RESEARCH PORTFOLIO

Towards Efficient and High-performance LLMs

Singon Kim

01 · TIMELINE

Content

- **[Lunit] Foundation Model Hackathon for Medical Science**
  [Hackathon] Foundation Model Hackathon· 2026.08.21~2026.08.22
- **[SKTelecom] A.X K2 Sovereign AI Foundation Model**
  [Intern] Sovereign AI Foundation Model Intern · 2026.03 ~ 2026.06
- **[NaverCLOUD] HyperCLOVA X Sovereign AI Foundation Model**
  [Intern] HyperCLOVA Foundation Model Research Intern · 2025.05 ~ 2025.10
- **ACoRN: Noise-Robust Abstractive Compression in Retrieval-Augmented Language Models**
  [paper] *International Joint Conference on Neural Networks (IJCNN)* · **2025**
- **[i-scream edu] AI assistant for elementary teachers using Large Language Models**
  [Industry-Academia R&D Project] · 2023.09 ~ 2024.01
- **[ADD] AI Agents for CBRN Operational Command and Guidance**
  [National R&D Project] · 2024.12~2025.05

01 · PROJECT ONE

Conquer Health: Foundation Model Hackathon for Medical Science

[Lunit] Foundation Model Hackathon

[Lunit] Conquer Health: Foundation Model Hackathon for Medical Science

Overview of the Hackaton

2026.08.21~2026.08.22

Title

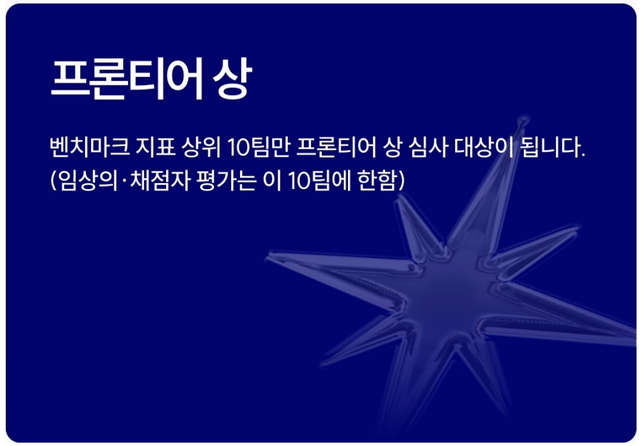
- Building a specialized public health management chatbot

Goal

- Outperform general-purpose frontier models on public health management tasks
  - By integrating specialized biomedical foundation models with MCP tools.

Evaluation Set up

- The conversation starts from a fixed set of N predefined questions
- Each initial question is extended into a multi-turn conversation of up to 3 turns
  - User fellow-up messages are generated by a simulated patient agent.
  - The evaluation harness was developed and provided by Lunit
- Blind pairwise comparison between frontier models and each team's chatbot
  - Chat bot: L2-preview developed by Lunit, based on Trillion Labs' Gravity (30B-A5B) model
- Evaluators voted for the better model, and the team with the most votes won the Frontier award



Frontier award(1st)

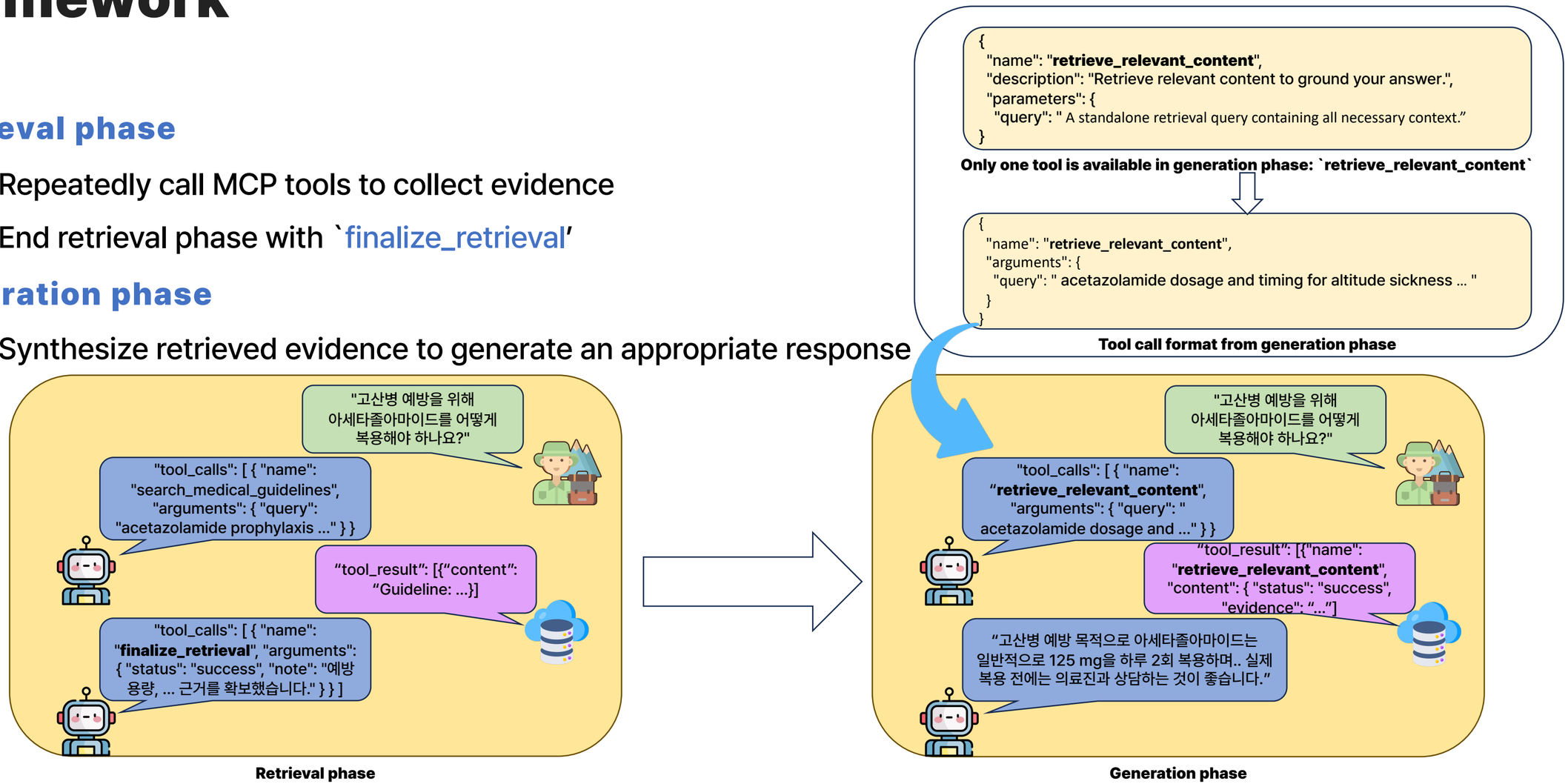
Framework

Retrieval phase

- Repeatedly call MCP tools to collect evidence
- End retrieval phase with `finalize_retrieval'

Generation phase

- Synthesize retrieved evidence to generate an appropriate response



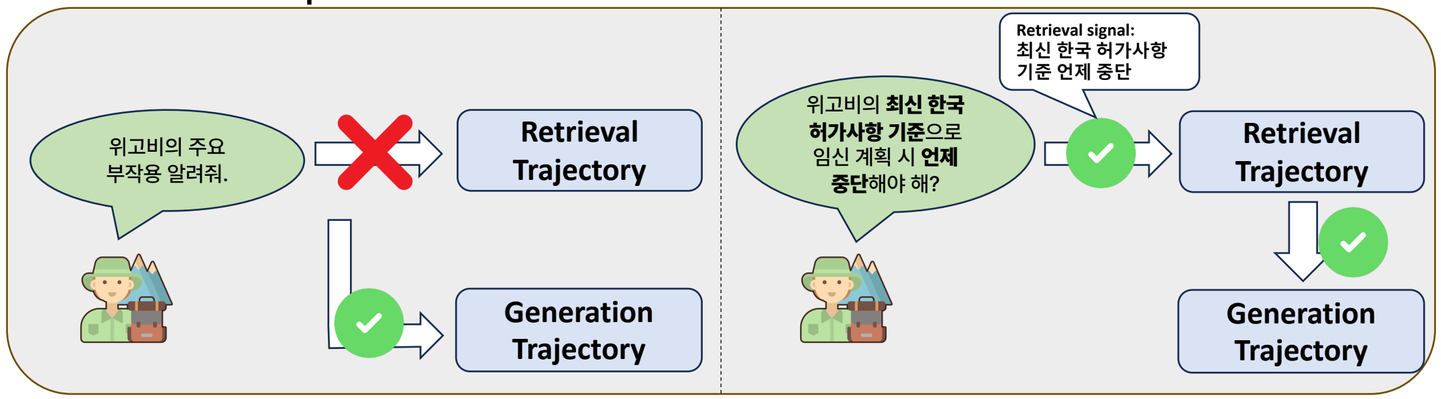
Strategy for the Retrieval phase (1/2)

Retrieval Usage Strategy

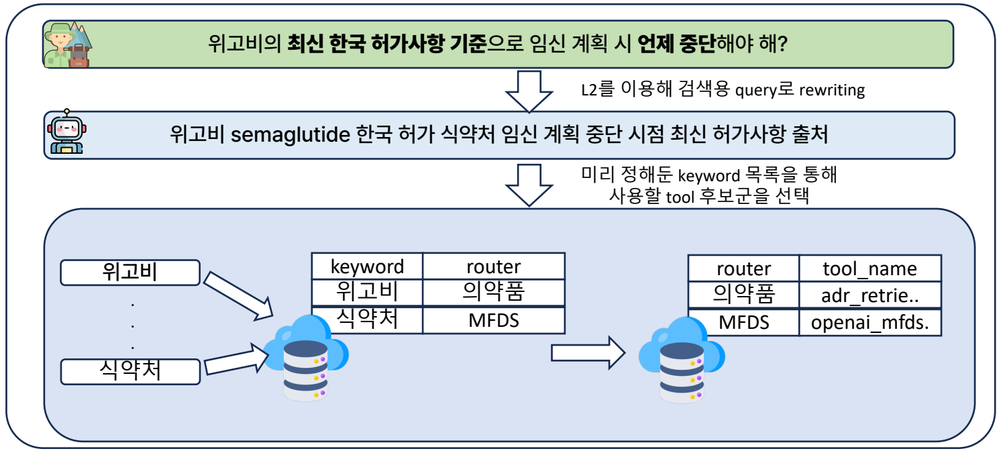
- Skip retrieval for general medical knowledge or basic consultation
  - Since L2-preview is already specialized in medical knowledge; *retrieval can degrade performance*
- Use retrieval when freshness or document-level evidence is needed

Retrieval Phase for Precise Tool Use

- Select a small tool subset based on query keywords
- Cap the number of retrieval turns



Rule-based gate for tool-use decisions



Tool candidate filtering

Strategy for the Retrieval phase (2/2)

`finalize_retrieval`

- End the retrieval phase with 'finalize_retrieval' and verify that sufficient evidence has been collected

Citation Validation

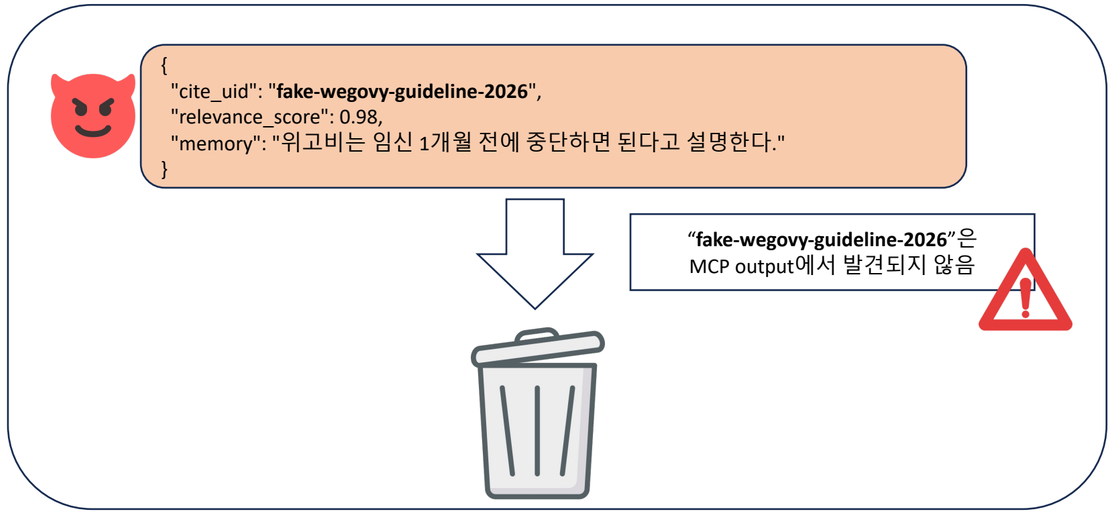
- Prevent the LLM from fabricating plausible-looking citations during tool use
- Validate each selected `cite_uid` against actual MCP tool outputs and discard invalid citations

Long History Compaction

- Compress history over 14000 tokens before generation

```
{
  "name": "finalize_retrieval",
  "arguments": {
    "status": "sufficient",
    "items": [
      {
        "cite_uid": "kr-label-wegovy-2024-pregnancy",
        "relevance_score": 0.94,
        "memory": "한국 위고비 허가사항은 임신 계획 시 semaglutide를 최소 2개월 전 중단하라고 설명한다."
      }
    ],
    "note": "한국 허가사항에 직접 해당하는 근거를 찾음"
  }
}
```

Example of using `finalize_retrieval `



Preventing Citation Hallucination

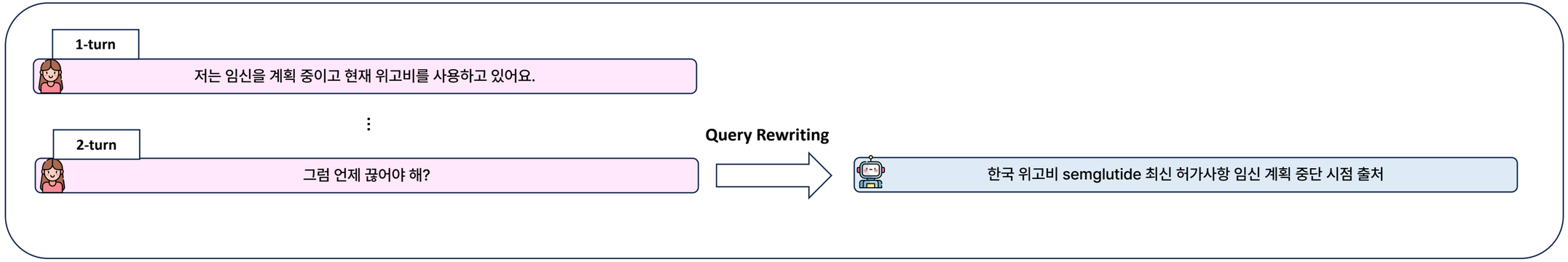
Additional Strategies

Query Rewriting

- Rewrite retrieval queries using prior conversation context
- Convert multi-turn context in a standalone query

Generation System Prompt

- Match the user's language and explain in clear, simple terms
- Prioritize safety guidance in urgent or high-risk situations
- Avoid unnecessary explanations and repetition



Rewriting query into standalone query using relevant context from previous turns

Evaluation

Blind Evaluation against GPT-5.6-Sol

- Evaluated by doctors and medical-domain developers
- Preferred over GPT-5.6 Sol in most test cases
  - Key Strengths
    - Clear and easy to follow structure
    - Detailed, situation-specific action guidance
    - Summary of key points at the end

TURN 1

선생님, 제가 며칠 전부터 아가트로반(아르가트로반)이라는 주사약을 맞기 시작했는데, 맞은 부위 혈관이 욱신욱신 아프고 줄기가 빨갛게 올라왔어요. 정맥염이라고 검색하니까 무섭게 나오던데, 이 약 자체가 원래 그런 부작용이 있는 건지, 아니면 저한테만 특별히 생긴 문제라서 지금 바로 응급실로 가야 하는 건지 모르겠어요.

User query in 1-turn

안녕하세요, 선생님. 말씀하신 상황을 바탕으로 정리해 드리겠습니다.

---

1. 결론부터: 이런 반응은 드문 게 아니라 알려진 주사부위 부작용입니다

말씀하신 **아가트로반(약품명: ARGATROBAN, 노바스탄)**은 직접 트롬빈 억제제라는 항응고제로, **정맥 주사로 투여**합니다.
항응고제 주사제에서는 **주입 부위가 붉어지고, 아프며, 혈관을 따라 줄기가 빨갛게 올라오는 증상(표재성 혈전정맥염 또는 정맥염)**이 **연관된 부작용으로 이미 알려져 있습니다.**

즉, *"나에게만 생긴 특별한 문제"*라기보다는 **의약품 자체로 인해 발생할 수 있는, 비교적 흔한 국소 반응**에 해당할 가능성이 높습니다.

Candidate (ours)

VS

❗ **아가트로반 주입 부위의 통증과 혈관을 따라 번지는 발적은 지금 바로 의료진에게 알려야 합니다.**
다만 증상만으로 곧바로 응급실이 필요한지는 동반 증상과 범위에 따라 달라집니다.

1. 약 때문일 수 있나요?

아가트로반(agatroban)은 정맥으로 투여하는 **항응고제(anticoagulant)**입니다. 주입 과정에서 정맥 자극, 카테터 위치 문제, 약액이 혈관 밖으로 샌 **혈관외유출(extravasation)** 또는 **정맥염(phlebitis)** 때문에 통증·붉은 선·부종이 생길 수 있습니다.

따라서 **약 자체의 국소 부작용일 수도 있지만, 단순한 정상 반응으로 넘기면 안 됩니다.** 혈관 감염이나 혈전성 정맥염(thrombophlebitis) 여부도 의료진이 확인해야 합니다.

GPT-5.6 Sol

3. 행동 지침 — "바로 응급실로?" vs "언제 누구에게 알려야?"

✅ **지금 당장 이렇게 하십시오**

1. **처방하신 주치의(주사 치료를 담당하는 의사) 또는 병동/진료과에 즉시 전화**해서

- 현재 증상("A정맥 주사를 맞은 부위가 붓고 아프며, 정맥을 타고 붉은 줄이 올라온다")을 설명한다.
- 약을 계속 맞아야 하는지, 주사 부위를 다른 혈관으로 변경할 수 있는지 의견을 듣는다.

1. 주사를 **임의로 중단하지 마십시오.** 아가트로반은 화학적 쇼크에 의한 헤파린 유발 혈소판감소증(HITT) 같은 심각한 상황에서 쓰는 약이므로, 치료 중단은 새 혈전·색전증 위험을 높일 수 있습니다.
2. 약물 일시 보류·중단 여부는 반드시 의사 확인 후 결정합니다. 의사는 중단보다는 **부위 변경, 주입 속도 조절, 추가적인 냉찜질 혹은 온찜질 등 지침**을 줄 것입니다.

⚠️ **이런 증상 발생 시에는 응급실(또는 바로 병원 외래)에 즉시 연락 및 방문**

- 붉은 줄기가 팔 위로 빠르게 확장되거나, 그 줄이 근육 깊숙이 죄는 듯한 통증으로 바뀔 때
- 손 끝으로 가는 혈류가 약해져 창백해지거나 차가움, 심한 부종
- 38°C 이상의 발열, 오한, 아픈 팔 혈관 주변 열감 증가
- 처음 맞은 부위가 벌겋게 달아오르고 국소적인 고름·농양 같은 병변(감염 시사)

Candidate (ours)

VS

3. 🚨 즉시 응급실로 가야 하는 경우

다음 중 하나라도 있으면 응급평가가 필요합니다.

- 붉은 범위가 빠르게 커지거나 팔 전체가 붓고 단단해짐
- 심한 통증, 피부색 변화, 물집 또는 피부가 검게 변함
- 발열, 오한, 고름, 심한 열감
- 숨참, 가슴통증, 갑작스러운 어지럼·실신
- 멈추지 않는 출혈, 검은 변, 피 섞인 소변, 심한 두통이나 신경학적 증상
- 입술·얼굴 부종, 전신 두드러기, 호흡곤란 같은 알레르기 증상

현재처럼 **혈관을 따라 빨갛게 올라오고 욱신거리는 증상은 최소한 즉시 담당 의료진에게 보고해야 하는 상황**입니다.

GPT-5.6 Sol

02 · PROJECT TWO

SK Telecom, Sovereign AI Foundation Model Intern

[SK TELECOM] Sovereign AI Foundation Model Project

Overview

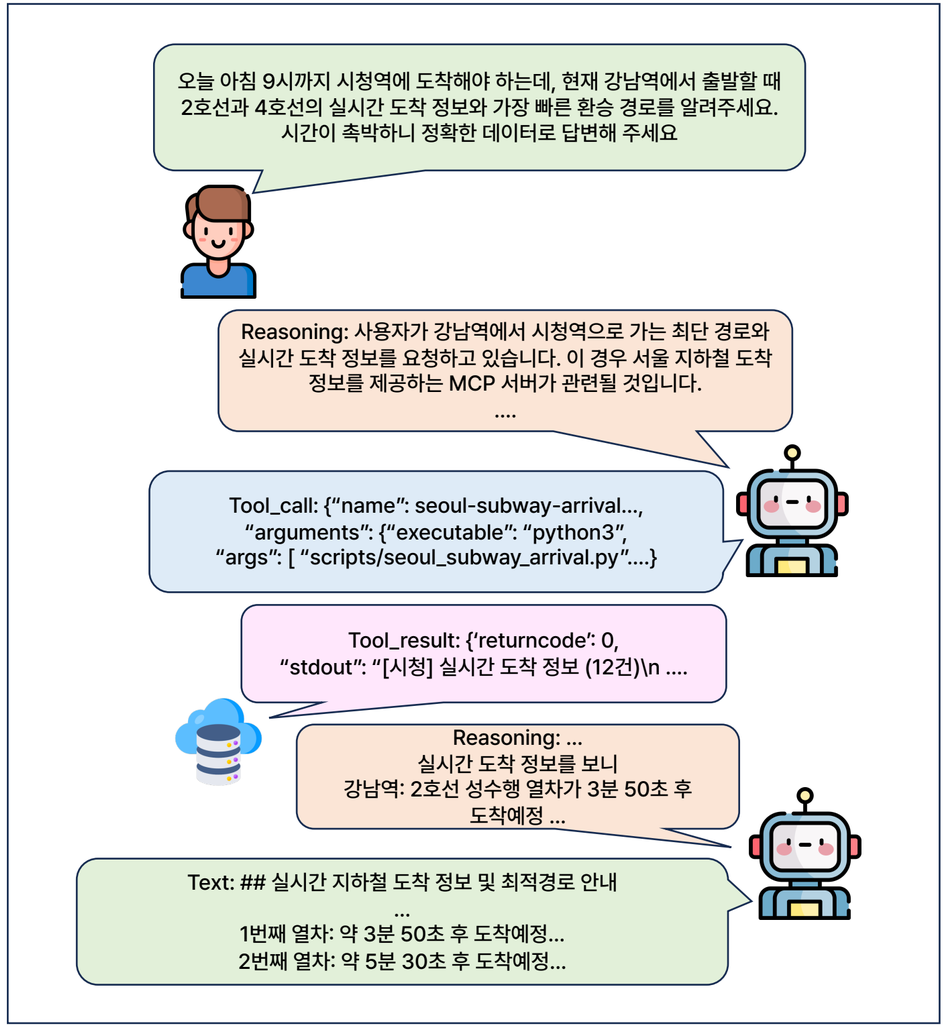
2025.03.16 – 2025.06.30

Goal

- Passing the second round of the K-sovereign AI Foundation model project, led by Korean government

Key Responsibilities

- Building synthetic data pipelines for Korean/English tool-use training
  - Generated synthetic data using LLMs via OpenRouter
  - Set up and operated pipelines on AWS EC2



Synthetic Data Example

Synthetic Data Pipeline

Approach

- Built tool-use SFT data by generating trajectories through execution in real environments
- Compared two synthetic data generation strategies
  - Task Synthetic: Generate the task first, then construct the trajectory
  - Trajectory Synthetic: Generate the trajectory from seed concepts, then derive the corresponding task
- Focused on real-world diversity rather than optimizing for a specific benchmark
- Optimized training data
  - Selecting the right tools from a large tool set
  - Using tool correctly and efficiently
  - Producing concise and complete trajectories
  - Reaching the correct final answer

[SK TELECOM] Sovereign AI Foundation Model Project

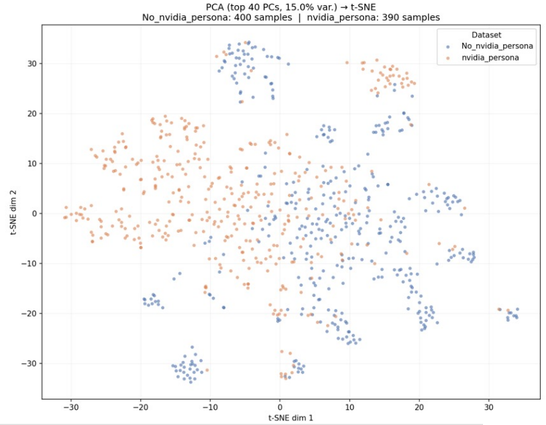
Improving Diversity in Task Synthetic

Limitations of LLMs on Open-Ended Questions[1]

- Task diversity remains limited when the available MCP tools and environments are similar
  - Despite using different LLMs, they often generate highly similar synthetic tasks under the same conditions

Addressing Task synthesis Diversity Limitations through Persona Injection

- Injecting diverse personas to generate varied synthetic tasks with persona-specific scenarios and request styles[2]
  - Korean personas, such as Nemotron-Personas-Korea[3], help generate synthetic tasks that better reflect Korean contexts

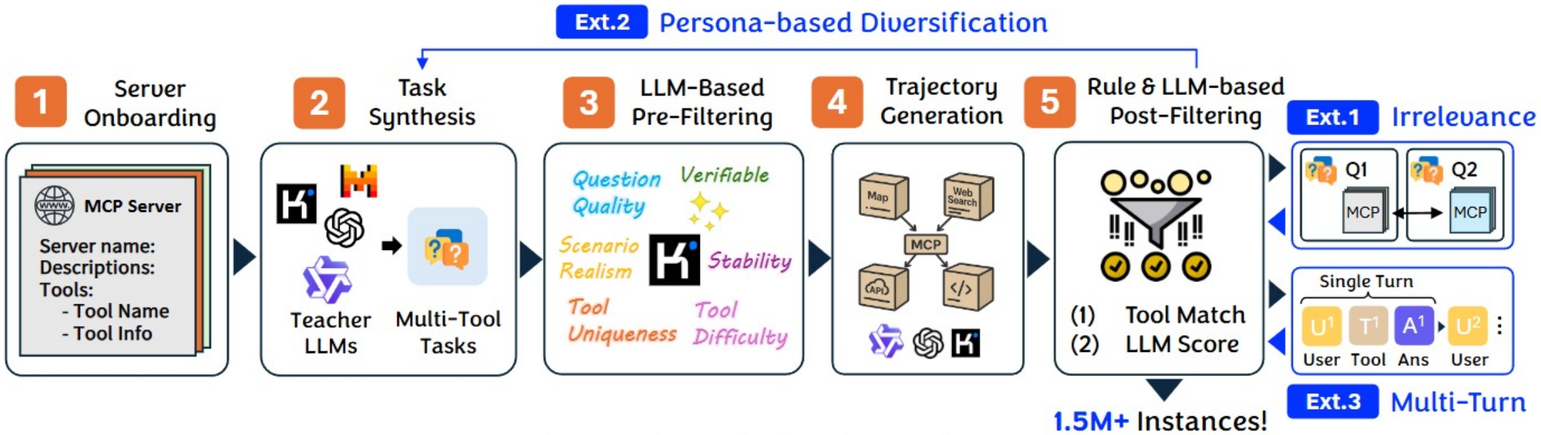


[1] L. Jiang et al., “Artifical Hivemind: The Open-Ended Homogeneity of Language Models (and Beyond),” in NeurIPS 2025 Datasets and Benchmarks Track (oral).
[2] T. Ge et al., “Scaling Synthetic Data Creation with 1,000,000,000 Personas,” arXiv: 2406.20094
[3] https://huggingface.co/datasets/nvidia/Nemotron-Personas-Korea

Task-First vs. Trajectory-First (1/2)

Limitation of Task-First Synthetic Generation[1]

- Some tasks may not be solvable with available tools
- Increasing failure rates when tasks require more tools
  - High Dependence on the LLM's agentic and actual problem-solving capabilities



Example of a synthetic data pipeline using task-first generation (TOUCAN)

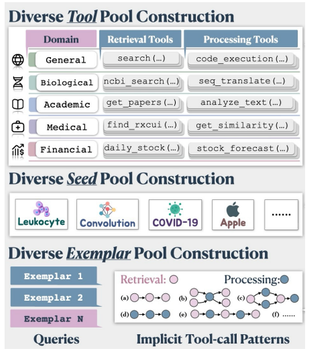
[1] Z. Xu et al., "TOUCAN: Synthesizing 1.5M Tool-Agentic Data from Real-World MCP Environments", arXiv: 2510.01179

Task-First vs. Trajectory-First (2/2)

Limitations of Trajectory-First synthetic generation[1]

- When reasoning using the same tool-call topology, solution patterns tend to become repetitive
  - Similar tool combinations with seed concept often produce similar patterns
- Task difficulty tends to be low in trajectory-first generation
  - Rewriting task to be more ambiguous [2] requires additional validation is needed

→Prioritized scale diversity and high-quality data over cost, and chose a task-first generation strategy



Resource Bank
Build domain-specific tool pools and collect seed concepts and examplars

Trajectory-First Synthetic Data Generation Pipeline: DIVE

Seed Concept: ATC A10BK (당뇨병 치료제 분류)
Tool Set (40개 제공): rxnorm_get_drugs, rxclass_get_class_members, icd10cm_diagnosis_codes, prescribable_get_spelling_suggestions 등 의학 API 모음

Config Sampling
Each synthesis run consists of 1 seed concept, 15-50 tools from the target domain, and 3-5 exemplars.

// Tool Call 1: ATC A10BK 분류에 속한 약물 검색
rxclass_get_class_members(class_id="A10BK")
→ Observation: ["Ertugliflozin", "Empagliflozin", "Dapagliflozin", ...]

// Tool Call 2: 해당 약물들의 기전(MOA) 및 부작용/금기 질환 조회
rxclass_get_sources_of_drug_class_relations(drug="Ertugliflozin")
→ Observation: MOA="N0000187058", Contraindication="Chronic Kidney Failure (MEDRT: D007676)"
...

Evidence Collection
The agent executes real APIs and collects actual observations as evidence

[Derived Query (역도출된 질문)] ATC A10BK 치료군에 속하고 특정 작용기전(MOA class N0000187058)을 공유하며 만성 신부전(MEDRT: D007676)에 금기사항이 있는 약물 중, UNII 코드가 '6C282481IP'인 약물을 찾으시오.
...

[Derived Answer (역도출된 정답)] Ertugliflozin (UNII 6C282481IP). (1) 10 MG (15 MG - 5 MG) (2) 2개 (canagliflozin anhydrous, dapagliflozin propanediol) (3) Empagliflozin (4개 브랜드: Jardiance, Glyxambi, Synjardy, Trijardy)

Task Derivation
A Task Generator LLM analyzes the collected observation traces and works backward to construct the corresponding question and answer

[1] A. Chen et al., "Dive: Scaling Diversity in Agentic Task Synthesis for Generalization Tool Use", in ICML, 2026.
[2] J. Liu et al., "WebExplore and Evolve for Training Long-Horizon Web Agents", arXiv:2509.06501

Direction for Synthetic Data Generation

Choosing a Task-First Synthetic Data Pipeline

- Aim to build training data that closely reflects real-world scenarios
  - Prioritize challenging, high-quality data grounded in realistic use cases
- Include examples to train the model to handle issues arising in specific environments
  - To train the model to identify the root cause of the issue
    - Internal errors (e.g., incorrect tool arguments)
    - External errors (e.g., API rate limit)

Key Challenges in Generating High-Quality Tool-Use Data

- Reasoning Hallucination
  - LLMs often rely on internal knowledge during reasoning, leading to factual hallucinations
    - The model completes the task using the necessary tools, but the trajectory still contains substantial hallucination
- Handling Environment-Specific Failures
  - Training data should include recovery behaviors, such as recognizing the failure and retrying appropriately

[SK TELECOM] Sovereign AI Foundation Model Project

Filtering to Reduce Hallucinated Data

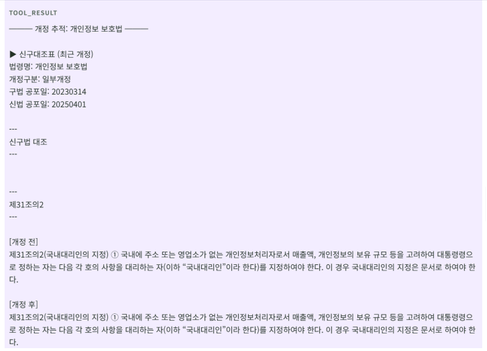
Method

- Apply rule-based filtering before using an LLM-as-a-Judge whenever possible
  - Using tool matching to filter samples for tasks with clearly required tools
- For open-ended questions, use explicit rubrics for evaluation and filtering
- For hallucination-prone domains (e.g., law), perform turn-level validation for every tool-use step

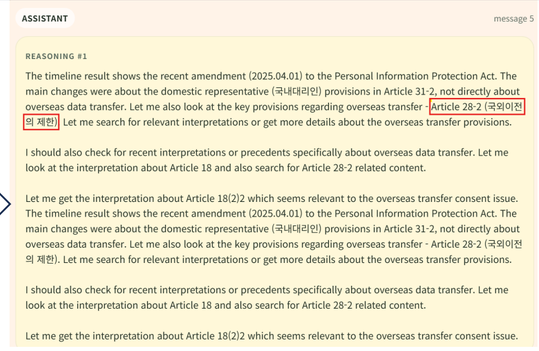
An Example of Hallucination in Legal Tool Use



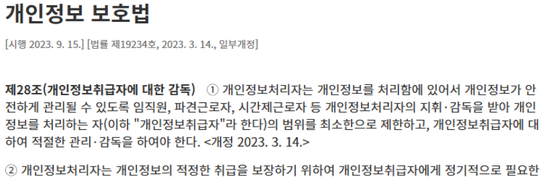
A legal question about whether separate consent is required for overseas transfer of personal data under **Korea's Personal information Protection Act (K-PIPA).**



The tool was used, but it returned only information related to **Article 31**, concerning domestic representatives.
(*Low relevance to the question*)



The model recognized that **Article 31 was not relevant** and relied on its internal knowledge to cite **Article 28** as a restriction on overseas transfer.



However, **Article 28 actually concerns supervision of personal information handlers**, making the retrieved reasoning factually incorrect.

Contributions and Limitations

Contributions

- Generated scale-diverse synthetic data across various domains, scenarios, and tool-use patterns
  - Included edge cases that may occur in real environments, such as no relevant data found
  - Built high-quality Korean tool-use data with a wide range of MCP tools despite limited Korean tool-use benchmarks
- Introduced extensive hallucination validation to reduce hallucinations reinforced through SFT

Limitations & Reflections

- A large portion of synthetic data was filtered out to maintain high quality
  - Only about 10%~20% of samples passed all filtering stages
- Couldn’t measure the impact of the synthetic data through model training and evaluation

[SK TELECOM] Sovereign AI Foundation Model Project

Publications (Technical Report)

arXiv:2608.30181 **A.X K2 Technical Report**

03 · PROJECT THREE

Naver Cloud, HyperCLOVA Foundation Model Research Intern

[NAVER CLOUD] HyperCLOVA Foundation Model Research Intern

Overview

2025.05.07 – 2025.10.31

Goal

- Built and pretrained LLMs from scratch

Key Responsibilities

- Large-scale multilingual and reasoning-focused pretraining data curation
- Explored efficient LLM architectures and training methods for context length generations
- Investigated distributed training frameworks to improve training efficiency

[NAVER CLOUD] HyperCLOVA Foundation Model Research Intern

Synthetic Data (Pre-training)

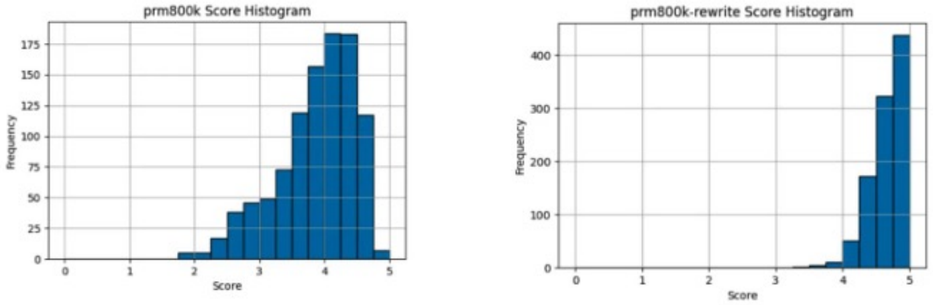
Limitations of Web Data[1]

- Simply scaling web data eventually hits a performance ceiling (Data Wall)
- Increasing need for high-quality, information-dense synthetic data

→ **Adopted the Source Rephrasing Paradigm to generate high-quality, scale-diverse synthetic data**

Synthetic Data Generation Strategies for Math & Code

- Swallow Math/Code[2] - style Rewriting
- Reverse Mathematical Reasoning[3]
- Long CoT Rewriting



Enhancing PRM800k with rewriting and data curation

Results related to Synthetic Data

- Improved 32B LLM MMLU (5-shot) from 79 to 81 with minimal trade-offs
- Achieved 84% GSM8K (5-shot) with an 8B LLM
- Built synthetic data covering specialized finance and engineering knowledge beyond standardized benchmarks

[1] Datology AI Team., “BeyondWeb: Lessons from Scaling Synthetic Data for Trillion-scale Pretraining”, arxiv:2508.10975
[2] K. Fujii et al., “Rewriting Pre-training Data Boosts LLM Performance in Math and Code”, arxiv:2505.02881
[3] P.Guo et al., “Exploring Reversal Mathematical Reasoning Ability for Large Language Models” in ACL, 2024.

[NAVER CLOUD] HyperCLOVA Foundation Model Research Intern

Insights from Building Synthetic Data

Scale & Diversity vs. High-quality

- Scale & Diversity become increasingly important once a certain quality threshold is reached
  - In our experiments, filtered MegaMath-Web-pro data had larger impact on model performance than rewritten PRM 800K data
  - Low-quality data may cause "Brain Rot" [1], while excessive use of even high-quality synthetic data can hurt model performance [2]

Longer Reasoning Does Not Always Mean Better Data

- On relatively simple math tasks, overly detailed reasoning or verification sometimes degraded performance
  - In our experiments on GSM8k, elementary and middle school level problems, data with standard step reasoning had the most positive impact on performance, even when generated by weaker models.

Solving Math Problems ≠ Understanding Mathematical Concepts

- Synthetic data pipelines often favor the (Question, Reasoning, Answer) format to improve mathematical benchmarks
  - However, it may provide limited coverage of underlying mathematical concepts
  - Interestingly, academic data from science, economics and other STEM domains substantially improved performance on some mathematical concept evaluations

[1] S. Xing et al., "LLM can get 'Brain Rot': A Pilot Study on Twitter/X", in COLM, 2026.
[2] Feiyang Kang et al., "Demystifying Synthetic Data in LLM Pre-training: A Systematic Study of Scaling Laws, Benefits, and Pitfalls", in EMNLP, 2025.

[NAVER CLOUD] HyperCLOVA Foundation Model Research Intern

Bottleneck in Large-Scale Synthetic Data Generation

Goal

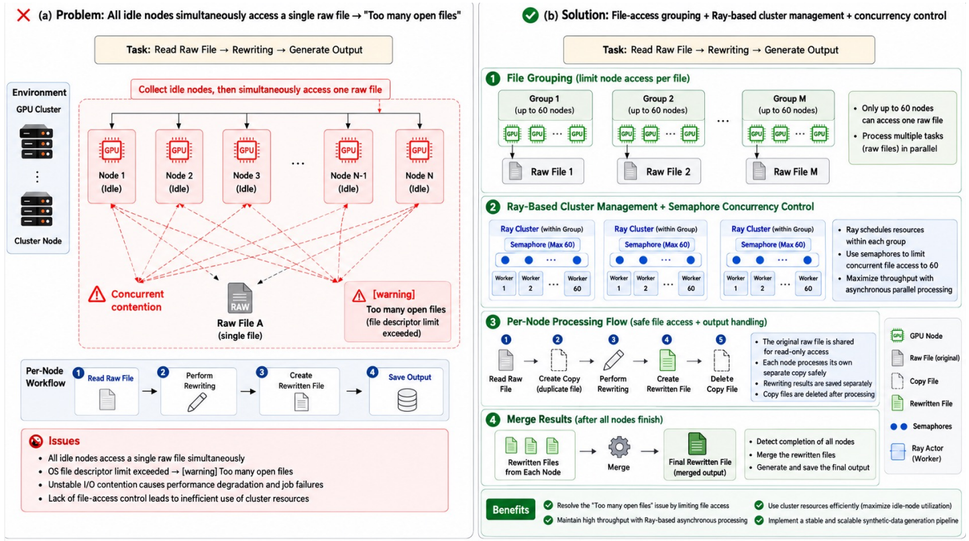
- Build a synthetic data pipeline that automatically utilizes all available idle cluster nodes for synthetic data generation

Issue

- Encountered the warning: `**Too many open files**`

Solution

- Limited the number of nodes accessing the same file simultaneously
- Managed cluster resources with Ray
  - Increased throughput by maximizing concurrency with semaphores



Issues and Solutions in Large-Scale GPU Utilization

[NAVER CLOUD] HyperCLOVA Foundation Model Research Intern

Reliability of Score-Based Quality Filtering

Issue

- Classifiers trained on labels annotated by Qwen3-223B-A22B (Model A) and GPT-OSS-120B (Model B) produced noticeably different scores

Analysis

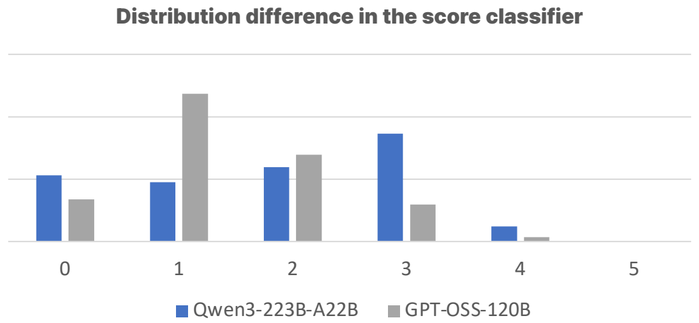
- Many samples (0.24%) were rated high (4+) by Model A but a low score (2-) from Model B
- In Contrast, relatively few samples (0.0025%) received a high score (4+) from Model B but a low score (2-) from Model A

Cause

- Model B focused heavily on the prompt criterion, “educational content suitable for elementary and middle school students” and penalized samples that were considered too difficult

Key Insight

- Different model prioritize different prompt elements, requiring careful evaluator selection and prompt validation



[NAVER CLOUD] HyperCLOVA Foundation Model Research Intern

Failure of Long Context Copying[1] with Noise Removal

Goal

- Build high-quality long-context training data from Korean articles and papers by removing irrelevant tokens such as [img]

Why Rewriting Is Needed

- Raw data often contains noise such as [Figure] or poorly extracted [Table] tokens
- Simply deleting these tokens can break the surrounding context
  - Example: [그림] 참고했을 때… → 참고했을 때…
  - However, many samples remain understandable even without the referenced figure or table

Issue

- Even frontier models struggle to follow rewriting constraints on ~60K tokens inputs

Solution

- Split documents by paragraph within a predefined context-length range
- Use a system prompt to preserve the original content while rewriting and removing unnecessary tokens
- For [Table] cases, apply heuristic filtering after rewriting and use LLMs to verify whether the rewritten context remains valid

[1] M. Meyer et al., “How Many Different Outputs Can a Transformer Generate?” in ICML 2026.

[NAVER CLOUD] HyperCLOVA Foundation Model Research Intern

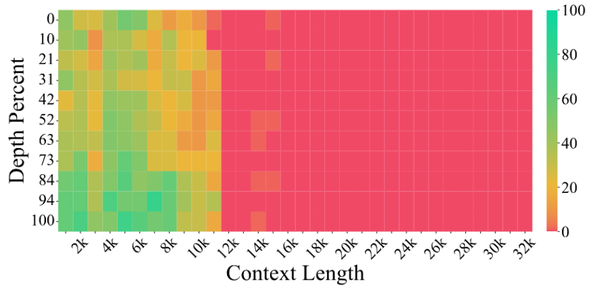
Exploring Architectures for Length Generalization

Motivation

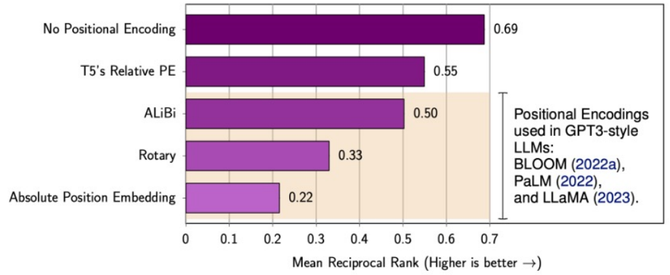
- Extending the context window with existing methods caused performance degradation on standard benchmarks
- Improve the "Lost in the middle" issue observed with Rope-based architecture[1],[2]
- Enable more efficient long-context processing

Research Direction

- Introducing a hybrid positional encoding to mitigate catastrophic forgetting while improving long-context capability
- Explore KV-cache compression or selection methods for efficient long-context inference



3B LLAMA-based model with 8K context on NIAH[3]



Compare the ability to retrieve and generate relevant information from long context[4]

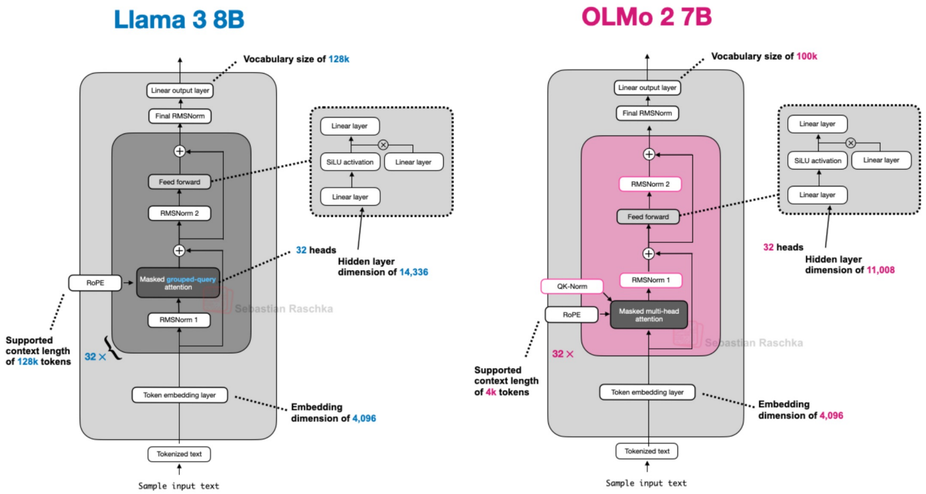
[1] N. F. Liu et al., "Lost in the Middle: How Language Models Use Long Contexts," TACL, 2024.
[2] M. Ravaut et al., "On Context Utilization in Summarization with Large Language Models," in ACL, 2024.
[3] Y. Chen et al., "HoPE: A Novel Position Encoding Without Long-Term Decay for Enhanced Context Awareness and Extrapolation," in ACL, 2024.
[4] A. Kazemnejad et al., "The Impact of Positional Encoding on Length Generalization in Transformers", in NeurIPS, 2023.

[NAVER CLOUD] HyperCLOVA Foundation Model Research Intern

RNoPE Architecture Study (1/2)

Experiment 1: Convert RoPE to RNoPE and Evaluate Benchmark performance

- Converting pretrained LLMs from RoPE to RNoPE caused knowledge forgetting
  - OLMO2, with QK-Norm, and post-sublayer RMS Norm, showed relatively better knowledge retention
  - Reasoning ability degraded more significantly than factual knowledge
- Evaluate Sliding Window attention (SWA) for inference efficiency and benchmark performance



Compare different RNoPE insertion intervals

X-interval: apply NoPE every X layers while keeping RoPE in the remaining layers

model_name	mmlu@acc	piqa@acc	gsm8k_cot@exact_match,fl	humaneval@pass@1	mbpp@pass_at_1
OLMo-2-1124-7B	0.6285	0.8101	0.6740	0.1402	0.2300
olmo2-RNope-8interval	0.5964	0.8041	0.3654	0.1341	0.0020
olmo2-RNope-4interval	0.5729	0.7992	0.2699	0.1220	0.0000
Llama-3-8B	0.6543	0.7943	0.5451	0.3598	0.4800
Llama-3-8B-RNope-8interval	0.5775	0.7922	0.1046	0.2500	0.0080
Llama-3-8B-RNope-4interval	0.4210	0.7840	0.0485	0.1768	0.0140
SmolLM2-1.7B	0.5027	0.772	0.3071	0.0061	0.0214
SmolLM2-1.7B-RNope-8interval	0.4608	0.765	0.0599	0.0061	0
SmolLM2-1.7B-RNope-6interval	0.3835	0.7557	0.0318	0	0

[NAVER CLOUD] HyperCLOVA Foundation Model Research Intern

RNoPE Architecture Study (2/2)

Experiment 2: Apply RNoPE during mid-training while extending the context window

- RNoPE had higher initial loss during mid-training but eventually converged to a similar level as RoPE
  - However, benchmark performance still declined slightly, and CoT capability was not fully recovered
  - *Why does language modeling loss converge while reasoning performance diverges after positional encoding adaptation?*

Experiment 3: Evaluating Sliding Window Attention(SWA) for inference efficiency

- As context length increased, SWA provided increasingly large inference-time savings

Evaluation after Long Context Training

model_name	mmlu@acc	leaderboard_bbh	kmmlu_direct	global_mmlu_en	global_mmlu_ko	gsm8k_cot@fle	gsm8k@flexible	humaneval@pa	mbpp@pass_at
hcx-8B-stage2-72000	0.6612	**0.4864**	0.4912	0.7075	0.6075	0.6778	0.6937	0.6402	0.5972
hcx-8B-stage3-v2-1000 (YaRN)	**0.6661**	0.4784	**0.5020**	**0.7200**	**0.6150**	**0.7005**	**0.6983**	0.6280	0.6280
hcx-8B-stage3-rnope-swa-100	0.6344	0.4338	0.4478	0.6775	0.5475	0.4344	0.6308	0.5549	0.5220
hcx-8B-stage3-rnope-swa-200	0.6360	0.4376	0.4578	0.6950	0.5750	0.4041	0.6543	0.5976	0.5360
hcx-8B-stage3-rnope-swa-300	0.6419	0.4432	0.4649	0.6900	0.5850	0.1721	0.6634	0.6159	0.5440
hcx-8B-stage3-rnope-swa-400	0.6388	0.4466	0.4544	0.6925	0.5925	0.4966	0.6611	0.6098	0.5600
hcx-8B-stage3-rnope-swa-500	0.6440	0.4508	0.4670	0.6925	0.5825	0.1918	0.6550	0.5976	0.5700
hcx-8B-stage3-rnope-swa-600	0.6409	0.4524	0.4684	0.6950	0.5875	0.2123	0.6763	0.6220	0.5540
hcx-8B-stage3-rnope-swa-700	0.6479	0.4463	0.4756	0.7100	0.5875	0.2335	0.6550	0.6280	0.5460
hcx-8B-stage3-rnope-swa-800	0.6422	0.4419	0.4677	0.7075	0.5900	0.3351	0.6611	**0.6341**	0.5560
hcx-8B-stage3-rnope-swa-900	0.6449	0.4522	0.4756	0.7150	0.5850	0.1243	0.6649	0.6159	0.5520
hcx-8B-stage3-rnope-swa-1000	0.6449	0.4541	0.4788	0.7100	0.5950	0.3351	0.6748	**0.6341**	0.5520

Measured the time required for context tokens to pass through 10 forward passes on HCX-8B
Unit: seconds(s) and Backbone context window: 128K tokens

model_name	10K	16K	32K	64K	128K
HCX-8B	7.41	15.67	37.67	103.69	328.77
HCX-8B-swa(8K)	7.3	15.12	31.99	71.82	175.33

[NAVER CLOUD] HyperCLOVA Foundation Model Research Intern

Publications (Technical Report)

arXiv:2601.03286 **HyperCLOVA X 32B Think**

arXiv:2601.01792 **HyperCLOVA X 8B Omni**

arXiv:2506.22403 **HyperCLOVA X THINK Technical Report**

04 · PROJECT FOUR

ACoRN: Noise-Robust Abstractive Compression in Retrieval-Augmented Language Models

International Joint Conference on Neural Networks (IJCNN), 2025

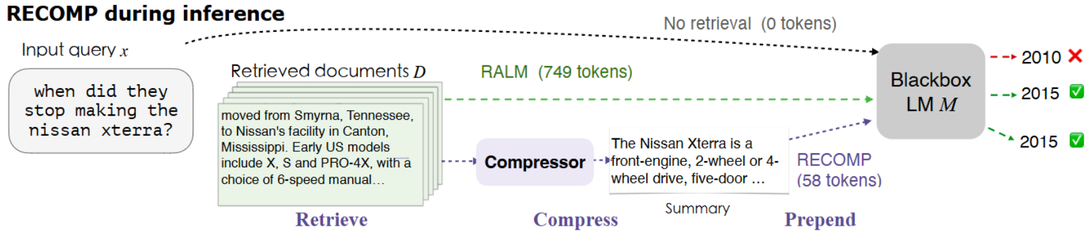
Research Overview

Research Background

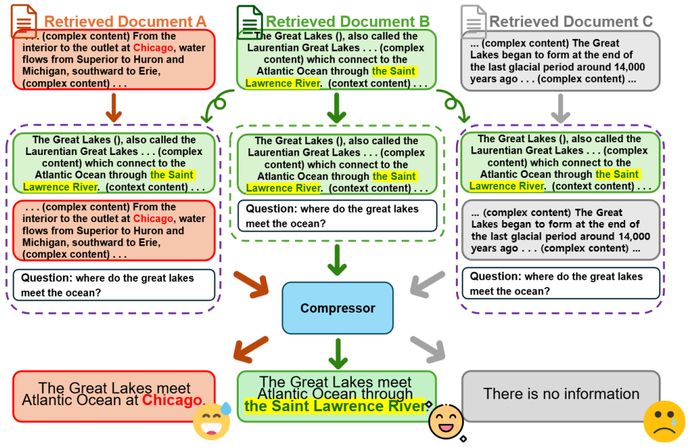
- Retrieval Augmented Generation (RAG)[1] introduces high computational cost when processing long retrieved context

Research Issue

- Existing abstractive compression methods [2] often omit information relevant to the query
  - High attention to irrelevant or incorrect information can cause useful evidence to be overlooked (attention dilution)
  - Long contexts are affected by position bias ("Lost in the Middle"), making relevant information harder to utilize depending on where it appears



A framework based on abstractive compression (RECOMP)



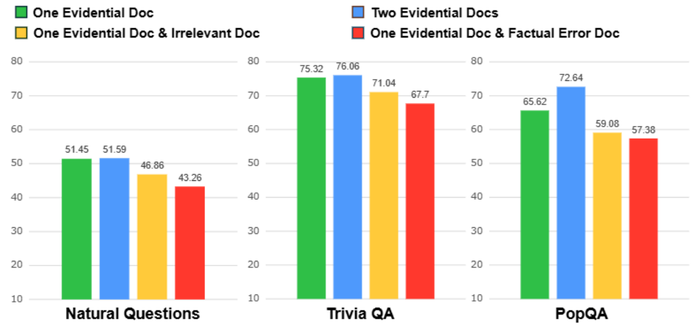
When retrieved documents contain noise, abstractive compression may omit answer-relevant information

[1] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in NeurIPS, 2020.
[2] F. Xu, W. Shi, and E. Choi, "RECOMP: Improving Retrieval-Augmented LMs with Compression and Selective Augmentation," in ICLR, 2024.

Analysis of Retrieval Documents

Classifying Retrieved Document

- Analyze how selecting documents for abstractive compression affects LLM performance
- Heuristically distinguish helpful documents from distracting ones
  - Define documents containing the answer string as Evidential Documents, since they are likely to provide direct evidence to the answer
  - Split non-evidential documents into:
    - Irrelevant Document: contain information unrelated to the query
    - Factual Error Document: contain misleading information that may lead to an incorrect answer



Exact Match (EM) scores for different types of noise documents, including irrelevant documents and factual error documents

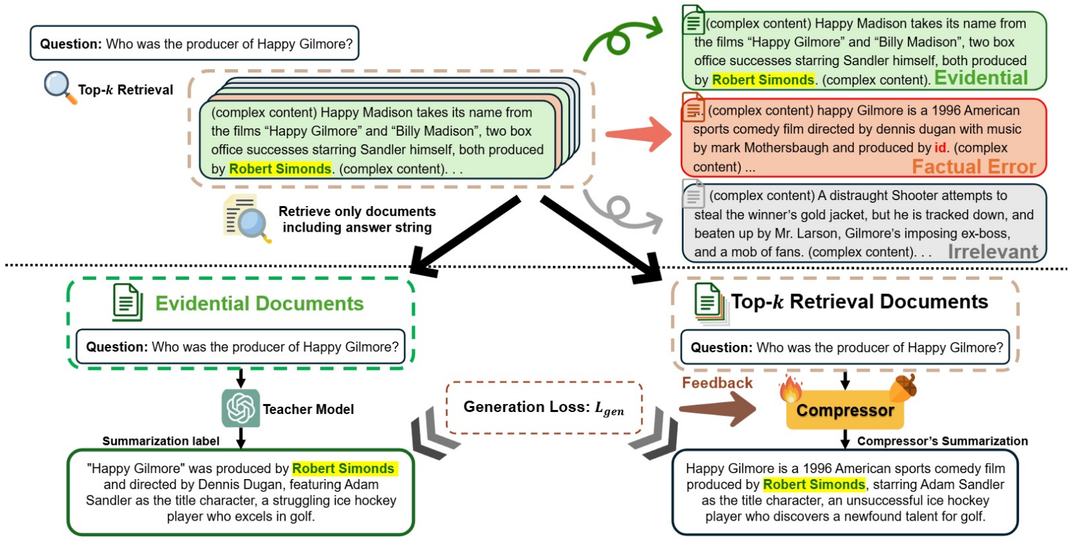
Method

Training Data Construction

- From retrieved documents, pass only Evidential documents to a teacher LLM to generate summarization labels

Supervised Fine-tuning(SFT)

- Fine-tune a T5 encoder-decoder model using the generated summarization labels



Overview of Abstractive Compression Robust against Noise (ACoRN)

Experiments & Results (1/2)

Strong Performance on Open-Domain Question Answering (ODQA) Benchmarks

- Higher QA accuracy than using uncompressed context (EM↑)
- Better compression efficiency by preserving only essential information (CR↓)
  - $CR = \frac{\#\ Compressed\ Tokens}{\#\ Retrieved\ Document\ Tokens}$
- Reduced input length to the LLM, resulting in lower inference time (Inference Time↓)

Results in Natural Questions, TriviaQA, PopQA Benchmarks

Method	NQ				TriviaQA				PopQA			
	EM (↑)	F1(↑)	CR(↓)	Inference Time(↓)	EM (↑)	F1(↑)	CR(↓)	Inference Time(↓)	EM (↑)	F1(↑)	CR(↓)	Inference Time(↓)
No Retrieval	19.94	31.46	-	0.249	49.65	59.59	-	0.162	20.51	23.68	-	0.132
Top-1 document	27.87	42.12	-	0.241	48.21	61.67	-	0.216	17.87	21.01	-	0.154
Top-5 documents	31.83	47.40	-	0.444	57.90	69.38	-	0.246	41.46	51.02	-	0.321
LongLLMLingua [22]	26.84	42.72	0.562	0.396	54.45	66.59	0.562	0.297	38.24	47.87	0.597	0.238
Quito [38]	29.94	44.04	0.484	0.316	56.61	67.61	0.484	0.258	41.89	50.36	0.523	0.241
RECOMP [12]	32.58	45.42	**0.052**	0.225	**58.64**	68.40	0.049	0.174	-	-	-	-
ACoRN (T5-large)	34.97	47.70	0.064	**0.205**	58.20	68.33	0.049	0.174	45.60	52.38	**0.059**	**0.137**
Flan-T5-large [29]	31.97	45.15	0.133	0.227	57.34	67.72	0.085	0.183	42.89	48.80	0.071	0.142
ACoRN (Flan-T5-large)	**35.56**	**48.48**	0.065	0.208	58.33	**68.58**	0.056	0.174	**45.75**	**52.82**	0.062	0.139

Experiments & Results (2/2)

Evaluation under Noisy Retrieval

- Training with fine-grained retrieval-noise categories minimizes performance degradation under noisy retrieval conditions

Reduced Keyword Loss

- Achieved a higher Preserving Answer String Ratio (PAR) than existing abstractive compression methods such as RECOMP
  - $PAR = \frac{\#\,(Compressed\ Outputs\ containing\ the\ Answer\ String)}{\#\,(Queries\ with\ at\ least\ one\ Evidential\ Document)}$
- Achieved preservation rates comparable to extractive compression methods such as token pruning
  - LongLLMLingua [1] : Removes low-importance tokens based on perplexity-driven token importance

Performance changes on Natural Questions, TriviaQA, PopQA with different types of injected noise

Method	NQ		TriviaQA		PopQA	
	EM (↑)	F1 (↑)	EM (↑)	F1 (↑)	EM (↑)	F1 (↑)
ACoRN (T5-large)	34.79 → **32.68** (**-2.11**)	**47.70** → **45.59** (**-2.11**)	58.20 → 57.31 (**-0.89**)	68.33 → **67.35** (**-0.98**)	**45.60** → **44.39** (**-1.21**)	**52.38** → **50.67** (**-1.71**)
w/o noise documents augmentation	**35.15** → 32.54 (-2.61)	47.65 → 44.88 (-2.77)	**58.57** → **57.38** (-1.19)	**68.54** → 67.23 (-1.31)	44.10 → 41.82 (-2.28)	50.37 → 47.88 (-2.49)

Compression Ratio (CR) and Preserving Answer String Ratio (PAR) on Natural Questions, TriviaQA, and PopQA with different types of injected noise

Method	NQ		TriviaQA		PopQA	
	CR(↓)	PAR(↑)	CR(↓)	PAR(↑)	CR(↓)	PAR(↑)
Token pruning						
LongLLMLingua	0.5607	0.6320	0.5625	0.7044	0.5963	0.7760
Abstractive compression						
RECOMP	**0.0557**	0.4966	0.0522	0.6283	-	-
ACoRN (T5-large)	0.0637	0.6425	0.0522	0.6380	0.0675	0.7521
ACoRN (Flan-T5-large)	0.0663	**0.6687**	0.0579	**0.6686**	**0.0674**	**0.7635**

[1] H. Jiang et al., "LongLLMLingua: Accerlating and Enhancing LLMs in Long Context Scenarios via Prompt Compression," in ACL, 2024.

Contributions & Limitations

Contributions

- Proposed robust abstractive compression considering fine-grained retrieval noise
  - Improved robustness to retrieval noise through noise-aware training
- Demonstrated the effectiveness of Evidential Document-based summarization labels
- Improved key-information preservation and ODQA performance over prior methods

Limitations

- Relies on a heuristic definition of Evidential Documents based on answer-string matching
- Scalability to multi-hop and multi-document reasoning remains unverified

05 · PROJECT FIVE

AI Assistant for elementary teachers using Large Language Models

[i-scream edu] Industry-Academia Research Project

[i-scream edu] Education LLM Assistant

Overview

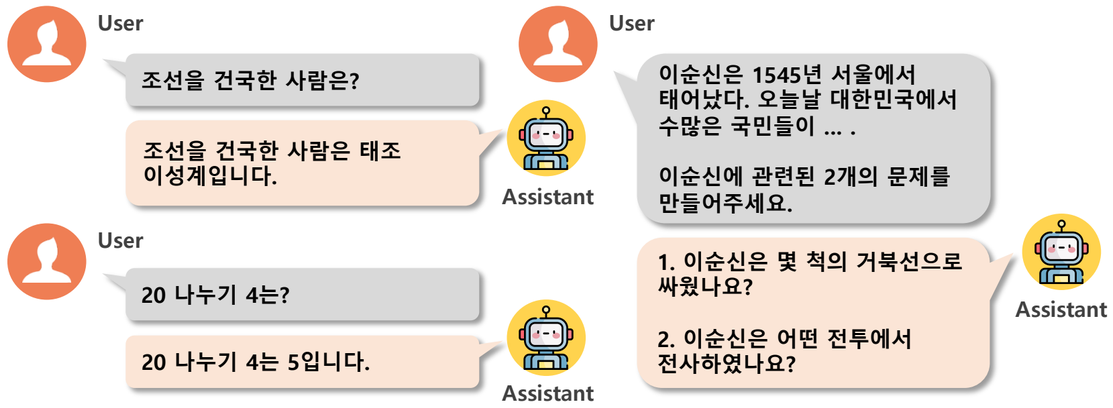
2023.09~2024.01

Research Goal

- Develop an LLM-based assistant that answers teachers' question based on the Korean elementary-school curriculum

Research Summary

- Injected knowledge of the Korean elementary curriculum, which was underrepresented in open-source LLMs
- Develop an assistant that generates curriculum-grounded responses to teachers' questions
- Supported question generation from user-provided content

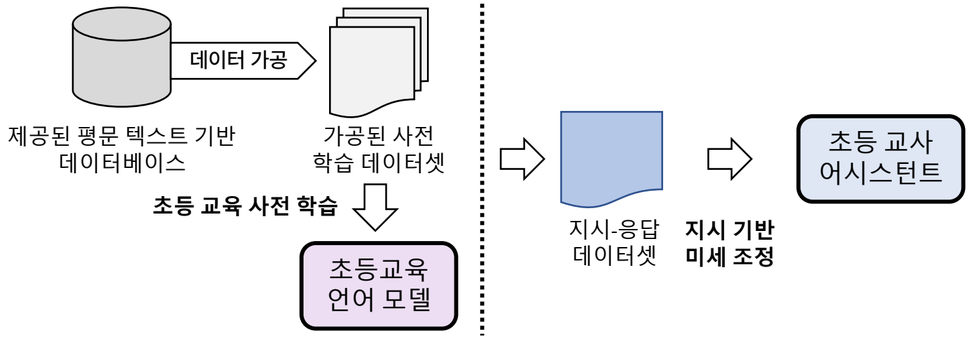


An Example of LLM based Assistant

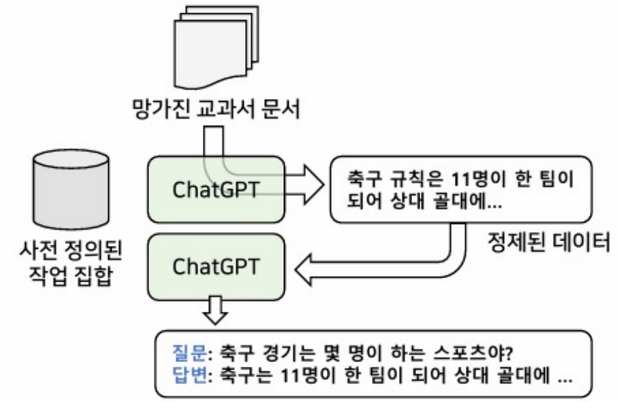
Framework

Model Training

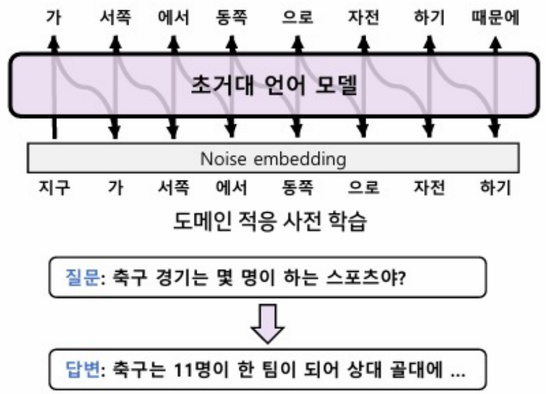
- Training using Poly-ko-12.8[1] as the base model
  - Despite relatively limited reasoning ability, the model has strong Korean-language and culturally aligned capabilities
- Domain Adaptive Pre-training (DAPT)[2]
  - To inject textbook knowledge not covered during backbone pretraining
- Instruction-following tuning (IFT)
  - Applied IFT for user-directed response generation
  - Used NEFTune[3] to improve robustness to input noise



Framework of language modeling



Process of building the IFT Dataset



Instruction-following tuning (IFT) with NEFTune

[1] H. Ko et al., “A Technical Report for Polyglot-Ko: Open-Source Large Scale Korean Language Models,” arXiv:2306.02254
[2] S. Gururangan et al., “Don’t Stop Pretraining: Adapt Language Models to Domain and Tasks,” in ACL, 2020.
[3] N. Jain et al., “NEFTune: Noisy Embedding Improve Instruction Finetuning,” in ICLR, 2024.

Pretraining Dataset

Pretraining Data Construction

- Built a pretraining corpus from provided plain-text databases containing textbooks and supplementary materials
  - Split documents by paragraph or problem unit
- Cleaned and curated data based on elementary curriculum knowledge
  - Conditioned training data on grade, subject, and curriculum unit

제목: 자전거 타기 성공. 오랜만에 날씨가 화창했다. 현관에 세워 놓은 자전거가 보였다. 어머니께서 내 마음을 아셨나 보다. 나는 어머니를 바라보며 기운차게 말했다.\n“오늘은 꼭 성공할 거에요!”\n어머니께서는 웃으며 고개를 끄덕이셨다. 점심을 먹은 뒤에 어머니와 함께 놀이터로 나갔다.\n … \n**국어 과목, 2학년 1학기, 1단원:[단원명]**

3. 나눗셈식을 이용하여 15의 약수를 알아봅시다. 1) 15의 약수를 모두 찾아보세요. $15 [빈칸] = 15$ ₩n … $15 / [빈칸] = 1$ ₩n 15의 약수를 작은 수부터 모두 써 보세요. ₩n …₩n **수학 과목, 3학년 2학기, 2단원:[나눗셈]**

An example of pretraining data

Instruction-following tuning Dataset

Instruction-following Tuning Data Construction

- Built instruction-following data for tasks such as question answering, and question generation[1]
  - Fine-tuned the model to better interact with users and follow diverse instructions

Cross-System Prompting with External Data

- Leveraged public Korean instruction datasets to improve general instruction-following ability
- Used perplexity to filter instructions or responses outside the elementary curriculum domain

USER: 6살 아이에게 축구 규칙에 대해 설명해줘.

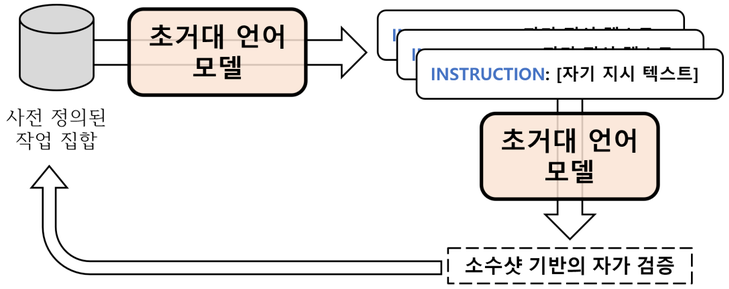
ASSISTANT: 축구는 11명이 한 팀이 되어 상대 골대에 ...

지시 기반 데이터셋 내 질의 응답의 예시

USER: 초등 학생용 축구 규칙에 대한 문제를 만들어줘.

ASSISTANT: 축구 경기는 몇 명이 하는 스포츠인가요?

지시 기반 데이터셋 내 문제 생성의 예시



Instruction-Following Data Construction Pipeline

SYSTEM: 초등 교육 교과과정을 기반으로 응답하는 어시스턴트입니다.

USER: 축구 경기는 몇 명이 하는 스포츠야?

ASSISTANT: 축구는 11명이 한 팀이 되어 상대 골대에 ...

초등 교과 과정 기반 지시 데이터 예시

SYSTEM: 지시나 질문에 응답하는 어시스턴트입니다.

USER: 라그랑주 역학에서 라그랑지언이 뭐야?

ASSISTANT: 계의 동역학을 나타내는 함수다. 라그랑주 역학에서는 계의 상태를 일반화 좌표와 일반화 속도로 나타내므로…

외부 지시 데이터 예시

SYSTEM: 초등 교육 교과과정을 기반으로 응답하는 어시스턴트입니다.

USER: 라그랑주 역학에서 라그랑지언이 뭐야?

ASSISTANT: 초등 교육 교과과정을 벗어나는 지식이 필요한 질문입니다.

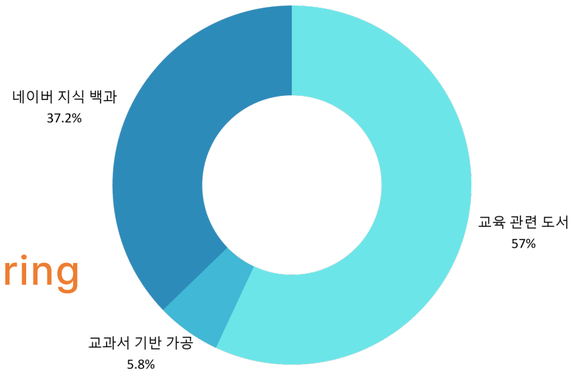
교차 시스템 프롬프팅

[1] L. Ouyang et al., “Training language models to follow instructions with human feedback,” in Proceedings of Advances in Neural Information Processing Systems (NeurIPS), 2022.

Dataset

Data Cleaning

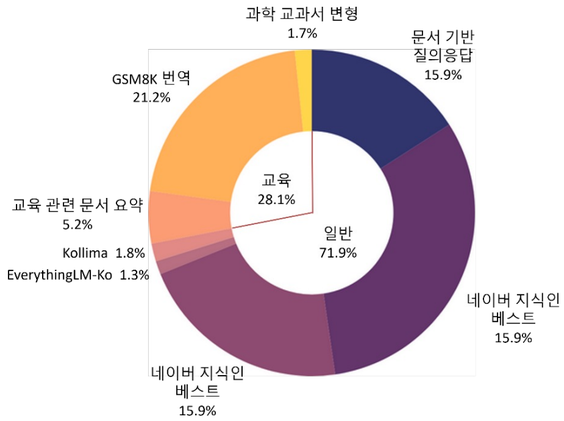
- Repaired OCR-corrupted textbook documents through LLM-based rewriting and filtering
  - OCR errors such as missing ‘/’ symbols for fractions and division
    - Applied manual filtering because the problems were relatively easy to review



Composition ratio of DAPT dataset

Dataset Composition

- DAPT data (57,129)
  - Curated textbook data: 3,325
  - Naver web-crawled data: 21,280
  - Education-related books: 32,524
- Instruction-following tuning data (58,143)
  - General-domain data: 41,850
  - Education-domain data: 16,293



Composition ratio of instruction data

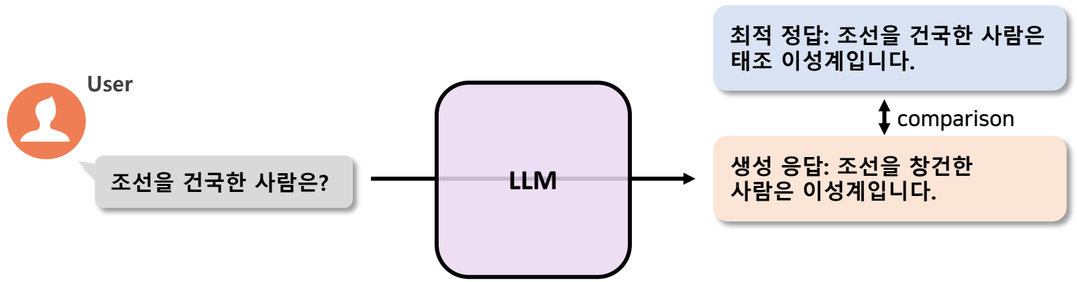
Evaluation of Problem-Solving Performance

Open-Ended Question Evaluation

- Evaluated answer-generation quality using BLEU (Bilingual Evaluation Understudy) Score[1]
  - Measured lexical overlap between generated and reference answers while considering sentence length and overlap
- Evaluated semantic similarity using BERT (Bidirectional Encoder Representations from Transformer) Score[2]
  - Measured precision and recall based on contextual embedding similarity between generated and reference answers
  - Evaluated the semantic and syntactic quality of generated response

Multiple-Choice Evaluation

- Provided the model with context, question, and answer choices
  - Used a standard multiple-choice QA format
  - Evaluated accuracy based on the selected answer choice



Analysis of response accuracy for generated open-ended questions

[1] K. Papineni et al., “BLEU: a method for automatic evaluation of machine translation,,” ACL, 2002.
[2] T. Zhang et al., “BERTScore: Evaluating Text Generation with BERT,” arxiv, 1904.09675

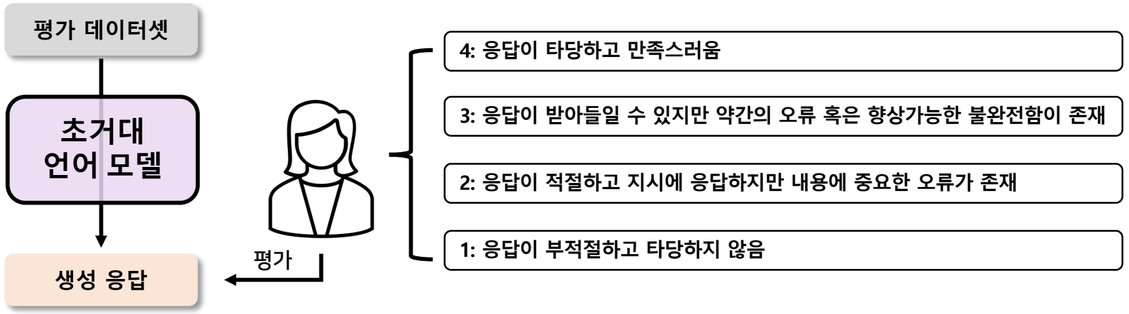
Evaluation

User Satisfaction Evaluation

- Conducted user evaluation of instruction-following response on a 1~4 scale
  - Built test instruction datasets covering diverse teacher-facing tasks
  - Users evaluated generated responses according to predefined criteria

Model Capability

- Strong performance in both core-subject problem solving and teacher administrative tasks



User Satisfaction Evaluation



Strong performance on notice-generation taks

06 · PROJECT SIX

AI Agents for CBRN Operational Command and Guidance

[Agency for Defense Development (ADD)] National R&D Project

[ADD] AI Agent for CBRN Operational Command and Guidance

Overview

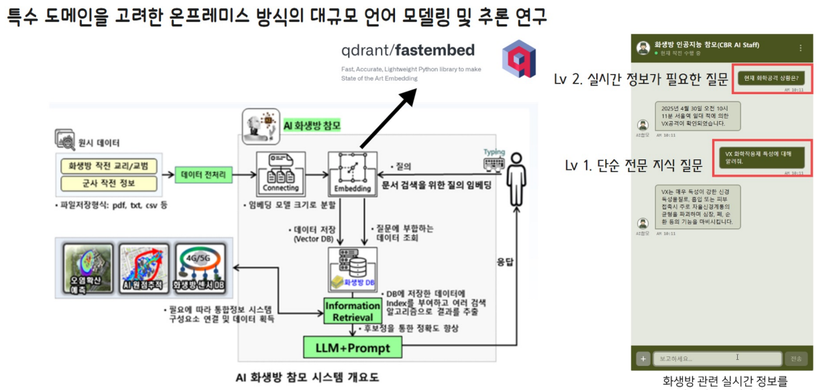
2023.09~2023.11 (Project proposal), 2024.12~2025.05 (Project development)

Research Goal

- Develop an AI agent that analyzes real-time CBRN detection data and provides decision support and relevant information

Research Summary

- Developed an agent framework for real-time CBRN situation analysis and detection support
- Explored efficient serving frameworks for on-premise deployment
- Built a demo visualizing retrieval and real-time NBC-RAMPS detection workflow



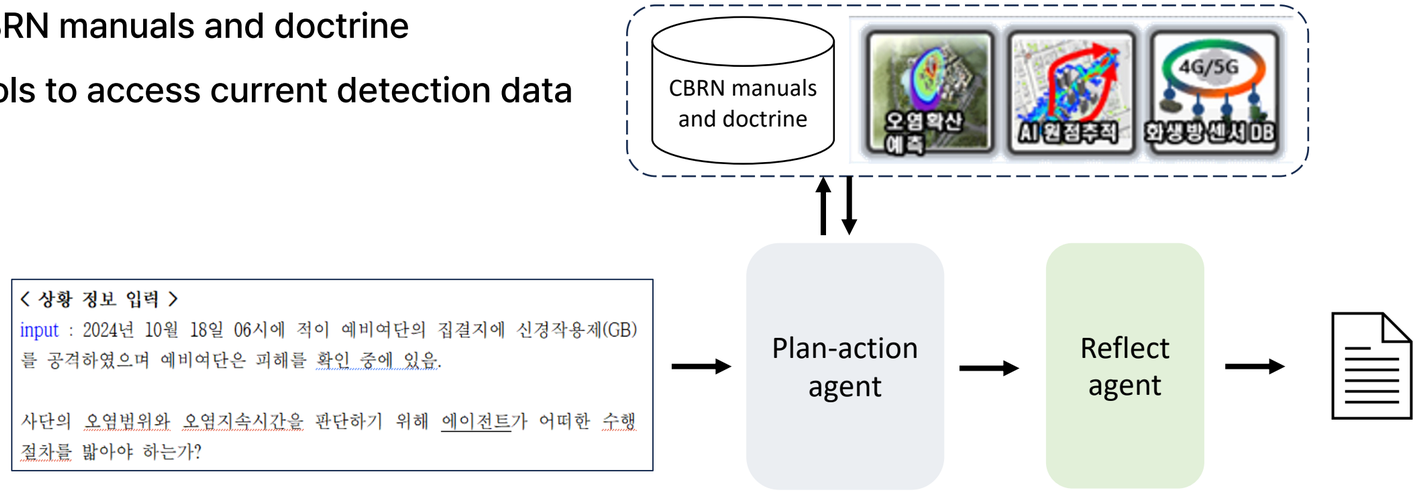
Framework

Key Requirement

- Fast token generation was critical, with TTFT (Time to First Token) as a key evaluation metric

Agent Architecture

- Built the initial baseline based on AutoAct [1]
- Routed queries based on information needs:
  - Domain knowledge: Retrieve relevant CBRN manuals and doctrine
  - Real-time information: Call dedicated tools to access current detection data



The overview of framework

[1] S. Qiao et al., "AutoAct: Automatic Agent Learning from Scratch for QA via Self-Planning," in ACL, 2024.