RAISE: 단계별 검색을 통한 거대 언어 모델의 과학적 추론 능력 향상

오민해, 이정우*
minhae.oh@cml.snu.ac.kr. *junglee@snu.ac.kr

RAISE: Enhancing Scientific Reasoning in LLMs via Step-by-Step Retrieval

Minhae Oh, JungWoo Lee*
Seoul National Univ.

요 약

과학적 추론은 복잡한 논리적 과정과 전문 용어에 대한 이해, 그리고 최신 지식의 적용을 필요로 한다. 거대 언어 모델(LLM)은 이러한 작업에서 환각문제를 겪거나 깊이 있는 추론에 한계를 보인다. 이를 해결하기 위해 본 논문에서는 단계별 검색 증강 프레임워크인 RAISE 를 제안한다. RAISE 는 문제를 하위 질문으로 분해하고, 추론 의도를 반영한 논리적 질의를 생성하며, 이를 통해 표면적 유사성이 아닌 논리적 연관성이 높은 문서를 검색한다. GPQA, MMLU 등 주요 벤치마크에서의 실험 결과, RAISE 는 기존의 RAG 및 추론 전략들보다 일관되게 우수한 성능을 보였으며, 특히 과학적 문제 해결에 필수적인 논리적 근거를 효과적으로 검색함을 확인하였다.

Ⅰ. 서 론

거대 언어 모델(LLM)은 다양한 분야에서 뛰어난 성능을 보이고 있으나, 대학원 수준의 생물학이나 화학과 같이 고도화된 과학적 추론이 필요한 작업에서는 여전히 어려움을 겪는다. 이러한 문제는 주로 모델이 도메인 특화 용어를 이해하지 못하거나 최신 지식이 부족하기 때문에 발생한다.

이를 보완하기 위해 검색 증강 생성기법이 널리 사용되고 있다. 그러나 기존의 RAG 방식은 질문과 문서 간의 표면적인 의미 유사도에 의존하여 검색을 수행하기 때문에, 복잡한 다단계 추론이 필요한 과학 문제에서는 정작 문제 해결에 핵심적인 논리적 연결고리를 놓치는 경우가 많다. 또한, 문제를 해결하는 각 단계마다 필요한 지식이 상이함에도 불구하고 단일 검색에 의존하는 것은 성능 저하를 야기할 수 있다.

이에 본 논문에서는 RAISE (Step-by-Step Retrieval-Augmented Inference for Scientific rEasoning) 프레임워크를 제안한다. RAISE 는 복잡한 과학 문제를 다수의 하위 문제로 분해하고, 각 단계에서 단순 키워드 매칭이 아닌 '추론 의도'를 포함한 논리적 질의를 생성하여 검색을 수행한다. 이를 통해 모델은 위키피디아와 같은 비정형 데이터셋에서도 문제 해결에 실질적으로 도움이 되는 문서를 효과적으로 찾아낼 수 있다.

Ⅱ. RAISE 프레임워크

RAISE 는 크게 문제 분해, 논리적 질의 생성, 논리적 검색의 세 단계로 구성된다. 문제 분해 단계에서는 원본 질문 x를 해결 가능한 단위의 하위 질문들($r_1,\cdots,r_n$)과 그에 상응하는 초기 검색 질의($q_1,\cdots,q_n$)로 분해한다. 이는 기존의 Single-query 접근법과 달리, 복잡한 문제를 구조화하여 순차적으로 해결할 수 있는 기반을 마련한다. 단순히 분해된 질의 q_i 를 그대로 검색에 사용하는 것은 비효율적이다. 초기 질의는 추론의 맥락이 결여되어 있고, 하위 질문 r_i 자체는 검색에 너무 구체적이거나 노이즈가 될 수 있기 때문이다. 따라서 RAISE 는 q_i와 r_i를 결합하여 해당 단계의 추론 의도를 포착하는 '논리적 질의(q_i^*)를 생성한다. 재구성된 논리적 질의 q_i^* 를 사용하여 외부 코퍼스에서 관련 문서 D_i 를 검색한다. 이때 유사도 임계값을 적용하여 관련 없는 문서를 필터링한다. 이후 모델은 검색된 문서 D_i, 원본 질문 x, 그리고 이전 단계의 맥락을 종합하여 해당 하위 질문에 대한 답변 a_i를 생성한다. 이 과정은 모든 하위 질문이 해결될 때까지 반복되며, 최종적으로 종합된 답변 y를 도출한다.

Ⅲ. 실험결과

3.1. 실험 환경

본 연구에서는 과학적 추론 능력을 평가하기 위해 GPQA , SuperGPQA, MMLU 벤치마크의 일부를 사용하였다. 검색기(Retriever)로는 Natural Questions 로 훈련된 DPR(Dense Passage Retrieval)을 사용하였으며, 생성 모델로는 Mistral Small 3.1 과 LLaMA 3.1-8B 등을 활용하였다.

	GPQA	SuperGPQA	MMLU
CoT	42.42	8.71	42.27
CoT+ RAG	45.96	9.21	40.85
Least-to-Most	44.95	10.22	41.31
RAISE	51.01	13.4	46.21

표 1 GPQA, SuperGPQA, MMLU 벤치마크에서의 실험결과

3.2. 실험 결과

RAISE 의 성능을 검증하기 위해 Direct CoT, CoT+ RAG, 그리고 분해 기반 방법론인 Least-to-Most+ RAG 과 비교 실험을 수행하였다.

실험 결과(표 1 참조), RAISE 는 모든 벤치마크에서 베이스라인 모델들을 일관되게 상회하는 성능을 보였다. 이는 RAISE 가 단순히 도메인 지식을 검색하는 것을 넘어, 문제 해결에 필요한 논리적 근거를 효과적으로 수집함을 시사한다.

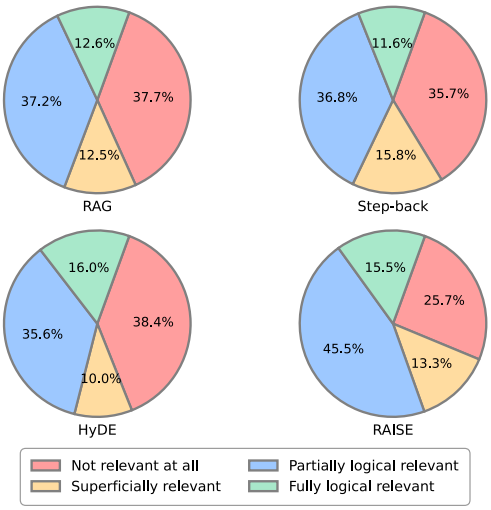

그림 1 정성적 분석(LLM-as-Judge)

3.3. 정성적 분석

LLM-as-a-judge 에 의한 정성 평가 결과, RAISE 는 기존 RAG 방식에 비해 '표면적으로만 관련된' 문서를 검색하는 비율이 낮고, '논리적으로 관련된' 문서를 검색하는 비율이 현저히 높았다. 예를 들어 화학 반응 문제에서 기존 RAG 가 단순 화합물 정의를 검색할 때, RAISE 는 반응 메커니즘이나 공식을 포함한 문서를 검색하여 정답 도출에 직접적인 도움을 주었다..

Ⅲ. 결론

본 논문에서는 과학적 추론을 위한 단계별 검색 프레임워크인 RAISE 를 제안하였다. RAISE 는 문제를 분해하고 논리적 질의를 생성함으로써, 비정형 데이터셋에서도 문제 해결에 필수적인 논리적 지식을 효과적으로 검색할 수 있음을 입증하였다. 향후 본 프레임워크는 수학적 문제 해결 등 정교한 다단계 추론이 요구되는 타 도메인으로 확장 가능할 것으로 기대된다.

ACKNOWLEDGMENT

This work is in part supported by the National Research Foundation of Korea (NRF, RS-2024-00451435(20%), RS-2024-00413957(20%)), Institute of Information & communications Technology Planning & Evaluation (IITP, RS-2025-02305453(15%), RS-2025-02273157(15%), RS-2025-25442149(15%) RS-2021-II211343(15%)) grant funded by the Ministry of Science and ICT (MSIT), Institute of New Media and Communications(INMAC), and the BK21 FOUR program of the Education, Artificial Intelligence Graduate School Program (Seoul National University), and Research Program for Future ICT Pioneers, Seoul National University in 2026.

참 고 문 헌

[1] Zilong Zhao, et al. Stepwise self-consistent mathematical reasoning with large language models. arXiv preprint arXiv:2402.17786, 2024.

[2] Patrick Lewis, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems, 33:9459-9474, 2020.

[3] Vladimir Karpukhin, et al. Dense passage retrieval for open-domain question answering.In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6769-6781, 2020.

[4] Jason Wei, et al. Chain-of-thought prompting elicits reasoning in large language models, 2022.

[5] Denny Zhou, et al. Least-to-most prompting enables complex reasoning in large language models. In The Eleventh International Conference on Learning Representations, 2023.

[6] David Rein, et al. Gpqa: A graduate-level google-proof q&a benchmark. In First Conference on Language Modeling, 2024.

[7] M-AP Team. Supergpqa: Scaling llm evaluation across 285 graduate disciplines. CoRR, 2025.

[8] Dan Hendrycks, et al. Measuring massive multitask language understanding, 2021.