한국어 검색 에이전트 환경에서 조사 단위 프레이밍이 미치는 영향 분석

최윤혁, 곽나연, 이현준, 이강원, 현진명
연세대학교 컴퓨터과학과
{choiyh592, nykwak0525, hyunjunlee, kangwon71, kid7ho}@yonsei.ac.kr

How You Mark It Matters: Particle-Level Framing Effects in a Korean Search Agent

Yoonhyeok Choi, Nayeon Kwak, Hyun June Lee, Kangwon Lee, Jin Myoung Hyun
Yonsei University, Department of Computer Science

요 약

생성형 검색 시스템들은 과학적인 질의에 대해서 판정과 그에 대한 인용으로 답한다. 기존 연구에서 해당 평가가 어휘 변용(Paraphrasing) 및 의미 보존 재작성 (Meaning-preserving rewriting)에 따라 변동될 수 있다는 점이 확인된 바 있으나, 이를 한국어 중심적으로 실제 상업 시스템 내에서 평가한 연구는 전무하다. 이를 해결하기 위하여 본 연구에서는 실제 상용 검색 에이전트(Liner)를 대상으로 한국어 조사 및 양태 표지 교체가 판정 층위와 문헌 층위에 미치는 영향을 정량적으로 분석하였다. 실험 결과, 의미 보존적 조사 변용은 판정을 안정적으로 유지하면서도 검색 문헌의 약 65%를 교체하는 반면, 방향성 반전 변용은 문헌 집합을 유지한 채 판정의 절반 이상(53.3%)을 뒤집는 이중 분리(Double Dissociation) 현상을 확인하였다. 이러한 결과는 판정의 안정성이 근거의 안정성을 보증하지 않음을 시사하며, 본 연구는 이를 바탕으로 학술적 인용을 위한 반복 호출 다수결($N^* = 5$) 등 실무적 검증 규칙을 제안한다. 연구에 사용된 분석 스크립트 및 데이터는 github.com/choiyh592/How_You_Mark_It_Matters에서 확인할 수 있다.

주제어: 정보검색, 담화 이해 및 대화 관리, 신뢰성

제 1. 절 서론

인공지능 검색 에이전트는 사용자의 질의를 바탕으로 근거가 될 자료를 모으고, 해당 자료를 인용하여 판정을 내리는 방법으로 작동한다. 이때 의미 및 정보 구조의 변동을 위해 통사 구조 재배치, 음성적 강세, 특수 구문 등 복합적인 통사·음운론적 기제를 동원해야 하는 영어와 달리, 한국어는 문법적 틀을 유지한 채 표지의 교체만으로 화제·초점 등 담화상의 정보 구조를 극적으로 전환함으로써 같은 명제에 관한 질의를 서로 다른 의미를 내포한 채로 제시할 수 있다.

중립적인 의미의 이/가, 대조 화제를 세우는 은/는, 첨가 전제를 얹는 도, 접한 적 있는 사실임을 함축하는 -는데와 같은 문법적 표지의 변화는 질문이 무엇을 전제하고 앞세울 것인지를 정한다. 그러나 질의의 핵심인 진리조건은 달라지지 않으며, 과학적 주장에 대한 인공지능 에이전트의 판정이 이 표지에 따라 갈린다면, 그 에이전트의 인식적 태도는 상당 부분 문법이 만들어 낸 산물이 된다. 이러한 인식적 태도의 형성 과정은 사용자가 명시적으로 볼 수도, 고를 수도 없다는 점에서 생성 결과물의 근거성과 신뢰성을 감소시킨다. 이는 검증에 대한 사용자의 부담을 가중시켜, 생성 결과물에 대한 심리적 거부감과 인용 기피 현상을 강화할 수 있다.

이러한 한계는 단순한 현상적 지적을 넘어, 상용 생성형 AI 검색 시스템을 학술·전문적 맥락에서 안심하고 인용하기 위한 정량적 기준과 문법적 편향으로부터 자유로운 검증 프레임워크의 필요성을 시사한다. 이에 따라 본 연구는 실제 상용 검색 에이전트인 Liner를 대상으로, 한국어 문법 표지 변용에 따른 인식적 가변성을 대규모로 측정한다. 특히 의미 보존적 변형과 의미를 명시적으로 바꾸는 변형을 분리하여, 문법적 프레이밍이 판정 자체를 변화시키는지, 아니면 판정을 유지한 채 근거 문헌과 설명 방식만 변화시키는지를 구분한다.

본 연구의 기여는 다음과 같다. 첫째, 한국어 조사 및 양태 표지 교체를 의미 보존적 실험 도구로 사용하여 영어 중심의 질의 변형 연구에서 포착하기 어려운 담화 프레이밍 효과를 측정한다. 둘째, 판정 층위와 문서 층위를 분리해 분석함으로써 동일한 결론이 서로 다른 문헌 집합에서 생성되는 이중 분리를 제시한다. 셋째, 반복 호출, 질의 변형, 문헌 다양화의 역할을 비교하여 상용 검색 에이전트의 실무적 검증 규칙을 제안한다.

제 2. 절 관련 연구

최근의 학술 질의응답 연구들은 주장과 인용의 대응 관계, 인용 가능한 근거 문장의 제시, 검색 과정의 재현성을 핵심 평가 대상으로 확장하고 있다. SQuAI는 복잡한 과학 질의를 하위 질문으로 분해하고, 표적 검색과 인라인 인용을 결합하여 각 주장에 대응하는 근거 문장을 제시한다[1]. SeekBench 역시 최종 정답만으로는 검색 에이전트의 인식적 역량을 충분히 평가할 수 없으므로, 근거에 기반한 추론, 낮은 품질의 검색 결과에 대한 질의 재구성, 근거 충분성 판단을 단계별로 평가해야 한다고 제안한다[2]. 이 흐름은 검색 에이전트를 답변 생성기가 아니라 근거 획득과 판단을 결합한 시스템으로 평가해야 한다는 점을

분명히 한다.

이때, 조사와 어미의 교체만으로 정보 구조를 조절할 수 있는 한국어에서는 변용 과정에서의 의미 변화와 표면형 변화를 서로 분리하여 분석하기 용이하지만[3, 4], 한국어의 의미 보존적 조사 · 양태 표지를 이용해 상용 검색 에이전트의 판정 층위와 문서 층위를 동시에 감사한 연구는 확인하기 어렵다. 또한 기존의 학술적 검색 에이전트 평가는 검색된 문헌의 품질과 최종 답변의 정확도를 주로 연결하지만[5, 6], 같은 판정이 서로 다른 문헌 집합에서 구성되는지, 또는 서로 다른 문헌이 같은 판정을 안정적으로 지지하는지는 일반적인 평가 축으로 체계화되지 않았다.

본 연구는 위의 공백을 실제 상용 서비스를 대상으로 한 실험으로 연결한다. 한국어의 은/는, 도, -는데를 이용해 명제의 진리조건은 유지하면서 정보 구조만 변형한 뒤, 동일한 질의를 대상으로 검색 문헌과 생성 판정을 함께 측정한다. 또한 진리조건을 완전히 바꾸는 긍정 · 부정 조건 및 동일 조건 반복에서 발생하는 변동을 활용하여 문법적 프레이밍 효과가 검색 단계의 문헌 교체인지 생성 단계의 판정 변화인지 구분한다. 이를 통해서 의미가 보존된 한국어 변용이 실제 상용 검색 에이전트의 문헌 선택과 과학적 판정에 각각 어떤 방식으로 전달되는지, 그리고 안정적인 판정이 불안정한 근거 위에 세워질 수 있는가를 실증적으로 분리하여 분석한다.

제 3. 절 연구 방법

3.1 질의 구성

질의의 기반이 될 명제 집합의 구성은 의학 및 과학 분야에서 학술 문헌에 기반을 두고 있는 60개의 주장의 수집을 통해 이루어졌다. 각 주장은 "요소 V가 결과 O에 대해 R의 관계를 가지고 있다"의 형태로, 각 주장의 증거적 논쟁성(evidential contesetedness)에 따라 확실한 기반을 두고 있는 $S_1\,(n=20)$, 연구가 진행 중이나 결과가 혼재되어 있는 $S_2\,(n=20)$, 충분한 검토가 이루어지지 않은 초기 연구 단계의 $S_3\,(n=20)$으로 분류하였다.

이후 각 명제에 적용하기 위하여 6가지의 변용 방법을 구성하였다. 변용은 4개의 의미 보존 변용(M 계열, $M_0/\ldots/M_3$) 및 2개의 방향성 반전 변용(D 계열, D_+/D_-)으로 이루어졌다. 각 변용 및 적용 예시는 표 1에서 확인할 수 있다. 이를 통해서 최종적으로 60개의 주장을 6가지로 변용한 360개의 질의를 구성하였다.

3.2 생성형 검색 시스템

생성형 검색에는 220여 개국에 걸쳐 실제 기업 및 교육기관에서 사용하는 상용 학술 검색 서비스인 Liner의 공개 API[7]를 활용하였다. 분석 대상 API는 Scholar Search(제목 · 학술지명 · 발행연도 · 인용 수 등 메타데이터가 포함된 상위 20개 학술 문헌 검색), Search Agent(학술 모드에서 실행되며 인용 정보를 포함한 최종 답변을 SSE 형식으로 스트리밍), 그리고 계층 축 분석을 위한 Quick Answer 및 Deep Research를 포함한다.

표 1. 각 변용 및 적용 예시

계열	설명	적용 예시 (R = 높인다)
M_0	중립	{V}와 {O}의 관계에 대한 연구 결과는?
M_1	대조(은/는)	{V}는 {O}를 높이는가?
M_2	가산(도)	{V}도 {O}를 높이는가?
M_3	공유(-는데)	{V}가 {O}를 높인다는데, 관련 연구는?
D_+	긍정	{V}가 {O}를 높인다는 증거는?
D_-	부정	{V}가 {O}를 낮춘다는 증거는?

Liner API를 비롯한 대다수의 생성형 검색 시스템은 입력된 자연어 질의를 시스템 내부에서 검색에 최적화된 질의어로 자동 재작성 후 전달한다. 이에 따라 본 연구에서 보고하는 제반 효과는 재작성기, 검색기, 생성기로 구성된 결합 파이프라인의 종합적 특성을 기반으로 하며, 이는 실제 사용자가 경험하는 상호작용 환경과 일치한다.

3.3 응답 데이터 수집

데이터 수집은 다음과 같은 순서로 진행되었다. 먼저 Scholar Search를 360개의 질의를 5번씩 반복 호출한 결과와, Search Agent를 동일한 방법으로 호출한 결과를 종합하여 총 3,600개의 질의-응답 쌍으로 이루어진 데이터셋을 구성하였다. 모든 데이터 수집은 단일 기간(2026년 8월 21일 KST, 1일간) 내에 진행되었다. 호출 순서는 무작위화되었으며, 초당 2회 이하의 호출 제한을 적용하였다. 계층 축 분석을 위해서는 논쟁성 계층별 10개씩 선정한 30개 주장에 M_0 변용을 적용하여 Quick Answer와 Deep Research를 각 5회 반복 호출한 300개의 응답을 추가로 수집하였다.

3.4 사용자 인지 및 수용성 데이터 수집

사용자 인지 데이터 수집을 위하여, 5인의 주석자(annotator)는 1,800개의 Search Agent 응답 중 1,200개에 대해 3자 판정(동의 / 반박 / 유보)을 부여하였다. Search Agent가 결과 분석에 필수적인 인용에 기반한 판정을 완전한 형태로 제공한다는 점과, Scholar Search의 판정 부재, Quick Answer의 짧은 응답 길이로 인한 근거 평가의 어려움, Deep Research의 높은 호출 비용을 종합적으로 고려하여 사용자 인지 평가 대상은 Search Agent 응답으로 한정하였다.

이때 판정 부담을 줄이면서도 주장 수를 유지하기 위해서 인용 가능성의 기준점이 되는 M_0 변용에 대해서는 5회 반복

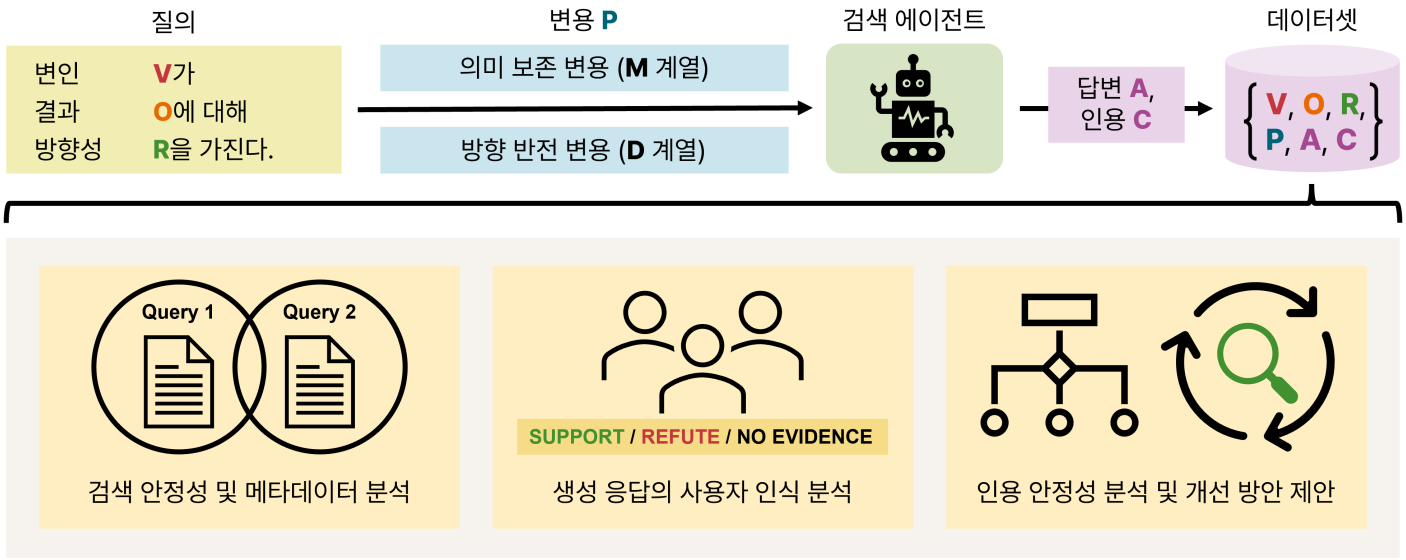


그림 1. 연구 파이프라인 요약

전체를, 나머지 조건에 대해서는 조건당 5회 중 3회를 무작위로 선택하여 평가 대상에 포함하였다. 응답은 라벨링된 항목의 15%(180개)에 대해 모든 주석자 쌍이 교차 오버랩되도록 분배되었으며, 평가 작업은 이중맹검(double-blind) 프로토콜 하에 조건과 반복 인덱스 및 질의 텍스트를 제거한 채 수행되었고, 방향성 판단을 위해 중립적 형태의 주장이 제시되었다. 이중 평가에서 판정이 불일치한 33개 항목은 사전 정의된 판정 규칙에 따른 조정을 거쳐 합의 라벨로 확정하였다.

3.5 측정 및 분석

생성형 검색 시스템은 검색 증강 생성(retrieval-augmented generation, RAG)을 기반으로 하여 질의 재작성, 검색, 재순위화, 문맥 구성, 생성, 검증의 과정으로 구성되어 있다[8, 9]. 이러한 시스템에서 검색 품질은 최종 답변 정확도와 분리해 평가할 필요가 있다. 따라서 본 연구에서는 변용이 검색(retrieval) 과정에 미치는 영향과 증강(generation) 과정에 미치는 영향을 분리하여 분석하였다.

검색 과정. 먼저 결과의 정량적인 분석을 위하여 정규화된 문헌 식별자를 바탕으로 조건 내(노이즈) 및 조건 간(프레이밍) 쌍체 Jaccard 계수[10]를 산출하였다. 또한 Scholar Search 호출 결과에서는 추가로 RBO@10 ($p = 0.9$)[11]을 측정하였다. 이후 변용에 따른 학술지/학회 발행국 비중, 평균 피인용수, 주제 정합도, 평균 발행연도의 변화를 측정하는 조건부 메타데이터 분석을 수행하였다.

증강 과정. 먼저 변용에 따른 사용자 해석의 일관성을 정량화하기 위하여 수집한 사용자 인지 데이터를 바탕으로 Krippendorff's α[12]를 측정하였다. 이후 동일 주장에 대한 변용 전후의 평정 반전(Verdict flip)의 통계적 유의성을 검정하기 위해 Δ(주장) = (조건 간 변동률 − 조건 내 변동률)를 M 계열 및 D 계열 변용에 적용한 Δ_M, Δ_D에 대하여 쌍체 순열 검정(paired permutation test, 10,000회 무작위 추출, two-tailed)을 수행하였다. 이때, 조건 내 변동률은 동일한 주장 및 변용을 적용하여 진행한 5회 반복 실행에 대한 모드(modality) 판정 비중의 여집합을 산출한 후 평균하여 측정하였으며, 조건 간 변동률은 주장별로 M-조건군 및 D-조건군 내에서 서로 다른 모드 판정을 보이는 조건 쌍의 비율로 산출하였다. 모든 다중 검정에는 통계적 유의수준을 제어하기 위해 Holm-Bonferroni 보정[13]을 적용하였다. 마지막으로 판정 전환의 발생 층위를 식별하기 위하여, 조건 쌍별 판정 전환 여부와 해당 쌍의 문헌 Jaccard 계수의 결합 분포를 분석하였다.

인용 안정성. 학술적 인용에 요구되는 재현 가능성을 정량화하기 위하여, 동일 질의를 N회 반복 호출해 다수결로 판정을 결정할 때 그 결과가 해당 조건의 최빈 판정과 일치할 확률 $P(N)$을 부트스트랩(주장별 M_0 5회 반복 라벨에서 비복원 추출, 동률 시 무작위 결정, 4,000회)으로 추정하고, $P(N) \geq 0.95$를 만족하는 최소 반복 수 N^*를 인용 가능성 기준으로 정의하였다. 또한 동일한 3회 호출 예산 하에서 (i) M_0 3회 반복 앙상블과 (ii) $M_1/M_2/M_3$ 각 1회로 구성한 조사 변용 앙상블의 자기 일관성 및 M_0 최빈 판정과의 일치율을 쌍체 부트스트랩(2,000회)으로 비교하였다. 생성 단계의 문헌 선택 편향은 각 응답에서 실제 인용된 문헌과 검색되었으나 인용되지 않은 후보 문헌 간의 평균 피인용수 및 발행연도 차이를 응답 단위 쌍체 순열 검정으로 측정하였다. 마지막으로 인용–본문 정합성을 확인하기 위해 240개 응답 표본에서 인용 표지가 가리키는 문헌이

해당 문장을 뒷받침하는지를 4단계(전면 뒷받침/부분 뒷받침/불일치/접근 불가)로 평가하였으며, 서비스 계층 축에서는 30개 주장에 대한 M_0 5회 반복을 4개 엔드포인트에 동일하게 적용하여 반복 간 참고문헌 일관성(쌍별 Jaccard)을 비교하였다.

제 4. 절 연구 결과

4.1 검색 과정

검색 결과 교체. Scholar Search는 동일 질의의 반복에 대해 완전히 결정적이었다(360개 셀 전부에서 5회 반복 결과가 동일, 조건 내 Jaccard 1.000). 의미 보존 변용에 대해서는 조사 하나의 교체만으로 상위 20개 검색 문헌의 평균 64.9%가 교체되었으며($p < 10^{-4}$), M 계열 조건 쌍의 평균 Jaccard는 0.14–0.16, RBO@10은 0.15–0.17에 그쳤다. 반면 진리조건을 반전시키는 D_+–D_- 쌍의 문헌 집합은 거의 동일하게 유지되었다(Jaccard 0.966, 문헌 교체율 3.5%). 검색 단계는 질의의 의미가 아닌 표면형에 반응하였으며, 의미를 보존한 변용이 의미를 반전한 변용보다 약 19배 많은 문헌을 교체하였다.

검색 결과 성질 전환. 교체된 문헌은 유의미한 성질의 차이를 보였다(표 2; 문헌 지표는 Scholar Search, 응답 지표는 Search Agent 응답 기준). M_0 대비 조사 변용은 국내 학술지 비율을 81.2%에서 54.2–68.2%로 낮추었고, 검색 결과의 평균 출판연도를 최대 1.5년(2014.0 → 2012.5) 앞당겼으며, 주장 변인이 제목에 나타나는 주제 정합도는 32.2%에서 46.1–48.5%로 높아졌다. 생성 단계에서 실제 인용되는 문헌을 대상으로 한 분석에서도 인용 문헌의 평균 피인용수는 M_0의 273.4회에서 M_1–M_3의 180.7–197.9회로 감소하였고 평균 발행연도는 최대 3.1년(2016.8 → 2019.9) 이동하였다. 조사 교체는 답변이 어떤 문헌 위에서 작성되는가를 학술지의 언어, 시기, 인용 규모의 수준에서 바꾸어 놓았다.

4.2 증강 과정

답변의 인식 전환. 검색 과정에서의 큰 변동과 대조적으로, 생성 결과는 의미 보존 변용에 대해 안정적이었다. 조건 내 잡음 하한(0.107)을 기준으로 한 M 계열의 초과 전환율은 $\Delta_M = 0.018$(95% 신뢰구간 $[-0.028, 0.071]$, $p = .490$, Cohen's $d = 0.09$)로 유의하지 않았던 반면, 방향 반전 변용은 $\Delta_D = 0.427$($p < 10^{-4}$)의 전환을 유발하였다. 사전 등록된 4개 조건 쌍 대비에서도 동일한 구조가 확인되었으며(표 3), M 계열의 어떤 쌍도 Holm–Bonferroni 보정 후 유의하지 않았고 D_+–D_-만이 유의하였다($p_{HB} < 5 \cdot 10^{-4}$). 판정의 동의 비율은 M 계열에서 77.7–81.7%로 변용 간 차이가 적었던 반면, D_- 조건에서는 동의 39.4%, 반박 40.6%로 양분되었으며 60개 주장 중 32개(53.3%)에서 D_+와 D_-의 최빈 판정이 일치하지 않았다.

검색과 생성의 분리. 두 과정의 상호작용을 확인해 본 결과 이중 분리(double dissociation)가 나타났다. 의미 보존 변용은 근거 문헌의 약 65%를 교체하고도 판정을 바꾸지 못하였고, 방향 반전 변용은 문헌 집합을 거의 그대로 유지한 채(판정이 전환된 32개 주장에서 D 쌍 문헌 Jaccard 평균 0.964) 판정만 뒤집었다. 판정 전환과 문헌 중복의 결합(Joint) 분포에서도 전환이 발생한 조건 쌍 77개 중 50개(64.9%)가 문헌 중복이 중앙값(0.33) 이상인 쌍에서 발생하여, 판정 전환이 검색 단계의 문헌 교체가 아니라 생성 단계에서 발생함이 확인되었다.

4.3 인용 안정성

판정 가독성. 5인 주석자가 교차 평가한 180개 응답에서 전체 Krippendorff's α는 0.601(원시 일치율 0.817)이었으나, 증거적 논쟁성 계층에 따라 S_1 $\alpha = 0.916$, S_2 $\alpha = 0.578$, S_3 $\alpha = 0.280$으로 가파른 기울기를 보였다. 방향 동조(Δ_D) 또한 같은 경향을 보였다(S_1 0.07, S_2 0.59, S_3 0.62). 이는 변용에 의한 시스템 판정의 불안정성과 인간 독자의 인식적 불안정성이 나타나는 주장이 겹침을 의미한다.

반복 다수결 및 앙상블. 단일 호출의 판정이 최빈 판정과 일치할 확률은 $P(1) = 0.895$였으며, 반복 시행을 통한 다수결로 $P(3) = 0.949$, $P(5) = 0.984$까지 상승하였다. 기준 0.95를 만족하는 최소 반복 수는 $N^* = 5$였으며, $P(3)$은 기준에 0.001 차이로 미달하였다. 동일한 3회 호출 예산에서 조사 변용 앙상블($M_1/M_2/M_3$ 각 1회)은 반복 앙상블(M_0 3회)보다 자기 일관성이 낮았고(0.901 대 0.948, 쌍체 차이 -0.047, $p = .041$), M_0 최빈 판정과의 일치율 또한 낮았다(0.889 대 0.958).

인용 문헌의 선택 편향과 정합성. 생성 단계는 검색된 후보 문헌들 중 피인용수가 높은 문헌을 뚜렷하게 선호하였다(인용 문헌 평균 218.8회 대 미인용 문헌 57.8회, 응답의 94.5%에서 인용 문헌이 더 높음, $p < 10^{-4}$). 반면 최신성 면에서는 인용 문헌의 출간 날짜가 후보 문헌보다 평균 0.75년 오래되었다(2017.8 대 2018.5, $p < 10^{-4}$). 인용–본문 정합성 평가에서는 240개 표본 중 77.1%가 전면 뒷받침, 14.2%가 부분 뒷받침이었고, 명백한 불일치는 4.2%, 원문 접근 불가는 4.6%였다. 서비스 계층 프로브에서는 호출 비용이 높아질수록 반복 간 참고문헌 일관성이 오히려 하락하여(Quick Answer 0.515, Search Agent 0.318, Deep Research 0.251; 결정적 검색인 Scholar Search는 1.000), 상위 계층의 비용이 근거의 재현성을 개선하지 않음을 확인하였다.

제 5. 절 논의

본 연구의 핵심 결과는 검색과 생성이 질의의 서로 다른 성분에 반응한다는 이중 분리이다. 검색 단계는 진리조건이 아닌 표면형에 반응하여 의미가 같은 질의에 서로 다른 문헌을, 의미

표 2. 변용별 요약 통계

항목	지표	M_0	M_1	M_2	M_3	D_+	D_-
데이터	평가 응답 수	300	180	180	180	180	180
판정 비율	동의 (%)	77.7	77.8	80.0	81.7	80.0	39.4
	반박 (%)	6.0	8.3	5.6	5.6	6.1	40.6
	유보 (%)	16.3	13.9	14.4	12.8	13.9	20.0
문헌 지표	국내 학술지 비율 (%)	81.2	54.2	54.5	68.2	56.5	57.0
	주제 정합도 (%)	32.2	48.3	46.1	48.5	52.3	52.8
	검색 결과 평균 출판연도	2014.0	2012.6	2012.5	2014.1	2012.6	2012.6
응답 지표	응답 길이 (중앙값, 자)	2,046	1,895	1,889	9,937	2,238	2,537
	인용 문헌 평균 피인용수	273.4	180.7	190.2	197.9	253.3	199.9
	인용 문헌 평균 연도	2016.8	2019.9	2019.2	2017.3	2016.9	2017.0

표 3. 조건 쌍의 판정 전환 통계

대비 조건 쌍	전환율	Δ	p	p_{HB}
M_0–M_1	0.100	-0.007	0.923	0.923
M_0–M_2	0.150	0.043	0.307	0.615
M_0–M_3	0.183	0.077	0.092	0.275
D_+–D_-	0.533	0.427	$< 10^{-4}$	$< 5 \cdot 10^{-4}$

가 반대인 질의에 같은 문헌을 반환하였다. 생성 단계는 반대로 문헌 교체에는 둔감하였으나 질의가 단정하는 방향에는 민감하여, 사실상 동일한 문헌 위에서 반대 판정을 산출하였다. 이는 질의가 단정하는 방향에 판정이 동조하는 영합적(sycophantic) 경향으로 해석할 수 있으며, 이 구조에서 사용자가 관찰하는 판정의 안정성은 근거의 안정성을 보증하지 않는다. 조사 교체만으로 기반 문헌의 3분의 2가 바뀌는 상황에서도 판정이 유지된다는 것은, 시스템이 근거를 견고하게 종합하기 때문이라기보다 상이한 문헌 집합들로부터 유사한 결론을 반복적으로 재구성하기 때문일 가능성을 시사한다.

방향 동조와 판정 가독성 저하가 논쟁적 주장(S_2, S_3)에 집중된다는 사실은 실무적으로 중요한 경고이다. 근거가 확립된 주장에서 프레이밍은 거의 무해하지만, 사용자가 검색 에이전트에 실질적으로 의존하게 되는 미확정 주장에서 시스템은 질의의 단정 방향에 동조하고, 그 답변은 인간 독자 사이에서도 일관되게 해석되지 않는다. 근거가 확립된 주장에 대한 응답의 판정은 독자 집단에게 일관되게 읽히지만, 근거가 혼재된 주장일수록 같은 응답이 독자마다 다르게 해석되는 것이다. 이는 판정이 독자에게 얼마나 일관되게 전달되는가를 드러내는 Krippendorff's α의 계층별 기울기에서도 분명하게 드러난다.

이상의 결과는 상용 검색 에이전트를 학술적 맥락에서 인용하기 위한 정량적인 검증 규칙으로 종합된다. 먼저 단일 호출의 판정을 그대로 인용해서는 안 되며, 최소 3회 반복 다수결(일치 확률 0.949), 엄격하게는 $N^* = 5$를 적용해야 한다. 또한 방향 반전 변용은 근거 문헌의 다양화 없이 판정을 해당하는 방향으로 유도하기에 질의는 중립적으로 구성해야 하며, 조사 변용은 문헌 다양화를 통한 강건성 점검 수단이 될 수 있으나, 판정 안정화 수단으로써는 단순 반복이 우월하다.

본 연구의 한계는 다음과 같다. 단일 상용 시스템을 단일 기간(1일)에 감사하였으므로 결과는 특정 시점의 결합 파이프라인(재작성, 검색, 생성)에 대한 것이며, 내부 재작성기가 조사를 정규화하거나 증폭하는 개별 기전은 분리하지 못하였다. 주장 집합은 의료 및 과학 분야 60개, 주석자는 5인으로 제한되었고, 변용은 조사 및 양태 표지 4종에 국한되어 한국어 정보 구조 전반으로의 일반화에는 주의가 필요하다. 또한 생성 단계에서 판정 전환을 일으키는 문체적, 수사적 기제에 대한 질적 분석은 후속 과제로 남긴다.

제 6. 절 결론

본 연구는 한국어 조사·양태 표지 교체를 의미 보존적 실험 도구로 사용하여 상용 검색 에이전트를 판정과 문헌의 두 층위에서 감사하였다. 진리조건을 보존하는 조사 변용은 판정을 바꾸지 않았으나($\Delta_M = 0.018$, 유의하지 않음) 근거 문헌의 약 65%를 교체하였고, 진리조건을 반전하는 변용은 문헌 집합을 유지한 채 절반 이상의 주장에서 판정을 뒤집었다($\Delta_D = 0.427$, $p < 10^{-4}$). 판정의 안정성과 근거의 안정성이 분리된다는 이 결과는 생성형 검색의 신뢰성 평가가 최종 답변의 정확도만으로는 불충분하며 문헌 층위를 독립적으로 감사해야 함을 보여준다. 아울러 반복 다수결 기준($N^* = 5$, 실무적 절충 $N = 3$)과 중립형 질의 원칙 등 인용을 위한 정량적 검증 규칙을 제안하였다. 후속 연구로는 다중 시스템·다중 시점 비교, 생성 단계 판정 전환의 질적 기제 분석, 조사 외 정보 구조 표지로의 확장을 계획한다.

감사의 글

본 연구는 연세대학교 인공지능학회 YAI와 주식회사 라이너의 산학협력 프로젝트의 일환으로, 주식회사 라이너로부터

Liner, Liner API 제품 제공 및 학술 활동 지원을 받아 수행되었다. 지원 기관은 본 연구의 설계, 자료 수집, 분석, 해석 또는 논문 작성 과정에 관여하지 않았다.

참고문헌

[1] I. Besrour, J. He, T. Schreieder, and M. Färber, "Squai: Scientific question-answering with multi-agent retrieval-augmented generation," *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, ser. CIKM '25, p. 6603–6608, Nov. 2025. [Online]. Available: http://dx.doi.org/10.1145/3746252.3761471

[2] J. Shao, Y. Lin, M. P. Lohani, Y. Miao, and B. Luo, "Do LLM agents know how to ground, recover, and assess? evaluating epistemic competence in information-seeking agents," *The Fourteenth International Conference on Learning Representations*, 2026. [Online]. Available: https://openreview.net/forum?id=r0L9GwlnzP

[3] 박철우, "한국어 정보구조에서의 정보 단위 획정 문제," *언어학*, pp. 131 – 163, 2017. [Online]. Available: http://dx.doi.org/10.17290/jlsk.2017..78.131

[4] S.-N. Kwon and A. Zribi-Hertz, "Differential function marking, case, and information structure: Evidence from korean," *Language*, Vol. 84, No. 2, pp. 258–299, 2008. [Online]. Available: https://muse.jhu.edu/pub/24/article/240484

[5] H. Zhang, M. Tang, K. Bi, and J. Guo, "Beyond relevance: Utility-centric retrieval in the llm era," *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '26, p. 5357–5361, 2026. [Online]. Available: https://doi.org/10.1145/3805712.3808641

[6] X. Sun, J. Chen, B. Zhou, and M.-J. Kuo, "Conrag: Contradiction-aware retrieval-augmented generation under multi-source conflicting evidence," *Journal of Advanced Computing Systems*, Vol. 4, No. 7, pp. 50–64, jul 2024. [Online]. Available: https://scipublication.com/index.php/JACS/article/view/257

[7] Liner, "Liner Developers Documentation," https://liner.com/ko/developers/docs, 2026, accessed: September 2, 2026.

[8] X. L. Dong, S. Sharma, K. Sun, J. Wang, Y. Xia, X. Yang, S. W.-t. Yih, and F. Zhang, "Retrieval-augmented generation (rag)— from modular to agentic systems," *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, ser. KDD '26, p. 13308–13313, 2026. [Online]. Available: https://doi.org/10.1145/3770855.3816460

[9] A. R. Faridi, F. Masood, Keshvi, and T. Yunus, "Retrieval-augmented generation for large language models: Evolution, architectures, applications, and challenges (2020–2025)," *WIREs Data Mining and Knowledge Discovery*, Vol. 16, No. 3, p. e70122, 2026, e70122 1014902. [Online]. Available: https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.70122

[10] P. Jaccard, "The distribution of the flora in the alpine zone," *The New Phytologist*, Vol. 11, No. 2, pp. 37–50, 1912. [Online]. Available: http://www.jstor.org/stable/2427226

[11] W. Webber, A. Moffat, and J. Zobel, "A similarity measure for indefinite rankings," *ACM Trans. Inf. Syst.*, Vol. 28, No. 4, Nov. 2010. [Online]. Available: https://doi.org/10.1145/1852102.1852106

[12] K. Krippendorff, *Content Analysis: An Introduction to Its Methodology*, fourth edition ed. Thousand Oaks, California: SAGE Publications, Inc., Sep. 2026. [Online]. Available: https://methods.sagepub.com/book/mono/content-analysis-4e/toc

[13] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, Vol. 6, No. 2, pp. 65–70, 1979. [Online]. Available: http://www.jstor.org/stable/4615733