참조 용어 | 미세조정 p.42

130 사후 훈련 기법

Post-training Techniques

이미 학습된 AI 모델을 추가로 개선해 성능·안전성·적응력을 높이는 기법

- 대규모 사전학습 이후 모델의 사용 목적에 맞게 추가 학습 및 보정 과정을 통해 품질을 제고하거나 위험을 줄이는 기법
- 추가 데이터·보상 신호·제약을 활용해 모델의 출력을 더 신뢰성 있게 만드는 과정

사후 훈련 기법 개요

사후 훈련 기법(Post-training techniques)은 이미 사전학습(Pre-training)을 마친 모델에 추가적인 학습·보정 과정을 적용해, 특정 목적에 맞게 기능을 강화하거나 위험을 줄이는 기술적 절차를 의미합니다. LLM이나 비전 모델은 방대한 데이터로 사전학습을 거쳐 기본적인 패턴·지식을 익히지만, 이 상태로는 실제 응용에 바로 사용하기에는 부적절하거나 안전성·일관성 면에서 한계가 존재하는 경우가 많습니다. 사후 훈련은 이러한 원시 모델을 실사용 환경에 적합하도록 다듬는 과정으로, 모델이 사용자 의도를 더 정확히 해석하고 사회적·윤리적 기준을 준수하도록 조정하는 역할을 합니다. 최근 '국제 AI 안전 보고서'에서는 AI 모델의 성능 도약이 모델 규모 확장뿐 아니라 사후 훈련 기법의 발전에 의해 촉진되었다고 강조하고 있습니다.

주요 사후 훈련 방식

사후 훈련에는 여러 방식이 포함되지만, 대표적으로는 미세조정(Fine-tuning), 지시 따르기 학습(Instruction tuning), 강화학습 기반 보정(RLHF, RLAIF), 안전성 보정(Safety tuning) 등이 활용됩니다. 미세조정은 구체적 업무(task) 해결을 위한 추가 데이터를 사용해 모델을 업무에 맞게 최적화하는 방식입니다. 지시 따르기 학습은 모델이 자연어 명령을 이해하고 응답하도록 예시 지시문과 출력 쌍을 학습시키는 과정입니다. RLHF·RLAIF는 인간 또는 AI가 평가한 보상 신호를 기반으로 모델이 바람직한 응답을 선택하도록 조정하는 기술로, 대형 모델의 일관성·선호도·유용성을 크게 개선합니다. 최근에는 안전성 보정을 통해 편향·유해성·환각을 줄이고, 모델의 사실성·책임성을 강화하는 연구도 활발히 이루어지고 있습니다.

사후 훈련 기법의 중요성

사후 훈련 기법은 AI 안전 보고서에서 AI 성능 향상의 핵심 단계로 평가되었는데, 이는 단순한 패턴 학습을 넘어 모델이 문제를 해결하는 사고 과정과 응답 구조를 재정렬해 정확도·일관성·추론 능력을 실질적으로 끌어올리는 역할을 하기 때문입니다. 특히 복잡한 수리·논리 문제나 다단계 작업에서 나타나는 성능 향상 상당수가 사후 훈련에서 비롯된 것으로 분석됩니다. 다만 LLM의 경우, 사후학습 없이도 여러가지 일이 가능하도록 방대한 데이터를 사용하여 학습되기 때문에 실무적으로 사후훈련을 하지 않는 경우가 증가하고 있습니다.