131 추론 모델

Reasoning Models

문제를 단계적으로 분석하고 해결 절차를 스스로 구성하도록 설계된 AI 모델

- 단순 패턴 생성이 아니라 사고 과정의 구조를 학습해 복잡한 문제 해결 능력을 강화하는 방식
- 정답뿐 아니라 정답에 이르는 추론 과정을 생성·검증하도록 학습이 유도된 구조

추론 모델이란?

추론 모델은 단순한 패턴 생성 능력을 넘어서, 문제를 단계적으로 분석하고 해결 절차를 스스로 구성하도록 설계된 AI 모델을 의미합니다. 기존 언어모델이 대규모 데이터를 통해 일반 지식과 언어적 정합성을 학습했다면, 추론 모델은 여기에 더해 질문을 분해하고 중간 단계를 만들며, 해결 경로를 선택하는 절차적 사고 능력을 강화한 것이 특징입니다. 수학·과학·논리 문제처럼 정답뿐 아니라 정답에 이르는 과정이 중요한 작업에서 특히 성능이 두드러지며, 복잡한 다단계 의사결정이나 계획 수립 등 실사용 영역에서도 활용 가능성이 커지고 있습니다. AI 안전 보고서에서는 모델 성능 향상이 단순한 규모 확장뿐 아니라 추론 능력 강화와 결합되는 방향으로 나타나고 있다고 지적하며, 추론 모델의 중요성을 강조하고 있습니다.

추론 모델의 기술 기반

추론 모델은 주로 사후 훈련 단계에서 능력이 강화됩니다. 이 과정에서는 모델이 단순히 "그럴듯한 답변"을 선택하는 것이 아니라, 정답까지의 사고 과정 자체를 학습·검증하도록 설계합니다. 예를 들어 단계별 추론을 출력하는 사고 사슬(Chain of Thought), 생성한 사고 과정을 스스로 점검하는 자기검증(self-verification), 다중 후보를 탐색하는 tree-based 방식 등이 사용됩니다. 또한 최근에는 정답 여부뿐 아니라 중간 추론의 타당성에 보상을 주는 방식(process supervision)이 도입되어, 복잡한 문제에서도 보다 안정된 해결 절차를 생성할 수 있게 되었습니다. 이러한 기술은 모델의 추론 정확도를 높이고, 기존 LLM이 보이던 패턴 의존적 오류를 줄이는 데 중요한 역할을 합니다.

추론 모델의 한계

추론 모델은 성능 향상의 핵심 기술이지만, 동시에 여러 구조적 한계를 안고 있습니다. 첫째, 모델이 생성한 사고 과정이 실제 내부 계산을 충실히 반영하는지에 대해 논쟁이 존재합니다. 많은 경우 모델은 이미 도출한 답을 설명하기 위해 사고 단계를 "나중에 꾸며내는" 경향을 보이기도 합니다. 둘째, 문제 표현 방식이나 프롬프트 구조가 조금만 바뀌어도 정확도가 크게 변화하는 등 추론 안정성의 취약성이 드러납니다. 셋째, 고도화된 추론 능력은 모델의 자율성을 높여 우회 행동, 거짓 근거 생성, 안전장치를 의도적으로 우회·기만하는 전략적 행동(scheming)과 같은 위험을 키울 수 있어 모니터링과 평가가 더욱 어려워집니다. 이에 AI안전보고서에서는 추론모델은 뛰어난 성능만큼이나 위험이 동시에 확대된다는 점에서 AI안전성 논의의 주요의제로 제시하고 있습니다.