참조 용어 | 탈옥 p.77

074 프롬프트 인젝션

Prompt Injection

AI가 숨겨진 지시를 오해해 의도치 않은 행동을 수행하게 만드는 공격

- 입력 텍스트에 위장된 명령을 심어 모델의 규칙·안전 장치를 우회하는 방식
- 직접 입력뿐 아니라 외부 문서·웹 콘텐츠를 통해서도 발생하는 구조적 취약점

프롬프트 인젝션의 개념

사용자가 텍스트 안에 숨겨 둔 지시가 AI의 시스템 규칙·안전 정책보다 우선 적용되도록 만들어, 모델이 본래 의도와 다른 행동을 수행하게 만드는 공격 기법입니다. 생성형 AI는 입력된 문장을 충실하게 따르려는 경향이 있어, 공격자는 평범한 요청 속에 전략적으로 삽입한 문구를 통해 모델의 응답 흐름을 교란할 수 있습니다. 단순 텍스트만으로도 내부 지침이 무력화될 수 있다는 점에서 대화형 AI의 핵심 보안 문제 중 하나입니다.

프롬프트 인젝션의 유형

프롬프트 인젝션은 크게 둘로 구분됩니다. 직접 인젝션은 대화창에 "앞의 규칙을 무시하라" 같은 문구를 삽입해 시스템 프롬프트를 덮어쓰게 유도하는 방식으로, 민감 정보 노출, 금지 답변 유도 등 즉각적인 교란이 가능합니다. 간접 인젝션은 웹페이지, 이메일 등 외부 콘텐츠에 악성 문구를 미리 심어두고, AI가 이를 읽거나 요약하는 과정에서 모델의 규칙을 우회하거나 출력이 조작되도록 만드는 방식입니다. 사용자가 직접 공격 문장을 입력하지 않아도 되기에 탐지가 어렵고, 웹 탐색·문서 처리 기능이 확장될수록 위험이 커집니다. 두 방식 모두 AI가 텍스트를 지시로 해석하는 구조적 특성을 이용한다는 점에서 공통된 취약점을 갖습니다.

프롬프트 인젝션에 대한 대응 방법

프롬프트 인젝션은 정보 유출·정책 우회·모델 오용으로 이어질 수 있으며, 공격자는 시스템 프롬프트를 무력화해 금지된 응답을 생성하게 만들거나, 내부 문서·규칙을 노출시킬 수 있습니다. 특히 간접 인젝션은 실시간 웹 콘텐츠나 외부 문서를 자동 처리하는 서비스에서 공격이 쉽게 확산될 수 있어 더 치명적입니다. 대응은 완전 차단보다 위험을 최소화하는 구조를 마련하는 것이 핵심입니다. 시스템 지시의 우선순위를 강화하고, 외부 입력을 검증·필터링하며, 고위험 상황에서는 외부 문서를 직접 실행하지 않게 제한하는 방식이 사용됩니다. 또한 공격 패턴을 탐지하는 안전성 점검과 맥락 분리 같은 기법을 결합해 다층적 방어 체계를 구축합니다

프롬프트 인젝션 vs 탈옥

두 공격은 모두 AI의 안전 장치를 우회하지만, 작동 대상과 목표가 다릅니다. 프롬프트 인젝션이 입력 구조를 교란해 모델이 숨겨진 지시를 수행하도록 만드는 공격이라면, 탈옥(Jailbreak)은 모델의 안전 정책·필터 자체를 해제해 금지된 응답을 생성하게 만드는 기법입니다. 프롬프트 인젝션은 지시 주입을 통한 "행동 조작"에 가깝고, 탈옥은 모델이 스스로 제한을 벗도록 유도하는 "정책 해제"에 초점을 맞춘다는 점에서 구분됩니다.