alignment and safety tuning for flagship foundation models at leading AI labs, including Google (Flamingo, Bard), Meta (Llama series), and OpenAI.
- Produced expert-level preference data for RLHF and instruction tuning, leveraging advanced platforms such as OpenAI's Feather tool to enhance model safety, factuality, and reasoning.
- Curated and annotated diverse data modalities, including image reasoning, coding solutions, mathematical queries, multi-turn conversations, and safety-critical content (PII and harmful data) to ensure robust evaluation and fine-tuning.

AR/VR Software Engineering Intern May 2023 – Nov 2023
Simacro *Cambridge, MA*

- Developed Unity-based VR/AR applications for transforming static P&IDs into interactive digital twins, streamlining operations for industrial clients including Hyundai Oil Bank.
- Built computer vision API (Python/C#) for symbol detection to automate 80% of manual labeling with 95% accuracy, accelerating the P&ID digitization workflow.
- Integrated Virnect's image tracking SDK into Unity apps, enabling robust, anchored AR overlays of diagrams onto physical industrial machinery.

High Performance Computing Intern Aug 2022 – May 2023
Dr. Edwin Michael's Lab, USF *Tampa, FL*

- Built digital twin platform (CitySEIRCast) for city-scale pandemic forecasting, processing large datasets with parallel and distributed computing (MPI, OpenMP, CUDA).
- Developed data pipelines in Python using `NumPy`, `Pandas`, and SQL to manage simulation I/O.
- Optimized C++ and Python simulation code for HPC clusters to enable faster data processing.

Mixed Reality Research Assistant Jan 2022 – May 2023
USF Mixed Reality Lab *Tampa, FL*

- Designed 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.