VERIFICATION WITHOUT SUFFICIENCY

멀티홉 RAG에서 Per-Chunk 검증은 왜 실패하는가?
-질문 분해를 통한 조건부 검증의 가능성

DeepShark Lab 이지언

목차

문제 제기

RAG 검증의 숨은 가정 - "조각 1개 = 충분한 근거"

표준 RAG 검증은 조각(chunk)마다 점수를 매겨 낮은 건 버린다 – 이는 "조각 1개가 답의 충분한 전제"라고 가정하는 것
그런데 멀티홉 질문은 애초에 어떤 조각 하나로도 답이 안 나오도록 구성되어 있다.

Hop 1 Bridge

"Kiss and Tell"문서
- 질문에 이름이 언급됨. 정답은 없고 브릿지 개체만 제공

Hop 2 Answer

"Shirley Temple" 문서
- 정답을 담고 있지만 질문 속 단어가 전혀 없음

-〉 검증 모델은 질문과 어휘가 겹치는 Hop1을 선호하고, 정작 정답이 담긴 Hop 2를 "관련 없다"고 오판하기 쉽다.

방법

세 가지 증거 검증 방식

PER-CHUNK
기존 방식

조각 하나하나를 독립적으로 점수화.
2번째 홉 조각은 애초에 답을 함의할 수 없음

SET-LEVEL

모든 조각 쌍을 점수화 (k=5 -> 10번 비교)
argmax 정답 착지율 44%

CONDITIONAL
제안 방식

임베딩으로 1홉 앵커 고정(92% 정확) -> 앵커 기준 나머지 검증.
4번 호출, recall 0.755

+ 검증 신호 2가지: Embedding (질문-조각 코사인 유사도)
Entailment (조각을 전제로, 질문에서 만든 명제가 참인지 NLI로 판단)

방법

검증 신호 수식 & 실험 셋업

$$s_{\text{emb}}(q, c) = \frac{1}{2}\left(1 + \cos(E(q), E(c))\right)$$

$$S_{ent}(q,c)=P(entail|c,H(q))$$

q = 질문, c = 조각, E() = 임베딩 함수, H() = 질문을 정답 자리 비운 평서문으로 바꾸는 함수

데이터 = HotpotQA 2WikiMultihopQA MuSiQue 각 500 문항 + 단일홉 대조군 SQuAD 300문항

평가지표 = AUC (정답 조각 vs 방해 조각 분리 능력, threshold-free)

핵심 발견

1. 모든 데이터셋에서 신호가 약하고, 2. 방향성까지 있다.

데이터셋별 Entailment AUC

Dataset	AUC
HotpotQA	0.643
2WikiMultihopQA	0.523
MuSiQue	0.560
SQuAD (단일 홉)	0.951

두 신호 모두 "질문에 이름이 언급된 조각"을 선호 – 양쪽 개체가 다 언급되는 comparison 질문에서는 이 편향이 사라짐 –〉 편향의 원인이 "질문과의 표면적 겹침"임을 증명

질문 유형별 방향성 편향 (HotpotQA)

질문 유형	Embedding deficit
bridge (정답 없는 조각 선호)	+0.131
comparison (양쪽 다 언급됨)	–0.003

ORACLE(정답 채운 상한선)조차 0.67을 못 넘김.
모델을 10배 키워도 개선 없음 - 구조적 문제

핵심 발견

3. 홉 수가 늘수록 심해진다

Hop 수	Embedding AUC	Entailment(SLOT) AUC
2-hop	0.819	0.590
3-hop	0.780	0.550
4-hop	0.677	0.542

질문에만 의존하는 신호는 홉이 늘수록(=질문-정답 거리가 멀수록) 급격히 약해짐

부가 실험 = gold 조각 2개를 합치면 AUC 0.881까지 오름. 반대로 "gold+distractor"는 글자 수가 더 많은데도

점수는 4배 낮음(0.127) -> 텍스트 길이가 아니라 "정답 정보가 충분한가"가 핵심

실전 영향

필터링을 안 하는 게 오히려 낫다

Selector	HotpotQA EM	2Wiki EM	MuSiQue EM
오라클 (정답 조각만 제공)	55.0	42.2	40.9
필터링 없음 (baseline)	47.0	34.0	20.9
Conditional (제안, ours)	45.8	32.6	14.3
Per-chunk gate (기존 방식)	33.6	27.8	9.7

세 데이터셋 모두 Per-chunk gating이 최하위.

원인 = 게이트가 평균 1.2개 조각만 남기고, 정답 조각의 35~46%만 생성 모델에 도달시킴.

게다가 생성 모델이 강력해질수록(0.5B->3B)이 손해는 더 커짐 (HotpotQA -4.6 -> -19.4).

해결책

01

1홉 앵커 고정

Embedding 요사도 최고 조각을 앵커로 고정 - 92%가 실제 정답 조각

02

질문 → 명제 변환

정답 자리를 비운 평서문으로 변환 (배포 가능, 정답 사전 지식 불필요)

03

앵커 기준 나머지 검증

앵커 + 나머지 조각을 합쳐 entailment 채점 → 최고점을 2홉 근거로 선택

비교 횟수 = Set-level C(k,2) (5개 -〉 10번) -〉 Conditional은 k-1 (5개-〉4번), 60% 절감

해결책

질문을 바꾸는 것만으로 성능이 극적으로 회복

하위 질문을 무엇으로 만드는가	AUC
원래 질문 그대로	0.546
분해된 하위 질문 (gold decomposition)	0.840

홉수에 따른 AUC도 원래 질문 기준으로는 거의 우연 수준이지만, gold 하위 질문으로 조건화하면 홉수와 무관하게 평탄해짐.

실제 배포형(Qwen2.5-7B, 검색 문서 참고 분해) = 0.637 (상한선의 31% 회복), 검색 없이 질문만 보고 분해하면 오히려 원래 질문보다 나쁨(0.533).

결론

1. Per-Chunk 검증의 한계

각 문단을 원래 질문과 독립적으로 검증 -〉 후속 홉의 근거를 놓칠 수 있음

-〉 필터링하지 않은 경우보다 EM 성능이 낮아짐

2. Set-Level 검증의 한계

여러 문단을 함께 검증 -〉 근거 충분성 판단 가능 -〉 문단 조합 비교로 계산 비용 증가

3. Conditional Evidence Selection

첫 번째 홉의 문단을 Anchor로 활용 -〉 적은 비교 횟수로 후속 근거 선택

4. Question Decomposition

질문을 홉별로 분해 -〉 후속 근거 식별 향상

-〉 AUC 0.546 -〉 0.840

감사합니다
Q&A