참조 용어 | 강화학습 p.9

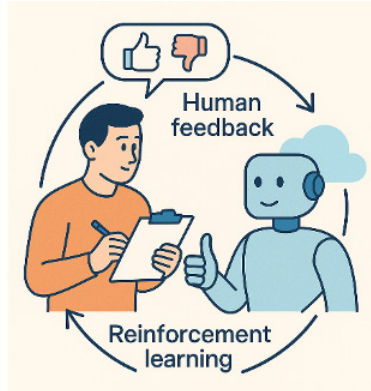
103 인간 피드백 기반 강화학습 / RLHF

Reinforcement Learning from Human Feedback

사람이 평가한 결과를 보상으로 삼아 AI의 출력을 개선하는 학습 방식

- AI가 생성한 여러 응답을 사람의 선호나 평가를 통해 비교·선정하고, 그 결과를 학습 보상으로 활용해 모델의 행동을 조정하는 강화학습 방법
- 인간의 가치와 판단 기준을 반영해 AI의 품질과 신뢰성을 높이는 학습 절차

인간 피드백 기반 강화학습 개요



RLHF는 사람이 직접 참여해 AI가 바람직한 출력을 내도록 조정하는 강화학습 기법입니다. 정답이 명확한 데이터를 사용하는 지도학습과 달리, RLHF는 생성형 AI처럼 정답이 불확실한 환경에서 인간의 평가를 통해 '무엇이 좋은 결과인지'를 학습합니다. 학습 과정은 기본 모델 학습, 보상 모델 학습, 강화학습 적용의 세 단계로 이루어집니다. 먼저 AI가 대규모 데이터로 언어 능력을 익히고, 이후 사람이 여러 응답 중 더 적절한 답을 선택해 보상 모델을 학습시킵니다. 마지막으로 AI는 이 보상 모델을 이용해 스스로 출력을 조정하며 인간의 평가 기준에 맞게 발전합니다. 이를 통해 AI는 단순한 언어 이해를 넘어 사람의 선호와 가치에 부합하는 행동을 학습합니다.

인간 피드백 기반 강화학습의 중요성

RLHF는 AI의 신뢰성과 사회적 수용성을 높이는 핵심 기술입니다. 인간의 평가를 학습 기준으로 삼음으로써 모델이 윤리적이고 사회적으로 바람직한 출력을 내도록 유도할 수 있습니다. 특히 생성형 AI에서 RLHF는 공격적·편향된 응답을 줄이고 사용자의 의도에 맞는 답변을 강화하는 데 활용됩니다. ChatGPT를 비롯한 주요 AI 모델은 RLHF를 핵심 절차로 채택해 '기술적 성능'보다 '인간과의 조화'를 목표로 발전하고 있습니다.

인간 피드백 기반 강화학습의 한계

RLHF는 사람의 참여가 필수적이기 때문에 비용과 시간이 많이 들며, 평가자의 주관이나 문화적 편향이 학습에 반영될 수 있습니다. 또한 사람의 선호가 반드시 논리적 정확성과 일치하지 않아 AI가 왜곡된 기준을 학습할 위험도 있습니다. 이를 보완하기 위해 AI가 직접 출력을 평가하는 RLAIF(AI 피드백 기반 강화학습)가 등장했으며, 향후 RLHF와 RLAIF를 결합한 하이브리드 방식으로 발전할 것으로 기대됩니다.