낙장불입

이동재

20240919

Abstract

현존하는 언어 모델들은 한 번 뱉은 말을 주워담을 수 없다. 이는 언어 모델이 이전 입력을 바탕으로 다음 입력을 '생성하는' 작업만 수행하도록 훈련되었기 때문이다. 즉, 한 번 거짓말을 하기 시작하면 끝도 없이 거짓말을 하게 되는 것이다. 이로 인해 언어 모델의 환각 (Hallucination) 현상은 더욱 심화된다. 본 글에서는 이러한 낙장불입 문제를 자세히 다루고, 이를 해결하기 위한 아이디어를 제시한다.

언어 모델의 환각 현상은 언어 모델의 대표적인 문제점 중 하나로, 최근에는 환각을 근본적으로 제거할 수 없다는 주장이 나오기도 한다. 이러한 환각 현상은 논리적 추론 과정에서 가장 큰 골칫거리다. 논리적 추론에서는 이전 추론을 바탕으로 다음 추론을 이어나가기 때문에, 이전 추론에서 환각이 발생하면 다음 추론에서도 자연스럽게 환각이 발생한다.

이러한 연쇄작용은 언어 모델에서만 발생하는 것이 아니다. 8년 전으로 알파고와 이세돌의 4번째 대국을 떠올려보라. 이세돌의 78수 이후 알파고의 예측 승률은 급격히 떨어졌으며, 그 후 알파고는 이해할 수 없는 행동을 하며 이세돌에게 패배했다. 이 역시 환각의 연쇄작용으로 간주할 수 있다.

환각의 연쇄작용을 막기 위해서는 환각 자체를 제거하는 것도 방법이지만, 연쇄의 연결고리를 끊는 것 역시 좋은 해결책일 것이다. 이를 위해서는 언어 모델이 현재 환각을 일으키고 있는지 판단하는 기능과, 어디까지 돌아갈 것인지 결정하는 말 주워담기 (Backtracking) 기능이 필요하다.

언어 모델은 지우개 토큰을 생성함으로써 이전에 생성했었던 토큰을 제거할 수 있어야 한다. 단순히 생성한 토큰을 제거하고 다시 생성하는 것만으로는 충분하지 않다. 핵심은 언어 모델이 잘못 생성한 토큰을 남겨두어 언어 모델이 스스로 어떤 실수를 범했는지 깨닫고, 이후의 생성에 이를 반영할 수 있어야 한다는 것이다. 토큰을 지워버리면 언어 모델도 자신의 실수를 잊어버리게 된다.

모델의 구현은 강화학습을 통해 가능할 것으로 보인다. 적절한 시점에 적절한 지우개 토큰을 생성하면 보상을 제공하고, 그렇지 않으면 처벌하는 방식으로 학습을 진행한다. 강화 학습을 통해 언어 모델은 환각의 영향력을 최소화하고, 좀 더 나은 논리적 추론 성능을 보일 수 있을 것으로 기대된다.