日本語の科学設問集の複数評価者による難易度付けと評価者間不一致の分析

東京学芸大学　江原遥

アノテータ：
但木 勇斗，渡邊 舜，村越 礼旺

ありがとうございました！

https://rebrand.ly/JMMLUsciencedifficulty

動機

STEM教育：教育AI分野でも結局多く研究者がやりたいと思っている

個別最適な教育，発達の最近接領域(Zone of Proximal Development, ZPD)＆足場かけ(Scaffolding)：

学習者が今できることの少し先のことができるように教育しよう

→難易度評価

LLM-as-a-Judge [Zheng et al.,NeurIPS2023]: LLMに評価させたい

難易度評価がLLMにどの程度正確にできるか？

→そもそも人間の難易度評価がどの程度正確か？

実はこれを調べられるデータセットは本当に少ない

BEA24 Shared Task→医学。受験者集合からの難易度の値1つ

CEFR-SP [Arase et al., EMNLP2022]（科目が英語）

だから今回作りました

データセットの内容

大規模言語モデルの評価用の多肢選択式設問データセットとしてMassive Multitask Language Understanding [MMLU, HendrycksらICLR21]が有名。STEM教育の内容を含む(著者は東京都のSTEM教育を担う技術科教員養成の国立大学の学科)。MMLUの一部を日本語に翻訳したデータセットとして日本語MMLU(JMMLU)が提案されている[尹ら言語処理学会24]

JMMLUの高校レベルの物理・化学・生物のデータセットについて、日本の学習指導要領に沿って4段階に難易度を付けたデータを作成した [江原, JLR25]

今回、**JLR25のデータを一切見ずに卒研生3名に改めて4段階の難易度付けをしてもらい、評価者間一致度が計算できるようにした**

- 0:中学までの知識で解ける
- 1:〇〇基礎までの知識で解ける
- 2:〇〇までの知識で解ける
- 3:高校を超える水準の問題である

アノテータ
弊学技術科教室の卒研生3名

理科の学習指導要領・学習指導要領の解説を読み込み、それに基づいてアノテートした

結果

- 0:中学までの知識で解ける
- 1:〇〇基礎までの知識で解ける
- 2:〇〇までの知識で解ける
- 3:高校を超える水準の問題である

約150問

2評価者間の一致

高校生物

κ係数：0.1066，QWK：0.2229

高校化学

κ係数：0.4260，QWK：0.5699

高校物理

κ係数：0.3333，QWK：0.3874

κ係数の解釈

0.0- 0.20:Slight

0.21 - 0.40:Fair

0.41 - 0.60: Moderate

0.61- 0.8:Substantial

[Landis and Koch,1977]

科目によって評価者間一致率に大きな違いがある

GPT-OSS-20Bと2026年の評価者との比較

- 0:中学までの知識で解ける
- 1:〇〇基礎までの知識で解ける
- 2:〇〇までの知識で解ける
- 3:高校を超える水準の問題である

約150問

2評価者間の一致

高校生物

κ係数：0.1066，QWK：0.2229

GPT-OSS-20B　κ係数：0.1411，QWK：0.2054

高校化学

κ係数：0.4260，QWK：0.5699

GPT-OSS-20B　κ係数：0.1455，QWK：0.2355

高校物理

κ係数：0.3333，QWK：0.3874

GPT-OSS-20B　κ係数：0.3172，QWK：0.4235

科目によって評価者間一致率に大きな違いがある

LLMとの一致率だけ低いことがある