언어 모델은 휴리스틱 백화점

이동재

20241114

Abstract

언어 모델은 연산하는 수의 크기가 커지거나 연산이 복잡해지는 경우 정확도가 떨어지는 모습을 보여 학습 과정에서 연산 결과를 전부 암기해 버린 것이 아니냐는 의심을 받았다. 오늘 소개할 논문에서는 언어 모델이 휴리스틱의 집합체를 통해 산술 문제를 해결한다는 것을 보여준다. 본 글에서는 이러한 주장과 근거를 간략히 소개하고 이러한 연산 방식을 정적 분석과 연관지어 살펴볼 것이다.

산술적 계산 능력은 언어 모델이 지닌 능력 중 하나로, 계산기만큼 완벽하진 않지만 대부분의 상황에서 산술적 계산을 잘 수행한다. 그러나, 문제가 복잡하지는 경우 오류를 발생시키는 경우가 종종 있어 단순히 답을 암기한 것이 아니냐는 의심을 받아왔다. 하지만, 아무리 방대한 데이터를 사용한다 하더라도 모든 종류의 산술 계산 문제가 들어있지는 않을 가능성이 높다. 하지만, 완벽히 알고리즘을 학습했다면 문제가 복잡해졌을 때 오류를 발생시키는 이유 또한 설명되지 않는다.

논문에서는 이러한 현상이 발생하는 이유는 언어 모델이 휴리스틱의 집합체를 통해 산술 문제를 해결하기 때문이라고 설명한다. 여기서 휴리스틱의 집합체라 함은 정적 분석의 요약 도메인과 유사하다. 예를 들어, 10으로 나눈 나머지가 8일때 활성화되는 뉴런이나, 수의 값이 25에서 50 사이일 때 반응하는 뉴런 등이 있다. 그리고, 어텐션 헤드에서는 뉴런 활성화 정보들을 등호 기호에 모아 최종 답을 도출해낸다.

정적 분석의 관점에서 이를 해석하면 다양한 종류의 요약 도메인을 사용하여 변수의 값을 분석하고 교집합을 적용하여 최종 답안을 도출해내는 것과 같다. 요약 도메인은 유한한 메모리만 사용하여 무한한 경우의 수를 표현할 수 있지만 정밀도가 떨어진다는 단점이 있다. 언어 모델 역시 무한한 메모리를 지닌 것이 아니기에 유한한 메모리를 사용하여 다양한 산술 연산을 최대한 많이 해결할 수 있는 방향으로 휴리스틱을 학습한 것이다.

역으로 언어 모델의 방식을 정적 분석에 적용해보는 것은 어떨까? 모듈로(Modulo) 도메인, 범위 도메인 등 다양한 도메인들을 무작위로 정의하고 각각에 대해서 코드를 정적 분석한 다음, 각 도메인의 분석 결과들을 교집합하여 변수의 값을 유추해낸다면 훨씬 정확한 분석이 가능할 것이다.