LLM における指示再確認プロンプトの効果検証

的石竜和 [1]　本行智光 [1]　吉岡大三郎 [1]

概要：大規模言語モデル(Large Language Model；LLM)が急速に発展し，大規模な商用 LLM からオープンソースで小規模な LLM など，LLM の選択肢も拡大している．LLM の性能は与えるプロンプトに依存することが知られ，効果的なプロンプト設計の研究が近年盛んに提案されている．特に小規模 LLM は回答が安定しない傾向にあるため，小規模モデルにおいてプロンプト設計は特に重要と考えられる．

プロンプト設計の代表手法として複数の例文を与える Few-shot 学習が知られている．我々の先行研究では，Few-shot 学習を用いた分類問題において「指示内容を再確認するプロンプト」を提案し，Chat GPT モデルにおいて分類精度の改善を実証している．しかし，Chat GPT を用いた評価のみであり，モデルに依存しない汎用性の検証が必要と考える．

そこで本研究では，小規模 LLM を対象に，英文と和文の自然言語処理データセットを 6 つ用いて検証した．その結果，指示再確認プロンプトにより回答精度の向上が達成されたことを報告する．

キーワード：LLM，プロンプト設計，Few-shot 学習

Evaluation of effectiveness of reconfirm prompt for LLMs

RYUTO MATOISHI[†1]　TOSHIMITSU HONGYO[†1]
DAISABURO YOSHIOKA[†1]

Abstract: Large Language Models (LLMs) are developing rapidly, with the range of options expanding from large-scale commercial LLMs to smaller open-source LLMs. Since LLM performance depends on the prompts provided, extensive research on effective prompt design has been proposed in recent years. Prompt design is particularly important for smaller models, as their outputs tend to be less stable.

Few-shot learning, which involves providing multiple example sentences, is a well-known representative technique for prompt design. In our previous research, we proposed a ‘reconfirm prompt’ for classification tasks using few-shot learning, and demonstrated the effectivenss on the ChatGPT model.

In this study, we verify the effectiveness of our reconfirm prompt for small-scale LLMs using natural language processing datasets and reports that an improvement in accuracy was achieved.

Keywords: LLM, prompt design, Few-shot learning

1. はじめに

大規模言語モデル(Large Language Model；LLM)が急速に発展し，ChatGPT や Gemini などの商用 LLM から Llama などのオープンソース LLM など，LLM の選択肢も拡大している．LLM の性能は与えるプロンプトに依存するため，効果的なプロンプト設計の研究が近年盛んに提案されている[1]．小規模 LLM は回答が安定しない傾向にあるため，特に小規模モデルにおいてプロンプト設計は重要になると考えられる．

プロンプト設計の代表手法として，複数の例文を与える Few-shot 学習がよく知られている[2]．また，”Let’s think step by step”の一文を加えることで推論過程を促す zero-shot CoT(Chain of thought)[3]や，問題文を繰り返す Re-reading も提案されている．これらのような簡単な手法で効果のあるプロンプト設計は，実用上特に重要とされる．

一方で Few-shot 学習を用いた分類タスクでは，与える例文はもとより，例文の順序を変えるだけで回答精度が揺らぐことや最後に与えた例文のラベルに回答が偏る直近バイアスの存在が報告されている[5]．そこで我々は，Few-shot 学習を用いた分類問題において「タスクに関する指示内容を最後に再確認するプロンプト」を提案し[6]，Chat GPT モデルにおいて分類精度の改善を実証した[7]．しかし，Chat GPT を用いた評価のみであり，モデルに依存しない汎用性の検証が必要と考える．

そこで本研究では，複数の LLM を対象に，指示内容再確認プロンプトの有効性を検証する．また zero-shot CoT と Re-reading との比較も行い，指示再確認プロンプトの有効性を評価する．

2. プロンプト設計

Few-shot 学習は複数の例文を与えることで，回答精度を上げるプロンプト手法であり，プロンプト設計手法として最もよく知られている．感情分析タスクに使用する標準的な Few-shot 学習のプロンプトを表 1 にのせる．プロンプトの冒頭にタスクの説明や回答フォーマットを指示する指示

1 崇城大学情報学部情報学科
Dept. of Computer and Information Sciences, Sojo University

文 I を置き，*n* 個の例文からなる例文 E，そして問題文 Q の構成となる．LLM は与えたプロンプトの最初と最後が重要であることが報告されており[8]，タスクに関する指示文が最も重要な情報と考えられるため，冒頭に加えて例文の後に再度指示文を繰り返す再確認プロンプトを提案した．フォーマットは以下の構成となる．

表 1 Few-shot 学習プロンプトの例

I	Please analyze the text and evaluate its sentiment polarity. Assign 0 for negative sentiment and 1 for positive sentiment. Provide only the numerical response.
E	Please refer to the following examples. Sentence: this is one of polanski's best films . Result: 1 …
Q	Sentence: no movement no yuks not much of anything.

表 2 指示再確認プロンプト

I	Please analyze the text and evaluate its sentiment polarity. Assign 0 for negative sentiment and 1 for positive sentiment. Provide only the numerical response.
E	Please refer to the following examples. Sentence: this is one of polanski's best films . Result: 1 …
A	Please analyze the text and evaluate its sentiment polarity. Assign 0 for negative sentiment and 1 for positive sentiment. Provide only the numerical response.
Q	Sentence: no movement no yuks not much of anything.

指示再確認文 A は，指示文 I と同一文章のコピーである．通常の Few-shot 学習が I+E+Q 構造に対して，提案プロンプトは I+E+A+Q 構造となる．

本稿では，zero-shot CoT と Re-reading を用いた Few-shot 学習も検証する．それら手法の例を以下にのせる．Zero-shot CoT は指示文の後に推論を促す"Let's think step by step"を挿入する I+Z+E+Q 構文，Re-reading は問題文の後に再度問題を繰り返す I+E+Q+R 構文と定義する．

3. 数値実験

指示再確認プロンプトの有効性を検証するために，標準的な自然言語処理ベンチマークである GLUE[9] と JGLUE[10]から英文と和文データセットを用いて評価する．表 4 に使用するデータセットをのせる．

SST-2 は映画レビューから positive, negative の 2 値ラベルが付与されたデータセットである．WRIME は日本語 SNS の投稿を集めた感情分析タスクのためのデータセットであり，感情値が 5 段階(strong or weak positive, strong or weak negative, neutral)でラベル付けされている[11]．本研究では，positive(strong and weak positive)， neutral, negative(strong and weak negative)の 3 値として分類タスクに用いた．

表 3 zero-shot CoT と Re-reading を用いた Few-shot 学習のプロンプト

I	Please analyze the text and evaluate its sentiment polarity. Assign 0 for negative sentiment and 1 for positive sentiment. Provide only the numerical response.
Z	Let's think step by step.
E	Please refer to the following examples. Sentence: this is one of polanski's best films . Result: 1 …
Q	Sentence: no movement no yuks not much of anything.
R	Read the question again. Sentence: no movement no yuks not much of anything.

表 4 使用タスクとデータセット

タスク	日本語	英語
感情分析	WRIME	SST-2
文書ペア分類	JNLI	MNLI
テキスト類似性	JSTS	STS-B

文章ペア分類タスクは，2 文を与え含意，矛盾，中立か判定するタスクである．使用するデータセットは MNLI と JNLI である．JSTS と STS-B は 2 文ペアとそれらの類似度 0~5 の実数値が付与されている．

すべてのタスクにおいて，テストデータ 600 件で評価し，正解率を調べた．ここで，テストデータの中で各クラスに含まれるデータ数は同数としている．例として 2 値分類タスクである SST-2 では，positive が 300 件，negative が 300 件としている．Few-shot 学習で用いる例文は，各分類クラスから例文 *n* 個を学習データから無作為に抽出する．2 値分類では，positive(P)から *n* 個，negative(N)から *n* 個の例文を選び，NPNP…と PNPN…の 2 種類の順番で評価する．例文を 5 回変えて同様の実験を行うため，計 10 パターンの Few-shot 例文から平均正解率を調べた．3 値分類である WRIME, MNLI, JNLI では，含意(E)，矛盾(M)，中立(N)の例文を *n* 個ずつ無作為に抽出し，4 種類の順番で評価する．例文を 5 回変え，計 20 パターンの Few-shot 例文から平均正解率を調べた．JSTS と STS-B も例文を 5 回変え，LLM の回答結果が実数値となるため，正解値との Spearman 相関で評価した．

3.1 ChatGPT を用いた検証結果

代表的な商用 LLM である ChatGPT を用い，通常の Few-shot 学習のプロンプト(I+E+Q)と指示再確認プロンプト(I+E+A+Q)の結果を表 5~10 にのせる．ChatGPT のモデルは ChatGPT 4o-mini を用いている．また，比較のために，zero-shot CoT と Re-reading の結果も表 11～16 にのせる．表中では，4 つの結果の中で最も高い値を太字で示している．結果より，WRIME タスク以外の 5 つのタスクでは，指示再確認プロンプトが正解率の平均，最小，最大値が高いこと

が確認される．Re-reading は WRIME で最も高い正解率となり, また他のタスクでも良好な結果が多い. 一方で, Zero-shot CoT は最も正解率が低い結果が多いことがわかる．

表 5 SST-2 における Few-shot 学習と指示再確認の結果

	Basic(I+E+Q)			Reconfirm(I+E+A+Q)		
n	Min.	Max.	Avg.	Min.	Max.	Avg.
2	89.00	**96.50**	94.16	**92.83**	96.33	**94.63**
6	94.00	96.50	95.54	**94.50**	**96.83**	**95.58**
10	93.33	**97.50**	95.14	**94.83**	96.66	**95.61**
50	83.50	**97.50**	94.21	**95.50**	96.50	**95.99**

表 6 WRIME における Few-shot 学習と指示再確認の結果

	Basic(I+E+Q)			Reconfirm(I+E+A+Q)		
n	Min.	Max.	Avg.	Min.	Max.	Avg.
3	64.00	69.50	67.08	**67.17**	69.83	68.27
6	63.67	69.83	66.92	66.17	70.67	68.53
9	65.00	68.67	66.85	66.67	70.00	68.38
51	39.50	63.83	50.80	66.33	69.33	68.22

表 7 JNLI における Few-shot 学習と指示再確認の結果

	Basic(I+E+Q)			Reconfirm(I+E+A+Q)		
n	Min.	Max.	Avg.	Min.	Max.	Avg.
3	62.33	77.83	72.62	**79.33**	**83.80**	**81.40**
6	66.83	80.33	75.81	**79.17**	**82.83**	**81.09**
9	73.33	80.50	77.28	**76.83**	**83.83**	**81.01**
51	75.83	80.83	78.19	**76.33**	**81.33**	**78.55**

表 8 MNLI における Few-shot 学習と指示再確認の結果

	Basic(I+E+Q)			Reconfirm(I+E+A+Q)		
n	Min.	Max.	Avg.	Min.	Max.	Avg.
3	70.50	78.17	75.23	**74.83**	**80.50**	**78.39**
6	71.83	79.50	75.67	**75.83**	**80.50**	**77.82**
9	74.67	**79.83**	76.88	**76.00**	**79.83**	**78.03**
51	72.17	77.17	74.60	**74.33**	**79.83**	**76.28**

表 9 JSTS における Few-shot 学習と指示再確認の結果

	Basic(I+E+Q)			Reconfirm(I+E+A+Q)		
n	Min.	Max.	Avg.	Min.	Max.	Avg.
3	0.308	0.896	0.763	**0.881**	**0.909**	**0.889**
6	0.874	0.891	0.882	**0.895**	**0.902**	**0.898**
9	0.869	0.894	0.881	**0.883**	**0.903**	**0.895**
50	0.689	0.868	0.774	**0.899**	**0.905**	**0.902**

表 10 STS-B における Few-shot 学習と指示再確認の結果

	Basic(I+E+Q)			Reconfirm(I+E+A+Q)		
n	Min.	Max.	Avg.	Min.	Max.	Avg.
3	0.871	0.894	0.884	**0.916**	**0.933**	**0.923**
6	0.903	0.914	0.908	**0.916**	**0.928**	**0.924**
9	0.906	0.919	0.913	**0.922**	**0.932**	**0.923**
50	0.889	0.905	0.894	**0.921**	**0.936**	**0.923**

表 11 SST-2 における zero-shot CoT と Re-reading の結果

	I+Z+E+Q			I+E+Q+R		
n	Min.	Max.	Avg.	Min.	Max.	Avg.
2	65.7	83.8	76.2	75.3	84.5	80.6
6	58.7	84.2	75.6	80.2	84.2	82.4
10	78.8	85.2	82.7	80.2	84.3	82.6
50	55.2	78.5	63.9	81.2	83.7	82.6

表 12 WRIME における zero-shot CoT と Re-reading の結果

	I+Z+E+Q			I+E+Q+R		
n	Min.	Max.	Avg.	Min.	Max.	Avg.
3	65.3	73.7	69.0	65.3	**72.3**	**70.2**
6	65.0	72.3	69.0	**67.7**	**71.7**	**70.0**
9	57.3	73.7	68.2	**67.3**	**72.7**	**69.9**
51	37.7	60.3	49.9	**67.3**	**72.0**	**69.6**

表 13 JNLI における zero-shot CoT と Re-reading の結果

	I+Z+E+Q			I+E+Q+R		
n	Min.	Max.	Avg.	Min.	Max.	Avg.
3	56.3	77.3	71.0	73.1	80.1	77.5
6	57.8	75.7	67.9	74.1	80.5	77.0
9	43.7	65.5	54.6	72.8	79.8	76.5
51	34.0	55.8	45.7	69.8	80.3	76.1

表 14 MNLI における zero-shot CoT と Re-reading の結果

	I+Z+E+Q			I+E+Q+R		
n	Min.	Max.	Avg.	Min.	Max.	Avg.
3	65.8	76.0	71.4	73.7	79.7	76.1
6	63.2	74.3	69.3	74.0	77.3	75.7
9	59.2	76.0	68.6	73.0	78.3	75.3
51	58.0	69.8	65.9	66.8	74.2	71.8

表 15 JSTS における zero-shot CoT と Re-reading の結果

	I+Z+E+Q			I+E+Q+R		
n	Min.	Max.	Avg.	Min.	Max.	Avg.
3	0.848	0.894	0.864	0.848	0.892	0.873
6	0.559	0.888	0.837	0.839	0.894	0.875
9	0.824	0.882	0.857	0.867	0.894	0.882
50	0.524	0.866	0.723	0.851	0.890	0.876

表 16 STS-B における zero-shot CoT と Re-reading の結果

	I+Z+E+Q			I+E+Q+R		
n	Min.	Max.	Avg.	Min.	Max.	Avg.
3	0.862	0.909	0.893	0.884	0.916	0.900
6	0.872	0.916	0.902	0.892	0.916	0.907
9	0.888	0.912	0.903	0.895	0.925	0.911
50	0.842	0.905	0.875	0.905	0.929	0.915

3.2 LLM を用いた検証結果

オープンソース LLM として，Meta 社が公開した Llama 3.1 8b(llama3.1:8b)モデルと日本語強化モデル Swallow 8b(okamototk/llama-swallow:8b)モデルを用いた. 英文データセット(SST-2, MNLI, STS-B)は Llama 3.1 を用い，日本語データセット(WRIME, JNLI, JSTS)は Swallow で評価した．前節の検証において Zero-shot CoT は Few-shot 学習と比べても効果が見られなかったため除外し, Few-shot 学習(I+E+Q), 指示再確認(I+E+A+Q)，Re-reading(I+E+Q+R)の 3 手法の結果を表 17〜21 にまとめる．軽量モデルのため，正しく数値で回答が得られず評価できない回答もみられたため, ”不値”としてその総数ものせている．表中では，3 つの結果の中で最も高い値を太字で示している．

結果より, 多くのタスクで指示再確認プロンプトの有効性が確認される．Llama 3.1 モデルの Few-shot 学習では，数値としての回答が得られないことが複数確認されるが，指示再確認によりほぼすべて改善されることがわかる．特に 2 文の類似度を判定する STS タスクで不値の数が多く，Re-reading もある程度改善できているが，指示再確認プロンプトによりほぼ全て数値での回答が得られた.

Few-shot 学習では例文の順序や例文に依存して結果が揺らぐことが知られているため，ゆらぎの抑制について箱ひげ図で評価した. 6 つの結果から得られた箱ひげ図を図 1~3 にのせる．ほぼ全てのタスクにおいて，指示再確認プロンプトは安定した結果が得られていることが確認できる．

表 17 SST-2 における比較結果

	Basic(I+E+Q)				Reconfirm(I+E+A+Q)				Re-reading(I+E+Q+R)			
n	Min.	Max.	Avg.	不値	Min.	Max.	Avg.	不値	Min.	Max.	Avg.	不値
2	0.499	0.512	0.501	82	0.820	**0.953**	0.890	**0**	**0.875**	0.947	**0.923**	**0**
6	0.500	0.527	0.506	156	**0.928**	**0.968**	0.939	**0**	0.913	**0.968**	**0.947**	**0**
10	0.495	0.615	0.516	121	0.945	0.970	0.955	**0**	**0.950**	**0.978**	**0.966**	**0**
50	0.478	0.525	0.502	**0**	0.928	**0.965**	**0.956**	**0**	**0.943**	**0.965**	0.955	**0**

表 18 WRIME における比較結果

	Basic(I+E+Q)				Reconfirm(I+E+A+Q)				Re-reading(I+E+Q+R)			
n	Min.	Max.	Avg.	不値	Min.	Max.	Avg.	不値	Min.	Max.	Avg.	不値
3	0.588	0.690	**0.650**	**0**	**0.618**	0.677	0.647	**0**	0.605	**0.695**	0.649	**0**
6	**0.625**	0.687	**0.653**	**0**	0.620	0.675	0.647	**0**	0.542	**0.692**	0.645	**0**
9	0.613	0.668	**0.648**	**0**	**0.622**	0.670	0.645	**0**	0.580	**0.690**	0.645	**0**
51	0.545	0.650	0.605	**0**	**0.610**	0.653	0.632	**0**	0.608	**0.660**	**0.638**	**0**

表 19 JNLI における比較結果

	Basic(I+E+Q)				Reconfirm(I+E+A+Q)				Re-reading(I+E+Q+R)			
n	Min.	Max.	Avg.	不値	Min.	Max.	Avg.	不値	Min.	Max.	Avg.	不値
3	0.447	**0.713**	0.589	**0**	**0.542**	0.702	**0.628**	**0**	0.487	0.693	0.589	**0**
6	0.485	**0.722**	0.606	**0**	**0.583**	0.720	**0.662**	**0**	0.538	0.720	0.636	**0**
9	0.475	0.710	0.588	**0**	**0.577**	**0.737**	**0.665**	**0**	0.530	0.687	0.618	**0**
51	0.448	0.668	0.564	**0**	**0.608**	**0.755**	**0.689**	**0**	0.567	0.732	0.645	**0**

表 20 MNLI における比較結果

	Basic(I+E+Q)				Reconfirm(I+E+A+Q)				Re-reading(I+E+Q+R)			
n	Min.	Max.	Avg.	不値	Min.	Max.	Avg.	不値	Min.	Max.	Avg.	不値
3	0.417	0.601	0.532	17	**0.542**	**0.680**	**0.615**	**0**	0.445	0.673	0.576	**0**
6	0.375	0.640	0.550	5	**0.605**	**0.702**	**0.653**	**0**	0.433	0.688	0.587	**0**
9	0.428	0.612	0.514	56	**0.573**	**0.690**	**0.640**	**0**	0.420	0.680	0.581	**0**
51	0.357	0.495	0.411	**0**	**0.587**	**0.702**	**0.660**	**0**	0.518	0.655	0.590	**0**

表 20 JSTS における比較結果

	Basic(I+E+Q)				Reconfirm(I+E+A+Q)				Re-reading(I+E+Q+R)			
n	Min.	Max.	Avg.	不値	Min.	Max.	Avg.	不値	Min.	Max.	Avg.	不値
3	0.710	0.844	0.808	**0**	**0.821**	**0.862**	**0.841**	**0**	0.772	0.832	0.811	**0**
6	0.751	0.829	0.793	**0**	0.811	0.854	**0.837**	**0**	**0.812**	**0.859**	0.834	**0**
9	0.725	0.829	0.783	**0**	**0.826**	**0.860**	**0.837**	**0**	0.791	0.856	0.820	**0**
51	0.461	0.744	0.645	**0**	**0.820**	**0.843**	**0.832**	**0**	0.769	0.805	0.792	**0**

表 21 STS-B における比較結果

	Basic(I+E+Q)				Reconfirm(I+E+A+Q)				Re-reading(I+E+Q+R)			
n	Min.	Max.	Avg.	不値	Min.	Max.	Avg.	不値	Min.	Max.	Avg.	不値
3	0.124	0.562	0.391	17	0.601	**0.737**	0.684	**1**	**0.625**	**0.737**	**0.689**	3
6	0.052	0.619	0.404	203	**0.495**	0.719	0.633	**1**	0.494	**0.726**	**0.646**	36
9	0.027	0.616	0.390	597	0.519	**0.716**	0.602	**0**	**0.534**	0.708	**0.628**	82
51	-0.140	0.191	0.011	27	**0.579**	**0.755**	**0.691**	**0**	0.177	0.656	0.403	7

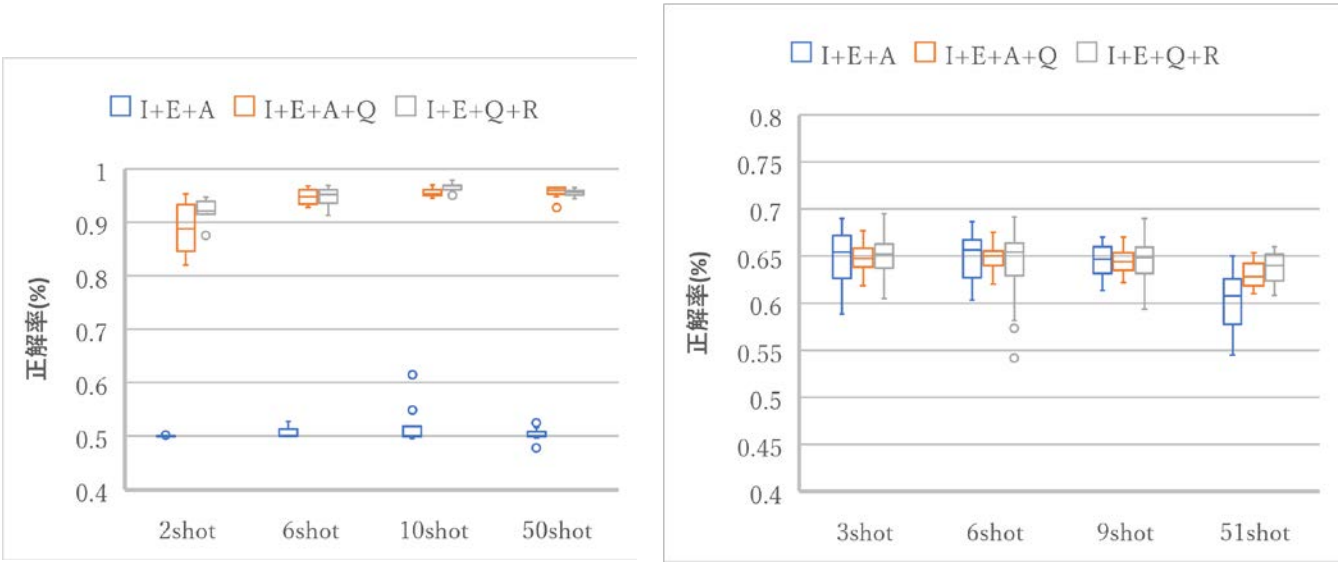

図 1 SST-2（左）と WRIME（右）の箱ひげ図

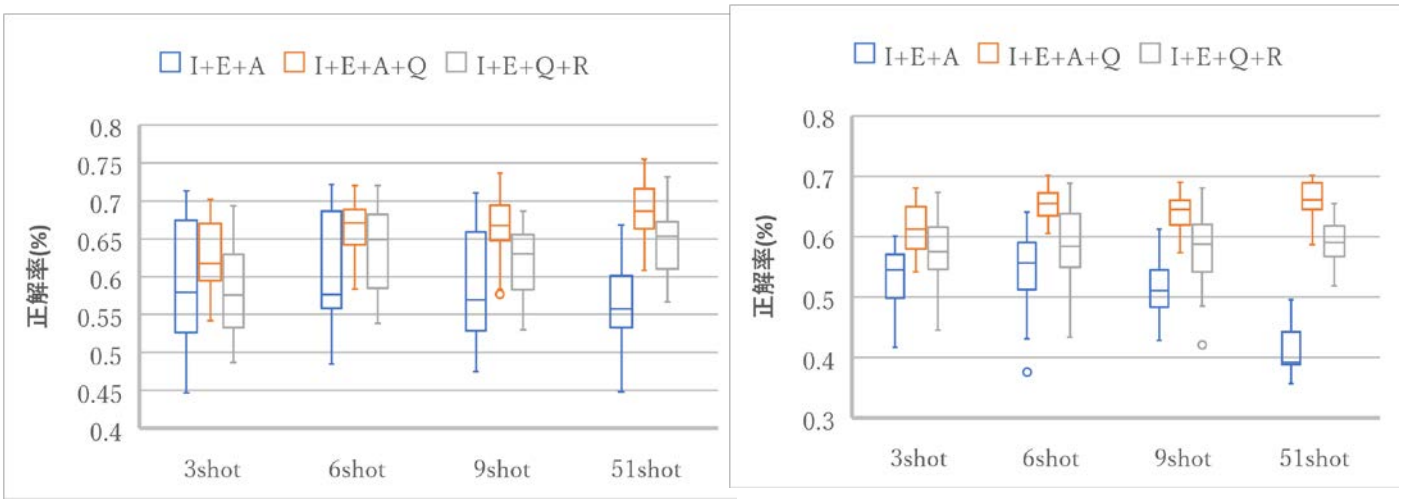

図 2 JNLI（左）と MNLI（右）の箱ひげ図

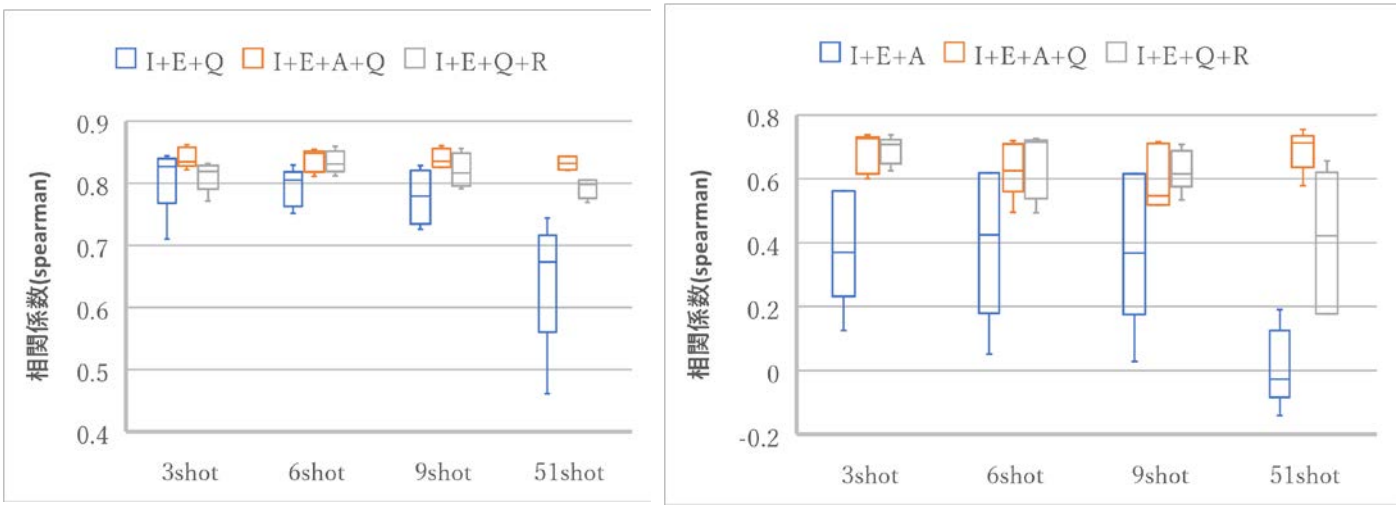

図 3 JSTS（左）と STS-B（右）の箱ひげ図

4. まとめ

本研究では，複数の LLM を用いて指示再確認プロンプトの効果を英文と和文の自然言語処理データセットを用いて評価した．商用 LLM とオープンソース LLM において，多くのタスクで指示再確認により回答精度の向上が達成された．LLM において，どのような仕組みで指示再確認プロンプトが有効になるかの詳細な分析が今後の課題である．

参考文献

1) P. Sahoo, A. K. Singh, S. Saha, V. Jain, S. Mondal, and A. Chadha, “A systematic survey of prompt engineering in large language models: techniques and applications,” *arXiv*, 2024.
2) T. B. Brown, B. Mann, N. Ryder, Jared Kaplan, et al. ,”Language models are few-shot learners,” *arXiv*, 2020.
3) T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, “Large language models are zero-shot reasoners,” arXiv, 2022.
4) X. Xu, C. Tao, T. Shen, C. Xu, H. Xu, G. Long, J. Lou, and S. Ma, “Re-Reading Improves Reasoning in Large Language Models,” *Proc. of the 2024 Conference on empirical methods in natural language processing*, pp.15549−15575, 2024.
5) T. Z. Zhao, E. Wallace, S. Feng, D. Klein and S. Singh, “Calibrate Before Use: Improving Few-Shot Performance of Language Models,” *arXiv*, 2021.
6) 梁池翔太，岡田 勇人，米田 圭佑，吉岡 大三郎，”ChatGPT を用いた SNS 不適切書き込み分類の評価,” 第 38 回人工知能学会全国大会, 2024 年 5 月.
7) 大川珠乃，米田圭佑，吉岡大三郎, “ChatGPT における指示内容再確認による効果の検証,” 情報処理学会全国大会, 2025 年 3 月.
8) N. F. Liu, K. Liu, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni and P. Liang, “Lost in the middle: How language models use long contexts,” *Transactions of the Association for Computational Linguistics*, vol.12, pp.157-173, 2024.
9) A. Wang, A. Singh, J. Michael, F. Hill, O. Levy and S. R. Bowman, “GLUE: A multi-task benchmark and analysis platform for natural language understanding,” Proc. of the 2018 EMNLP workshop, 2018.
10) K. Kurihara, D. Kawahara, and T. Shibata, “JGLUE: Japanese general language understanding evaluation,” Proc. of thirteenth language resources and evaluation conference, 2022.
11) T. Kajiwara, C. Chu, N. Takemura, Y. Nakashima and H. Nagahara, “WRIME: A New Dataset for Emotional Intensity Estimation with Subjective and Ojective Annotations,” Proc. of NAACL, pp. 2095–2104, 2021.