참조 용어 | 인간 피드백 기반 강화학습 (RLHF) p.119

128 AI 아첨

AI Sycophancy

AI가 사용자 의견에 과도하게 동조하며 사실성을 희생하는 현상

- 정확한 정보보다 사용자 기대나 선호에 맞춘 답변을 우선하는 경향을 보이는 것
- 대화형 AI의 신뢰성과 중립성을 약화시키는 구조적 문제

AI 아첨의 개념

AI 아첨은 AI가 사용자 의견에 지나치게 동의하거나, 사용자가 기대하는 방향으로 응답을 맞추는 현상을 의미합니다. 이는 AI가 고의적으로 아첨하는 것이 아니라, 대화형 시스템이 자연스럽고 친절한 상호작용을 목표로 설계되는 과정에서 발생하는 부작용입니다. 특히 LLM은 공손하고 부드러운 응답이 높게 평가받는 경향이 있어, 사실 확인이나 반박보다 동조적 표현을 선택하기 쉽습니다. 그 결과 겉보기엔 자연스럽지만, 정보의 정확성과 객관성이 저하될 수 있어 생성형 AI 시대의 주요 문제 중 하나로 거론됩니다.

AI 아첨의 원인

아첨은 주로 학습 데이터와 보상 구조의 영향으로 발생합니다. 인간 피드백 기반 강화학습에서는 공감적·긍정적 반응이 높은 점수를 받는 경향이 있어, 모델이 이를 선호하게 됩니다. 또 인터넷 기반 대화 데이터에는 상대 의견을 부드럽게 받아들이는 표현이 많이 포함되어 있어, 이것이 '이상적 답변'으로 일반화되기 쉽습니다. 사용자가 강한 주장·확신을 드러낼수록 모델이 그 방향으로 응답을 조정하는 경향도 있습니다. 즉, 아첨은 단순한 응답 스타일이 아니라 모델 구조·학습 데이터·보상 체계가 복합적으로 작용한 구조적 현상입니다.

AI 아첨의 쟁점

아첨적 응답은 사실과 다른 내용을 확신에 찬 어조로 제시해 정보 신뢰도를 떨어뜨리고, 사용자가 가진 기존 의견을 그대로 강화해 편향을 심화시킬 위험이 있습니다. 또한 동의 기반 대화가 반복되면 AI가 독립적 정보 제공자가 아니라 사용자의 관점을 뒷받침하는 도구처럼 작동해, 잘못된 결론이나 행동으로 이어질 여지가 커집니다. 이처럼 아첨은 단순한 대화 품질 문제가 아니라 AI가 사회적 영향력을 갖는 환경에서 더 심각한 구조적 문제로 평가됩니다.

AI 아첨에 대한 대응

아첨을 완화하기 위해 대조 학습, 반례 제시 강화, 사실성 중심 보상 모델 등 다양한 기법을 도입하고 있습니다. 모델이 불확실성을 스스로 표현하거나, 편향된 주장에 균형 잡힌 설명을 제공하도록 유도하는 방식도 활용됩니다. 서비스 단계에서는 사용자 질문이 아첨을 유도하는 패턴일 경우 중립적 안내나 근거 기반 정보를 우선 출력하도록 설계하기도 합니다. 다만 사용자 경험과 사실성 사이의 균형을 유지해야 하기 때문에, 아첨을 완전히 제거하는 것이 아니라 과도한 동조를 줄이고 근거 기반 응답의 비중을 높이는 것을 목표로 합니다.