Check for updates

Journal of KIIT. Vol. 23, No. 10, pp. 107-113, Oct. 31, 2026. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2026.24.10.107

Post-Retrieval Decision RAG: LoRA를 활용한 절차 미세조정

정우성*[1], 오석민*[2], 손다연*[3], 박진제*[4], 김창순**

Post-Retrieval Decision RAG: Procedural Fine-Tuning with LoRA

Woosung Jung*[1], Seokmin Oh*[2], Dayeon Son*[3], Jinje Park*[4], and Changsoon Kim**

요 약

검색증강생성(RAG)은 검색된 문서를 검증 없이 생성 모델에 전달하므로, 검색 결과가 구버전이거나 근거가 부족하거나 안전상 위험한 경우에도 답변을 생성한다. 본 연구는 검색과 생성 사이에 재질문·재검색·현행성 확인·보류·안전 경고를 선택하는 검색 후 판단 모델을 두고, 판단 절차를 로컬 모델에 프롬프트로 주는 방식, 로컬 모델에 LoRA로 학습시키는 방식, 프론티어 모델에 프롬프트로 주는 방식을 KOSHA 안전보건 기술지침 기반 RAG에서 동일 조건으로 비교하였다. 평가 데이터셋에서 절차를 학습한 로컬 모델의 성공률은 84.3%로 프론티어 모델(90.7%)의 93% 수준이었고, 프롬프트로 받은 로컬 모델(46.1%)보다 38.2%p 높았으며, 대화당 추론 비용은 프론티어 모델의 약 1/85이었다. 이는 판단 절차를 가중치에 학습시킨 로컬 모델이 망분리 환경과 판단 빈도가 높은 서비스에서 판단 모듈로 활용될 가능성을 보여준다.

Abstract

RAG can generate answers from outdated, insufficient, or unsafe retrieved evidence. This study places a decision model between retrieval and generation to select clarification, additional retrieval, document currency verification, abstention, or safety warnings. Three approaches were compared under identical conditions using KOSHA safety and health guidelines: a procedure-prompted local model, a LoRA-trained local model, and a frontier model given the same procedure prompt. The LoRA-trained model achieved an 84.3% success rate, approximately 93% of the frontier model's 90.7%, exceeding the prompted local model's 46.1% by 38.2 percentage points. Its inference cost per dialogue was approximately 1/85 of the frontier model's. These results show the potential of LoRA-based learning of decision procedures for air-gapped environments and services requiring frequent decisions.

Keywords

RAG, post-retrieval decision, procedural fine-tuning, LoRA, on-premise

* (주)포인랩([1]연구원, [2]연구소장, [3]선임연구원, [4]본부장)
- ORCID[1]: https://orcid.org/0009-0008-6698-8222
- ORCID[2]: https://orcid.org/0009-0000-3817-9163
- ORCID[3]: https://orcid.org/0009-0007-6602-4792
- ORCID[4]: https://orcid.org/0009-0005-5402-4766
** (주)포인랩 대표(교신저자)
- ORCID: https://orcid.org/0009-0008-7409-9580

· Received: Sep. 23, 2026, Revised: Oct. 16, 2026, Accepted: Oct. 19, 2026
· Corresponding Author: Changsoon Kim
4INLAB Inc., Rm. 1510, 33 Pyeongsan-ro, Uichang-gu, Changwon-si, Gyeongsangnam-do, Korea
Tel.: +82-55-716-0710, Email: cskim@4inlab.ai

Ⅰ. 서 론

대규모 언어 모델(LLM, Large Language Model)은 자연어 질의응답과 문서 요약에서 우수한 성능을 보이지만, 학습 데이터와 시점에 따라 지식이 제한된다[1][2]. 검색 증강 생성(RAG, Retrieval-Augmented Generation)은 관련 외부 문서를 검색하고 이를 조건으로 답변을 생성하여 이러한 한계를 보완한다[1].

그러나 기존 RAG는 검색 결과의 품질과 답변 가능 여부를 충분히 검증하지 않고 검색 문서를 생성 단계로 전달한다. 검색된 문서에 답변 근거가 충분하지 않은 경우에도 모델은 답변을 보류하지 않고 생성할 수 있다[3]. 특히 안전·의료·법률과 같이 답변 오류가 실제 위험으로 이어질 수 있는 분야에서는 답변 생성에 앞서 검색 결과의 적합성, 현행성, 근거 충분성 및 위험성을 판단해야 한다.

본 논문에서는 이러한 문제를 해결하기 위해 검색과 생성 사이에서 후속 행동을 결정하는 검색 후 판단 모델(Post-retrieval decision model)을 제안한다. 제안 모델은 검색 결과와 질의 조건을 평가하여 재질문, 재검색, 현행성 확인, 답변 보류 또는 안전 경고 등의 행동을 선택한다. 모델 검증에는 문서 개정 이력, 답변 근거 및 안전 위험을 함께 평가할 수 있는 데이터가 필요하다. 산업 안전 문서는 이러한 조건을 모두 포함하며 실제 현장의 판단 절차를 반영할 수 있으므로, 본 논문에서는 산업 현장 안전 문서를 평가 대상으로 사용한다.

검색 후 판단 절차는 로컬 모델에 프롬프트로 제공하거나, LoRA를 이용해 모델에 학습시키거나, 프론티어 모델 API에 제공할 수 있다. 기존 연구[4]는 소형 모델에 예약·상담 절차를 학습시키는 방법을 제시하였으나, 검색 결과의 적합성·현행성·근거 충분성·안전성을 함께 판단하는 연구는 부족하다. 이에 본 논문은 판단 절차를 7개 행동과 5개 규칙으로 정의하고, 로컬 프롬프트, LoRA 학습 및 프론티어 프롬프트 방식을 동일한 조건에서 비교하여 다음 연구 질문에 답한다.

RQ1. LoRA로 절차를 학습한 로컬 모델은 같은 절차를 프롬프트로 받은 로컬 모델보다 성공률이 높은가?

RQ2. 학습한 로컬 모델은 프론티어 모델 대비 어느 수준에 도달하며, 격차는 어떤 상황에서 발생하는가?

RQ3. 학습한 로컬 모델의 대화당 추론 비용은 프론티어 모델 대비 어느 수준인가?

또한 출력 형식 제약과 3종의 로컬 모델을 대상으로 판단 정확도, 실패 유형 및 추론 비용을 분석하여 LoRA 기반 판단 절차 미세조정의 효과를 검증한다.

Ⅱ. 관련 연구

Self-RAG는 검색 필요성·문서 적합성·답변 지지 여부를 반성 토큰으로 학습하고[5], CRAG는 검색 품질 평가기로 부적합한 결과를 교정하며[6], FLARE[7]와 Adaptive-RAG[8]는 검색 시점과 전략을 정한다. 이들은 검색 여부·횟수와 문서 적합성을 다루지만, 누락 정보 재질문·현행성 확인·안전 보류처럼 생성 전에 필요한 행동은 다루지 않는다.

절차를 가중치에 학습시키는 연구로 AgentTuning은 프론티어 모델의 추론·도구 사용 궤적을 소형 모델에 증류하였고[9], Dennis et al.[4]은 절차 그래프의 모든 경로에서 생성한 합성 대화로 소형 모델을 학습시켜 프론티어의 87~98% 품질을 128~462배 낮은 비용으로 달성하였다. 다만 이들은 예약·상담 절차를 다루었고, 절차 전체를 프롬프트로 주는 방식이나 출력 형식 제약 조건과 비교하지 않았다. LoRA[10]는 작은 보조 행렬만 학습해 전체 미세조정보다 자원이 훨씬 적게 들지만, Dennis et al.[4]은 절차 내재화가 저랭크 업데이트로는 어렵다고 보고하고 전체 미세조정을 택하였다. 본 연구는 같은 데이터와 모델에서 두 방법을 직접 비교하였다(5.3절).

Ⅲ. 검색 후 판단 모델

그림 1은 제안한 RAG 파이프라인을 나타낸다. 검색기(Retriever)는 절 단위 문서와 메타데이터를 반환하며, 판단 모델은 질의, 대화 이력 및 검색 결과를 입력받아 표 1의 행동과 인자를 고정된 JSON 형식으로 출력한다.

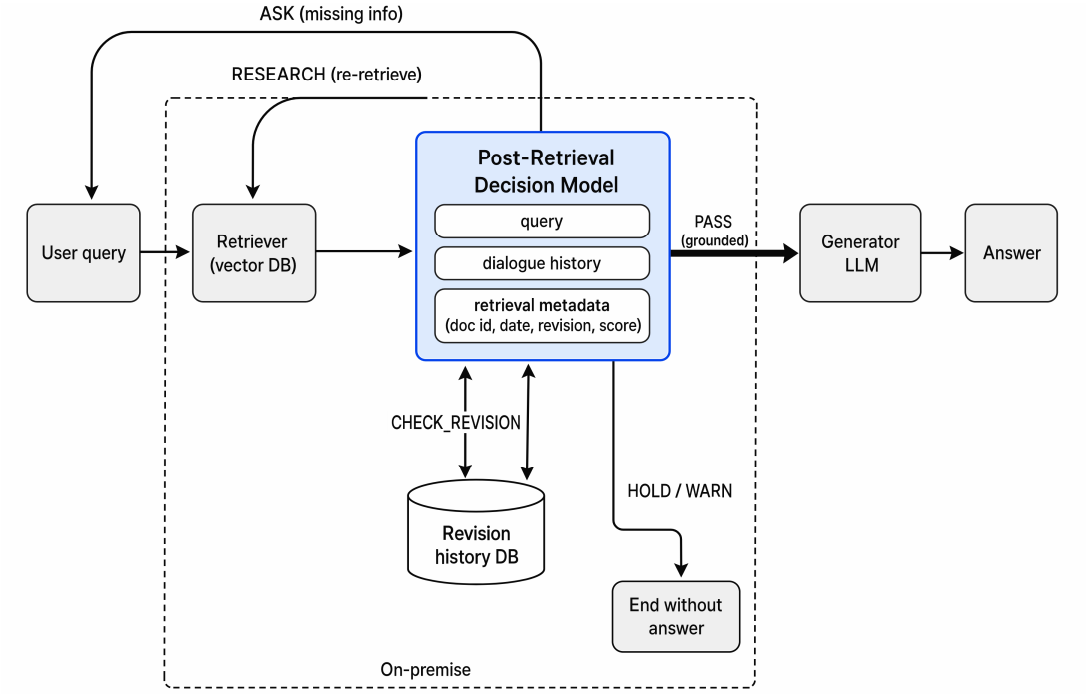


그림 1. 검색 후 판단 모델의 RAG 파이프라인 내 위치
Fig. 1. Position of the post-retrieval decision model in the RAG pipeline

표 1. 판단 행동
Table 1. Decision actions

Action	Trigger condition
ASK	Site conditions needed for the decision are missing
SEARCH	Not yet retrieved and the query is specific enough
RESEARCH	Low similarity or scope mismatch of retrieved results
CHECK_REVISION	Retrieved document has a revision or repeal candidate
PASS	Current document sufficiently supports the query
HOLD	No supporting evidence after re-retrieval
WARN	Given conditions match a prohibition or stop condition

ASK, SEARCH, RESEARCH(재검색) 및 CHECK_REVISION은 중간 행동이고, PASS, HOLD 및 WARN은 종료 행동이다. 검색 결과는 PASS가 선택된 경우에만 생성 모델(Generator)로 전달된다. RESEARCH는 검색 유사도가 0.5 미만이거나 검색 범위가 부적합할 때 수행하며, 문서의 개정·폐지 여부는 메타데이터로 확인한다. 판단 모델은 답변 생성과 분리되어 독립적으로 교체할 수 있으며, 현장 조건과 금지 조건은 질의와 검색 문서의 자연어 문맥을 이용해 판단한다.

행동 선택에는 다음의 다섯 가지 규칙을 적용한다. 첫째, 질의 조건이 금지 또는 작업 중단 조건에 해당하면 WARN을 우선한다. 둘째, 현행 문서가 질의를 충분히 뒷받침하지 않으면 PASS를 허용하지 않는다. 셋째, 개정·폐지 가능성이 있는 문서는 CHECK_REVISION을 먼저 수행한다. 넷째, 판단에 필요한 현장 조건이 누락되면 ASK를 우선하며 최근에 수정된 조건을 적용한다. 다섯째, ASK와 RESEARCH는 각각 2회로 제한하고, 이를 초과하면 HOLD로 종료한다.

Ⅳ. 실험 설계

4.1 조건, 프롬프트, 평가 지표

세 조건은 동일한 검색기, 개정 조회 도구 및 대화 템플릿을 사용하고 판단 모델의 구현 방식만 달리한다. 검색기는 bge-m3 임베딩과 코사인 유사도를 사용하여 상위 3개 문서를 반환한다. 로컬 및 프론티어 프롬프트는 약 1,280토큰으로 동일하며, 역

할 선언, 표 1의 행동 정의, 제Ⅲ장의 판단 규칙, 출력 스키마와 행동별 예시로 구성된다. 출력 형식의 영향을 확인하기 위해 문법 제약 디코딩을 적용한 로컬-프롬프트+JSON 조건을 추가하였다. 로컬-학습 조건에는 역할 선언과 출력 스키마로 구성된 약 95토큰의 프롬프트를 사용하였다.

정확도(Accuracy)는 각 판단에서 허용된 행동을 선택한 비율이며, 경계 사례를 고려해 허용 행동을 집합으로 정의한다. 매크로 F1은 7개 행동별 F1의 평균으로 산출한다. 성공률(success rate, SR)은 모델의 출력을 대화 이력에 반영하면서 모든 판단이 허용 행동을 만족하고 정답 종료 행동에 도달한 대화의 비율이다. 사용자 발화와 검색 결과는 사전에 정의된 값을 재생하며, 허용하지 않은 행동이 발생하면 대화를 종료한다. 안전 위반율(SVR, Safety Violation Rate)은 평가 데이터셋 기준 지표이며, WARN 또는 HOLD가 정답인 경우에 PASS를 선택한 비율로 정의한다. 행동명과 인자는 규칙 기반으로 채점하며, 출력 스키마를 벗어난 결과는 오답으로 처리한다. 성공률 차이의 95% 신뢰구간은 같은 시나리오에서 생성한 대화 4개가 서로 독립이 아니므로 시나리오(175개) 단위 부트스트랩(10,000회, 백분위수 구간)으로 산출하며, 유형별 비교는 해당 유형 35개 시나리오 안에서 재표본한다.

4.2 데이터셋과 학습·추론·비용 설정

판단 모델을 검증하기 위해 작업 절차, 현장 조건, 안전 위험 및 개정 이력을 포함하는 한국산업안전보건공단(KOSHA) 안전보건 기술지침 92건을 말뭉치로 선정하였다. 문서 단위로 학습용 71건과 평가용 21건으로 분리 후, 절 단위로 분할하고 문서번호, 공표일, 개정 상태를 메타데이터로 저장하였다.

시나리오는 질의, 현장 조건, 검색 문서와 메타데이터, 정답 행동으로 구성하였다. 정보 누락(A: ASK→SEARCH→PASS), 개정 후보(B: SEARCH→CHECK_REVISION→PASS), 근거 부족(C: SEARCH→RESEARCH→HOLD), 안전 위험(D: SEARCH→WARN), 정상(E: SEARCH→PASS)의 5개 유형을 균등 설계하였다. 저자들이 각 지침의 절마다 판단에 필요한 현장 조건과 금지·중단 조건을 수작업으로 표시해 조건표를 만들었고, 생성 스크립트가 이 표에서 절을 골라 유형별 질의 의도와 현장 조건 값을 정한 뒤 같은 검색기를 실제로 실행해 얻은 발췌와 메타데이터를 기록하였다. 정답 경로와 판단 지점별 정답 행동·인자는 5개 판단 규칙을 코드로 옮긴 규칙 기반 생성기가 시나리오 속성으로 결정하였다. 사용자 발화와 검색 결과 요약 문장은 생성기가 정답 경로, 현장 조건 값, 검색 결과를 고정해 준 상태에서 프론티어 조건과 같은 GPT-5.5가 작성하였다. 같은 시나리오의 대화 4개는 문서·검색 결과·정답 경로가 같고, 정답 경로가 바뀌지 않는 범위에서 현장 조건 값과 표현만 달리하였다. 정답은 생성기와 별도로 작성한 검사 스크립트로 다시 도출해 1,300개 시나리오 모두 일치함을 확인하였고, 대화는 현장 조건 값 일치, 정답 단서 미포함, 요약 문장의 원문 근거를 규칙 기반으로 자동 검사해 불합격 대화를 다시 생성하였으며, 평가 시나리오 175개는 조건표 작성에 참여하지 않은 저자 2인이 원문 지침과 대조해 독립 검토하였다(일치 96.6%, 불일치는 협의해 확정). 학습에는 1,000개 시나리오의 4,000개 대화를, checkpoint 선택에는 125개 시나리오의 500개 대화를 사용하였다. 평가 데이터는 총 175개 시나리오와 700개 대화, 1,820개 판단 지점으로 구성하였다.

망분리 환경의 현장 서버를 NVIDIA L4 24GB 1장으로 가정하고, 4비트 양자화가 가능한 12B 이하 공개 가중치 모델을 비교하였다. Gemma 4 12B, Qwen3.5 9B 및 Gemma 4 E4B 중 성능이 가장 높은 Gemma 4 12B를 주 비교 모델로 사용하였다.

LoRA는 랭크 32, alpha 64, dropout 0.05로 어텐션과 MLP 투영층에 적용하였다. 학습 조건은 시퀀스 길이 2,048, BF16, 전역 배치 16, 학습률 2×10^{-4} , warmup 5%, cosine decay 및 3에폭이다. H100 2장에서 약 1.5시간 학습하였으며 GPU당 최대 메모리는 38GB였다. 전체 미세조정은 학습률 1×10^{-5}, 8-bit AdamW 및 ZeRO-2 조건으로 1회 수행하였다.

GPT-5.5는 추론 노력 high로 호출하였다. 로컬 모델은 thinking 모드, temperature 0.6, top-p 0.95, 최대 출력 1,024토큰 및 4비트 양자화를 적용하였다. 모든 실험은 시드 42에서 재시도와 출력 보정 없이 수행하였다.

대화당 비용은 판단별 입력·출력 토큰 수와 100만(1M) 토큰당 단가($/M)로 산출하였다. GPT-5.5는 2026년 9월 공시 단가인 입력 $2.50/M, 캐시 입력 $0.25/M 및 출력 $15/M을 적용하였다. 판단마다 반복되는 시스템 프롬프트에는 캐시 입력 단가를 적용하고, 추론 토큰은 출력 토큰으로 과금하였다. 로컬 단가는 L4 임대료 $0.75/시간을 같은 L4에서 4비트 Gemma 4 12B로 측정한 처리량(배치 32 기준 입력 약 2,700토큰/s, 출력 약 480토큰/s)으로 나누어 입력 $0.077/M, 출력 $0.434/M으로 산출하였다. 지연시간은 동일한 L4에서 배치 1로 측정하였으며, 가동률 100%를 가정하고 학습 비용은 제외하였다.

V. 결 과

5.1 정확도, 성공률, 안전 위반율

표 2는 주 결과다. 로컬-학습의 정확도는 91.4%로 로컬-프롬프트(68.4%)보다 23.0%p, 매크로 F1은 0.90 대 0.66, 성공률은 84.3% 대 46.1%로 38.2%(95% CI [32.9, 43.4]) 높았다. 두 조건은 모델·검색기·평가 데이터가 같고 절차를 프롬프트로 주는지 4,000개 대화로 학습시키는지가 다르므로, 이 차이는 두 구현 설정의 성능 차이로 해석한다. 안전 위반율의 차이는 더 뚜렷했다. WARN 또는 HOLD가 정답인 280개 판단 중 로컬-프롬프트는 32건(11.4%)에서 PASS했고 로컬-학습은 6건(2.1%)이었다. 로컬-프롬프트의 위반 32건 중 23건은 유사도 0.8 이상인 판단으로, 규칙 1을 프롬프트로 받았음에도 유사도가 높으면 안전 조건을 무시하고 전달하는 경향이 있었다. 절차를 읽는 것과 따르는 것은 같지 않다.

표 2. 평가 데이터셋 주 결과
Table 2. Main results on the evaluation dataset

Condition	Accuracy	Macro-F1	SR	SVR	Cost ($)	Latency
Local prompt	68.4%	0.66	46.1%	11.4%	0.00064	13.0s
Local prompt+JSON	76.2%	0.74	58.4%	9.6%	0.00064	-
Local tuned	91.4%	0.90	84.3%	2.1%	0.00036	9.5s
Frontier	95.7%	0.95	90.7%	1.1%	0.0308	10.5s

로컬-학습의 성공률은 프론티어보다 6.4% 낮았다(95% CI [2.6, 10.1]). 유형별 격차는 정보 누락(A)에서 15.0%로 가장 컸으며, 해당 유형의 로컬-학습 실패 37건 중 30건은 ASK 대신 SEARCH를 선택한 경우였다(표 3). 정상(E)에서는 로컬-학습의 성공률이 더 높았으나 차이의 신뢰구간은 0을 포함하였다. 프론티어의 과잉 HOLD 30건 중 23건은 이 유형에 집중되었다.

표 3. 시나리오 유형별 성공률
Table 3. Success rate (%) by scenario type

Type	Local prompt	Local tuned	Frontier
A Missing info	28.6	73.6	88.6
B Revision	43.6	86.4	93.6
C Insufficient	38.6	85.0	91.4
D Safety	50.0	90.0	96.4
E Normal	70.0	86.4	83.6

5.2 비용 효율과 로컬 모델 규모

로컬-학습은 절차 프롬프트의 반복 입력을 줄여 대화당 입력 토큰을 로컬-프롬프트와 프론티어의 4,121개에서 1,040개로 감소시켰고, 대화당 출력 토큰은 로컬-프롬프트 753개, 로컬-학습 636개, 프론티어 1,865개였다. 가동률 100%에서 판단 모델의 추론 비용은 프론티어의 약 1/85이었으며, 가동률 50%와 25%에서는 각각 약 1/43과 1/21이었다. 로컬-학습과 프론티어의 지연시간은 유사했고, 로컬-프롬프트가 가장 길었다(표 2).

Gemma 4 E4B와 Qwen3.5 9B의 성공률은 각각 62.7%, 78.0%, 안전 위반율은 8.2%, 3.9%, 지연시간은 4.5초, 7.0초였다. 두 모델은 Gemma 4 12B보다 지연시간이 짧았으나 성능은 낮았다. 모델 계열이 달라 이를 일반적인 규모 효과로 해석하기는 어렵다.

5.3 실패 유형과 학습 방법

로컬-프롬프트의 실패 377건 중 196건(52.0%)은 형식 오류였다. 문법 제약으로 형식 오류를 제거한 로컬-프롬프트+JSON에서도 성공률은 58.4%로 로컬-

학습보다 25.9% 낮았으며, ASK 대신 SEARCH를 선택한 오류 108건과 현행성 확인 누락 60건이 남았다. 로컬-학습에서도 형식 오류는 없었으므로, 성능 격차는 출력 형식 교정만으로 설명되지 않는다.

전체 미세조정의 성공률은 85.7%로 LoRA보다 1.4% 높았으나, 차이의 95% 신뢰구간은 [−1.9, 4.6]으로 0을 포함하였다. 학습 시간은 6.0시간으로 LoRA의 1.5시간보다 4배 길었으며, GPU당 최대 메모리는 약 78GB였다. 본 실험에서 명확한 성능 향상이 확인되지 않아 학습 자원이 적게 드는 LoRA를 선택하였다.

Ⅵ. 결론 및 시사점

본 연구에서는 RAG의 검색과 생성 단계 사이에 검색 후 판단 모델을 적용하고, 세 가지 구현 방식을 동일한 KOSHA 안전보건 기술지침 기반 RAG 환경에서 비교하였다. LoRA 기반 로컬-학습 모델은 로컬-프롬프트보다 성공률이 38.2% 높았다(RQ1). 또한 프론티어 모델 성능의 93% 수준에 도달하였으며, 격차는 주로 정보 누락 상황의 재질문 판단에서 발생하였다(RQ2). 대화당 추론 비용은 가정한 처리량과 공시 단가 기준으로 프론티어 모델의 약 1/85이었다(RQ3). 이는 LoRA 기반 로컬 학습 방식이 망분리 환경과 판단 요청이 많은 서비스에 활용될 가능성을 보여준다.

다만 로컬-학습 모델의 안전 위반율은 2.1%로 프론티어 모델의 1.1%보다 높았다. 따라서 높은 안전성이 요구되는 환경에서는 금지 조건을 규칙 기반 차단기로 재검증하거나 판단 임계를 WARN 방향으로 조정할 필요가 있다.

본 연구는 규칙 기반 절차 준수만 평가하였으며, 현장 타당성과 최종 답변 품질은 다루지 않았다. 또한 단일 도메인·시드·프론티어 모델을 사용하였고, 학습·평가 대화 문장을 프론티어 조건과 같은 GPT-5.5가 작성하여 그 문체에 익숙한 로컬-학습·프론티어 조건이 유리했을 수 있으며, 비용은 토큰 단가와 가동률에 따라 달라질 수 있다. 향후에는 정보 누락 시나리오의 학습 비중을 조정하고, 종단 간 답변 품질과 다른 문서·언어 환경에서의 재현성을 평가할 필요가 있다.

References

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks", Advances in Neural Information Processing Systems (NeurIPS 2020), Online, Vol. 33, pp. 9459-9474, Dec. 2020.

[2] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of Hallucination in Natural Language Generation", ACM Computing Surveys, Vol. 55, No. 12, pp. 1-38, Dec. 2023. https://doi.org/10.1145/3571730.

[3] H. Joren, J. Zhang, C.-S. Ferng, D.-C. Juan, A. Taly, and C. Rashtchian, "Sufficient Context: A New Lens on Retrieval Augmented Generation Systems", Proc. 13th International Conference on Learning Representations (ICLR 2025), Singapore, Apr. 2025.

[4] S. Dennis, R. Patil, K. Shabahang, and H. Guo, "Compiling Agentic Workflows into LLM Weights: Near-Frontier Quality at Two Orders of Magnitude Less Cost", arXiv preprint, arXiv:2605.22502, pp. 1-19, May 2026. https://doi.org/10.48550/arXiv.2605.22502.

[5] A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, "Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection", Proc. 12th International Conference on Learning Representations (ICLR 2024), Vienna, Austria, May 2024.

[6] S.-Q. Yan, J.-C. Gu, Y. Zhu, and Z.-H. Ling, "Corrective Retrieval Augmented Generation", arXiv preprint, arXiv:2401.15884, Jan. 2024. https://doi.org/10.48550/arXiv.2401.15884.

[7] Z. Jiang, F. F. Xu, L. Gao, Z. Sun, Q. Liu, J. Dwivedi-Yu, Y. Yang, J. Callan, and G. Neubig, "Active Retrieval Augmented Generation", Proc. 2023 Conference on Empirical Methods in Natural

Language Processing (EMNLP 2023), Singapore, pp. 7969-7992, Dec. 2023. https://doi.org/10.18653/v1/2023.emnlp-main.495.

[8] S. Jeong, J. Baek, S. Cho, S. J. Hwang, and J. C. Park, "Adaptive-RAG: Learning to Adapt Retrieval -Augmented Large Language Models through Question Complexity", Proc. 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2024), Mexico City, Mexico, pp. 7036-7050, Jun. 2024. https://doi.org/10.18653/v1/2024.naacl-long.389.

[9] A. Zeng, M. Liu, R. Lu, B. Wang, X. Liu, Y. Dong, and J. Tang, "AgentTuning: Enabling Generalized Agent Abilities for LLMs", Findings of the Association for Computational Linguistics: ACL 2024, Bangkok, Thailand, pp. 3053-3077, Aug. 2024. https://doi.org/10.18653/v1/2024.findings-acl.181.

[10] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models", Proc. 10th International Conference on Learning Representations (ICLR 2022), Online, Apr. 2022.

저자소개

정 우 성 (Woosung Jung)



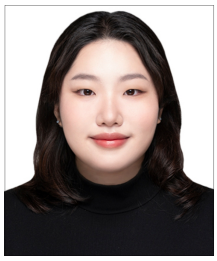
2026년 2월 : 경남대학교 컴퓨터공학과(공학사)
2025년 12월 ~ 현재 : (주)포인랩 연구원
관심분야 : 기계학습, 딥러닝, 자연어 처리, 인공지능

오 석 민 (Seokmin Oh)



2021년 8월 : 경남대학교 미디어커뮤니케이션학과(법학사)
2026년 2월 : 경상국립대학교 정보통계학과(이학석사)
2024년 7월 ~ 현재 : (주)포인랩 연구소장
관심분야 : 산업 AI, LLM, AI-Agent, Agentic AI

손 다 연 (Dayeon Son)



2023년 2월 : 동의대학교 컴퓨터공학과(공학사)
2025년 7월 ~ 현재 : (주)포인랩 선임연구원
관심분야 : 검색증강생성, 대규모언어모델, 소프트웨어 아키텍처

박 진 제 (Jinje Park)



2020년 2월 : 국립창원대학교 전기공학과(공학사)
2022년 2월 : 국립창원대학교 전기공학과(공학석사)
2024년 6월 ~ 현재 : (주)포인랩 본부장
관심분야 : 인공지능, 스마트 제조, 디지털트윈

김 창 순 (Changsoon Kim)



2014년 2월 : 국립창원대학교 전기공학과(공학사)
2016년 2월 : 국립창원대학교 전기공학과(공학석사)
2025년 8월 : 국립창원대학교 전기공학과(공학박사)
2024년 6월 ~ 현재 : (주)포인랩 대표
관심분야 : 스마트 제조, AI-Agent, LLM