

AI끼리 검증시켰더니, 검증도 틀렸다

세 답을 비교하고, 결과를 취합하고, 내가 확인하는 'AI 의심법'

출발점

교차검토를 시킨 뒤에도 '그래서 결과가 무엇인지'를 한 번 더 정리해야 했다.

세 AI의 답을 따로 받은 뒤 서로 검토시키고, 공통 결론·충돌·확인 필요 항목을 한 화면에 취합했다. 마지막에는 인용·숫자·출처를 원문과 대조하고 내가 결론을 썼다.

3

실험 사례

150

실제 기록 발언

38

입장값 변경

15

판정 보류

AI가 바꾼 것은 답이 아니라 나의 사용 습관이었다

비개발자 학습자가 '질문하는 사람'에서 '검증 구조를 설계하는 사람'으로

BEFORE

- AI 한 개의 첫 답을 정답처럼 수용
- 유장한 문장을 정확한 정보로 착각
- 결과만 복사하고 판단 과정은 남기지 않음
- 오류가 나면 다시 질문하고 이유는 기록하지 않음



AFTER

- 반대 역할 다른 모델로 첫 답을 공격
- 사실 해석·가치 판단을 분리
- 발언 근거 입장 변화 합의 실패까지 기록
- 오류와 한계를 공개하고 사람이 최종 판단

3회 사례 탐색 · 설정 160턴 · 기록 150개 · 21,934토큰

결과를 복사하는 데서 멈추지 않고, 근거를 확인한 뒤 내 판단을 남겼다

1. 실제로 무엇을 했나

답이 다르면 누가 맞는지 궁금했다

여러 AI가 역할을 나눠 토론하도록 만든 Panelogue를 실제로 운영했다. GPT·Claude·Grok 계열 모델에 주장자·반대자·검증자 역할을 주고, 결론만이 아니라 발언 원문·근거·입장값·사회자의 선택을 함께 기록했다.

중요한 구분

Panelogue는 출력의 주제가 아니라 관찰 도구다.

이 사례의 주제는 AI 토론 서비스가 아니다. 토론 결과까지 다시 의심하고 원문과 계산으로 확인하면서 생긴 개인의 학습 습관이다.

세 번의 실행

사례	설정	관찰 목적
일상 판단 1	100턴 설정	반복되는 논거와 입장값 변화
일상 판단 2	30턴 설정	사회자 질문 뒤의 변화
일상 판단 3	30턴 설정	대안·검증 절차에 대한 반응

설정 합계는 160턴이지만 실제 전사록에 남은 발언은 150개였다. 사례 1에는 기록 누락 10턴과 응답 오류 1건이 있었다. 따라서 ‘160턴의 완전한 대화’라고 부르지 않았다.

사례 1의 생성 조건도 결과의 일부였다

조건	값	왜 공개했나
입장변화 확률	45	변화가 설득만으로 생겼다고 오해하지 않도록
유연성 / 작업기억	50 / 20턴	같은 문맥을 얼마나 유지했는지 알리기 위해
응답 / 도전·다양성	300토큰 / 100·100	반복 응답과 합의 실패가 설정의 영향도 받았음을 밝히기 위해

2. 원본을 다시 보니

숫자는 맞았지만 설명은 틀렸다

38/38	2/34	약 20	15
변경값 체인 일치	축자 인용 일치	자기 출력을 원인에 배치	원인 판정 보류

처음에는 38번의 입장값 변화에서 '어떤 문장이 AI를 설득했는가'를 찾았다. 그러나 전사록·입장 이력·설정 JSON을 다시 대조하자, 기존 분석은 수치는 정확해도 원인을 과도하게 설명하고 있었다.

처음 분석	원본 대조 후 판단
따옴표가 붙은 34개 축발 문구	원문과 그대로 일치한 것은 2개
직전 문장이 입장을 바꿨다고 해석	일부는 변화한 AI 자신의 출력 또는 이동 방향과 반대인 문장
반복 압력·방어 반응으로 분류	근거가 약한 15건은 판정 보류
반복 문구가 40회 이상이라는 AI 주장	원문에서 실제로 센 횟수는 26회

38건 재검증: 숫자는 맞았고, 해석은 고쳤다

원본 전사록 · 입장 이력 · 설정 JSON · 검증 CSV를 다시 대조한 결과



주의: 사례 2·3의 설정 파일은 확보되지 않았다. 이 자료는 연구가 아니라 검증 가능한 활용 기록이다.

변화한 값과 그 이유를 분리해 다시 코딩한 결과

3. 내가 바꾼 사용법

교차검토에서 끝내지 않고, 결과를 취합한다

01	02	03	04	05
질문 정리	독립 답변	교차검토	결과 취합	최종 판단
질문과 확인된 사실을 적는다	세 AI가 서로 보지 않고 답한다	서로의 오류와 약점을 찾는다	공통·충돌·확인 필요를 모은다	원문 확인 뒤 내가 결론 낸다

세 AI의 검토 결과를 따로 읽어서는 결론과 남은 의문을 한 번에 보기 어려웠다. 그래서 4단계에서 공통 결론, 끝까지 충돌하는 주장, 원문 확인이 필요한 숫자·인용·출처, 잠정 결론, 아직 말할 수 없는 것을 함께 모았다. 이 취합 결과도 AI가 만든 잠정 요약이어서 5단계에서 원문을 다시 확인했다.

짧게 따라 해 보는 방법

순서	할 일
1	판단이 필요한 질문 하나와 내가 직접 확인한 사실을 적는다.
2	같은 질문을 세 AI에게 주고 서로의 답을 보지 않은 채 답하게 한다.
3	각 AI가 나머지 두 답의 강점·오류·확인 불가 주장을 검토하게 한다.
4	공통 결론·남은 충돌·원문 확인 필요·잠정 결론을 한 번에 취합한다.
5	따옴표·숫자·출처를 원문에서 확인한 뒤 내 결론과 남은 의문을 쓴다.

복사해 쓰는 요청문

교차검토 뒤에는 반드시 결과를 취합한다.

‘세 답과 세 교차검토를 함께 읽고 ①공통 결론 ②끝까지 충돌하는 주장 ③원문 확인이 필요한 숫자·인용·출처 ④잠정 결론 ⑤아직 말할 수 없는 것을 구분해 줘. 다수 의견을 사실로 확정하지 말고, 확인이 필요한 항목은 그대로 남겨 줘.’

4. 그래서 결과가 무엇인가

다섯 칸으로 정리하고, 그대로 믿지는 않았다

검토문을 차례로 읽다 보니 결론과 남은 의문을 놓치기 쉬웠다. 그래서 4단계에서는 결과를 아래 다섯 칸으로 다시 정리했다. 다수결을 하려는 표가 아니라, 지금 확인할 수 있는 것과 아직 확인하지 못한 것을 나누기 위한 표다.

취합 항목	남기는 내용	읽는 원칙
공통 결론	세 검토가 같은 근거로 동의한 부분	확정 사실이 아니라 우선 확인 후보
남은 충돌	결론·근거·전제가 끝까지 엇갈린 부분	역지로 하나의 답으로 합치지 않음
원문 확인 필요	숫자·인용·출처·날짜·고유명사	실제 문서·계산·페이지에서 대조
잠정 결론	현재 확인된 사실만으로 말할 수 있는 범위	조건과 한계를 함께 적음
아직 말할 수 없는 것	자료 부족으로 판단을 보류한 주장	추측으로 빈칸을 채우지 않음

사람이 말하는 마지막 일

취합 결과도 다시 의심한다.

공통 결론에 여러 AI가 동의했더라도 원문 대조 전에는 사실로 확정하지 않는다. 사용자는 확인한 사실, 남은 의문, 자신의 최종 결론을 서로 다른 칸에 남긴다.

확인된 범위와 남은 한계

이 사례가 확인한 것은 한 사용자의 행동 변화와 세 실험의 원본 대조 결과다. 다른 학습자에게 같은 효과가 있는지, 어떤 수업에서 학습 성과가 높아지는지는 아직 측정하지 않았다. 그래서 검증된 교육 프로그램이 아니라 누구나 시험해 볼 수 있도록 공개한 실제 활용법으로 제시한다.

체험·증빙: AI 교차검증 5단계 MVP · Panelogue 실행 기록