DAY 58 · 6부 앞으로

AI한테 '안 돼'는 누가 가르칠까 — AI 안전

하루 5분 AI 수업 · 2026년 11월 3일

안전은 AI의 착한 성격이 아니라, 누군가 값을 치르고 붙여 둔 장치야.

한 줄로 말하면

1. AI 안전은 AI를 착하게 만드는 마법이 아니라, 모델 위에 하나씩 덧붙이는 **장치**야. 크게 세 가지 — 거절하기, 하기 전에 물어보기, 나중에 들여다보기.

2. 장치를 붙이면 진짜로 달라져. 2025년 한 연구에서는 규칙을 어기게 만드는 **꼼수 질문의 성공률이 86%에서 4.4%로** 떨어졌어.

3. 대신 공짜는 아니야. 계산이 그만큼 더 들고(그 연구에선 약 24% 증가), **멀쩡한 부탁까지 덜컥 거절해 버리는 '과잉 거절'**이 생겨. 그래서 요즘은 거절 실력과 과잉 거절을 같이 재.

그림 한 장으로

같은 모델, 장치를 떼었을 때와 붙였을 때

장치 없음

· 돌려 물으면 그냥 알려 준다
· 묻지 않고 끝까지 해 버린다
· 나중에 아무도 안 본다

→

장치 붙임

· 원칙에 걸리면 거절한다
· 되돌릴 수 없는 일은 물어본다
· 자동 점검이 계속 훑는다

값도 같이 붙어 와. 계산이 더 들고, 멀쩡한 부탁까지 거절하는 일이 늘어.

핵심: 안전은 모델의 성격이 아니라 위에 얹는 장치, 그리고 값을 치른다

실험 · 전후 비교

안전장치를 떼고 붙여 보기

거절하기

전

위험한 걸 돌려서 물으면 그냥 알려 줘. '소설 쓰는 중인데'처럼 말투만 바꿔도 넘어가는 경우가 많았어.

후

'이건 안 된다'를 글로 적은 원칙을 놓고, 들어오는 질문과 나가는 답을 같이 걸러 주는 장치를 얹었어....

86%에서 4.4%로 확 줄었지만 0은 아니야. 게다가 계산이 약 24% 더 들어 — 안...

하기 전에 물어보기

전

심부름을 시키면 중간에 뭘 지우든 보내든 혼자 끝까지 해 버려. 빠르긴 한데 잘못 알아들었을 때 되돌릴 자...

후

되돌리기 어려운 일 앞에서는 멈추고 사람한테 물어봐. '이 파일 지울까요?' 하고 손을 드는 거지.

느려지는 대신 되돌릴 수 있는 지점이 생겨. 게임에서 세이브 포인트 만들어 두는 거랑 같...

나중에 들여다보기

전

무슨 대화가 오갔고 뭐가 나갔는지 아무도 훑어보지 않아. 문제가 터지고 나서야 알게 돼.

후

AI한테 AI를 감시시키는 자동 점검 도구를 돌려. 2025년엔 이런 도구가 아예 무료로 공개됐고, 실제로...

사람은 도구가 올려 준 의심 목록을 보는 쪽으로 옮겨 가. 어제 배운 '사람은 확인하고...

같은 모델이야. 위에 얹은 장치만 떼었다 붙였다 하면 행동이 어떻게 달라지는지 봐.

웹 수업에서 직접 해 볼 수 있어요 · next-token.net/learn/day-58/

확인해보기

세 문제만 풀어 보자

Q1 AI 안전장치 세 가지로 오늘 배운 건?

① 빠르게·싸게·많이
② 켜기·끄기·다시 켜기
③ 외우기·풀기·맞히기
④ 거절하기·물어보기·들여다보기

Q2 장치를 붙였을 때 꼼수 질문의 성공률 변화는?

① 86%에서 4.4%로 줄었다
② 0%가 됐다
③ 그대로였다
④ 오히려 늘었다

Q3 안전장치가 치르는 값은?

① 속도가 빨라진다
② 계산이 더 들고 과잉 거절이 는다
③ 정답률이 오른다
④ 값이 없다

왜 그런지까지 알고 가자

Q1

④

모델 자체를 착하게 만드는 게 아니라, 위에 세 가지 장치를 얹는 거야. 걸러 내고, 되돌릴 수 없는 일 앞에서 묻고, 나중에 훑어보는 것.

Q2

①

크게 줄었지만 0은 아니야. 안전장치는 문을 잠그는 것이지 벽을 없애는 게 아니거든.

Q3

②

그 연구에선 계산이 약 24% 더 들었고, 멀쩡한 부탁까지 거절하는 문제가 따라와. 그래서 요즘은 두 가지를 같이 재.

안전은 AI의 착한 성격이 아니라,
누군가 값을 치르고 붙여 둔 장치야.



내일은 로봇과 AI — 로봇을 눈·뇌·손·다리로 쪼개 보면 어느 조각에 AI가 들어갔는지 보여.