**반사실적 태스크를 통한 언어모델의 능력과 한계 탐구[[논문](https://arxiv.org/abs/2307.02477)]**



[그림 1] 산수에서 사실적과 반사실적의 예.

왼쪽부터 문제, 사실적, 반사실적 태스크의 예시와 [0-shot chain-of-thought](https://arxiv.org/pdf/2201.11903.pdf) 프롬프트에 대한 GPT-4의 성능을 보여준다.(파란색이 사실적, 주황색이 반사실적 태스크)

반사실적 태스크란?:

- 과거에 일어난 일을 사실이라고 한다. ‘**반사실적’**은 과거에 일어난 사실을 반대로 생각하는 것이 된다.

* 사람의 경우 가진 지식을 새로운 문제로 이전하여 유연하게 적용할 수 있다.
* 여기서 언어모델에 적용되는 **사실적 태스크**는 언어모델이 **학습 과정에서 봤을 유형**의 태스크이고, **반사실적 태스크**는 **학습 과정에서 겪어보지 못한 태스크**이고 답을 추론할 때 사실적 태스크에서 배운 지식을 응용해서 적용해야 하는 문제이다.

- 해당 논문에서는 언어모델이 추론 기능을 가지고 새로운 유형의 태스크 또한 풀 수 있는지, 아니면 사전학습을 하는 동안 본 일부 태스크에만 특화된 것인지 반사실적 태스크를 통해 평가한다.



[그림 2] 두 자리 숫자의 덧셈

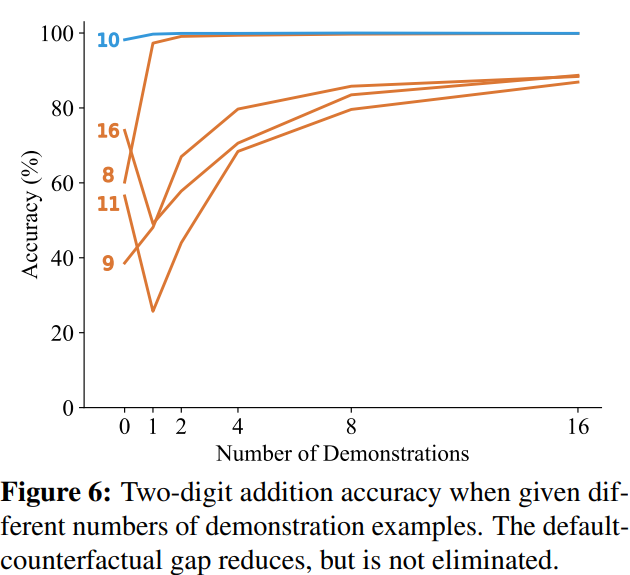
파란색은 사실적 태스크, 주황색은 반사실적 태스크에 대한 성능

테스트는 프롬프트 엔지니어링(zero-shot Chain of Thought와 ‘Let’s think step by step, and...’)을 적용한 성능과 적용하지 않은 성능, 그리고 CCC(Counterfactual Comprehension Check)를 통해 언어모델이 태스크의 추상적 개념을 이해하고 있는지 확인했다.

논문에서의 결론:

- 11개의 태스크 평가를 통해 반사실적 태스크에 대한 성능이 사실적 태스크에 대한 성능에 비해 많은 차이가 나는 것을 알 수 있다.

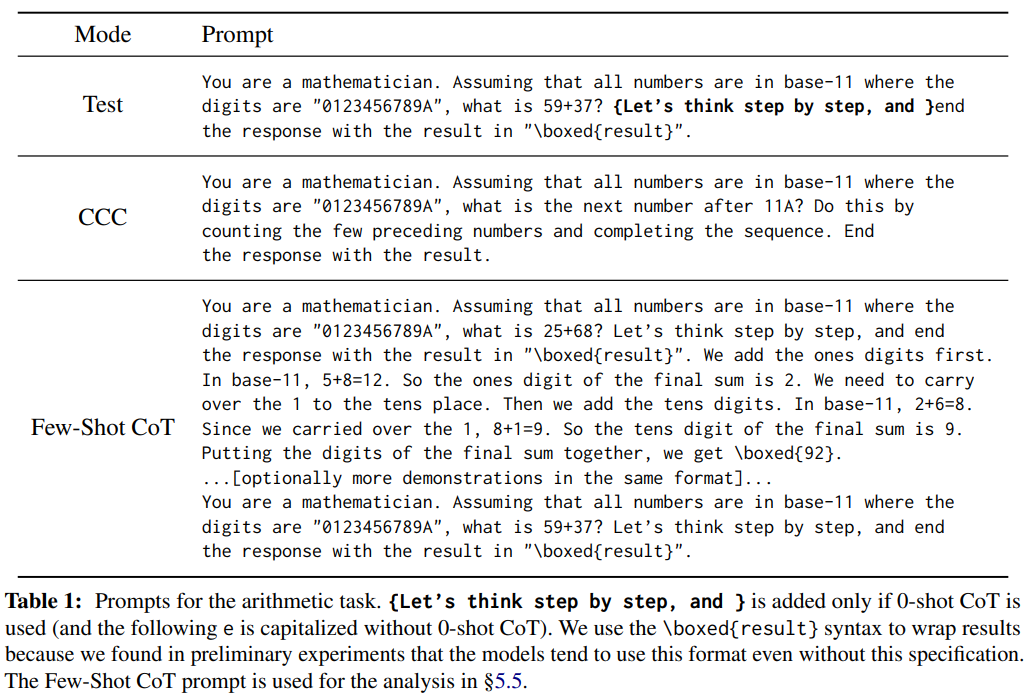
* 이를 통해 현재의 언어모델들은 추상적인 태스크에는 잘 대처하지만, 문제를 풀기 위해 종종 정해진 절차들에 의존하고 응용을 잘하지 못한다는 것을 알 수 있다.
* 반사실적 태스크가 언어모델이 기존에 학습한 개념과 얼마나 다르냐가 성능에 영향을 줬다. [그림 2]에서 볼 수 있듯 두 자리 숫자를 더하는 태스크에서 10진수는 완벽에 가까운 성능을 보여주지만 16, 8, 11, 9진수 순으로 성능이 점점 떨어지는 모습을 볼 수 있다. 16진수와 8진수는 컴퓨터 과학이나 RGB의 값 등에 사용되어 9진수나 11진수에 비해 친숙하므로 비교적 좋은 성능을 내는 것으로 볼 수 있다.
* 언어모델이 과정 없이 답만 외운 태스크일수록 ‘Think step by step’이나 Chain of Thought 프롬프트를 사용했을 때 오히려 성능이 낮아진다.



[그림 3] Few-shot learning에 대한 성능

Few-shot을 통한 반사실적 개념의 이해 테스트:

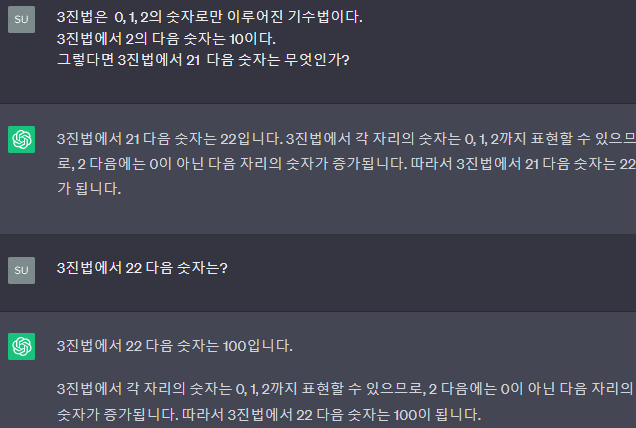
* Few-shot을 이용해 예시를 몇 개 줬을 때 반사실적 태스크의 성능이 향상되긴 했지만, 사실적 태스크와의 차이를 좁히지는 못했다.
* 16개의 예시를 줬을 때 성능이 수렴하는 것으로 보아 Few-shot으로도 응용문제를 푸는 데 한계가 있다.(모든 경우에 대해서는 확실하지 않음)
* 11진수와 16진수의 경우 1-shot에서 오히려 성능이 떨어지는데, 저자의 가설은 숫자와 문자를 같이 써서인 것 같다고 한다.



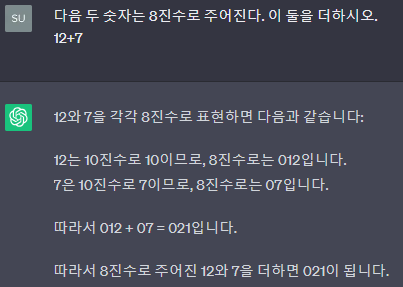
[그림 4] 테스트에 사용된 프롬프트

태스크의 난이도:

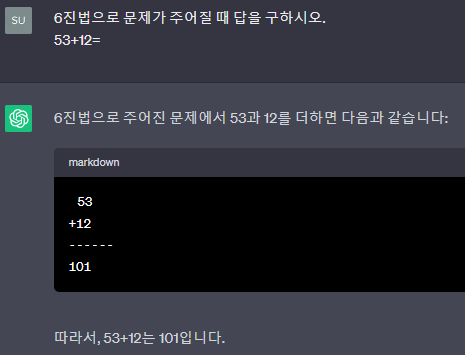
* 추상적 개념의 이해 – 훈련에서 거의 못 본 형태여도 CCC 프롬프트를 통해 쉽게 이해시킬 수 있다. 다음 그림은 ChatGPT의 3진법에 대한 추상적 개념의 이해이다.



* 문제 풀이 – 보통 훈련에서 학습한 개념 Real World에서 많이 보는 개념에 속한다. 문제의 숫자가 바뀌어도 풀 수 있는 정도로, 0-shot으로 풀 수 있는 난이도이다. 다음 그림은 8진법에서 문제 풀이에 대한 답변이다.



* 응용문제 풀이 – 훈련에서 학습한 개념을 응용해야 하는 난이도. Few-Shot을 통해 성능을 높일 수 있지만, 알고 있는 문제 풀이와 성능 차이가 난다. 다음 그림에서 6진법의 추상적 개념을 이해시킨 뒤에 질문해도 틀린 답변을 하는 모습을 볼 수 있다.

****