참조 용어 | 강화학습 p.9

105 추론 기반 강화학습

Inference-Based Reinforcement Learning, IBRL

AI가 외부 피드백 없이 자체 추론 결과를 기반으로 학습을 강화하는 방식

- 보상 신호 대신 내부 추론의 일관성과 논리를 학습 기준으로 삼아, 자율적 성능 향상을 도모하는 강화학습 기법
- 인간·다른 AI 피드백에 의존하지 않고 AI가 스스로 판단 근거를 검증하는 자기 개선형 학습 구조

추론 기반 강화학습 개요

추론 기반 강화학습(IBRL)은 AI가 사람의 피드백이나 외부 정답에 의존하지 않고 자신의 추론 결과를 학습의 피드백 신호로 활용해 스스로 개선해 나가는 방식입니다. 기존 강화학습에서는 사람이 "이 답이 더 좋아"라고 평가해야 학습이 진행됐지만, IBRL에서는 모델이 스스로 "내 답이 얼마나 신뢰할 만한가"를 판단합니다. 이를 가능하게 하는 것은 쿨백-라이블러 발산(KL Divergence)으로, 모델이 기존에 학습한 기준적인 행동 방식에서 지나치게 벗어나지 않도록 조정하는 역할을 합니다. 즉, IBRL은 외부의 정답 없이도 AI가 자기 출력을 검토하며 점차 더 나은 방향으로 발전하도록 돕는 자기 평가 기반 학습 구조입니다.

추론 기반 강화학습의 장점

IBRL은 LLM의 학습 효율성과 일반화 능력을 동시에 높일 수 있는 새로운 패러다임으로 주목받고 있습니다. 고품질 피드백 데이터를 대량으로 구축하는 데 비용과 시간이 많이 드는 인간 피드백 기반 강화학습(RLHF)과 달리 IBRL은 비지도 학습이 가능해 데이터 수집 부담을 줄이고, 특정 정답에 맞추지 않아 새로운 문제나 낯선 분야에도 잘 대응합니다. 또한 모델이 자기 추론의 신뢰도를 학습하면서 결과의 안정성과 설명 가능성을 함께 높일 수 있습니다. 이러한 특성 덕분에 IBRL은 AI 피드백에 의존하는 RLAIF의 한계를 보완하는 자율적 정렬 기술로 평가됩니다.

추론 기반 강화학습의 의의

IBRL을 적용해 수학 문제 풀이 등 다양한 벤치마크에서 RLHF 수준 혹은 그 이상의 성능을 보여주었던 사례가 있습니다. 특히 외부 보상 없이도 높은 정확도와 일반화 능력을 달성해 IBRL의 실용성과 확장 가능성을 입증했습니다. 또한 IBRL은 자기 일관성을 강화하는 데에도 중요한 역할을 합니다. 모델이 동일한 입력에 대해 여러 추론 결과를 비교·평균함으로써, 단순히 정답을 맞히는 것을 넘어 왜 그 답을 선택했는지에 대한 사고 과정 자체를 학습하게 됩니다. 이러한 구조는 장기적으로 AI가 단순한 도구를 넘어, 자기 설명적이고 신뢰할 수 있는 지능으로 발전하는 기반이 됩니다.