言語処理学会第31回年次大会 併設ワークショップ JLR2025
日本語言語資源の構築と利用性の向上

日本語プロンプトにおける zero-shot-CoT の効果

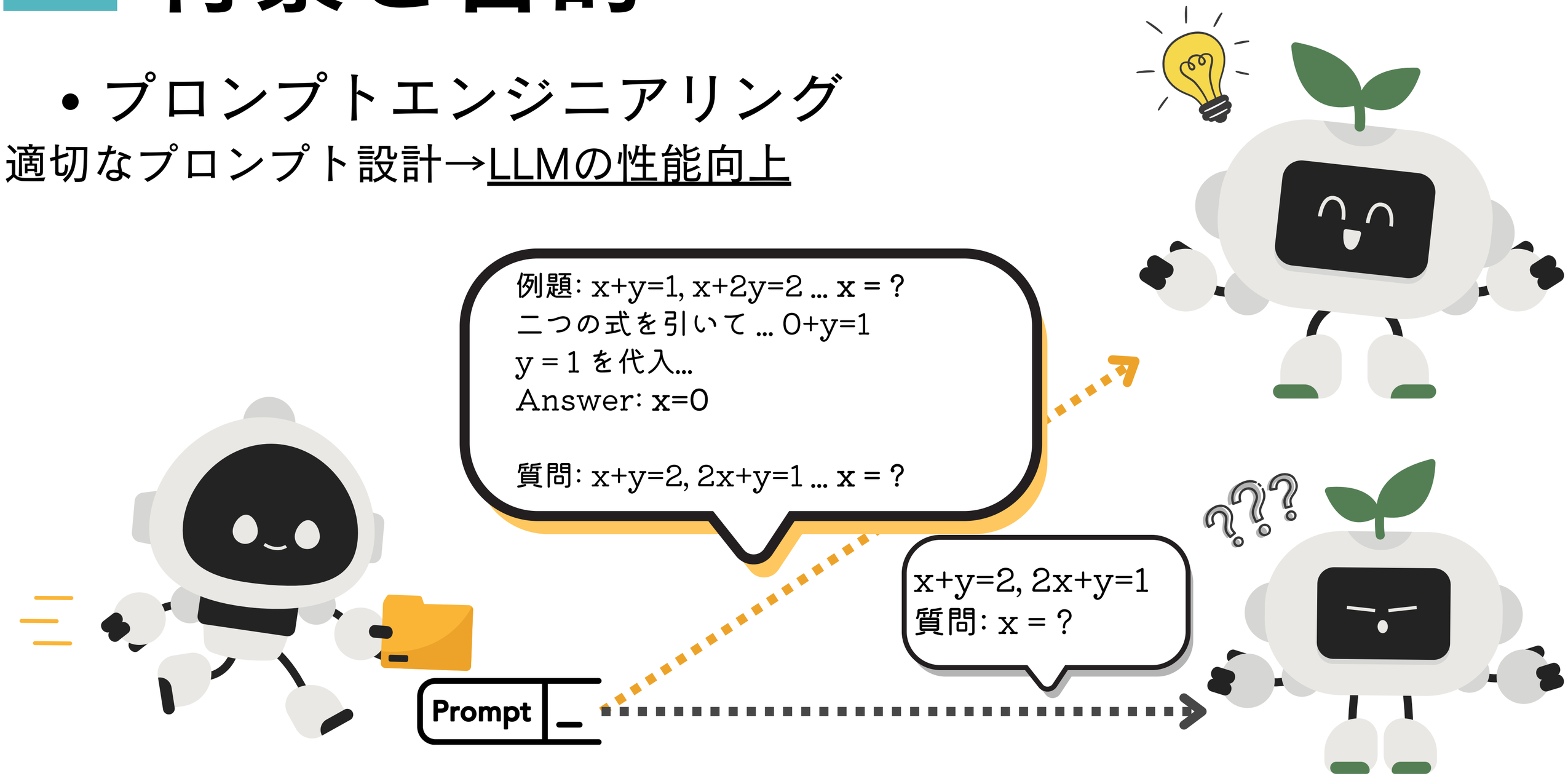
Effectiveness of zero-shot-CoT in Japanese Prompts

公立はこだて未来大学
高山 修輔
Ian Frank

背景と目的

- プロンプトエンジニアリング

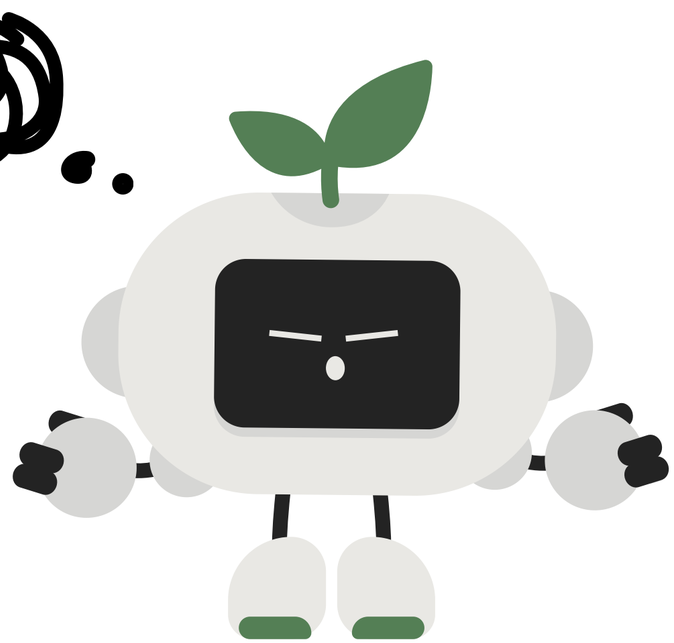
適切なプロンプト設計→LLMの性能向上



背景と目的

- 日本語プロンプトの課題

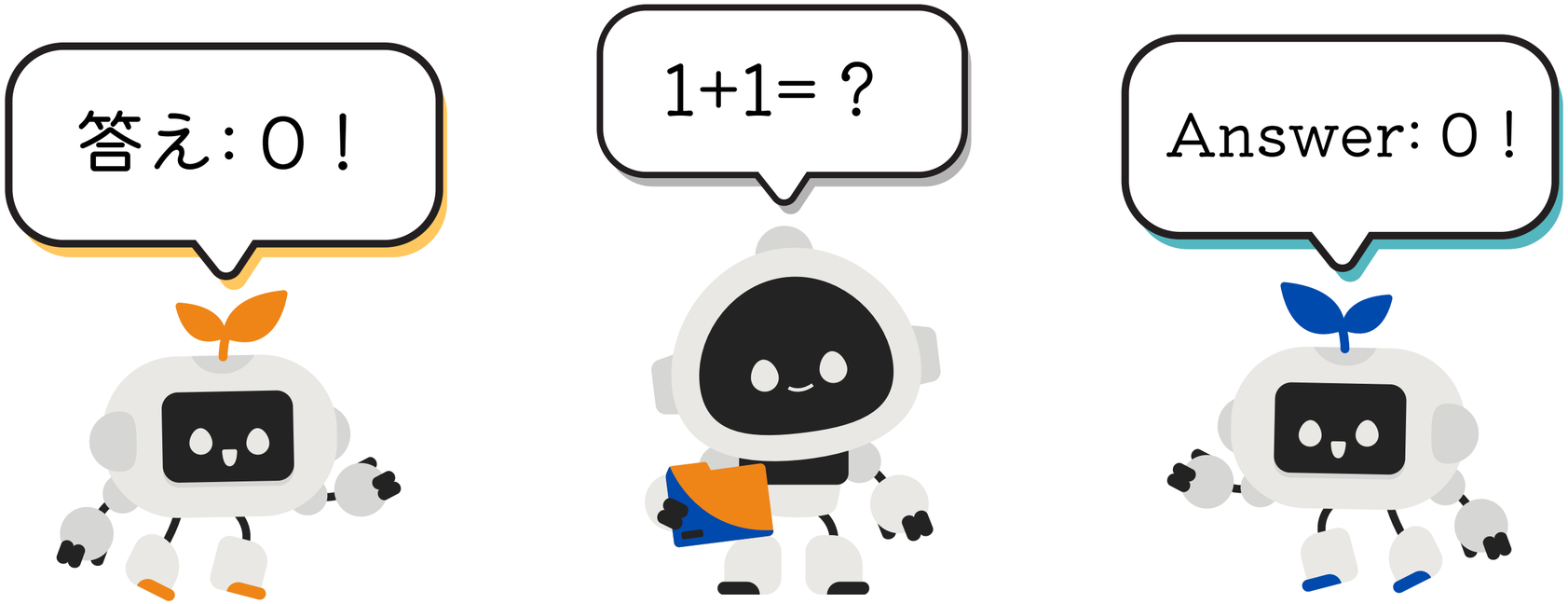
既存研究の多くは英語ベース（1710件 vs 112000件）



背景と目的

- 目的

プロンプトの効果の検証: 日本語と英語で比較



提案手法

- 概要

JMMLU
MMLU

zero-shot-CoT

GPT-3.5
GPT-4o-mini

Prompt _

Dataset

Prompt

Model

実験と評価

- GPT-3.5

英語・日本語ともに、CoTなしの方がスコアが高い
スコアの低下幅：MMLU -0.022、JMMLU -0.059
スコア上昇タスク：MMLU 13、JMMLU 8

JMMLU：t値=2.29、p値=0.02（有意差あり）
MMLU：t値=0.79、p値=0.46（有意差なし）

	ja-cot	ja-no cot	COT change J	en-cot	en-no cot	COT change E
GPT3.5	0.469	0.528	-0.059	0.580	0.602	-0.022
GPT4o-mini	0.332	0.666	-0.334	0.258	0.682	-0.424

実験と評価

- GPT-4o-mini

CoTによるスコア低下が顕著（3.5よりも大幅に低下）
スコアの低下幅：MMLU -0.424、JMMLU -0.334
スコア上昇タスク：MMLU 0、JMMLU 2

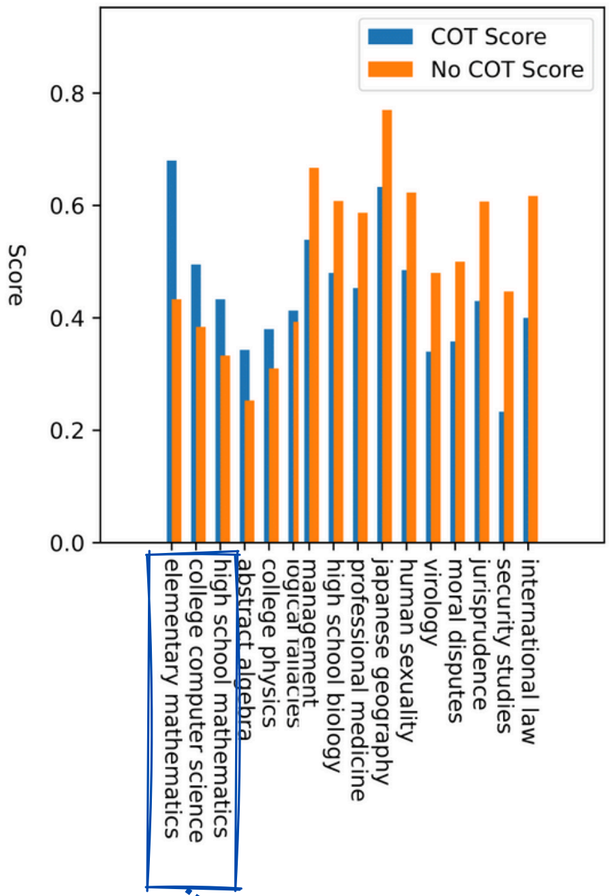
JMMLU：t値=17.0、p値=3.61×10^{-25}（有意差あり）
MMLU：t値=14.1、p値=1.60×10^{-25}（有意差あり）

	ja-cot	ja-no cot	COT change J	en-cot	en-no cot	COT change E
GPT3.5	0.469	0.528	-0.059	0.580	0.602	-0.022
GPT4o-mini	0.332	0.666	-0.334	0.258	0.682	-0.424

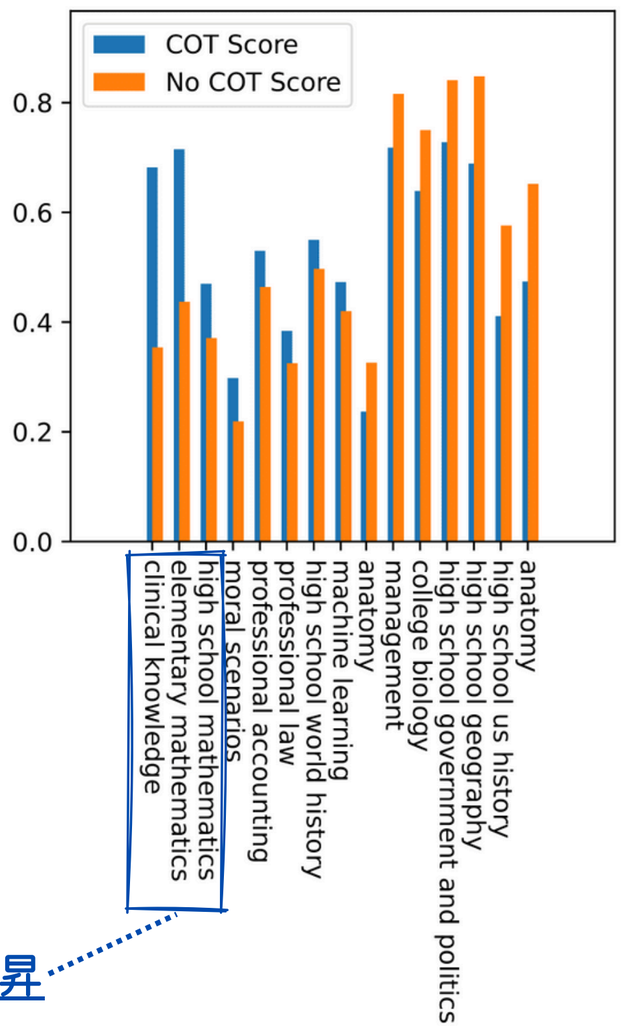
実験と評価

青色が長い: CoTによってスコア上昇
オレンジが長い: CoTによってスコア下降

GPT-3.5 J

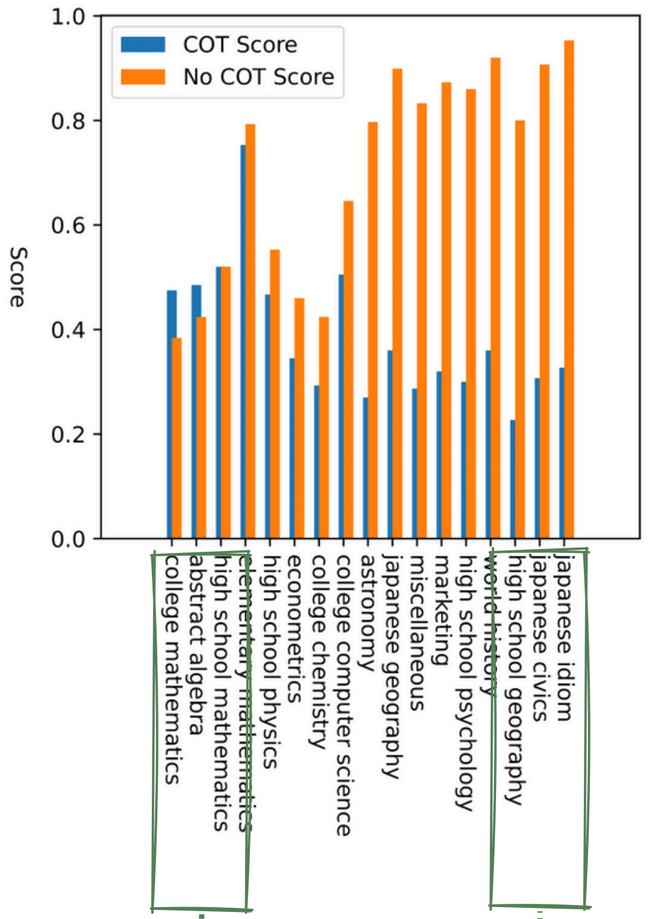


GPT-3.5 E



数学タスク: 上昇

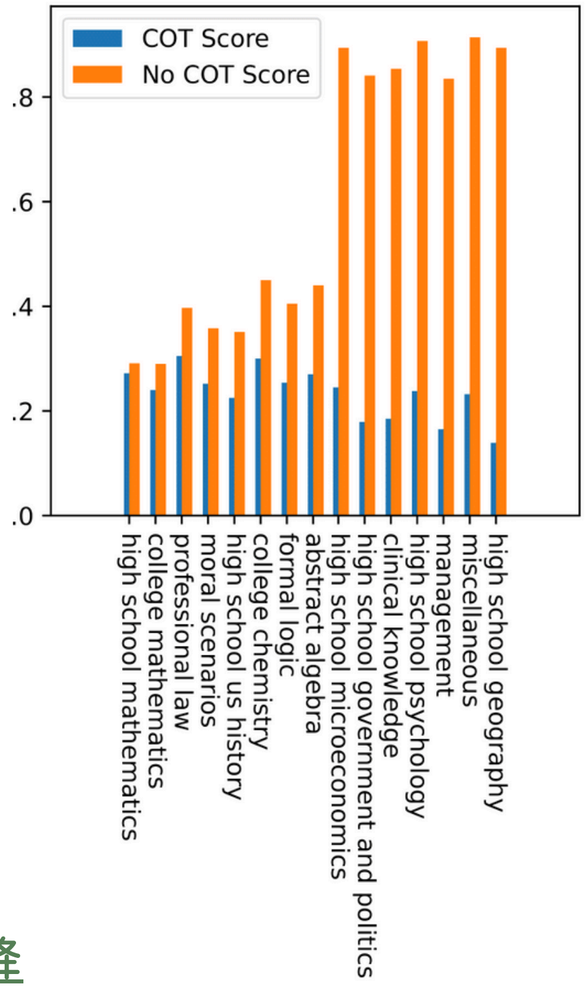
GPT-4o J



日本語タスク: 下降

数学タスク: 上昇

GPT-4o E



タスク全てで低下

実験と評価

- 英語と日本語の比較

GPT-3.5:

- 共に数学でスコア上昇
- スコア上昇タスク数 J<E

GPT-4o-mini:

- スコア上昇タスク J→2 E→0
- 低下幅 J<E

英語で有効

日本語で有効

実験と評価

- GPT-3.5:

CoTなし: 生成された回答が答えだけの場合も多い
CoTあり: ほぼすべての回答に推論ステップが追加

- GPT-4o-mini:

CoTなし: 推論ステップを最初から含む
CoTあり: 推論ステップを含む回答数には変化なし

- CoTが余計な推論を追加?

→節の増加による悪影響[1]

[1] Mirzadeh, Iman, Alizadeh, Khashayar, Shahrokhi, Hamid, Tuzel, Osman, Bengio, Yoshua, & Farajtabar, Mohammad. "GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models." arXiv:2410.05229, 2024.

結論

- 英語と日本語

GPt-4oでは日本語の方が有効な効果
→特に数学タスク

- スコア低下の原因

推論の増加
→余計な推論を追加してしまっている？