OpenSchloar

Synthesizing scientific literature with retrieval-augmented language models

Akari Asai et al.

Nature, 2026

Deep Shark Lab 인턴

양현민

Contents

1. **Introduction**
   1) 연구 배경
   2) OpenScholar 개요

2. **Methods**
   1) Retrieval
   2) Reranking
   3) Self-feedback & Generater

3. **Experiment**
   1) scholarQABench
   2) 주요실험 결과 4가지

4. **Discussion**
   1) 실험결과 요약
   2) 한계점

Introduction

연구 배경

- **과학 논문의 증가**
  최신 연구 파악
  여러 논문의 결과를 비교·종합하는 부담 증가

- **기존 RAG 시스템의 한계**
  비공개 API·범용 모델 의존
  과학 분야 특화 DB 부족

- **기존 LLM의 한계**
  Hallucinatio
  학습 시점 이후의 최신 연구 반영 어려움

- **평가 체계의 한계**
  단일 분야·단순 질의응답 집중
  여러 논문을 종합한 긴 답변의 평가 부족

OpenSchloar

OSDS

peS2o v3의 오픈 액세스
24년 10월까지
약 4500만 편의 논문
256단어 고정 길이 청크
+ 논문제목 결합
검색용 벡터DB

평가 실험은 peS2o v2사용
(23년 1월까지)

Adaptive retrieval modules

Retrieval
OSDS 벡터검색 + 외부 API

Reranker
Qurery - 검색된 청크
관련성 평가 및 재정렬
상위 N개 선택

Self-feedback

인용 포함 초안작성
피드백 생성
필요시 추가 근거 검색
답변 수정 및 인용 검증

Method

Retrieval

기반 모델

Contriever
Bi-encoder

Query, 청크 개별 인코딩

Openscholar Retriever

학습데이터

peS2o + CCNews
+ Proofpile2

학습 설정

Learning rate : 5e-5
Training steps : 500,000
Batch size : 4,096

Retrieval

OSDS

Query 인코딩

DB에서 최근접 이웃 검색
(내적)

상위 70개의 청크 선택

Semantic Scholar API

생성기로 키워드 추출

키워드 별로
피인용수 기준 논문 정렬

상위 10개의 논문 선택

HTML 확보 → 본문 전체
확보 x → Abstract

You.com API

Query 원문으로
학술플랫폼 검색
(Ex. arXiv, PubMed)

상위 10개 검색 결과 선택

Open Access → 본문 전체
Closed Access → Abstract

Reranking

기반 모델

BGE-reranker
Cross-encoder

Query + 후보 청크
공동 인코딩

학습

Llama-3-70B-Instruct
peS2o 초록 기반 질의 생성
후보 청크 10개 검색

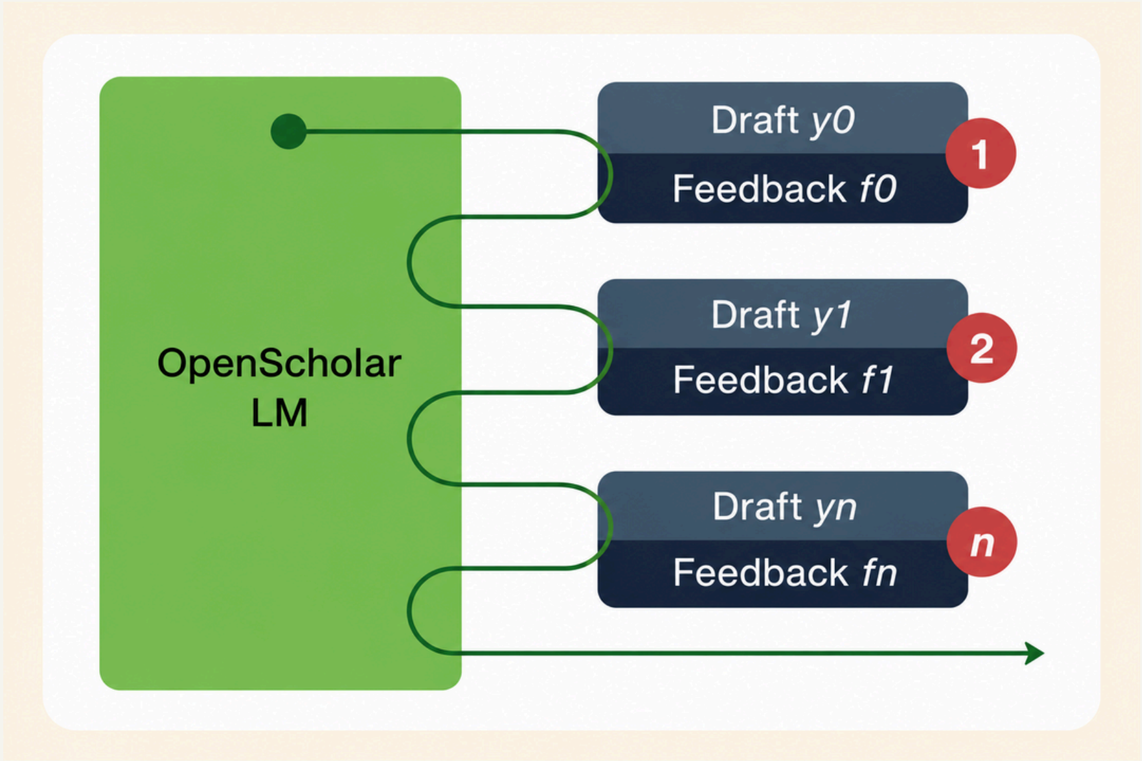
관련성 1~5점 평가
4~5점 : Positive
1~2점 : Negative
3점 : 제외

학습 설정
Epochs : 5
Learning rate : 6e-5

Reranking

- 검색 후보 통합
  OSDS + 외부 API 청크
- 관련성 평가
  Query, 청크 공동 인코딩
  → 점수 산출
- 메타 필터링
  피인용 수 보정
  논문별 청크 제한 3개
- 최종 선택·전달
  상위 N개 청크 → 생성기
  단일 논문 과제 N=5
  다중 논문 과제 N=10

Self-feedback



1. 초안·피드백 생성

Query + 상위 N개 청크

→ 인용 포함 초안 작성

피드백 최대 3개 생성

필요시 추가 검색 질의 생성

2. 반복적 답변 수정

피드백 순차 반영

필요시 추가 검색 및 근거 보충

이전 답변 수정

모든 피드백 처리 후 종료

3. 인용 검증

답변을 문단별로 분할

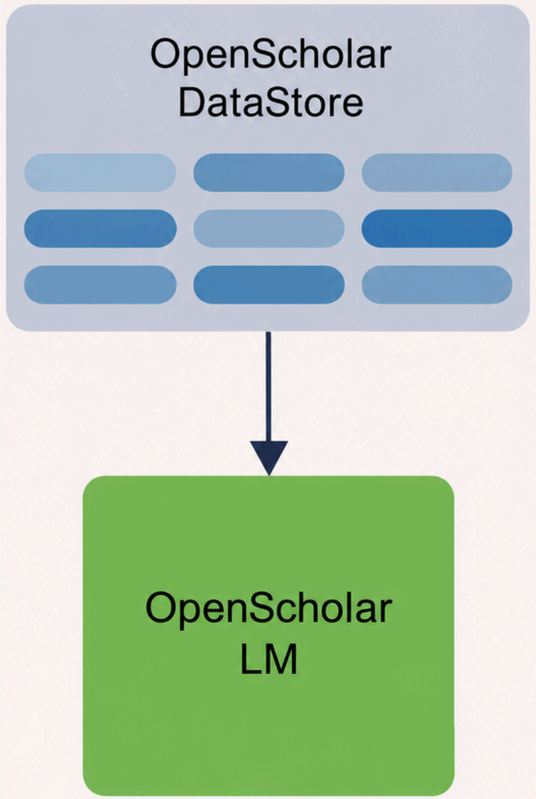
문단-근거 대조

근거가 있는 누락된 인용 삽입

근거 부재 시 문장 유지

generator

합성 데이터 생성



질문 생성(Llama 3.1 70B)

OSDS에서 가장 많이 인용된 논문 선택

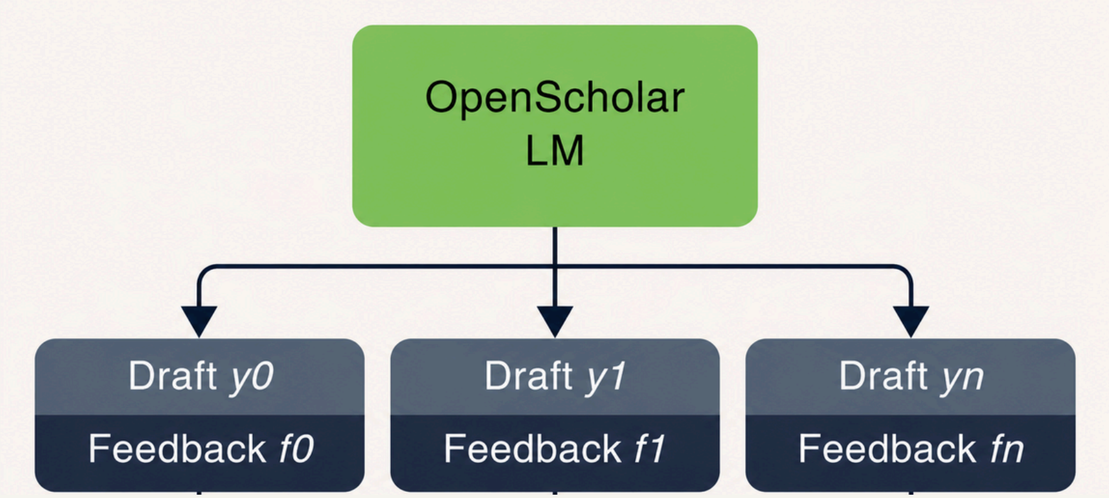
pe2So 약 100만개의 논문 초록 샘플링
메타데이터 수집(Ex. 피인용 수, 출판 연도)

초록을 기반으로 질의 생성

2017년 이후 논문 무작위 선택
질의 생성 하도록 프롬포트 입력

generator

합성 데이터 생성

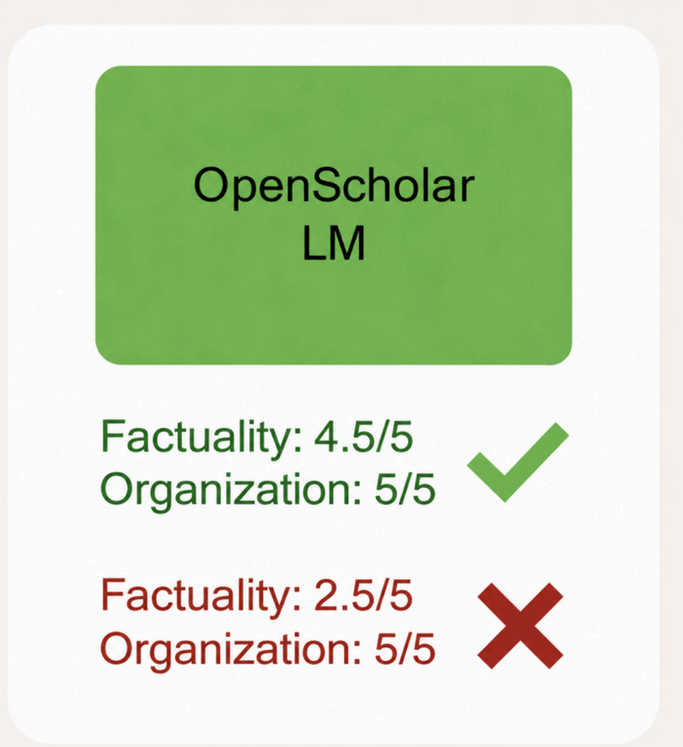


답변 생성(Llama 3.1 70B)

Openscolar 파이프라인 사용하여
초기 답변, 피드백, 최종 답변 생성

generator

데이터 필터링(Llama 3.1 70B)



Pairwise 필터링

초기 답변 ↔ 최종 답변 비교
더 좋은 답변을 정답 답변으로 선택

rubric 필터링

사실성 : 답변 내용의 사실적 정확성
구성 : 답변의 구조와 정리 수준
통과 기준 : 두 범주 각각 4.5점 이상

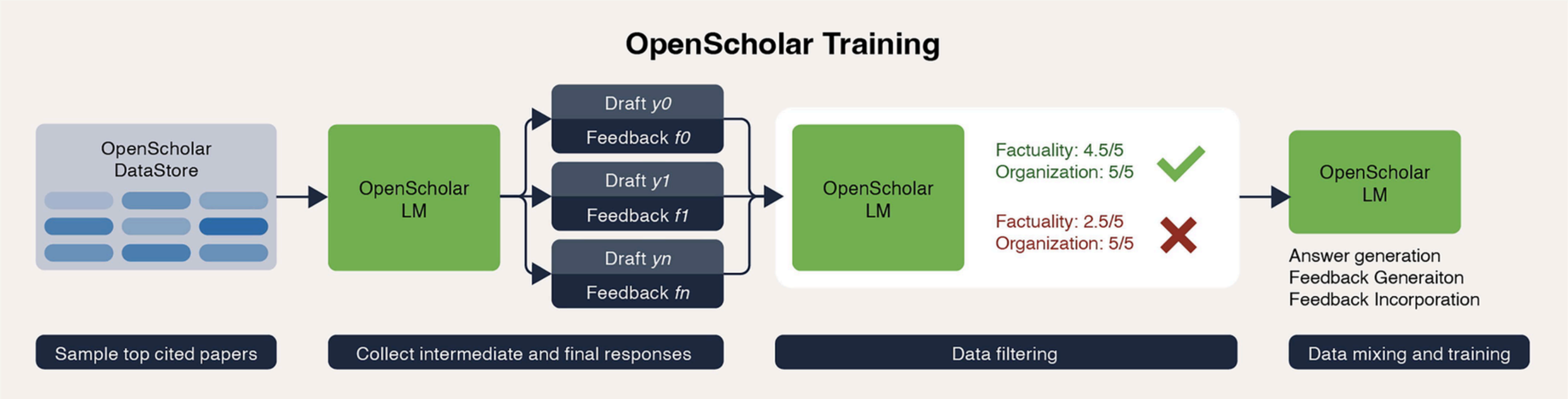
generator

학습 데이터 유형

구분	데이터 구성	학습 목적
합성 — 답변 생성	질문 → 답변	근거 기반 답변 작성
합성 — 피드백 생성	초기 답변 → 피드백	답변의 개선점 파악
합성 — 피드백 통합	이전 답변 + 피드백 → 수정 답변	피드백을 반영한 답변 개선
기존 — 일반 도메인 지시 데이터	다양한 분야의 질문·지시와 응답(50%)	일반적인 지시 수행 능력 유지
기존 — 과학 지시 데이터	과학 문헌 관련 작업과 응답(50%)	과학 분야 작업 수행 능력 강화
추가 합성 — 사실 검증·불리언 QA	peS2o 피인용 수 상위 10만 편의 초록 활용	주장 검증 및 예·아니오 질문 응답

Llama-3.1-8B 학습

generator



- Base Model：Llama-3.1-8B-Instruct
- Epochs：2
- Learning rate：5e-6 or e-4
- Batch size：1
- Optimizer：AdamW
- GPU：A100 80GB x 8

Experiment

ScholarQABench

	데이터셋	도메인
단일 논문	SciFact	생의학
	PubMedQA	생의학
	QASA	컴퓨터과학—AI·ML
다중 논문	Scholar-CS	컴퓨터과학
	Scholar-Bio	생의학
	Scholar-Neuro	신경과학
	Scholar-Multi	컴퓨터과학·생의학·물리학

- **단일 논문 과제 : 하나의 논문을 근거로 질의에 응답 또는 참, 거짓 판단**
- **다중 논문 과제 : 여러 논문의 내용을 비교·종합하여 문헌 검토 질문에 답변**
- **전문가 관여 : 기존 단일 논문 데이터셋은 전문가 주석 활용 새로운 다중 논문 데이터셋은 박사급 전문가들이 질문·답변·평가 기준 작성에 참여**

ScholarQABench 결과

	Single-paper performance						Multi-paper performance						Cost
	Pub		Sci		QASA		CS		Multi		Bio	Neu	CS
Model	**Acc.**	**Cite**	**Acc.**	**Cite**	**Acc.**	**Cite**	**Rub.**	**Cite**	**LLM**	**Cite**	**Cite**	**Cite**	**USD per question**
Llama 3.1 8B	61.5	0	66.8	0	14.3	0	41.9	0	3.79	0	0	0	0.0001
+RAG^{OSDS}	75.2	63.9	75.5	36.2	18.6	47.2	46.7	26.1	4.22	25.3	38.0	36.8	0.0001
OpenScholar-8B	76.4	68.9	76.0	43.6	23.0	56.3	51.1	47.9	4.12	42.8	50.8	56.8	0.003
Llama 3.1 70B	69.5	0	76.9	0	13.7	0	44.9	0	3.82	0	0	0	0.0004
+RAG^{OSDS}	77.4	71.1	78.2	42.5	22.7	63.6	48.5	24.5	4.24	41.4	53.8	58.1	0.0004
OpenScholar-70B	79.6	74.0	82.1	47.5	23.4	64.2	52.5	45.9	4.03	54.7	55.9	63.1	0.01
GPT-4o	65.8	0	77.8	0	21.2	0	45.0	0.1	4.16	0.7	0.2	0.1	0.006
+RAG^{OSDS}	75.1	73.7	79.3	47.9	18.3	53.6	52.4	31.1	4.03	31.5	36.3	21.9	0.01
OpenScholar-GPT-4o	74.8	77.1	81.3	56.5	18.7	60.4	57.7	39.5	4.51	37.5	51.5	43.5	0.05
PaperQA2	–	–	–	–	–	–	45.6	48.0	3.82	47.2	56.7	56.0	0.3–2.3
Perplexity	–	–	–	–	–	–	40.0	–	4.15	–	–	–	0.002*

- Acc : Accuracy
- Cite : 인용 F1
- Rub : Rubric Accuracy

→ 핵심 내용 충족도, 일반 평가기준

- LLM(Prometheus) : 평가용 LLM

→ 관련성, 포괄성, 조직력

- LM 단독: 검색 없이 답변
- +RAG^{OsDs}: 기존 LM + OSDS 검색
- OpenScholar: 검색·재순위화·자기 피드백·인용 검증
- 생성기: 학습된 8B / 기존 70B·GPT-4o
- 경쟁 시스템: PaperQA2·Perplexity

Hallucination Evaluation

	Computer science			Biomedicine		
Model	**Total no.**	**No. of hallucinated (↓)**	**Ratio (↓)**	**Total no.**	**No. of hallucinated (↓)**	**Ratio (↓)**
OpenScholar-8B	9.65	0	0	6.25	0	0
Llama 3.1 8B	5.20	4.79	92.1%	5.58	5.46	97.6%
Llama 3.1 70B	6.14	4.78	78.1%	6.98	6.74	96.6%
GPT-4o	5.74	4.52	78.7%	5.24	4.97	94.8%

- Total no. : 인용한 전체 논문 수
- No. of hallucinated : 존재하지 않는 인용 논문 수
- Ratio : 전체 인용 논문 중 존재하지 않는 논문의 비율

Comparison with Human Experts

	Scholar-CS					Scholar-Multi				
	Length	No. of cit.	Rubric	Prec.	Rec.	Prec.	Rec.	Org.	Cov.	Rel.
Human	424.3	7.1	54.0	43.2	40.1	44.4	41.5	–	–	–
OpenScholar-GPT-4o	1,447.3	11.8	57.7	41.1	38.1	38.0	37.1	4.63	4.50	4.23
OpenScholar-8B	706.3	10.1	51.1	40.6	44.1	42.5	43.2	3.92	4.44	4.02
PaperQA2	288.7	2.7	45.6	45.9	48.0	53.1	42.2	3.67	3.46	4.20
GPT-4o	281.5	0.5	45.0	1.0	0.8	0.8	0.6	4.06	3.94	4.21

- Length : 평균 답변 길이
- No. of cit. : 평균 인용 논문 수
- Prec : 인용 정밀도
- Rec : 인용 재현율
- Org : 구성
- Cov : 내용의 포괄성
- Rel : 관련성
- Org.·Cov.·Rel : LLM(Prometheus) 평가 (1~5)

Expert vs Model Responses

	Fine-grained (1–5, average)			Overall usefulness (%)	Relative to human (%)		
	Organization	Coverage	Relevance		Win	Tie	Lose
GPT-4o	4.63 (+0.4)	4.06 (−0.2)	4.50 (−0.1)	69.7 (−13.9)	31.9	13.8	54.2
OpenScholar-8B	3.82 (−0.3)	4.30 (+0.7)	4.00 (−0.4)	72.1 (+8.7)	50.8	12.3	36.9
OpenScholar-GPT-4o	4.47 (+0.8)	4.38 (+0.9)	4.30 (0)	80.0 (+22.5)	70.0	6.8	23.2

괄호안의 값

- 양수(+): 모델이 더 높게 평가됨
- 음수(−): 인간 답변이 더 높게 평가됨
- 0: 동일한 평균 점수

실험 결과 요약

- **작은 모델의 경쟁력:** OpenScholar-8B가 검색 없는 GPT-4o와 경쟁 시스템인 PaperQA2·Perplexity보다 높은 루브릭 점수 달성
- **전문가 이상의 루브릭 점수:** OpenScholar-GPT-4o가 전문가보다 높은 점수 기록(57.7 vs 54.0)
- **전문가 수준의 인용 정확도:** OpenScholar-8B는 루브릭 점수는 전문가보다 낮지만, 인용 정확도는 전문가 수준
- **포괄적인 문헌 종합:** OpenScholar는 더 많은 논문을 인용하며, 더 폭넓고 상세한 답변을 생성하는 경향

한계점

- **평가 데이터 규모와 편향**: 전문가 주석에 비용·시간이 많이 들어 평가 규모가 작으며, 평가자의 전문성에 따른 편향 가능성
- **검색의 불완전성**: 질문에 가장 관련성이 높거나 해당 분야를 대표하는 논문을 항상 검색하지 못함
- **인용 내용의 정확성**: 실제로 존재하는 논문을 인용하더라도, 내용을 잘못 설명하거나 해당 논문이 뒷받침하지 않는 주장을 할 수 있음
- **평가자의 전문성 한계**: 평가자의 세부 전문 분야 밖에 있는 질문에서는 미묘한 오류를 놓쳤을 가능성
- **일반화의 한계**: 전문가 평가가 16명·108개 질문·세 과학 분야에 한정되어 다른 분야로 일반화하는 데 제약