P0-01

数学多肢選択式問題における誤概念との一貫性を考慮した誤答生成

渡邊 裕介[1], 松林 優一郎[1,2], 畔柳 直旺[1,2], 坂口 慶祐[1,2]
([1] 東北大学, [2] 理化学研究所)

概要

・数学MCQsの誤答生成において正しい計算過程, 誤概念, 誤り計算過程, 誤答を同時に生成

・誤概念と誤答の一貫性の評価では性能向上が見られた一方で, 誤答の評価は低下

背景

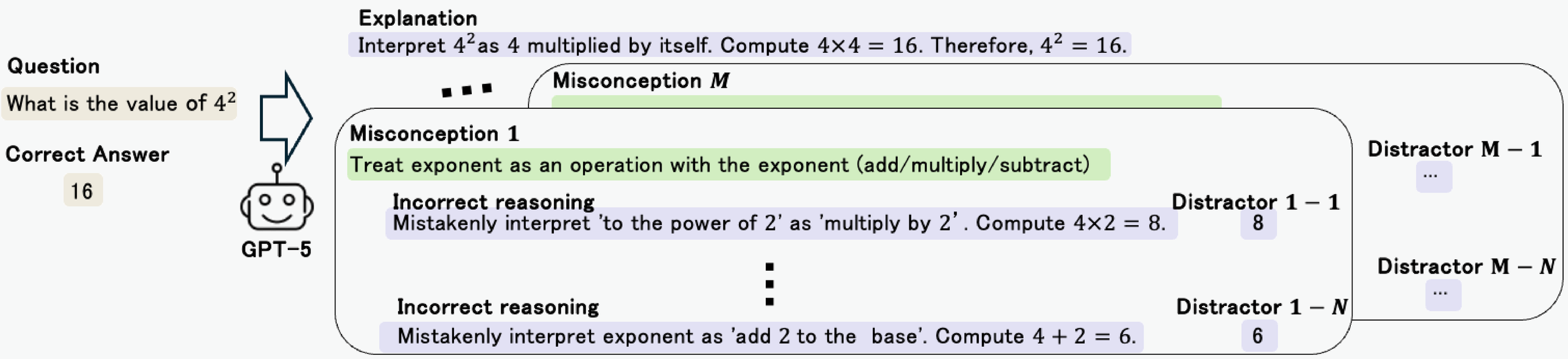
・数学の多肢選択問題(MCQs)の誤答選択肢生成において, 誤概念を中間生成させると, 生成される誤答選択肢の質が向上[Fernandez+'24]

・既存手法では誤概念と誤答の一貫性が低いことが課題[Fernandez+'24]

・誤概念と誤答の一貫性を向上させ, 誤答選択肢の質を向上させたい

提案手法

入力{問題文,正答} 出力{正答の計算過程, 誤概念(M個), **誤り計算過程**($M \times N$個), 誤答選択肢($M \times N$個)}



実験設定

データセット: Eedi Mining Misconceptions in Mathematics (各問 = 問題文, 正答, 誤答選択肢×3, 誤概念×3)

モデル : ・GPT-5 ・DiVERT (Mistral-7Bを基盤とするベースライン)

誤概念-誤答選択肢一貫性評価 (LLM-as-a-Judge)

{問題文, 生成した誤概念, 生成した３つの誤答選択肢}を入力. 誤概念に対応しそうな選択肢を選ばせる

```
Which answer choice a student with the following misconception would be most likely to choose?
Question: What is the value of 4^2
Misconception: Treat exponent as an operation with the exponent
Choices: A) 44 B) 8 C) 42 D) Not A–C
```

Score
✓ 1
✗ 0

誤答選択肢評価

人が作成した３つの誤答のうち, モデルが何個同じ誤答を出力できたか割合(Proportional match)を計算

結果

表 1: 誤概念-誤答選択肢の一貫性

手法	Accuracy	Incorrect	Incorrect(chose D)
DiVERT(Baseline)	0.5326	143	87
GPT-5	0.9346	20	6
GPT-5(w/o incorrect reasoning)	**0.9673**	10	5

表 2: 誤答選択肢評価

手法	K=3	K=15
DiVERT(Baseline)	**45.45**	**63.64**
GPT-5	40.22	56.20
GPT-5(w/o incorrect reasoning)	40.50	**60.06**

Kは生成する誤答の数

・誤概念-誤答選択肢の**一貫性**が向上 ⇨ **同時生成**の効果, 使用モデルの能力の差

・誤答選択肢の**一致率**が低下 ⇨ 学習の有無, **Match評価の限界**

・誤り計算過程を出力しない方が性能が向上 (考察中 計算過程がノイズになった？)

Match評価の限界

データセットの選択肢と**一致＝良い誤答？**

今後の展望

- 誤り計算過程が生成する誤答選択肢に与える影響の考察
- 人やLLMに解かせることによる誤答選択肢評価