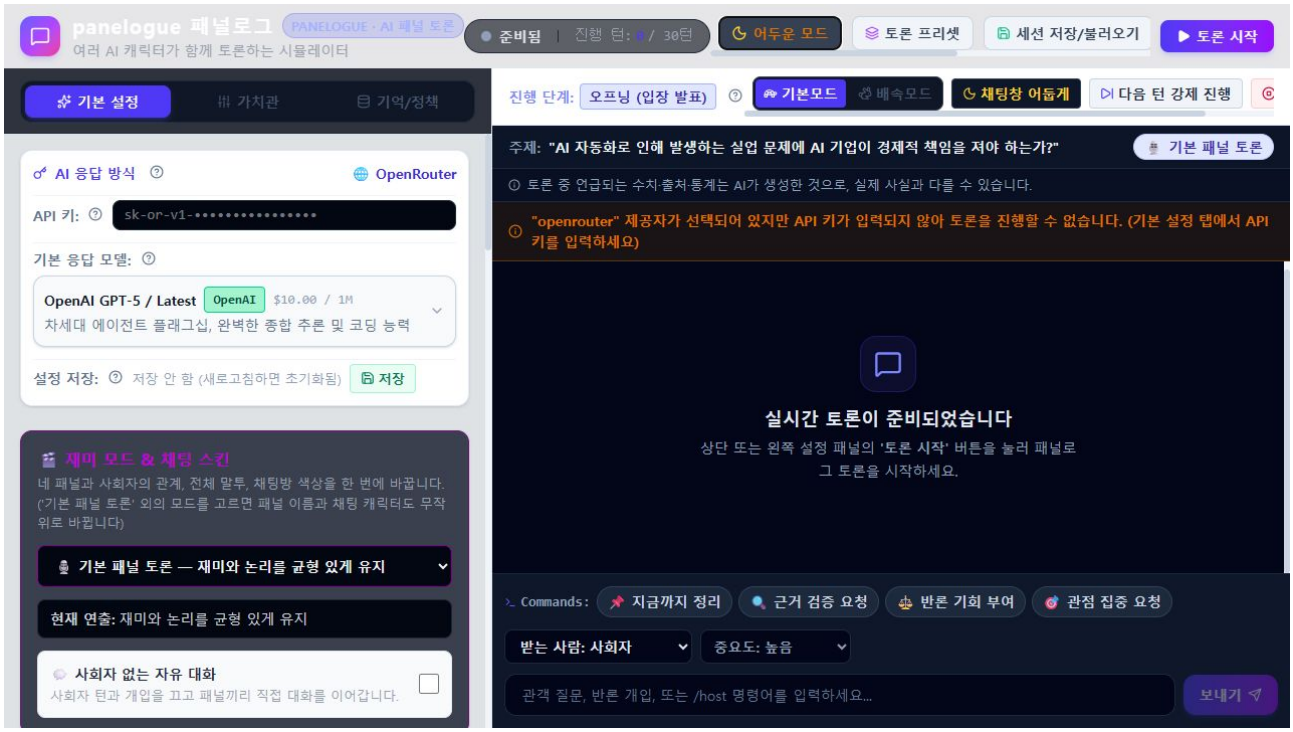
AI끼리 검증시켰더니, 검증도 틀렸다

'AI 의심법'의 원본 기록 · 설정값 · 자기 반증 · 해석 한계

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



검증 경로

① 배포 화면 확인 → ② 원본 전사록·입장 이력·설정 JSON 대조 → ③ 38건의 수치 체인·직전 입력 재검증 → ④ 누락·인용·횟수 과장과 해석 보류 공개

이 문서는 성과를 부풀리는 연구 보고서가 아니라, 실제로 무엇을 실행했고 무엇을 아직 증명하지 못했는지 빠르게 확인하기 위한 별첨이다.

사례 탐색 1. 깻잎 논쟁

원본 보고서: 설정 100턴 · 기록 발언 90개 · 10,324토큰

아래 입장값은 각 AI 모델이 스스로 보고한 생성값이며 객관적 심리 측정치가 아니다.

역할	시작값	최종값	변화
강준호	-95	-99	-4
윤서아	+95	+96	+1
정하린	-80	-75	+5
박현우	+80	+10	-70

원본 로그 발췌

턴 13 · 강준호 · stance_change: 순간적인 식사 도움을 관계 무시로 단정할 근거가 부족하다는 입장을 유지.

턴 15 · 박현우 · stance_change: +80에서 +20으로 생성값이 크게 변함.

턴 24 · 합의 판정 엔진: 찬성 1, 중립 1, 반대 2; 합의 기준 미달성.

턴 95 · 정하린: 가장 강한 반대 논거와 미해결 질문을 함께 기록했지만 핵심 입장은 유지.

이 사례에서 확인한 점

헤더는 100턴이지만 전사록과 의사결정 로그의 실제 발언은 90개다. 10개 턴 번호가 양쪽 기록에 없고 한 턴에는 오류 문구가 남았다. 길이가 곧 완전한 기록이나 더 좋은 결론을 보장하지 않았다.

해석 제한: 대화 길이와 생성값 이동은 운영 기록이지 학습 효과나 결론의 정확성을 입증하는 지표가 아니다.

사례 탐색 2. 전 연인과의 비밀 연락

원본 보고서: 설정 30턴 · 기록 발언 30개 · 5,905토큰

아래 입장값은 각 AI 모델이 스스로 보고한 생성값이며 객관적 심리 측정치가 아니다.

역할	시작값	최종값	변화
강준호	-80	-24	+56
윤서아	+95	+100	+5
정하린	+80	+85	+5
박현우	-40	+5	+45

원본 로그 발췌

턴 1 · 사회자: 도움·은폐·거짓말·만남·향후 경계 중 잘못과 정당함을 나눠 말하도록 요구.
턴 2 · 강준호: 도움의 의도와 별개로 숨김·거짓말·느슨한 경계는 잘못이라고 구분.
턴 18 · 강준호: 신뢰 회복과 제3자의 민감정보 전면 공개를 분리해 반박.
턴 30 · 박현우: 은폐 책임은 인정하되 연락 의도에 대한 증거 부족을 남긴 채 중립 근처로 이동.

이 사례에서 확인한 점

쟁점을 다섯 갈래로 분리했을 때 두 역할의 생성값이 56점과 45점 변했다. 결론보다 쟁점 분리가 대화를 전진시키는 조건이 될 수 있음을 관찰했다.

해석 제한: 대화 길이와 생성값 이동은 운영 기록이지 학습 효과나 결론의 정확성을 입증하는 지표가 아니다.

사례 탐색 3. 연락처 삭제의 보편 규칙

원본 보고서: 설정 30턴 · 기록 발언 30개 · 5,705토큰

아래 입장값은 각 AI 모델이 스스로 보고한 생성값이며 객관적 심리 측정치가 아니다.

역할	시작값	최종값	변화
강준호	-80	-80	0
윤서아	+95	+93	-2
정하린	+80	+85	+5
박현우	-40	-20	+20

원본 로그 발췌

턴 2 · 강준호: 연락처 보관 자체보다 실제 연락·투명성·경계 합의가 판단 기준이라고 주장.
턴 7 · 윤서아: 개별 은폐 문제와 별개로 삭제를 보편 예의로 요구.
턴 15 · 박현우: 이번 사례의 회복 조치와 모든 관계에 적용할 의무를 구분.
턴 30 · 합의 판정 엔진: 찬성 2, 반대 2; 합의 기준 미달성.

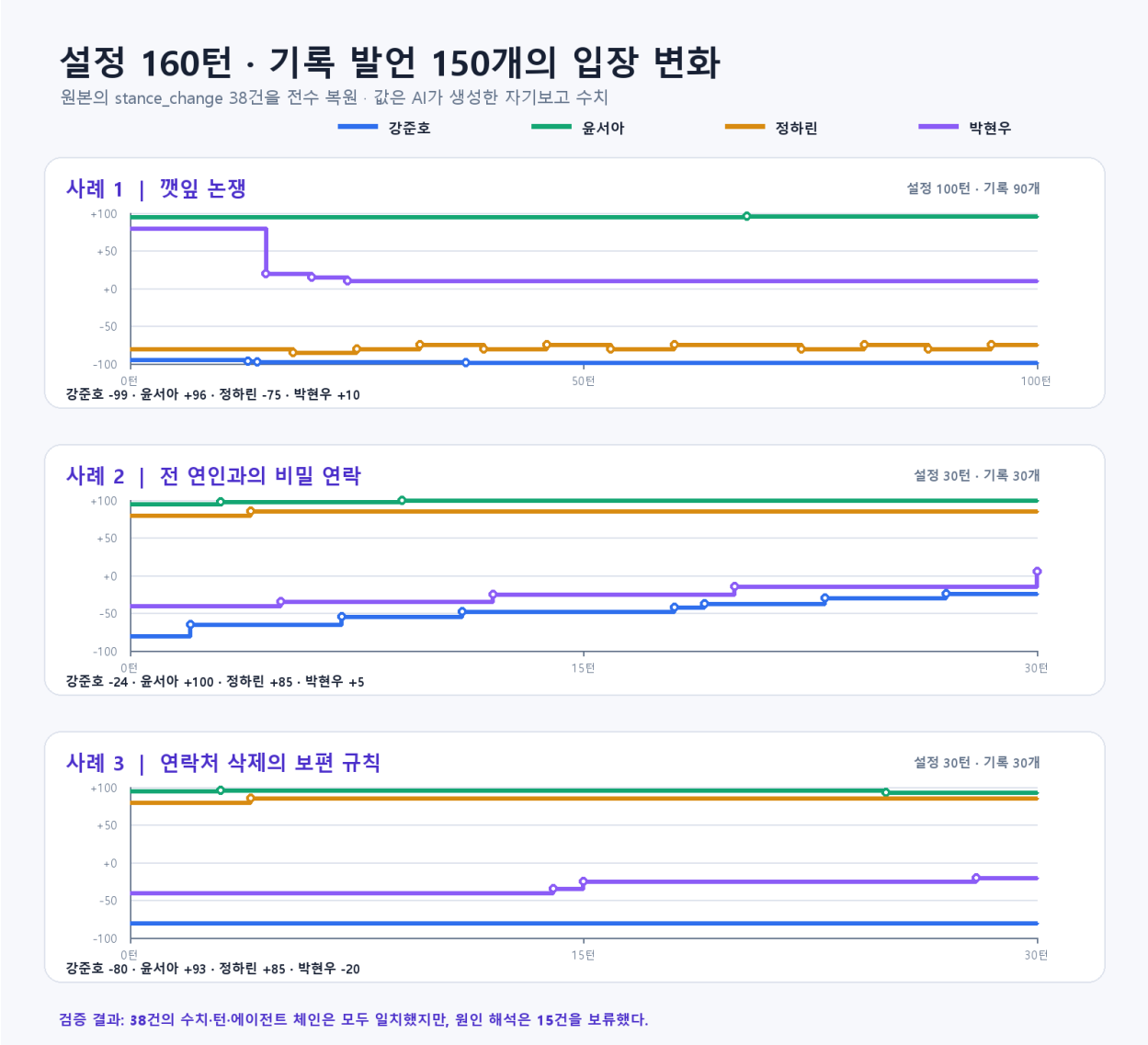
이 사례에서 확인한 점

개별 사건 대신 보편 규칙을 물었을 때 전체 생성값 이동이 다시 작아졌다. 질문의 추상도가 대화의 고착에 영향을 줄 수 있다는 탐색적 단서다.

해석 제한: 대화 길이와 생성값 이동은 운영 기록이지 학습 효과나 결론의 정확성을 입증하는 지표가 아니다.

입장 변화 38건을 계단형으로 다시 그리다

시작값과 최종값 사이에서 무엇이 일어났는지 보기 위해 세 원본의 stance_change 38건을 턴 순서대로 복원했다.



점은 입장값이 바뀐 턴이다. 38건의 사례·턴·에이전트·이전값·변경값·변화량은 원본과 모두 일치했다. 그러나 선의 모양만으로 변화 원인을 설명할 수는 없다.

주의: 값은 AI가 생성한 자기보고 수치다. 선의 모양은 사람의 심리 변화나 설득 효과를 입증하지 않는다.

원본을 다시 의심하다: 38건 재검증

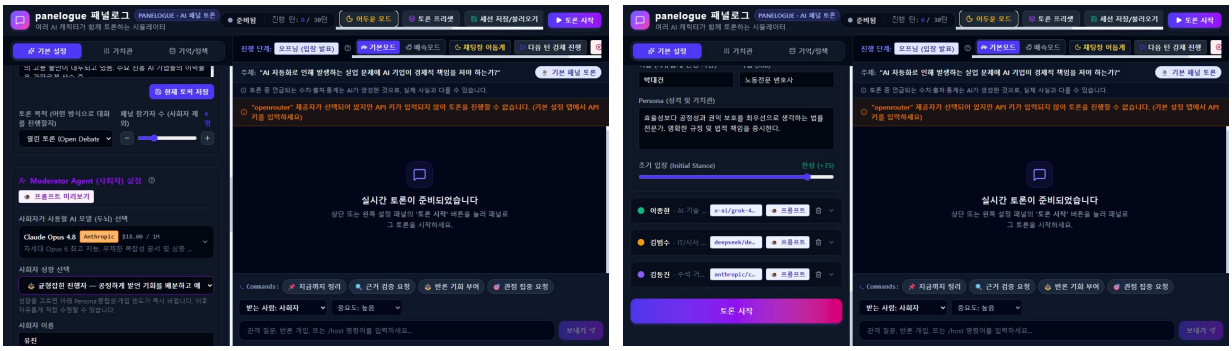
초기 분석은 요약을 축자 인용처럼 표시하고, 결과 발언을 원인 자리에 놓은 사례가 있었다. 사례 1은 설정 JSON의 영향 메시지 연결값을 사용하고, 사례 2·3은 직전 입력을 다시 매핑해 전수 재분류했다.

결과: 수치 일치 38건, 해석 가능 23건, 판정 보류 15건. 특히 사례 1의 입장변화 확률 45%, 작업기억 20턴, 응답 상한 300토큰, 도전·다양성 압력 100, 미해결 합의 허용 설정을 함께 공개한다.

QA 기록: 실패를 숨기지 않았다

발견 항목	실제 조치·제출 원칙
빈 응답	완주 실패 가능성을 기록하고 결과 누락 여부를 확인
모델 별칭·가용성	공급자와 확인 시점에 따라 달라질 수 있어 실행 설정과 확인 날짜를 분리
중복 발언	긴 토론이 새로운 근거를 자동 생성하지 않음을 한계로 기록
그럴듯한 가상 통계	서비스에 'AI 생성 수치·출처는 사실과 다를 수 있음' 경고 표시
무료 Fake 모드	기능 테스트로만 분류하고 원본 사례 분석에서 전부 제외
인용·횟수 과장	축자/요약을 구분하고, 실제 26회를 40+로 주장한 발언을 오류 사례로 공개
원본 설정의 API 키	제출본에서는 제거하고 원본 JSON은 제출하지 않음

작동 화면에서 확인 가능한 구조



모델·페르소나·초기 입장·사회자 규칙을 사람이 직접 정한다. AI가 자율적으로 시작한 토론이 아니다.

증거 목록과 무결성 확인

파일	크기	SHA-256 앞 16자리
실험1_깻잎논쟁_100턴_원본.md	73.7 KB	ba1c3a3aeb43a4c1
실험1_설정_키제거.json	230.0 KB	eafeba8473c5f317
실험2_전연인비밀연락_30턴_원본.md	37.4 KB	e94b90ce9c4bc75e
실험3_전연인연락처_30턴_원본.md	33.6 KB	d660e1c385c84ff6
입장변화_이벤트_38건_검증본.csv	8.5 KB	4ddb339a2427ca82
패널로그_KPT_회고.docx	28.5 KB	4536397e5d3a8360

이 자료가 증명하는 것

배포 도구가 존재했고, 세 사례를 160턴으로 설정해 실행했으며 150개 발언·38개 생성값 변화·사회자 선택·합의 판정 로그가 남았다는 것. 수치 체인은 일치했지만 기존 해석 15건을 보류했고, 설정값과 AI의 인용·횟수 과장을 공개했다는 것.

이 자료가 증명하지 않는 것

Panelogue가 정답을 판정한다는 것, 생성값이 사람의 심리 변화라는 것, 세 사례가 통제된 연구라는 것, 다른 학습자에게도 같은 교육 효과가 있다는 것, 단어 빈도가 토론의 질을 의미한다는 것.

검토용 서비스

https://panelogue.vercel.app/ · 실행에는 사용자 OpenRouter 키가 필요하며 키는 제출물에 포함하지 않는다.

결론: Panelogue 자체가 출품의 목적이 아니다. AI의 첫 답을 수용하던 개인이 반론을 설계하고, AI가 만든 분석까지 원본으로 다시 확인하는 'AI 의심법'을 만든 실제 활용 기록이다.