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 (신규 편입) 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

LangSAE Editing: Improving Multilingual Information Retrieval via Post-hoc Language Identity Removal
Kim, D., Yoon, J., Park, C., & Lim, H.
arXiv Preprint, 2026 ACL 2026 Under Review

CrossDocVQA: A Benchmark for Multi-Hop Cross-Document Visual Question Answering
Shim, G., Shin, J, **Kim, D.**, Park, C., & Lim, H.
ACL 2026 Under Review

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

Towards Computational Comprehension:
A Non-Anthropocentric Framework for Evaluating LLM Understanding
Kim, D., Shim, G., Kim, M., Park, C., & Lim, H.
arXiv Preprint, 2025 ACL 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

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

WBL Independent AI Foundation Model Project (VAETKI) *Collab with NC AI, ETRI*

- Owned engineering of the large-scale evaluation pipeline, standardizing **50+ benchmarks** into a single reproducible runtime with automated regression tracking across checkpoints.
- Implemented end-to-end checkpoint-to-evaluation automation, triggering evaluations immediately after checkpoint creation to shorten the training feedback loop.
- Integrated **vLLM** to accelerate inference throughput and improve evaluation turnaround time at scale.
- Built *W&B* dashboards and **W&B Weave** trace analysis to debug reasoning failures, visualize inference traces, and support slice-level comparisons.
- Implemented **data-mixture contribution analysis** to quantify which dataset combinations drove metric gains, translating results into actionable data and recipe updates.
- Introduced contamination defenses and runbooks, including deduplication and overlap scans, to support fair and comparable evaluations.

KULLM Reasoning Model Training

`nobrand/KULLM-R` *NLP&AI Lab, Korea University*

- Implemented **GRPO** reinforcement learning in **VERL** with multi-rollout group scoring, enabling critic-free policy optimization for Korean reasoning tasks.
- Designed **verifiable custom reward functions** optimizing correctness, Korean final-answer consistency, and an **adaptive length penalty** to reduce verbosity on easy problems while preserving depth on hard ones.
- Tuned reward weights and RL hyperparameters (e.g., KL coefficient, rollout settings, max response length) to balance accuracy, compute cost, and response quality.
- Ran iterative evaluations on Korean math and reasoning benchmarks (Pass@1 and length diagnostics), using results to refine reward shaping and training stability.

KULLM 3 & Ko Gemma Model Training

`nlpai-lab/KULLM3-20240604`, `nlpai-lab/ko-gemma-7b-v1` *NLP&AI Lab, Korea University*

- Contributed to **post-training** in a 10-person team, including instruction tuning, training framework improvements, and multilingual plus code-switch dataset development.
- Curated coding and math corpora, established quality gates, and ran capability evaluations for coding and mathematics.

KULLM 3 & Ko Gemma Model Training

`nlpai-lab/KULLM3-20240604`, `nlpai-lab/ko-gemma-7b-v1` *NLP&AI Lab, Korea University*

- Contributed to **post-training** in a 10-person team, including instruction tuning, training framework improvements, and multilingual plus code-switch dataset development.
- Curated coding and math corpora, established quality gates, and ran capability evaluations for coding and mathematics.

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.

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*

- **Research:** Co-authored 10+ papers including top-tier ACL venues (ACL, EMNLP, NAACL).
- **Evaluation:** Drove benchmark design and benchmark profiling analyses, established reusable evaluation recipes, regression policies, and dataset, prompt versioning used across lab and consortium model iterations (incl. WBL).
- **LLM Training:** Supported Korean LLM post-training and reasoning improvements, including GRPO in VERL for KULLM Reasoning, and instruction tuning plus data curation for KULLM3 and Ko Gemma.
- **Mechanistic Interpretability:** Developed SAE-based feature discovery, feature steering, and causal intervention workflows to diagnose behaviors, quantify feature effects, and connect internal signals to evaluation outcomes.
- **Inference Optimizations:** Improved serving and evaluation throughput with vLLM and cache-aware decoding settings, balancing latency, determinism, and large-batch benchmarking stability.
- **Agent Systems:** Prototyped RAG and agentic pipelines for domain tasks, including retrieval calibration, tool routing, and trace-based debugging to localize failure modes.
- **MLOps & GPU Infrastructure Management:** Operated shared infrastructure (40+ A100/H100 GPUs) for 30+ researchers, maintaining Kubernetes, Docker environments, experiment tracking, artifact hygiene, and reproducible runbooks.
- **Demo & Safety:** Built interactive evaluation demos and safety audit harnesses, implemented jailbreak and refusal checks, PII and toxicity screening, and regression monitoring for risk-sensitive deployments.

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

- Generated high-difficulty **preference and instruction data** for alignment workflows (PPO, DPO), with emphasis on **coding and mathematics** and other reasoning-heavy domains.
- Performed **QA and peer review** to maintain rubric consistency and data reliability across trainers, delivering structured feedback using client tooling (e.g., OpenAI Feather).
- Collaborated directly with client engineers to identify emerging alignment gaps and rapidly translate requirements into updated production guidelines and examples.
- Covered multi-modal and safety-sensitive scenarios (image reasoning, multi-turn conversations, PII, harmful content) to support robust fine-tuning and evaluation.

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.