

AI에게 “네가 만든 규칙을 스스로 반증해봐”라고 시켰더니, 제 판단이 5.7배 틀렸다는 걸 찾아냈습니다

정답지 없이 과적합을 잡아내는 법 — AI에게 판단이 아니라 계산을 시키기

5.7배

규칙이 과학계
발동한 정도

17.1% → 0.4%

실제 데이터
발동률

0.625 → 0.833

정확도(F1)

① 상황 — 진짜 무서운 건 틀린 답이 아닙니다

그렇듯해서 틀린 줄도 모르는 답입니다. 저는 공공 문서를 읽고 규정 위반 여부를 판정하는 작업을 하고 있었습니다. 정답이 붙은 예시 200건으로 AI와 함께 판정 규칙을 만들었고, 정확도는 계속 올라갔습니다. 여기서 멈췄으면 “잘 만들었다”고 믿은 채 틀린 걸 제출했을 겁니다.

② 활용 — AI에게 반대 일을 시켰습니다

“네가 방금 만든 규칙을, 정답이 없는 2만 건에 그대로 돌려봐.
규칙이 ‘위반’이라고 판정하는 비율을 세어줘.
그 비율이 예시 200건의 위반 비율과 얼마나 차이 나는지 비교해줘.”

핵심은 정답이 없어도 된다는 점입니다. “얼마나 자주 발동하는가”만 세면 되니까요. 예시에서 위반이 3%인데 실제 데이터에서 17%에 발동한다면, 그 규칙은 과학계 잡고 있는 겁니다.

③ 결과 — 제 판단이 5.7배 틀렸습니다

구분	비율
예시 200건의 실제 위반 비율	3.0%
규칙이 2만 건에서 발동한 비율	17.1%

원인도 AI가 찾았습니다 — 안내문 한 줄이었습니

조건을 이리저리 좁혀봤지만 17.1%가 15.4%로밖에 안 줄었습니다. 조건 문제가 아니었습니다. 그래서 다시 시켰습니다.

“예시 데이터와 실제 데이터의 구성 자체를 비교해줘.
어떤 항목의 분포가 크게 다른지 찾아줘.”

예시 200건에서 특정 계약 유형이 17.5%였는데, 실제 2만 건에서는 44.8%였습니다. 예시가 실제 세상을 대표하지 못했던 겁니다. 그리고 결정적인 것 — 제가 아침에 읽고 그냥 넘긴 안내문 비교란 한 줄에 “이 유형에서는 해당 제한이 허용된다”고 적혀 있었습니다.

고친 결과

지표

이전

이후

실제 데이터 발동률	17.1%	0.4%
예시 데이터 정확도(F1)	0.625	0.833

발동률은 40분의 1로 줄었는데 정확도는 오히려 올랐습니다. 틀린 걸 덜 잡으니 맞는 게 선명해진 겁니다.

얻은 방법 — 세 문장

1. AI에게 결론을 묻지 말고 절차를 시킨다.

“이거 맞아?”라고 물으면 AI는 맞다고 합니다. “세어봐 / 비교해봐 / 반증해봐”라고 시키면 숫자를 가져옵니다.

2. 작은 데이터로 만든 것은 반드시 큰 데이터에 돌려본다.

정답이 없어도 됩니다. 발동 빈도만 비교해도 과적합은 드러납니다. 검증에 정답지가 필요 없다는 게 이 방법의 핵심입니다.

3. 예시 데이터와 실제 데이터의 구성을 비교하게 한다.

“왜 안 맞지”를 사람이 고민하는 대신 AI에게 분포를 비교시키면 원인이 대개 거기서 나옵니다.

어떤 일이나 적용됩니다

- 고객 문의 100건으로 분류 기준을 만들었다면 → **전체 문의에 돌려서** 각 분류에 몇 %가 떨어지는지 세어보게 하기
- 지난달 지출로 예산 규칙을 세웠다면 → **작년 전체에 돌려보고** 몇 번 걸리는지 비교하기
- 보고서 검토 기준을 만들었다면 → **예전 보고서 전부에 적용해** 지적 비율이 상식적인지 보기

공통점은 하나입니다. **AI가 만든 것을 AI에게 검증시키되, 판단이 아니라 계산을 시키는 것.**

왜 이게 중요한가

AI로 일하면 결과물이 빨리 나옵니다. 그래서 **검증할 시간은 생겼는데 검증할 방법을 안 쓰는** 상태가 됩니다. 저는 그날 “잘 만들었다”고 생각하고 있었습니다. AI가 제 판단을 뒤집어 주지 않았다면 그대로 갔을 겁니다.

AI를 잘 쓴다는 건 AI를 믿는 게 아니라, AI에게 나를 의심하게 시키는 것이라는 걸 배웠습니다.

출처 및 윤리 · 사용한 데이터는 공개 경진대회에서 참가자에게 배포된 자료이며 대회 규정이 허용하는 범위에서만 활용했습니다. · 진행 중인 대회라 대회명·도메인·구체적 판정 규칙은 밝히지 않았습니다. 공유하는 것은 **검증 방법**입니다. · 본문의 모든 수치(17.1% / 0.4% / 0.625 / 0.833 / 17.5% / 44.8%)는 실제로 측정한 값이며 지어낸 숫자가 없습니다. · 개인정보가 포함된 데이터는 다루지 않았습니다.