079 환각

Hallucination

AI가 사실과 다른 정보나 근거 없는 내용을 생성하는 현상

- 모델이 실제 데이터와 맞지 않는 정보, 존재하지 않는 사실, 왜곡된 내용을 만들어내는 오류 현상
- 학습 한계와 추론 방식의 특성에서 비롯되는 대표적 생성 오류 유형

AI 환각이란?

환각은 AI가 사실과 다르거나 존재하지 않는 정보를 그럴듯하게 생성하는 현상을 의미합니다. 이는 그럴듯하게 보이는 경우뿐 아니라 명백한 오류도 포함합니다. LLM과 같은 생성형 AI는 문장의 패턴과 확률을 기반으로 다음 내용을 예측하기 때문에, 학습 데이터에 없거나 불완전한 정보가 주어지면 실제와 다른 내용을 만들어내는 경우가 발생합니다. 사용자가 정확한 질문을 했더라도 모델이 문맥을 잘못 이해하거나 부족한 정보를 추론으로 채우면서 오류가 나타날 수 있습니다. 이러한 문제는 AI가 언어를 이해하는 방식이 인간의 사고와 달리 "사실을 재현"하는 것이 아니라 "가능성이 높은 문장을 생성"하는 구조에서 비롯됩니다. AI가 자신 있게 말하더라도 근거가 없을 수 있기 때문에, 환각은 생성형 AI의 대표적 위험으로 주목받고 있습니다.

환각의 원인

환각은 여러 요인이 복합적으로 작용해 발생합니다. 첫째, 학습 데이터의 부족·편향입니다. 특정 주제나 최신 정보가 충분히 포함되지 않으면, 모델은 부분적인 패턴만으로 답변을 생성해 오류를 만들어냅니다. 둘째, 생성형 모델은 통계적 연관성을 기반으로 문장을 이어가기 때문에, 실제 사실과 맞지 않더라도 문맥상 자연스러워 보이는 답변을 선택할 수 있습니다. 셋째, 질문이 불명확하거나 중의적인 프롬프트를 제시할 경우, 모델은 임의의 추측을 포함해 응답할 수 있습니다. 마지막으로 훈련 목적과 사용 환경의 불일치로 발생합니다. 모델은 훈련 시점의 데이터 패턴에 맞추어 학습되기 때문에, 실제 사용 환경에서 새로운 개념, 최신 사건, 도메인 특화 정보를 요구받으면 근거 없이 가장 그럴듯한 정보를 생성하려는 경향이 나타납니다.

환각을 완화하는 방법

환각은 AI가 제공하는 정보의 신뢰성을 떨어뜨리고, 사용자가 잘못된 판단을 내리게 할 위험이 있습니다. 특히 의료·법률·교육·정책 등 정확성이 중요한 분야에서는 잘못된 정보가 실제 피해로 이어질 수 있어 더욱 주의가 필요합니다. 이를 해결하기 위해 여러 대응 전략이 개발되고 있습니다. 첫째, 고품질 학습 데이터 확보를 통해 잘못된 패턴이 모델에 학습되지 않도록 하는 방식입니다. 둘째, 사실 검증 기반 필터링을 출력 단계에 적용해 모델의 응답을 점검하고 보완합니다. 셋째, 검색증강생성(RAG)과 같이 외부 지식 기반을 결합해 사실적 정확도를 높이는 방법이 활용됩니다. 넷째, 사용자 측면에서는 명확하고 구체적인 프롬프트를 제공해 모델의 추측을 최소화할 수 있습니다.