Jaehee Kim

Ph.D. Student, Seoul National University

jaehee_kim@snu.ac.kr | KimJaehee0725.github.io | KimJaehee0725 | jaehee-kim-hoyang | Google Scholar

Research Summary

Ph.D. student working on **information retrieval**, **LLM evaluation**, and **LLM post-training**. First-author papers on memory-efficient dense retriever training (NeurIPS 2024) and privacy for black-box LLM APIs (ICML 2026); co-authored work on LLM-as-a-judge (EMNLP 2025) and RLVR. As project lead on industry and government-funded projects, built retrieval, evaluation, and multi-agent systems end to end.

Education

Seoul National University — *Seoul, Korea*
Ph.D. Student, Industrial Engineering (Advisor: Prof. Pilsung Kang) — *Sep 2024 - Aug 2028 (Expected)*

Korea University — *Seoul, Korea*
M.S., Industrial and Management Engineering (Advisor: Prof. Pilsung Kang) — *Mar 2022 - Aug 2024*

Sungkyunkwan University — *Seoul, Korea*
B.A., Consumer Science — *Mar 2014 - Feb 2021*

Publications

Conference Papers

[C5] **Jaehee Kim**, Pilsung Kang. AlienLM: Alienization of Language for API-Boundary Privacy in Black-Box LLMs. In *International Conference on Machine Learning* (**ICML**), 2026. [paper]

▷ *First text-only privacy method for black-box LLM APIs that applies to general tasks, using only standard fine-tuning APIs. Retains >81% of plaintext-oracle performance across 4 backbones and 7 benchmarks; recovery attacks reconstruct <0.22% of tokens.*

[C4] Yukyung Lee, JoongHoon Kim, **Jaehee Kim**, Hyowon Cho, Jaewook Kang, Pilsung Kang, Najoung Kim. CheckEval: A Reliable LLM-as-a-Judge Framework for Evaluating Text Generation Using Checklists. In *Empirical Methods in Natural Language Processing* (**EMNLP**), 2025. [paper]

▷ *Checklist-based LLM-as-a-judge; correlates strongly with humans and raises agreement across 12 evaluator models by 0.45.*

[C3] Joonwon Jang, **Jaehee Kim**, Wonbin Kweon, Seonghyeon Lee, Hwanjo Yu. Verbosity-Aware Rationale Reduction: Sentence-Level Rationale Reduction for Efficient and Effective Reasoning. In *Findings of the Association for Computational Linguistics* (**ACL Findings**), 2025. [paper]

[C2] **Jaehee Kim**, Yukyung Lee, Pilsung Kang. A Gradient Accumulation Method for Dense Retriever under Memory Constraint. In *Neural Information Processing Systems* (**NeurIPS**), 2024. [paper]

▷ *With 11GB of memory, beats 80GB DPR training on 18 of 24 metrics (up to +4.9) across 5 IR benchmarks; 34% faster than GradCache.*

[C1] Yookyung Kho, **Jaehee Kim**, Pilsung Kang. Boosting Prompt-Based Self-Training With Mapping-Free Automatic Verbalizer for Multi-Class Classification. In *Findings of Empirical Methods in Natural Language Processing* (**EMNLP Findings**), 2023. [paper]

Under Review

[P2] Keonwoo Kim, **Jaehee Kim**, Dongyoon Han, Haanju Yoo. Belief-Shift Reward: Redistributing RLVR Credit with Policy-Internal Target Support. *Under review*, 2026.

▷ *Reward transformation for RLVR; improves the math macro-average in all 8 model–objective combinations (up to +3.7pp) and agentic search by +4.6pp (NQ) / +3.9pp (HotpotQA).*

[P1] Jihun Nam, **Jaehee Kim**, Pilsung Kang. A Domain-Specialized Framework for Semantic Embedding and Retrieval of Semiconductor Equipment Logs. *Under review*, 2026.

Workshop Papers

[W2] Yukyung Lee, **Jaehee Kim**, Doyoon Kim, Yookyung Kho, Younsun Kim, Pilsung Kang. Painsight: An Extendable Opinion

Mining Framework for Detecting Pain Points Based on Online Customer Reviews. *WASSA Workshop at ACL*, 2023. **(Oral)**

[W1] Jaehyuk Heo, YongGi Jeong, Sunwoo Kim, **Jaehee Kim**, Pilsung Kang. REVECA: Rich Encoder-decoder framework for Video Event CAptioner. *LOVEU Challenge Workshop at CVPR*, 2022. **(3rd place)**

Domestic Journals (Korean)

[J1] Siyeol Sung, **Jaehee Kim**, Jaewon Cheon, Junyoung Son, Pilsung Kang. FindCoResearcher: A Retrieval-based Collaborative Researcher Recommendation System. *Journal of the Korean Institute of Industrial Engineers*, 52(1).

Research Projects

Autonomous Operation Agents for Semiconductor Wafer Manufacturing (Project Lead) *Apr 2026 - Present*
Korea Evaluation Institute of Industrial Technology (KEIT)

- Led the SNU team in a 5-organization industry–academia consortium (3 companies, 2 universities; 2026–2028).
- Designing the multi-agent framework and orchestrator for monitoring, prediction, scenario, and control agents.
- Building a virtual wafer-process simulator and logging system to test orchestration before testbed deployment.

Hierarchy-Aware Retrieval over Insurance Policy Documents (Project Lead) *Mar 2025 - Mar 2026*
Samsung Fire & Marine Insurance

- Built an LLM parser (PDF $\rightarrow$ hierarchical JSON) recovering clause hierarchy, tables, and cross-references.
- Generated 24.6K synthetic customer queries; filtered the test split with an LLM relevance verifier for reliable offline evaluation.
- Trained hierarchy-aware retrievers (Qwen3-Embedding 0.6B–8B): Top-10 accuracy 41–64% $\rightarrow \geq$90%.
- Served retrievers with vLLM for up to 9.7$\times$ higher throughput (8B: 8 $\rightarrow$ 77 QPS on A100) behind a sub-second search UI.

Worker-Friendly AI Agents for Autonomous Manufacturing, Phase 1 (Project Lead) *Apr 2025 - Dec 2025*
Institute for Information & Communications Technology Planning & Evaluation (IITP)

- Led the host-institution (SNU) team in a 7-organization consortium (5 universities, 2 companies).
- Built an LLM orchestrator (intent $\rightarrow$ RAG instruction retrieval $\rightarrow$ task decomposition $\rightarrow$ constraint check) over 3 agents.
- 99% intent-recognition accuracy on 100 in-house worker requests; built 20 validation scenarios (5 industries $\times$ 4 workflows).

Researcher Matching System for Industry–Academia Collaboration (Project Lead) *Dec 2024 - Jul 2025*
College of Engineering, Seoul National University

- Synthesized 530K training queries over a 52K-document corpus via LLM augmentation (query–doc overlap 23.6% $\rightarrow$ 7.4%).
- Fine-tuned and weight-merged BGE-M3 retrievers: avg. Top-5 accuracy 17.5% $\rightarrow$ 22.0%, above BM25, mE5, KURE.
- Designed a lab-level evaluation protocol aligned with the matching objective, and delivered a Gradio demo.

LLM Evaluation Framework for the Korean Financial Domain (Project Lead) *Nov 2023 - Aug 2024*
KakaoBank

- Designed a 5-topic, 16-ability framework (MCQA, free-form, LLM-as-a-judge); benchmarked 21 LLMs from 7 providers.
- Built intent-recognition benchmarks from real customer inquiries (2.5K items); controlled MCQA position bias.

Customer Review Retrieval and Trend Analysis, Stage 2 (Project Lead) *Dec 2022 - Nov 2023*
LG Electronics

- Framed the client's problem as retrieval and clustering tasks; built the end-to-end pipeline and visualization tools.
- Trained a phrase-level contrastive retriever on $\sim$400K reviews: accuracy 0.39 $\rightarrow$ 0.59 vs. sentence-level.
- Built a trend-analysis framework (clustering and keyword graphs) that surfaced event-driven shifts in customer interest.

Technical Skills

- **Frameworks:** Python, PyTorch, Hugging Face, vLLM, Docker, Linux, Gradio, WandB
- **Scale:** multi-node RL post-training (GRPO-family) on up to 32 nodes $\times$ 8 H100 GPUs
- **Evaluation:** offline retrieval metrics (Top-k accuracy, MRR), LLM-as-a-judge, benchmark and test-set construction

Professional Service

- **Conference Reviewer:** ICLR (2026), ICML (2026), NeurIPS (2026)
- **Workshop Reviewer:** LASS Workshop at CIKM (2026)