CUAI 하계 컨퍼런스 NLP 1팀

2025.07.08

발표자 : 홍성빈

스터디원 소개 및 만남 인증

스터디원 1 : 이나현

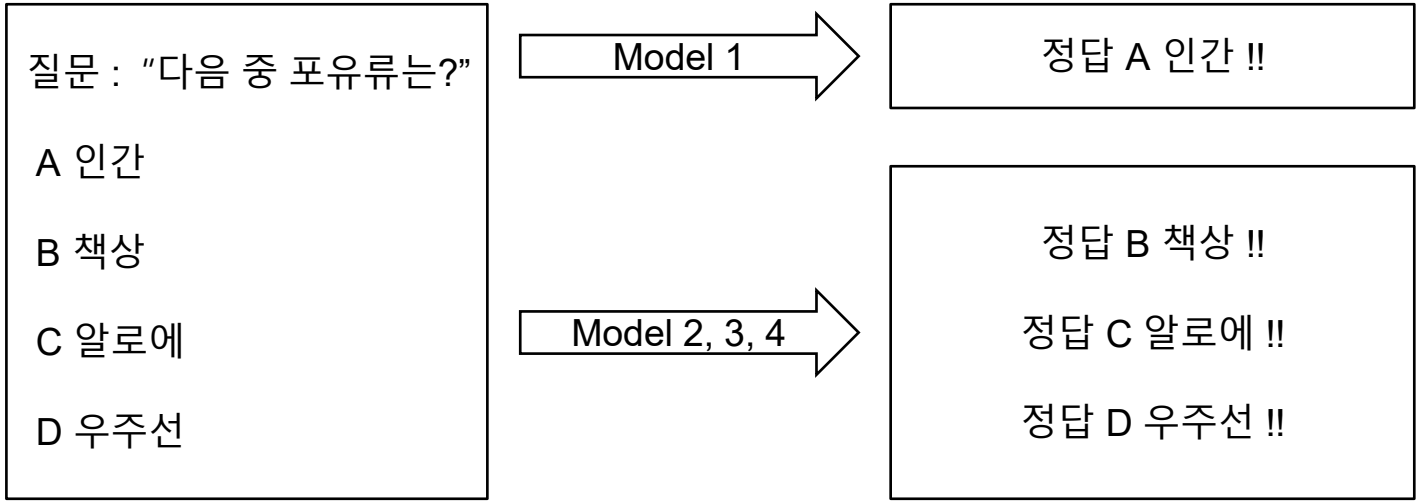
스터디원 2 : 조민지

스터디원 3 : 홍성빈

목차

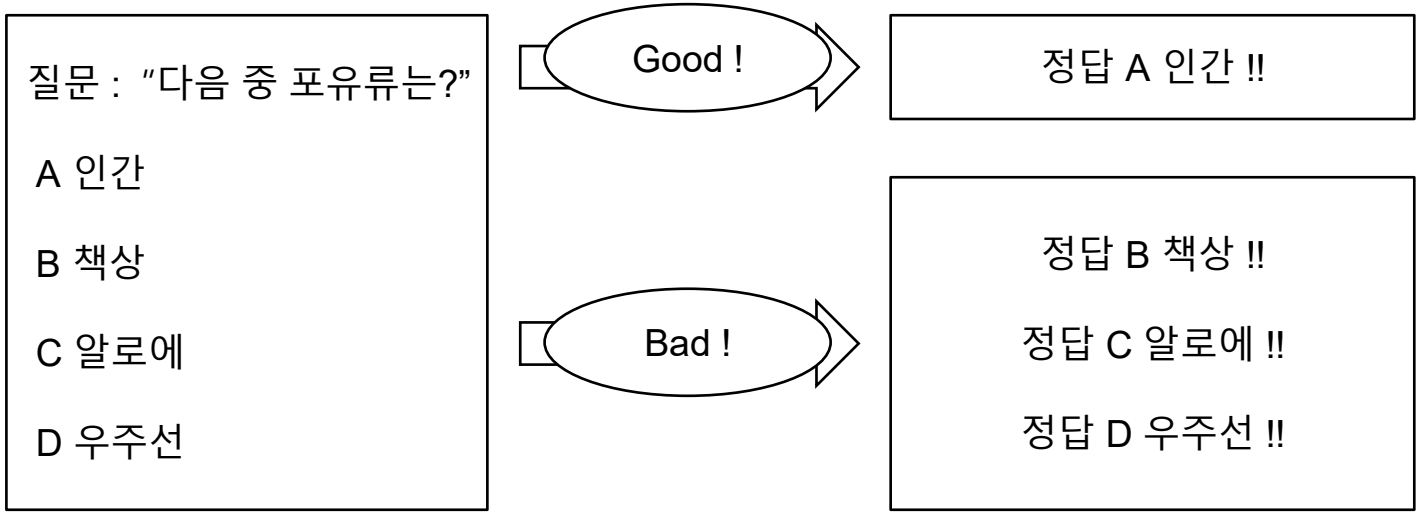
배경

1. MCQA(Multiple Choice Question Answering)



배경

1. MCQA(Multiple Choice Question Answering)



배경

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

배경

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