



한국어 질의 표현과 생성형 검색 응답: 체계적 문헌고찰 프로토콜

표현 변형과 응답 차이를 함께 다룬 문헌을 모아 무엇이 측정됐고 무엇이 비어 있는지 가리기 위한 프로토콜입니다. 고찰의 결과는 아직 담기지 않았습니다.

한지훈 · 발행 2026-06-30 · 최종 수정 2026-06-30

수록	제2권 제1호 · pp. 1-26
저자	한지훈
투고	2026-03-14
게재 확정	2026-05-22
주제어	체계적 문헌고찰 · 한국어 질의 · 표현 변형 · 프로토콜 · 근거 지도

초록

같은 의도의 질문을 다르게 쓰면 생성형 검색의 응답이 달라진다는 말은 실무에서 널리 쓰이지만, 그 말이 어디까지 확인된 것인지는 정리된 적이 없다. 이 글은 질의의 표현 변형과 응답 차이를 함께 다룬 문헌을 모아 무엇이 측정됐고 무엇이 비어 있는지를 가리기 위한 체계적 문헌고찰 프로토콜이다. PRISMA-P 가 요구하는 항목에 따라 연구 질문과 검색 전략, 선정과 배제의 기준, 선별 절차, 자료 추출 항목, 질 평가 등급, 종합 방법을 수행 전에 정해 공개한다. 마지막 절은 예비 범위 검토에서 관찰한 문헌의 형태만을 보고하며 어떤 결과 수치도 담지 않는다.

1. 배경

생성형 검색을 쓰는 사람은 같은 것을 알고 싶을 때도 매년 다른 문장으로 묻는다. “전세 계약 갱신 거절 사유”라고 적기도 하고 “집주인이 전세 계약 갱신을 거절할 수 있나요”라고 적기도 한다. 두 질문의 의도는 같고 형태는 다르다.

이 형태 차이가 답을 바꾼다는 말은 실무에서 전제처럼 쓰인다. 문제는 그 전제가 어디까지 확인된 것인지 아무도 모른다는 데 있다. 표현을 바꾸는 개입을 다룬 글은 여러 갈래에 흩어져 있고, 갈래마다 무엇을 바꾸고 무엇을 했는지가 다르다.

흩어진 채로는 다음 연구가 어디서 출발해야 하는지 알 수 없다.

이 글은 그 흩어진 것을 한 번 훑기 위한 프로토콜이다. 검색식과 선별 기준, 추출 항목과 질 평가 등급을 수행 전에 정해 공개함으로써 결과를 본 뒤에 기준을 고치는 일을 막는다. PRISMA-P 가 요구하는 항목을 순서대로 채웠고, 마지막에 예비 범위 검토에서 관찰한 문헌의 형태를 덧붙였다.

2. 연구 질문

주 질문은 하나다. 같은 의도를 유지한 채 질의의 표현만 바꾸었을 때 생성형 검색의 응답이 달라지는지를, 지금까지의 문헌은 무엇을 근거로 말하고 있는가.

하위 질문은 넷이다.

- 1. 질의의 어떤 속성이 응답 차이와 연결지어 측정된 적이 있는가 — 문형, 존대 수준, 조사와 어순, 조건의 제시 순서, 어휘 가운데 무엇이 다루어졌는가
- 2. 응답의 “차이”를 무엇으로 조작적 정의했는가 — 인용된 출처의 집합인가, 답에 담긴 주장인가, 제시 형식인가
- 3. 시스템 자체의 기저 변동과 표현 변형의 효과를 분리한 연구가 있는가
- 4. 한국어 질의를 대상으로 삼은 연구가 있는가

PICO 는 다음과 같이 변형해 적용한다.

구분	이 고찰에서의 범위
대상(P)	검색을 거쳐 답을 생성하는 시스템과 그 응답. 검색 없이 답하는 모델은 비교군으로만 다룬다
개입(I)	의도를 유지한 질의 표현의 변형 — 문형·존대·조사와 어순·조건 순서·어휘
비교(C)	변형하지 않은 기준 질의, 또는 같은 질의를 반복했을 때의 기저 변동
결과(O)	출처 집합의 차이, 주장의 차이, 제시 형식의 차이 가운데 명시적으로 정의된 것
연구 유형(S)	측정 절차를 밝힌 실증 연구, 벤치마크 논문, 재현 보고. 주장만 있는 글은 배제

3. 검색 전략

데이터베이스는 국외와 국내를 나누어 검색한다. 국외는 ACM Digital Library, IEEE Xplore, Scopus, ACL Anthology, arXiv 다섯이고, 국내는 한국학술지인용색인, RISS, DBpia, ScienceON 넷이다. 검색식은 제목·초록·키워드 필드를 대

상으로 하며, 필드 지정 문법이 데이터베이스마다 다르므로 원문 검색식을 그대로 기록해 부록에 실는다.

검색어는 세 묶음을 교차한다. 첫째는 대상 묶음으로 generative search, retrieval-augmented, answer engine, 생성형 검색, 검색 증강이다. 둘째는 개입 묶음으로 query variation, paraphrase, query rewriting, prompt sensitivity, phrasing, 질의 변형, 표현 민감도다. 셋째는 결과 묶음으로 citation, source attribution, answer consistency, 출처, 인용, 응답 일관성이다. 묶음 안은 논리합으로, 묶음 사이는 논리곱으로 잇는다.

대상 기간은 2021년 1월부터 검색 실행일까지로 한다. 검색을 거쳐 답을 만드는 상용 시스템이 일반에 열린 시점을 기준으로 잡되, 그 이전의 질의 재작성 연구가 배경으로 필요하므로 인용 추적으로만 앞쪽을 본다. 언어는 한국어와 영어로 제한한다.

포함된 문헌의 참고문헌을 거슬러 올라가는 후방 추적과 그 문헌을 인용한 글을 훑는 전방 추적을 각각 한 차례 수행한다. 회색문헌은 별도 대장에 기록하되 본 종합에는 넣지 않고 부록에서 형태만 서술한다. 기업 기술 블로그와 백서, 학회가 아닌 자리의 발표자료가 여기에 해당한다.

4. 선정·배제 기준

구분	기준
포함	질의의 표현을 의도적으로 바꾸고 그 전후를 비교한 연구
포함	응답의 차이를 측정 가능한 형태로 정의하고 그 정의를 밝힌 연구
포함	검색을 거쳐 답을 생성하는 시스템을 대상으로 한 연구
포함	동료 심사를 거친 논문, 또는 절차와 자료를 함께 공개한 사전공개본
배제	측정 없이 요령이나 권고만 제시한 글
배제	표현이 아니라 내용을 바꾼 개입을 다룬 연구
배제	응답 차이의 정의를 밝히지 않은 채 결과만 보고한 글
배제	원문을 구할 수 없거나 초록만 공개된 문헌

내용을 바꾼 개입을 배제하는 이유는 이 고찰의 개입 정의가 의도 유지이기 때문이다. 조건을 하나 더 붙이거나 범위를 좁힌 질의는 다른 질문이므로 같은 저울에 올릴 수 없다.

5. 선별 절차

선별은 두 단계로 나눈다. 1단계는 제목과 초록만 보고 명백한 비해당을 걸어내는 단계이고, 2단계는 전문을 읽고 기준을 적용하는 단계다. 두 단계 모두 두 사람이 서로의 판정을 보지 않고 독립으로 수행한다.

선별자는 연구위원 한 명과 측정·평가 분과에서 지명한 한 명으로 구성한다. 시작 전에 무작위로 뽑은 문헌 스무 편으로 예행 선별을 하고, 기준 해석이 갈린 지점을 문서로 정리한 뒤 본 선별에 들어간다. 이 정리본은 선별 도중에 고치지 않는다.

불일치는 두 선별자가 근거를 적어 교환하고 합의한다. 합의되지 않으면 편집위원회가 지명한 제3 검토자가 판정한다. 선별자 간 일치도는 단계별로 코헨 카파를 계산해 보고하며, 값이 낮으면 낮은 채로 보고하고 원인이 기준 해석의 어느 항목에 있는지 함께 적는다. 배제된 문헌은 2단계에 한해 배제 사유를 문헌별로 남겨 공개한다.

흐름도는 PRISMA 형식으로 그린다. 다만 검색 결과 건수와 단계별 잔존 편수는 검색을 실행한 뒤에 채우며, 이 글에는 빈 틀만 있다.

6. 자료 추출 항목

추출은 두 사람이 각각 하고 대조한다. 숫자 항목이 어긋나면 원문으로 돌아가 확인한다.

항목	기록 내용
서지	저자, 연도, 출처, 문헌 유형
대상 시스템	검색 결합 여부, 접근 표면이 화면인지 인터페이스인지
질의 언어	대상 언어와 번역 경유 여부
변형 축	바꾼 속성. 한 축만 바꾸었는지 여러 축을 동시에 바꾸었는지
변형 생성 방식	사람 작성인지 자동 생성인지, 의도 동일성의 확인 방법
반복 설계	같은 질의의 반복 횟수, 수집 시점의 분산
기저 변동 처리	변형 없는 반복의 변동을 따로 켜는지
결과 정의	차이를 무엇으로 정의했는지, 총위를 나누었는지
판정 주체	자동 판정인지 사람 판정인지, 일치도 보고 여부
자료 공개	질의 목록과 원자료의 공개 여부와 위치
수집 시점	수집 기간과 그 사이 대상 시스템의 변경 고지 여부

7. 질 평가

이 분야에는 무작위 배정이 없어 기존 임상 도구를 그대로 쓸 수 없다. 대신 **재현에 필요한 정보가 얼마나 공개됐는가로** 네 등급을 매긴다. 등급은 결론의 참거짓이 아니라 보고의 충실도를 가린다.

등급	기준
A	질의 목록·수집 시점·반복 횟수·판정 기준을 모두 공개하고, 기저 변동을 따로 추정했다
B	절차는 서술했으나 질의 목록이나 원자료가 공개되지 않았다
C	결과는 제시했으나 반복 설계나 판정 기준이 밝혀지지 않았다
D	추정 절차 없이 사례와 주장만 제시했다

두 사람이 독립으로 등급을 매기고 불일치는 5절과 같은 방식으로 조정한다. D 등급은 배제 기준과 겹치므로 본 종합에서 빠지지만, 그런 글이 어느 갈래에 얼마나 몰려 있는지는 형태로 보고한다.

8. 종합 방법

양적 통합은 하지 않는다. 이유는 셋이다. 결과 지표가 통일돼 있지 않아 같은 효과크기로 환산되지 않고, 대상 시스템이 시점마다 달라 같은 모집단으로 묶이지 않으며, 기저 변동을 보고하지 않은 연구의 값은 표현의 효과와 시스템의 흔들림이 섞인 상태이기 때문이다. 섞인 값을 평균 내면 없던 정밀도가 생긴다.

그러므로 서술적 종합을 택하고, 그 뼈대로 근거 지도를 만든다. 세로축은 변형 축이고 가로축은 결과 정의 층위다. 칸마다 해당 문헌의 수와 질 등급의 분포를 적고, 빈칸은 빈칸으로 남긴다. 이 지도의 목적은 어느 조합이 이미 다뤄졌고 어느 조합이 비어 있는지를 한눈에 보이게 하는 것이다.

질 등급에 따른 민감도 확인도 함께 한다. A·B 등급만 남겼을 때 지도의 모양이 크게 달라진다면 그 사실 자체를 결과로 적는다. 이질성이 심해 결론이 한 문장으로 모이지 않으면 갈래별로 따로 적는다.

9. 예비 범위 검토에서 본 것

본 검색에 앞서 검색식을 다듬으려고 소규모 예비 검토를 했다. 여기서 보고할 수 있는 것은 문헌의 형태뿐이며 어떤 결과 수치도 옮기지 않는다.

첫째, 같은 것을 여러 이름으로 부른다. 질의 재작성, 패러프레이즈 강건성, 프롬프트 민감도, 표현 변형이 서로 겹치는 뜻으로 쓰이고, 한쪽 이름으로만 검색하면 다른 쪽이 통째로 빠진다.

둘째, 문헌이 두 계열로 갈려 있고 서로를 잘 인용하지 않는다. 한쪽은 검색 없이 답하는 모델의 입력 민감도를 다루고, 다른 쪽은 검색 결과의 인용과 출처를 다룬다. 표현 변형과 출처 변화를 한 연구 안에서 같이 본 글은 눈에 잘 띄지 않았다.

셋째, 결과의 정의가 제각각이다. 답변 문장의 표면 유사도로 차이를 말한 글과 인용된 출처의 집합으로 말한 글이 같은 결론처럼 읽히는 자리가 있었다.

넷째, 반복 측정을 밝힌 글이 드물다. 한 번 물어 보고 답이 달랐다는 서술이 적지 않으며, 그 경우 표현의 효과와 시스템의 기저 변동이 분리되지 않는다.

다섯째, 한국어 질의를 대상으로 삼은 글은 예비 검토 범위에서 확인되지 않았다. 조사와 어순, 존대 수준처럼 한국어에서 두드러지는 속성이 변형 축으로 등장한 사례도 마찬가지다. 다만 예비 검토는 범위가 좁으므로 이를 부재의 증거로 삼지 않는다.

여섯째, 회색문헌에는 요령을 권하는 글이 많고 절차를 밝힌 글이 드물다. 이 비대칭 자체가 기록할 만한 형태다.

10. 한계

- **색인 편향.** 선택한 아홉 개 데이터베이스에 색인되지 않은 문헌은 후방·전방 추적으로만 잡히며, 그 물을 빠져나간 분량은 알 수 없다
- **언어와 기간의 절단.** 한국어와 영어로 제한했고 2021년을 시작점으로 삼았다. 다른 언어권의 논의와 그 이전의 질의 재작성 연구가 부분적으로만 들어온다
- **대상의 비영속성.** 관측된 시스템은 시간이 지나면 바뀐다. 문헌이 보고한 상태를 다시 확인할 수 없는 경우가 많으므로 종합은 과거 시점들의 모음이지 현재의 진술이 아니다
- **질 등급의 성격.** 등급은 보고의 충실도만 가린다. A 등급 연구의 결론이 참이라는 뜻이 아니고, C 등급 연구의 관찰이 거짓이라는 뜻도 아니다
- **회색문헌의 분리.** 실무의 축적이 본 종합에서 빠진다. 부록의 형태 서술로는 그 손실을 메우지 못한다
- **선별자의 관점 중복.** 두 선별자가 같은 학회에서 일하므로 기준 해석이 같은 방향으로 치우칠 수 있다. 예행 선별 기록과 배제 사유를 모두 공개해 외부에서 되짚을 수 있게 하는 것이 유일한 완화책이다

11. 공개와 갱신

이 글은 프로토콜이며 고찰의 결과를 담고 있지 않다. 따라서 공개의 대상도 결론이 아니라 절차다.

검색을 실행하기 전에 이 프로토콜을 학회 홈페이지에 고정 주소로 올린다. 실행 뒤에는 데이터베이스별 원문 검색식과 실행일, 흐름도의 채워진 숫자, 2단계 배제 문헌의 사유 목록, 추출표 전체를 함께 공개한다. 추출표는 표 형식의 원자료로 내려받을 수 있게 하고 출처 표시 조건으로 재사용을 허용한다.

프로토콜을 고치면 무엇을 왜 언제 고쳤는지 변경 이력에 남긴다. 검색은 12개월마다 같은 검색식으로 다시 실행하며, 갱신 회차마다 새로 들어온 문헌과 근거 지도의 달라진 칸을 따로 표시한다. 같은 절차를 다른 연구자가 다시 밟아 다른 지도를 얻는다면, 그 차이가 어디서 생겼는지를 이 기록으로 짚을 수 있어야 한다.

참고문헌

- Aggarwal, P. et al. (2023). *GEO: Generative Engine Optimization*. arXiv:2311.09735. — 질의와 문서의 문면을 바꾸는 개입을 다룬 초기 연구로, 이 고찰의 개입 정의와 검색어 묶음을 잡는 출발점으로 삼았다.
- 한국AI검색학회 측정·평가 분과. 반복 측정 설계 초안. 내부 검토 중. — 7절의 기저 변동 관련 등급 기준과 6절의 반복 설계 추출 항목이 이 초안을 따른다.