CUAI 하계 컨퍼런스 NLP 1팀

2025.07.08

발표자 : 홍성빈

스터디원 소개 및 만남 인증

이나현	
데스트로 성명하려니 어렵네요 결국에 개가 하고싶은건 mccp의 대체제를 갖고 싶은건데 이게 아우리 고민에봐도 모르겠어요 ㅋㅋㅋ 그래서 계속 생각하고 있던 건 기존 객관식 벤지마크에서, 오답의 품질을 끌어올려보고 (누가 봐도 '오답'인 선지를 한두개정도 교체)	
교체하는건 '그림듯한' 선지가 되어야 할 거구요 그렇게 해서 기존 벤지마크와의 성능 비교를 해보는 것이 어떨까	
이나현	
이거는 오픈소스 모델로도 할 수 있지 않을까 싶어요 고성능 모델이 필요한게 아니니까	
근데 결국에 mcqa의 대안을 제시하지 못한다는 한계가 있져 그냥 제안이	었습니다 요후 608
이나현	
아니면 또 해보고싶었던게 멀티모달 mcqa에서 (사진 + 객관식문제) 정보가 충돌할 때 llm이 어디에 의존해서 답변을 주는지도 궁금했는데 이거는 디벨롭 방향이 안보여서 그냥 훌덩상태입니다	
조민지	
예를 들어서 잘문이 '다음 중 포유류는가의 잘문이라면 원래 선지 A 안간 B 책정 C 얼굴에 D 우두선 플	
A 인간 8 개구리 C로마에 D 도마에 이런 식으로 교체에서 중 더 성세한 평가가 가능한 벤치마크를 만들어보자	
같은 말씀이신가요???	오후 6:11
이나현	
오 네네 맞아요 오후 6:11	
조민지	
오재밌을 것 같아요 요후 6:11	

스터디원 1: 이나현

스터디원 2: 조민지

스터디원 3: 홍성빈

- 1. 배경
- 2. 주제
- 3.데이터셋

1. MCQA(Multiple Choice Question Answering)

질문 : "다음 중 포유류는?" A 인간 B 책상 C 알로에 D 우주선 Model 1 정답 A 인간 !!

정답 B 책상 !!

정답 C 알로에 !!

정답 D 우주선 !!



1. MCQA(Multiple Choice Question Answering)

질문 : "다음 중 포유류는?"

A 인간

B 책상

C 알로에

D 우주선

Good!

정답 A 인간 !!

Bad!

정답 B 책상 !!

정답 C 알로에 !!

정답 D 우주선 !!





2. MCQA(Multiple Choice Question Answering) 한계

- 1. 위치에 따른 차이
- 2. 실제 정답이 아닌 가장 틀리지 않는 답 선택
- 3. 데이터셋 품질 문제 등



3. 선행 연구 파악

- 1. Can multiple-choice questions really be useful in detecting the abilities of LLMs?
 - → 순서 편향 및 LFGQ의 답변 일치도 낮음의 문제점 발견
- 2. LLMs May Perform MCQA by Selecting the Least Incorrect Option
 - → 실제 정답이 아닌 가장 틀리지 않는 답 선택 경향 발견
- 3. A Study on Large Language Models' Limitations in Multiple-Choice Question Answering
 - → 정답 선택, 선택지 순서 독립성 등에서 모델의 근본적 한계 발견



주제

Plausible distractor 만들기

Plausible distractor 생성 → 모델 추론 과정에서 '헷갈림' 겪는지 증명

MCQA 다양한 한계 → 'llm as a judge' 이용

But, 여전히 객관식 벤치마크 여러 성능 평가에 쓰임

따라서, 기존 벤치마크에서 오답 하나를 그럴듯한 오답으로 변경 후 성능의 하락 폭 측정

데이터셋

기준

- 1. 친근한 주제
- 2. 한국어

변경 방법

- 1. 직접 생성
- 2. 오답 generator 모델 생성
- → 1단계 (Ranker 훈련) : '순위 평가기(Ranker)' 모델 우선 훈련
- → 2단계 (Generator 훈련) : '순위 평가기' 선호 방향으로 더 매력적인 오답 생성하도록 '오답 생성기(Generator)' 모델을 DPO로 훈련