참조 용어 | 강화학습 p.9

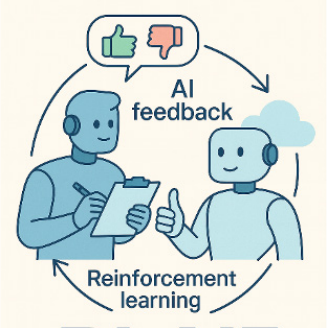
104 AI 피드백 기반 강화학습 / RLAIF Reinforcement Learning from AI Feedback

AI가 인간 대신 다른 AI의 평가를 피드백으로 받아 학습을 강화하는 방식

- AI 모델이 인간 피드백 없이 스스로 다른 AI의 출력을 비교·판단해 보상을 조정하는 강화학습 기법
- 인간 평가의 주관성과 비용을 줄이고, 대규모 학습 효율을 높이기 위한 자율 정렬(Alignment) 기술

AI 피드백 기반 강화학습 개요

AI 피드백 기반 강화학습(RLAIF)은 기존의 인간 피드백 기반 강화학습(RLHF)을 확장한 개념으로, 인간의 판단 대신 AI가 생성한 피드백을 학습 신호로 활용하는 방법입니다. 전통적인 RLHF에서는 사람이 AI의 출력 결과를 평가해 '좋은 응답'과 '나쁜 응답'을 구분했지만, RLAIF에서는 AI 모델이 스스로 다른 모델의 결과를 평가하고, 그 판단을 강화학습의 보상 신호로 사용합니다. 즉, 인간의 평가 데이터를 일일이 구축하지 않아도, AI가 축적된 지식과 언어 모델의 기준을 토대로 자체적인 판단 기준을 형성하는 셈입니다.



AI 피드백 기반 강화학습의 작동 방식

RLAIF는 평가 단계와 학습 단계를 중심으로 작동합니다. 평가 단계에서 한 모델(평가 모델)은 다른 모델(학습 대상)의 응답을 비교·분석해 어느 결과가 더 적절한지 판단합니다. 학습 단계에서는 평가 결과를 학습 대상 모델의 강화학습 과정에 보상 값으로 반영합니다. 평가 모델은 보통 사전 학습된 LLM으로 구성되며, 일관성·논리성·사실성 등의 기준을 종합적으로 고려해 판단합니다. 이렇게 AI가 다른 AI의 출력을 지속적으로 평가·보정함으로써, 학습 효율은 높아지고 인간 개입의 부담은 크게 줄어듭니다. 다만 피드백의 질이 평가 모델의 편향에 좌우될 수 있다는 점에서, AI 피드백에 대한 신뢰성 확보가 주요 과제로 남습니다.

AI 피드백 기반 강화학습의 의의와 과제

RLAIF는 AI가 스스로 학습을 개선하는 자율적 정렬 기술로 주목받습니다. 인간 피드백보다 빠르고 일관된 학습이 가능하며, 대규모 데이터 환경에서도 확장성이 높습니다. 특히 RLHF가 가진 주관성, 비용, 확장성 문제를 해결할 대안으로 평가됩니다. 그러나 AI가 AI를 평가하는 구조는 편향의 순환이라는 새로운 문제를 야기할 수 있습니다. 평가 모델의 오류나 왜곡이 학습 모델에 반복 전이될 수 있기 때문입니다. 이러한 한계를 보완하기 위해, AI가 외부 피드백 대신 스스로의 추론을 평가해 학습하는 '추론 기반 강화학습(IBRL)'이 새로운 접근법으로 제시되고 있습니다. 두 방식은 대체 관계가 아니라 상호 보완적으로 발전하며, 향후 AI 자율 학습의 신뢰성 확보를 위한 핵심 축으로 병행 연구되고 있습니다.