HanJun Cho

AI Research Engineer · Seoul, South Korea

gkswns0531@gmail.com · GitHub

WEBSITE

hanjun.alphajo.ai

SUMMARY

Engineer specializing in Retrieval-Augmented Generation (RAG), large-scale LLM inference optimization, and search relevance. Experience building and operating multi-regional inference servers handling ~30M monthly requests. Contributor to NVIDIA TensorRT-LLM(Qwen3 Dense & MoE support) and vLLM. Research background in retrieval and numerical reasoning with top-tier conference publications (ACL & TACL). Strong track record of measurable impact in latency reduction, accuracy improvement, and system reliability.

EXPERIENCE

Allganize Korea - AI Research Engineer

Seoul, South Korea · Sep 2024 - Present

- **Multi-tenant agentic AI platform** for financial institution, deployed as a hybrid on-premise architecture (network-segregated). (installation & ops)
- **Single-tenant RAG** on **AWS EKS** (K8s) for enterprise client: **300 RPS** in production, validated up to 10K RPS in load tests. (deploy & ops)
- **GraphRAG** over interconnected enterprise documents for portfolio-level strategic queries. (deploy)
- **Multi-Region** embedding inference servers across KR/US/JP, **30M** req/month. (deploy & ops)
- **TensorRT** Engine & **Triton Server**-based priority/distributed scheduling system for inference serving: (deploy & ops)
- **Prometheus** & **Grafana** based monitoring system (deploy & ops)
- Designed strategies for **ambiguous queries** and **factual-conflict resolution** across document versions.
- Developed framework to automatically generate high-quality RAG evaluation datasets from client's documents (**RARE**).

OPEN-SOURCE CONTRIBUTOR

- **NVIDIA TensorRT-LLM** - Merged PRs (#5650, #6470) adding/expanding support for Qwen3 Dense & MoE engines.
- **vLLM** - Merged PR (#35849) fixing FP8 / NVFP4 quantization bug for sequence classification models (Qwen3 Reranker).

PROFESSIONAL SERVICE

- **Reviewer**, *Conference on Neural Information Processing Systems (**NeurIPS**), 2026*.

RESEARCH & PUBLICATIONS

- **Hanjun Cho**, Jay-Yoon Lee. "RARE: Redundancy-Aware Retrieval Evaluation Framework for High-Similarity Corpora." *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (**ACL**), 2026* .
- **Hanjun Cho**, Gahyun Yoo, Hanseong Kim, Jay-Yoon Lee. "Generalizing Numerical Reasoning in Table Data through Operation Sketches and Self-Supervised Learning." *Transactions of the Association for Computational Linguistics (**TACL**), 2026*.

OTHERS

- **Core Part Lead**: Led the Core team (3 engineers, ~6 months) driving technical differentiation: OCR & VL embedding serving (63% lower latency), Agentic RAG (accuracy 45% -> 82%).
- **Ralli**: enterprise RAG product, sole developer: built end-to-end in ~2 months, shipped to first paying customer (Hyundai Motor Securities).

EDUCATION

- **Seoul National University (SNU)** - M.S. Data Science · Mar 2022 - Aug 2024
- **Hanyang University** - B.A. Economics & Finance · Mar 2017 - Feb 2022

SKILLS

- **Programming:** Python, C/C++, SQL, R, PHP, HTML
- **ML/DL:** PyTorch, TensorFlow, scikit-learn; PEFT, QLoRA, FlashAttention, DeepSpeed
- **Inference / Serving:** TensorRT-LLM, TensorRT, vLLM, Triton, CUDA, OpenCL, MPI
- **Retrieval:** Elasticsearch, Milvus
- **Data / Infra:** MongoDB, PostgreSQL, Redis, Kafka, RabbitMQ
- **MLOps / DevOps:** Docker, Kubernetes, AWS, Nginx, Prometheus, Grafana

PROJECTS

AI Grand Challenge (Top-5, 2023)

- Developed table-text QA system with Docker deployment (Kobigbird, KLUE-RoBERTa, pko-T5).
- Built evaluation pipeline; achieved competitive performance.

Securities Research Assistant (RA) - Sep 2023 - Dec 2023

- Automated US market/finance news QA system (BM25 + SBERT, Pegasus/BART, FinBERT, PrimeQA).
- Improved retrieval benchmark 47% -> 83%.

Disease Diagnosis via LMs - Sep 2022 - Dec 2022

- Built Top-5 prediction model; utilized Neo4j graph reasoning for transparency.

Korean LM Compression - Sep 2022 - Dec 2022

- Achieved 30-40% size reduction with minimal performance loss via pruning, quantization, and distillation.

XR Nursing Training (Contract) - Aug 2022 - Jan 2023

- Implemented real-time NPC dialogue; reduced latency 6-10s -> ~3s; showcased at CES 2024.