# 080 AI 가드레일

AI Guardrails

**AI가 위험한 행동이나 부적절한 출력을 하지 않도록 제한하는 안전 장치**

- 잘못된 정보, 유해 콘텐츠, 개인정보 노출, 범죄 조장 등 다양한 위험을 줄이기 위한 보호 체계
- AI가 규범·정책·윤리 기준을 벗어나지 않도록 입력·출력·추론 과정을 조정하는 기술·운영적 장치

## AI 가드레일의 개념

AI 가드레일은 생성형 AI가 안전 기준을 벗어난 응답을 생성하거나 위험한 행동을 유발하지 않도록 경계를 설정하는 안전 장치를 의미합니다. 생성형 AI는 사용자의 요구를 유연하게 수용하는 특성이 있어, 적절한 제한이 없다면 잘못된 사실, 유해 표현, 편향된 판단, 개인정보 노출, 불법·유해 행위 조장과 같은 문제가 발생할 수 있습니다. 이를 방지하기 위해 가드레일은 AI가 어떤 질문에 어떻게 응답해야 하는지, 어떤 범위에서는 응답을 제한해야 하는지를 미리 정의해 모델이 안전한 규칙 내에서 동작하도록 유도합니다. 이 과정은 단순 차단이 아니라, 위험 상황을 인식해 적절한 대체 정보 제공이나 표현 조정 등을 수행함으로써 활용성과 안전성의 균형을 동시에 확보하는 데 목적이 있습니다.


출처 : THE AI

## AI 가드레일의 종류

AI 가드레일은 적용 목적과 단계에 따라 여러 유형으로 나뉩니다. 입력 가드레일은 사용자가 위험한 질문이나 규범을 벗어난 요구를 할 경우 이를 감지해 적절히 거부하거나 안전한 형태로 재구성합니다. 다음으로 출력 가드레일은 모델이 부정확한 정보, 유해 표현, 개인정보 등을 생성하지 않도록 결과물을 점검하고 필요 시 수정·차단합니다. 또한 모델 자체의 행동 규칙을 정의하는 시스템 가드레일이 있어, 모델이 준수해야 할 목적과 제한 범위를 내부적으로 설정합니다. 여기에 프롬프트 인젝션 등 공격을 차단하는 보안 가드레일, 특정 산업 규제나 윤리 기준을 반영하는 정책 가드레일이 함께 적용되어 AI의 안전성을 다층적으로 보호합니다.