인사이트

AI에 남의 프롬프트를 그대로 넣어도 같은 결과가 나오지 않는 이유는 공개되지 않은 시행착오에 있습니다

방동걸 (주)웨이스 대표이사 2026년 10월 12일

양자컴퓨팅 이론의 석학 스콧 애런슨 교수는 2025년 9월 GPT-5의 도움으로 증명 하나를 끝냈습니다. 그런데 그가 블로그 댓글에 남긴 기록을 보면 AI에게 던진 질문은 9개였고, 쓸 만한 답은 8번째 질문 뒤에야 나왔습니다. 첫 답은 그럴듯했지만 틀린 답이었고, 애런슨은 그것이 틀렸다는 것을 알아봤습니다.

제조 AI 운영체제 VEXPLOR를 만드는 웨이스가 이 기록에서 주목한 것은 증명이 아니라 그 앞의 8번입니다. SNS에는 AI로 난제를 풀고 게임을 만들었다는 결과가 매일 올라옵니다. 같은 프롬프트를 옮겨 넣어도 같은 결과가 나오지 않는 이유는, 결과물에 실리지 않은 시행착오와 그것을 걸러 낸 판단에 있습니다. 웨이스는 이 글에서 그 근거를 연구로 확인하고, 저희가 실제 업무에서 쌓은 시행착오가 어떤 결과를 냈는지도 함께 적습니다.

01

양자컴퓨팅 석학이 AI에게 받은 첫 답은 틀린 답이었습니다

애런슨 교수와 프레이크 비테베인 연구원은 양자 복잡도 이론의 한 문제(QMA의 블랙박스 증폭 한계)를 풀고 있었습니다. 애런슨은 2025년 9월 27일 블로그 글 「The QMA Singularity」에서 그 과정을 이렇게 적었습니다.

- **5분 뒤 받은 첫 답** — 「자신감 있고, 그럴듯해 보이고, (내가 보기에) 틀린 답」이었습니다.
- **틀린 이유를 알려 준 뒤** — AI가 사과하고 다시 시도했고, 답이 조금씩 나아졌습니다. 애런슨은 이 과정을 「대학원생이나 동료와 주고받는 것과 비슷했다」고 적었습니다.
- **30분 안에 나온 핵심 아이디어** — 증명의 열쇠가 된 함수 하나를 AI가 제안했고, 연구진은 그것이 맞는지를 AI 없이 직접 확인했습니다.

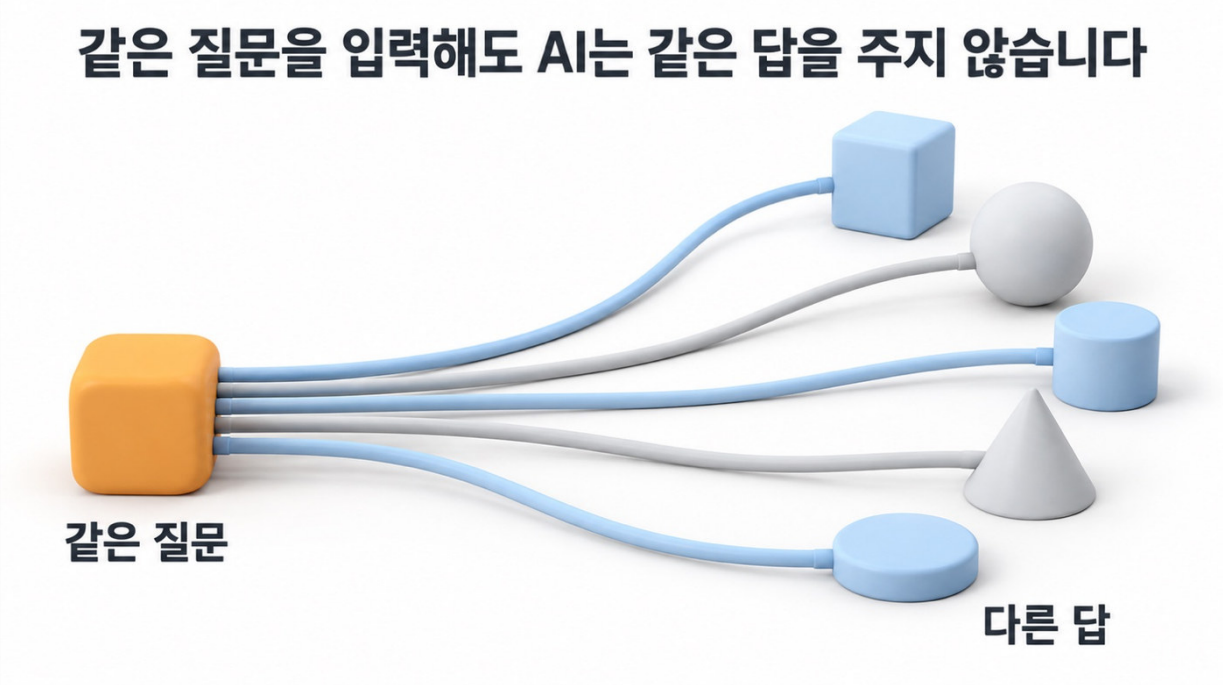
애런슨은 이 아이디어가 「돌이켜 보면 당연하다」는 비판도 받아들였습니다. 그는 「우리가 더 많이 알았거나, 문헌을 더 오래 공부했거나, 전문가에게 물었다면 당연했을 것」이라고 답했습니다. 같은 질문을 입력했더라도 첫 답이 틀렸다는 것을 알아볼 지식이 없었다면, 그 사람은 틀린 증명을 들고 끝났을 것입니다.

필즈상 수상자 테렌스 타오 교수도 같은 점을 짚습니다. 그는 2026년 1월 수학계의 난제 모음인 에르되시 문제에서 AI가 실제로 풀어 낸 비율을 1~2% 수준으로 보았고, 실패한 시도는 잘 보고되지 않아 편향이 크다고 적었습니다. 타오가 자신에게 둔 원칙은 단순합니다. AI의 결과를 수업 시간에 직접 설명할 수 없다면 그 일에는 AI를 쓰지 않습니다. SNS에 올라오는 성공 사례 뒤에는 올라오지 않은 실패가 훨씬 많습니다.

02

같은 질문을 입력해도 AI는 같은 답을 주지 않습니다

결과를 따라 하기 어려운 이유는 사람에게만 있지 않습니다. AI부터가 같은 입력에 같은 출력을 내지 않습니다.



같은 질문 하나를 넣어도 AI는 매번 다른 답을 낼 수 있습니다

연구	실험	결과
Thinking Machines (2025.9)	같은 질문을 온도 0으로 1,000번 입력	서로 다른 답 80가지
Sclar 외 (ICLR 2024)	뜻이 같은 프롬프트의 형식만 변경	정확도 차이 최대 76%포인트

Thinking Machines 연구소의 호레이스 허는 「리처드 파인먼에 대해 알려 줘」라는 같은 문장을 무작위성을 끈 설정(온도 0)으로 1,000번 입력했습니다. 답은 80가지로 나뉘었고, 첫 102개 토큰까지 같다가 103번째 토큰부터 달라졌습니다. 원인은 서버에 동시에 몰린 요청 수에 따라 계산 묶음 크기가 바뀌는 데 있었습니다. 내가 입력한 시각에 서버가 얼마나 바빴느냐에 따라서도 결과가 달라집니다.

멜라니 스클라 연구진은 뜻이 같은 프롬프트에서 구분 기호, 띄어쓰기, 대소문자만 바꿨습니다. LLaMA-2-13B 모델에서 정확도 차이는 최대 76%포인트였습니다. 모델을 키우거나 예시를 늘려도 이 차이는 사라지지 않았습니다.

공개된 프롬프트는 그 사람이 여러 번 고쳐 쓴 끝에 남은 마지막 판입니다. 그 판이 그 사람의 모델·시점·작업에서 잘 맞았다는 기록이지, 다른 사람의 작업에서도 맞는다는 보장은 아닙니다.

03

프롬프트 요령을 배운 사람이 오히려 더 많이 틀렸습니다

하버드 경영대학원과 보스턴컨설팅그룹(BCG)은 2023년 컨설턴트 758명을 대상으로 실험했습니다. GPT-4를 쓰지 않는 집단, GPT-4를 쓰는 집단, GPT-4와 프롬프트 작성 교육을 함께 받은 집단으로 나눴습니다.

과제 구분	AI 사용 집단의 변화	비고
AI가 잘하는 과제 18개	처리량 12.2%↑ · 속도 25.1%↑	품질 40% 이상↑
AI가 못하는 과제 1개	정답률 19%포인트↓	AI 미사용 84.5% 정답
같은 과제, 교육 받은 집단	정답률 24%포인트↓	교육 없는 집단은 13%포인트↓

연구진은 이것을 「들쭉날쭉한 기술 경계」라고 불렀습니다. 난이도가 비슷해 보여도 AI가 잘하는 일과 못하는 일이 섞여 있고, 그 경계는 겉으로 보이지 않습니다. 눈여겨볼 대목은 마지막 줄입니다. 프롬프트 요령을 배운 집단이 AI를 더 믿었고, AI가 못하는 과제에서 더 많이 틀렸습니다. 요령은 결과를 보장하지 않습니다. 어디까지 맡겨도 되는지 아는 판단이 결과를 좌우합니다.

AI 연구기관 METR의 2025년 실험도 같은 쪽을 가리킵니다. 숙련 개발자 16명이 자기 프로젝트의 실제 과제 246개를 처리했습니다. 개발자들은 AI가 속도를 24% 높여 줄 것으로 예상했고, 끝난 뒤에도 20% 빨라졌다고 느꼈습니다. 측정해 보니 AI를 쓴 과제가 19% 더 오래 걸렸습니다. METR은 2026년 2월 이 결과가 최신 도구에는 맞지 않는다고 스스로 밝혔습니다. 그래도 남는 사실이 있습니다. 쓰는 사람의 느낌과 실제 결과는 다를 수 있고, 그 차이는 직접 측정해야만 보입니다.

04

논문만 보고 만든 레이저는 빛을 내지 못했습니다

이 현상은 AI 이전에도 있었습니다. 과학사회학자 해리 콜린스는 1970년대 초 여러 연구소가 새로운 레이저(TEA 레이저)를 따라 만드는 과정을 추적했습니다. 학술지에는 만드는 법이 자세히 실려 있었습니다. 그런데 논문만 보고 만든 연구소의 레이저는 겉모양은 같아도 빛을 내지 못했습니다.



학술지의 설명대로 만든 레이저도 작동하는 연구소의 경험 없이는 빛을 내지 못했습니다

콜린스는 2021년 인터뷰에서 「이미 작동하는 레이저가 있는 연구소에서 시간을 보낸 적이 없는 사람은 아무도 레이저를 작동시키지 못했다」고 말했습니다. 처음 만든 사람들조차 무엇이 성패를 좌우하는지 정확히 몰라서 글로 적을 수 없었습니다. 철학자 마이클 폴라니가 1966년에 남긴 문장 그대로입니다. 「우리는 말할 수 있는 것보다 더 많이 안다.」

AI로 결과를 낸 사람들의 프롬프트와 컨텍스트 파일은 그 논문과 같습니다. 내용은 정확하지만, 그 문서가 나오기까지 버린 시도와 걸러 낸 판단은 담기지 않습니다.

05

웨이스는 업무 노하우를 규칙과 도구로 바꿔 쓰고 있습니다

웨이스는 제안서, 견적, 메일 발송, 회사 정보 관리 같은 일반 사무 업무를 AI와 함께 처리합니다. 2026년 8월부터 두 달 남짓 동안 AI 작업에서 막힌 일을 어떻게 풀었는지 한 줄씩 적었고, 그 기록이 800건을 넘었습니다. 이 기록은 규칙과 도구가 됐고, 지금 매일 업무에 쓰입니다.



웨이스는 막힌 일을 푼 기록을 규칙과 도구로 바꿔 매일 업무에 사용합니다

업무	지금 쓰는 방식	결과
메일 발송	첨부를 원문으로 검사한 뒤 발송	첨부가 확인된 메일만 나감
회사 숫자	회사 사실 27개를 한 파일로 관리	문서마다 같은 숫자로 맞춤
견적	정본 하나에서 자동 전파	가격 일치를 매일 점검
새 업무 착수	작업 기록 먼저 검색	3,000건 넘는 기록에서 시작

- **메일 발송** — 메일 프로그램이 「보냄」이라고 응답해도 그대로 믿지 않습니다. 먼저 저희 앞으로 시험 발송한 메일의 원문을 열어 첨부 파일이 실제로 들어 있는지 검사하고, 통과해야만 받는 사람에게 보냅니다.
- **회사 숫자** — 인원·연락처·특허·매출 같은 회사 사실 27개를 한 파일에 모으고, 쓰면 안 되는 옛 값 6개를 따로 적어 둡니다. 매일 아침 자동 점검이 문서와 대조해, 제안서·소개서·홈페이지가 같은 숫자를 사용하도록 맞춥니다.
- **견적** — 단가는 한 곳에서만 고칩니다. 공개 가격표, 홈페이지 가격대, 라이선스 프로그램은 도구가 그 값을 받아 생성하고, 가격이 서로 맞는지 매일 점검합니다.
- **새 업무 착수** — 새 일을 시작하기 전에 3,000건이 넘는 작업 기록과 해결 기록을 먼저 검색합니다. 누군가 이미 풀어 둔 방법이 있으면 거기서 시작하므로 같은 일을 처음부터 다시 하지 않습니다.

웨이스의 업무 규칙 문서에는 규칙이 언제, 왜 바뀌었는지 적은 기록이 139개 있습니다. 이 규칙 문서는 누구나 복사할 수 있습니다. 하지만 그 뒤에 쌓인 800여 건의 노하우는 복사되지 않습니다. 남의 프롬프트를 옮겨도 같은 결과가 나오지 않는 이유를 저희는 직접 쌓아 보며 확인했습니다.

06

공장에서도 AI의 결과는 검증 장치가 있어야 쓸 수 있습니다

이 원칙은 사무 업무에만 해당하지 않습니다. 타오 교수는 구글 딥마인드의 AlphaEvolve 같은 시스템이 제대로 작동하는 이유로 사람이 만든 검증기를 꼽았습니다. AI가 답을 내면, 환각이 끼어들 수 없는 검증기가 그 답을 확인하는 구조입니다.

공장도 마찬가지입니다. 웨이스는 AI의 명령이 PLC로 가기 전에 가상 공장에서 먼저 실행되도록 VEXPLOR를 만들고 있습니다. AI가 내놓은 답을 그대로 설비에 보내지 않고, 가상 공장에서 결과를 먼저 확인하려는 것입니다. 메일 첨부를 보내기 전에 검사하는 것과 같은 방식입니다.

07

남의 결과를 따라가기 전에 자기 검증 방법부터 만들어야 합니다

결과만 보고 따라 해서는 같은 결과가 나오지 않습니다. 결과를 만든 과정은 각자 쌓아야 합니다. 지금까지의 근거를 모으면 할 일은 세 가지입니다.

- **결과를 직접 확인할 수 있는 일부터 맡깁니다** — 애런슨은 첫 답이 틀렸다는 것을 알아봤고, 타오는 설명할 수 없는 결과는 쓰지 않습니다. 맞는지 판별할 수 없는 일을 AI에게 먼저 맡기면 틀린 답이 그대로 남습니다.
- **막힌 일을 어떻게 풀었는지 한 줄씩 적습니다** — 무엇이 막혔고, 어떻게 해결했는지를 남깁니다. 이 기록이 다음 작업의 출발점이 되고, 남이 따라 할 수 없는 자기 노하우가 됩니다.
- **한 번 푼 방법은 장치로 남깁니다** — 기억에 맡기지 않고 검사 도구나 규칙으로 바꿉니다. 웨이스의 메일 첨부 검사, 숫자 대조, 견적 전파, 착수 전 검색이 모두 이렇게 생겼습니다.



이 글을 한 장으로 요약하면 이렇습니다

AI가 주는 결과물은 누구에게나 열려 있지만, 그 결과를 쓸 수 있게 하는 시행착오의 기록은 각자 쌓아야 합니다. 그 기록이 쌓인 만큼 같은 AI에서 다른 결과가 나옵니다.

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

08

함께 읽어보세요

- AI에게 주는 규칙 파일이 길어지면 무엇이 달라지는가 — 취리히 연방공대 실험과 규칙 파일에 남길 것 4가지
- 웨이스는 업무 AI 에이전트를 직원 한 사람이 아니라 회사에 도입하도록 만듭니다 — 기준 문서·규칙·기록·권한은 개인 계정이 아니라 회사에 남습니다
- 웨이스가 VEXPLOR에서 AI의 판단을 PLC로 보내기 전에 같은 스캔의 가상 공장에서 먼저 실행하려는 이유

원문 blog.vexplor.com/post/ai-results-hidden-trial-and-error
발행 2026년 10월 12일 · 방동걸 (주)웨이스 대표이사

회사 주식회사 웨이스 (WACE Inc.) · 대표이사 방동걸 · 사업자등록번호 439-87-02852
주소 (18462) 경기 화성시 동탄구 동탄영천로 150, 에이동(A동) 1702-1708호 (영천동, 실리콘앨리동탄)
연락처 전화 070-4730-0629 · contact@wace.me
홈페이지 wace.me · linkedin.com/company/wace-inc

원문 홈페이지 링크드인