

AI/ML

그래프RAG는 정말 지고 있을까

문서를 그래프로 엮어 AI에게 읽히는 방식이 몇 해 동안 당연한 다음 단계처럼 이야기됐습니다. 그런데 실증 연구가 쌓이면서 이런 보고가 이어졌습니다. 여러 실증 연구가, GraphRAG가 많은 과제에서 일반 RAG를 넘지 못한다고 보고했다.

방동걸 (주)웨이스 대표이사

2026년 9월 2일

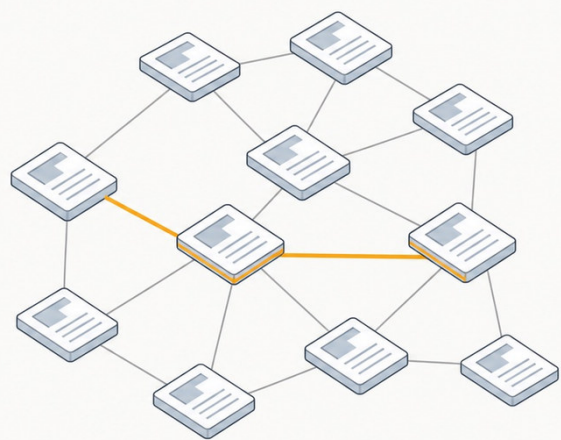
문서를 그래프로 엮어 AI에게 읽히는 방식이 몇 해 동안 당연한 다음 단계처럼 이야기됐습니다. 그런데 실증 연구가 쌓이면서 이런 보고가 이어졌습니다. VEXPLOR에서 제조 데이터를 지식그래프로 다루는 웨이스가 그냥 넘길 수 없었던 보고들입니다.

여러 실증 연구가, GraphRAG가 많은 과제에서 일반 RAG를 넘지 못한다고 보고했다.

올해 이 상황을 정면으로 다룬 논문이 두 편 나왔습니다. 한 편은 제목이 「RAG에서 그래프를 언제 쓸 것인가」(ICLR 2026), 다른 한 편은 아예 질문입니다. 「우리에게 아직 GraphRAG가 필요한가」.

둘 다 읽었습니다. 그리고 예상과 다른 것을 확인했습니다. 두 논문 어느 쪽도 "그래프는 필요 없다"고 말하지 않습니다. 오히려 한 편은 그동안의 시험 방식 자체를 문제 삼습니다.

조각을 꺼내거나, 지도를 따라가거나



조각을 꺼내 오는 방식과 지도를 따라가는 방식 — 같은 문서를 두고 검색 단위가 다릅니다

01

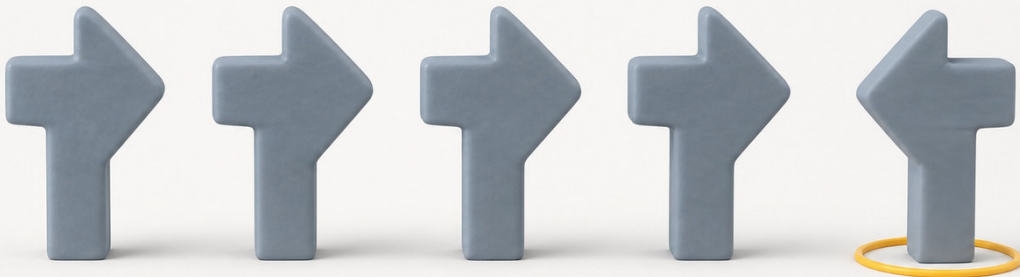
판이 다섯 방향으로 갈라졌습니다

먼저 올해 나온 연구를 늘어놓으면 어디로 가고 있는지가 보입니다.

방향	무엇을 바꿨나
값싸게	그래프 짓는 비용 절감
기억하게	여러 AI가 기억 공유
분야별로	법률 추론 전용
돌아다니게	지도가 아니라 활동 공간
되묻기	그래프가 필요한가

앞의 넷은 그래프를 더 잘 쓰려는 쪽이고, 마지막 하나는 필요 여부를 되묻는 쪽입니다. 그런데 되묻는 쪽도 결론은 "쓰지 말자"가 아니었습니다. 뒤에서 보겠습니다.

넷은 더 잘 쓰려 하고, 하나는 되물습니다



넷은 그래프를 더 잘 쓰는 쪽으로, 하나는 필요 여부를 되묻는 쪽으로 갔습니다

값싸게 쪽이 올해 가장 붐볐습니다. 이름만 들어도 방향이 보입니다. 검색을 두 단으로 줄이거나, 그래프 구조 자체를 단순하게 만들어 짓는 값을 깎거나, 관계 추출을 아예 빼고 가벼운 이름 인식만 남기는 쪽입니다.

돌아다니게 쪽은 성격이 다릅니다. 지금까지 그래프는 꺼내 보는 표였는데, 이제는 AI가 그 안을 걸어 다니며 여러 번 찾고 스스로 되짚는 공간으로 다릅니다. 한 번 찾고 끝내지 않습니다.

경쟁의 질문이 바뀌고 있습니다. *****그래프를 쓸 것인가**가 아니라 **"어떤 그래프를 얼마나 싸게 지을 것인가"*****입니다.

02

"필요 없다"는 쪽의 논리

되묻기 쪽 주장은 이렇습니다.

AI가 스스로 여러 번 찾고 되짚을 수 있게 되면서, **미리 지어 둔 지도가 하던 일을 그때그때 해낼 수 있게 됐다**는 것입니다. 지도를 짓는 값을 안 치르고도 비슷한 답이 나온다면, 지을 이유가 줄어들 겁니다.

여기까지는 두 논문이 겹칩니다. 그래프를 쓴다고 항상 나아지지 않는다는 것입니다.

일리가 있습니다. 그래프는 공짜가 아니고, 공짜가 아닌 것은 값을 하는지 매번 물어야 합니다.

다만 이 논문도 절반만 그렇게 말합니다. 같은 저자들이 결론에 이렇게 적어 뒀습니다. **여러 단계를 이어야 하는 추론에는 그래프 검색이 여전히 중요하고, 조건이 복잡해질수록 그래프 쪽이 일관되게 더 나은 성능과 안정성을 보였다**. 실제로 멀티홉 질문에서는 그래프 쪽이 평균 27점 앞섰고, 여섯 시험 세트 중 셋이 그 유형이었습니다.

제목이 질문형인 것은 **"버려도 되나"**가 아니라 *****어디까지 필요한가*****에 가깝습니다.

03

ICLR 논문이 실제로 지적한 것

여기서 짚어야 할 것이 있습니다. 이 논문은 **"그래프가 못하다"**를 결론으로 낸 논문이 **아닙니다**. 그 보고들을 출발점으로 삼아, 왜 그런 결과가 나왔는지를 파고든 논문입니다.

그리고 저자들이 내린 진단은 이렇습니다.

기존 벤치마크들이 GraphRAG를 제대로 평가하지 못하고 있다 — 분야별 자료가 없고, 과제 구분이 지나치게 단순하기 때문이다.

시험지가 문제였다는 것입니다. 그래서 이 논문은 새 시험지를 만들었습니다. 소설과 의료 두 분야에서, 난이도를 네 층으로 나눠서.

그 결과 이득이 나는 구간과 손해인 구간이 갈렸습니다.

질문 유형	결과
단순 사실 검색	이득 없거나 손해
깊은 맥락 추론	이득
계층을 따라가는 검색	이득

단순 사실 검색에서 손해라는 대목이 중요합니다. 답이 한 곳에 있는 질문에 지도를 지어 두면 **값만 치르고 얻는 게 없습니다**. 그러니 앞서 쌓였던 "일반 RAG보다 못하다"는 보고들은, 시험 문제 상당수가 그 유형이었다는 뜻이기도 합니다.

거꾸로 읽으면 이렇습니다. **계층을 따라가야 하는 질문과 깊게 파고드는 질문에서는 이득이 난다고 같은 논문이 적었습니다**. 배경 문장만 옮기면 이 절반이 사라집니다.

04

새 시험지에도 우리 문제는 없습니다

ICLR 논문이 "분야별 자료가 없다"를 문제로 짚은 것은 정확합니다. 그런데 그 논문이 새로 고른 분야가 소설과 의료입니다.

연구	시험 범위
ICLR 논문	소설·의료
되물기 논문	위키 기반 6종
공통	산업 현장 없음

되물기 논문이 쓴 여섯 세트는 Natural

Questions·PopQA·TriviaQA·HotpotQA·Musique·2WikiMultiHopQA입니다. 위키백과에서 답이 나오는 질문들이고, "그 배우가 출연한 영화의 감독은 누구인가" 같은 형태입니다.

이런 질문에는 특징이 있습니다. 등장하는 이름이 이미 세상에 적혀 있습니다. AI가 학습할 때 수없이 봤던 이름입니다.

그래서 같은 지적을 한 번 더 할 수 있습니다. 시험지를 문제 삼은 논문의 새 시험지에도, 산업 현장은 여전히 없습니다.

05

갈리는 곳은 "몇 개를 이어야 하는가"였습니다

여기에 참고할 만한 실측이 하나 있습니다. GPT-4로 보험 도메인 질문 43개를 풀리면서, 같은 데이터를 관계형 DB로 물었을 때와 지식그래프로 물었을 때를 비교한 연구입니다.

질문 유형	SQL	지식그래프
전체	16.7%	54.2%
단순한 구조	25~37%	67~71%
복잡한 구조	0%	35~38%

여기서 "복잡한 구조"는 답 하나를 만드는 데 테이블 다섯 개가 넘게 걸리는 질문입니다.

두 가지가 동시에 읽힙니다.

- 구조가 복잡해질수록 격차가 벌어집니다. 단순한 질문에서는 둘 다 그럭저럭 되지만, 다섯 개를 넘어가면 한쪽은 0%가 됩니다
- 그래프를 붙여도 절반 넘게 틀립니다. 35~38%는 자랑할 숫자가 아닙니다

두 번째를 빼고 첫 번째만 인용하면 광고가 됩니다. 그래서 같이 적습니다. 그리고 이 연구를 낸 저자들이 지식그래프 회사 소속이라는 점도 함께 적어 둡니다.

06

시험지에 없는 문제

이제 앞의 두 벤치마크로 돌아가 보겠습니다.

소설과 의료와 위키 상식 질문은 대체로 짧은 연결로 풀립니다. 그리고 짧은 연결이면 AI가 여러 번 찾아보는 것으로 대신할 수 있습니다. 되묻기 논문이 관찰한 것이 정확히 그것입니다.

공장 질문은 생김새가 다릅니다.

"지난달 그 불량률이 늘어난 라인에서, 같은 자재를 쓰는 다른 공정은 어디입니까"

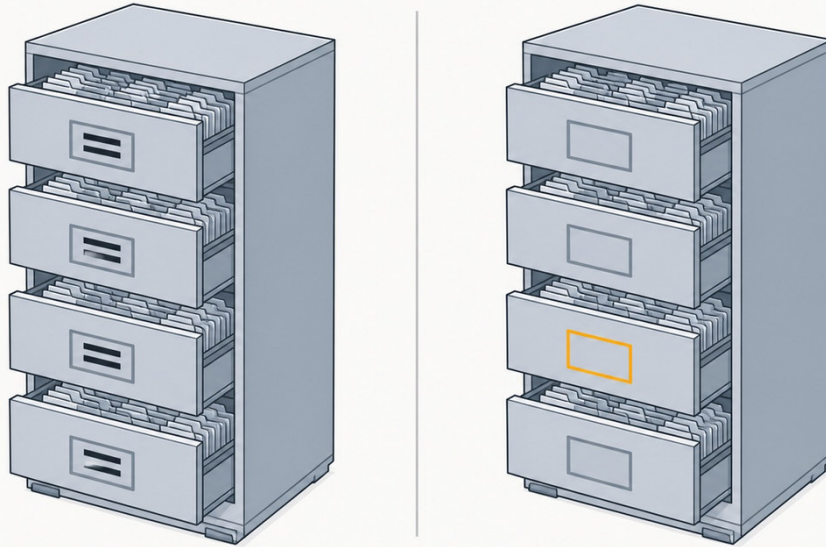
이 한 문장에 설비·라인·자재·공정·기간·품질 기록이 한꺼번에 걸립니다. 위 실측이 말한 "테이블 다섯 개가 넘는 질문"이 특별한 경우가 아니라 기본값인 영역입니다.

그리고 이런 질문은 현장에서 특별한 요청이 아닙니다. 불량률이 늘면 누구나 묻는 것이고, 답이 늦으면 그날 라인이 그대로 돕니다. 시험지에 안 나온 유형이라고 해서 드문 질문인 것은 아닙니다.

여섯 세트 전부 위키백과에서 만든 문제입니다(하나만 웹이 섞여 있습니다). 공장에만 있는 이름으로 시험한 항목은 하나도 없습니다.

그러니 이 논문을 "그래프는 이제 필요 없다"로 읽으면 저자들이 쓰지 않은 말을 읽는 것입니다. 저자들이 쓴 말은 *****위키 기반 일반 질의응답에서는 이점이 줄었다*****입니다.

AI가 아는 이름과 모르는 이름



안에 든 것은 양쪽이 같습니다. 한쪽만 이름표가 비어 있습니다

07

값을 깎아도 남는 일이 있습니다

올해 경량화 연구가 왜 그렇게 많이 나왔는지도 짚고 넘어갈 만합니다.

지금까지 그래프를 지으려면 문서를 통째로 AI에게 읽혀 "이것과 저것은 이런 관계"를 뽑게 했습니다. 문서가 많으면 그만큼 값이 듭니다. 올해 나온 연구들은 이 단계를 여러 방식으로 줄였습니다. 검색을 두 단으로 나누거나, 그래프 구조 자체를 성글게 만들거나, 태그로 계층을 유도하거나, 관계 대신 더 작은 단위로 쪼개는 쪽입니다.

방향은 하나입니다. **그래프를 포기하지 않으면서 짓는 값을 깎는 것**. 이 진영이 커진 것 자체가 "그래프가 필요 없다"보다는 "그래프가 비쌌다"에 가까운 진단입니다.

그런데 값을 깎아도 안 없어지는 일이 하나 있습니다. **같은 것을 하나로 묶는 일**입니다.

관련해 자주 인용되는 시연이 있습니다. 규칙 없이 AI에게 관계를 뽑게 하면 노드 34개에 관계 종류가 15가지로 흩어지는데, 종류를 먼저 정해 주면 23개에 4가지로 줄고, 그다음 각 대상에 식별자를 붙여야

비로소 중복이 합쳐집니다. 정리하면 식별자가 없으면 그래프를 지어도 의미가 없다는 것입니다. (가상의 회사를 예로 든 시연이고, 실제 도입 규모와는 다른 이야기입니다.)

여기가 경량화로 안 사라지는 자리입니다. 관계를 뽑는 값은 깎을 수 있어도, **무엇과 무엇이 같은 것인지 정하는 일은 사람이 답을 줘야** 합니다. 도면의 정식 명칭과 현장 줄임말과 시스템 코드가 같은 물건인 줄은 데이터 어디에도 안 적혀 있는 경우가 많습니다.

08

그때그때 만든 길은 되짚을 수 없습니다

벤치마크에 안 나오는 것이 하나 더 있습니다.

미리 지은 지도에는 관계가 적혀 있습니다. 답이 이상하면 어느 줄이 잘못됐는지 찾아가 고칠 수 있고, 고치면 다음부터 그 답이 달라집니다.

AI가 그때그때 길을 찾는 방식은 다릅니다. 답이 이상해도 고칠 자리가 없습니다. 다시 물으면 다른 길로 갈 수도 있습니다. 좋은 날은 맞고 나쁜 날은 틀립니다.

사무실 질의응답이면 이 정도는 건넵니다. 다시 물으면 되니까요.

설비를 움직이는 판단에서는 다릅니다. 왜 그 자재를 그 라인에 넣기로 했는지 나중에 설명해야 하는 자리가 있고, 같은 조건이면 같은 결정이 나와야 하는 자리가 있습니다. 여기서는 정확도 몇 퍼센트보다 **되짚을 수 있는가**가 먼저입니다.

정확도만 놓고 두 방식을 비교하면 이 항목이 통째로 빠집니다. 벤치마크는 이걸 재지 않습니다.

고칠 자리가 있습니까



왼쪽은 다시 물으면 다른 길로 갑니다. 오른쪽은 어느 줄이 틀렸는지 찾아가 고칠 수 있습니다

09

그래서 어떻게 읽어야 하는가

정리하면 이렇습니다.

맞는 부분 — 그래프를 기본값으로 두는 것은 틀렸습니다. 짧은 연결로 풀리는 질문에서는 값만 치르고 얻는 게 없습니다. 올해 경량화 연구가 쏟아진 것도 그 값이 실제로 아팠다는 뜻입니다.

두 논문이 실제로 한 말 — 한 편은 기존 시험지가 GraphRAG를 제대로 못 잤다고 했고, 다른 한 편은 여러 단계를 잇는 질문에서는 그래프가 여전히 앞선다고 했습니다. 어느 쪽도 "쓰지 말라"가 아닙니다.

여전히 빠진 부분 — 시험지를 문제 삼은 그 논문의 새 시험지도 소설과 의료였습니다. 한 답에 대여섯 개가 걸리는 질문이 기본값인 영역은 아직 아무도 안 잤습니다.

아예 없는 부분 — 되짚을 수 있는가는 어느 벤치마크에도 없습니다.

그러니 이 논문들을 근거로 "그래프는 끝났다"고 말하기는 어렵고, 반대로 "그래도 그래프가 낫다"고 뭉뚱그리기도 어렵습니다. 각자의 질문이 몇 개를 이어야 답이 나오는지부터 세어 보는 편이 빠릅니다. 그건 남의 벤치마크가 아니라 자기 데이터로만 알 수 있습니다.

세는 방법은 간단합니다. 최근에 실제로 받은 질문 스무 개를 적어 놓고, 하나씩 답을 만들려면 어느 표를 몇 개나 열어야 하는지만 세어 보십시오. 도구도 예산도 필요 없고 한 시간이면 끝납니다.

- 대부분이 한두 개에서 끝나면 — 그래프를 지어도 값을 못 뽑습니다. 위 벤치마크의 결론이 그대로 적용되는 자리입니다
- 다섯 개를 넘는 질문이 세 개에 하나꼴이면 — 그 영역은 저 시험지 밖에 있습니다. 남의 결론을 그대로 가져오면 안 됩니다
- 세다가 "이건 표에 없고 사람한테 물어야 한다"가 자주 나오면 — 순서가 다릅니다. 그래프보다 그 사실을 어디에 적어 둘지가 먼저입니다

세 번째가 가장 자주 나옵니다. 그리고 이건 어느 방식을 고르든 남는 일입니다.

10

저희도 아직 답을 찾는 중입니다

이 글이 남의 논문 읽고 판정만 하는 글로 끝나면 절반입니다. 저희가 어디에 서 있는지도 적어 두겠습니다.

아직 못 푼 것부터 적습니다.

- **관계를 뽑는 값** — 경량화 연구들이 이 값을 깼습니다. 저희가 그 방식들을 아직 다 따라가 보지 못했습니다
- **같은 것을 하나로 묶는 일** — 도면 이름과 현장 줄임말과 시스템 코드를 잇는 작업입니다. 아직 사람 손이 제일 많이 들어가는 자리이고, 이걸 줄이는 방법을 계속 찾고 있습니다
- **우리 영역의 시험지** — 산업 현장에서 그래프가 언제 값을 하는지 제대로 켜 벤치마크가 아직 없습니다. 남이 만들어 주기를 기다리기보다 우리가 재야 할 몫이라고 봅니다

바꾼 것도 적습니다. 그래프를 기본값으로 두지 않기로 했습니다. 질문이 한 문서 안에서 끝나면 지도를 짓지 않습니다. 이건 이번에 읽고 정한 것이고, 다음 논문을 읽으면 또 바뀔 수 있습니다.

부정하는 연구가 나올 때마다 방어할 논리를 찾기보다, **어디까지 맞는지 확인하고 그만큼 고치는 쪽으로** 가려고 합니다. 반년이면 지형이 바뀌는 분야입니다. 지금 쥐고 있는 답이 계속 맞을 거라고 보지 않습니다.

11

제목만 보고 넘기지 않는 편이 낫습니다

쓰고 있는 방식을 부정하는 것처럼 보이는 연구가 나오면 불편합니다.

그런데 이번에 읽으면서 얻은 것이 하나 있습니다. **두 편 다 제목이 주는 인상과 결론이 달랐습니다.** 한 편은 "언제 쓸 것인가"를 묻는 것처럼 보였지만 실제로는 기존 시험 방식을 문제 삼는 논문이었고, 다른 한 편은 "아직 필요한가"로 시작해서 "여러 단계 질문에는 여전히 필요하다"로 끝났습니다.

무엇을 잤고 무엇을 안 잤는지만 확인하면, 어디까지 내 이야기이고 어디부터 아닌지가 갈립니다.

벤치마크는 결론이 아니라 시험지입니다. 시험지를 볼 때는 점수보다 문제부터 보는 편이 낫습니다. 어떤 문제가 나왔는지를 알아야 그 점수가 나에게 무슨 뜻인지도 정해집니다.

그쪽에서는 이 판정을 어떻게 하고 계신지 궁금합니다.

WACE(웨이스)는 공장이 스스로 판단하고 멈추지 않게 하는 제조 AI 운영체제 VEXPLOR를 만드는 회사다.



원문 blog.vexplor.com/post/graphrag-benchmarks
발행 2026년 9월 2일 · 방동걸 (주)웨이스 대표이사

회사 주식회사 웨이스 (WACE Inc.) · 대표이사 방동걸 ·
사업자등록번호 439-87-02852
주소 (18462) 경기 화성시 동탄구 동탄영천로 150, 에이동(A동)
1332·1333·1334·1335호 (영천동, 실리콘엘리동탄)
연락처 전화 070-4730-0629 · contact@wace.me
홈페이지 wace.me · linkedin.com/company/wace-inc



원문



홈페이지



링크드인