大規模言語モデルにおける ICL バイアスの選択的補正

酒井 祐介　Natthawut Kertkeidkachorn　白井 清昭
北陸先端科学技術大学院大学 先端科学技術研究科
{yusuke.sakai,natt,kshirai}@jaist.ac.jp

概要

大規模言語モデルを用いた質問応答システムが広く注目されているが, 生成される回答にはさまざまなバイアスが生じることが指摘されている. 本研究では, LLM が In-Context Learning(ICL) における例示の与え方によって引き起こすバイアスに対し, キャリブレーション手法と Retriever-Augmented Generation(RAG) が与える影響を検証する. 特に, 与えた ICL 例示によってモデルの予測が大きく変動する問題を ICL 依存問題と定義し, それらを選択的に補正するキャリブレーション手法を提案した上で, RAG との併用効果を評価した. 実験の結果, 従来手法よりも効果的に ICL によるバイアスを抑制し, 特定条件下では精度でも従来手法を上回ることを確認した.

1　はじめに

大規模言語モデル (LLM) は様々なタスクで高い性能を示しているが, In-Context Learning(ICL) 時の例示の与え方によって出力が大きく変動するという問題（バイアス）が指摘されている. ここで, ICL とは, モデルのパラメータを更新することなく, プロンプト中に示された少数のデモンストレーション（例示）からタスクを推論する LLM の能力を指す [1]. 具体的には, ICL の形態として, 例示を与えない場合を Zero-shot Learning, 1 件だけ与える場合を One-shot Learning, 複数のデモを与える場合を Few-shot Learning と呼ぶ. このようなバイアスを適切に抑制する必要性が高まる一方, 既存の研究では, 例示やタスク形式に依存しやすいバイアスへの対策として, いくつかのキャリブレーション手法が提案されている [2, 3, 4].

また, Retriever-Augmented Generation(RAG)[5] は, 外部知識を活用して, ICL における回答精度を向上させる効果的な手法として注目されている. しかし, RAG が ICL におけるバイアスに与える影響や, 既存のキャリブレーション手法との相互作用については十分な検証がなされていない.

本研究では, ICL 例示によるバイアスの抑制を対象に, キャリブレーション手法と RAG を組み合わせた際の効果を検証する. また, この検証過程で, ICL 例示を与えた際，例示の違いによって回答が大きく変動する問題を「ICL 依存問題」と定義し, それらを選択的に補正する手法の有効性について評価する. 検証は医療系タスクを題材に行う. 医療領域は専門性が特に高く, RAG による外部知識の活用効果を測定する事例として適していると考えられる.

2　関連研究

2.1　Retriever-Augmented Generation

本研究では, Xiong らが提案した医療特化型の RAG フレームワーク MedRAG[6] を使用する. MedRAG は, 医療系を中心とする複数コーパスと多様な LLM を組み合わせ可能な RAG のフレームワークである. あわせて, Xiong らは MIRAGE と呼ばれる医療 QA のベンチマークを作成し, RAG の多角的な性能評価を行っている. MedRAG や MIRAGE ベンチマークの詳細は付録 A に記す.

2.2　In-Context Learning とバイアス

近年の研究では，プロンプトのフォーマットや ICL において提示するタスク例，例示の順序によってモデルの精度が大きく変動する現象が詳細に報告され, その改善手法もいくつか提案されている. Zhao らは，このような変動の一例として，例示のうち最後に示された回答例や最も多く登場した選択肢に回答が偏るというバイアスを指摘しており，その対策として，出力確率へのキャリブレーション手法を提案している [7].

一方，RAG は，外部知識を活用して ICL における回答精度を向上させる効果的な手法であると考えられるが，RAG が ICL におけるバイアスに与える

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

影響や，既存キャリブレーション手法との相互作用は十分に検証されていない．先ほど述べた Zhao らの研究では，キャリブレーション単独を対象とし，RAG との併用は取り上げていない．そこで本研究では，RAG とキャリブレーションを組み合わせることで生じる効果を総合的に評価し，ICL バイアスの抑制にどの程度寄与するかを明らかにすることを目指す．

3 提案手法

3.1 ICL 依存問題

本研究では，ICL 例示を与えた際，例示の違いによって回答が大きく変動する問題を「ICL 依存問題」と定義する．具体的には，テスト問題の集合 D に含まれる任意の問題 r に対して，n 通りの ICL 例示で推論を行い，$\mathrm{ICL}(r,i)$ を問題 r に対するラベル i の予測回数としたとき，

$$D_{\mathrm{ICL}} = \left\{ r \mid \frac{\max_i \mathrm{ICL}(r,i)}{n} \leq \alpha,\ r \in D \right\} \quad (1)$$

で定義された D_{ICL} を ICL 依存問題の集合とする．

本研究では $n = 10, \alpha = 0.7$ を採用したが，これらの値は調整可能である．

3.2 キャリブレーション手法

3.2.1 Zhao らの手法

Zhao らによるキャリブレーションでは，モデルが各選択肢に割り当てる確率分布を取得した上で，以下のアフィン変換を施すことでバイアスを補正している．

$$\hat{\mathbf{q}} = \mathrm{softmax}\left(W\hat{\mathbf{p}} + \mathbf{b}\right) \quad (2)$$

ここで，$\hat{\mathbf{p}}$ は元の出力確率ベクトル（例：クラスラベルごとの確率）を再正規化したものである．また，W は対角行列，$\mathbf{b}$ はバイアス項を表す．

学習データが十分に得られない状況では，W と $\mathbf{b}$ を直接学習するのが難しいため，`N/A` のような無意味な文字列である「コンテンツフリー入力」を用い，モデルが各ラベルを出力する確率 $\hat{\mathbf{p}}_{\mathrm{cf}}$ を取得し，W と $\mathbf{b}$ を以下のように設定する．

$$W = \mathrm{diag}\left(\hat{\mathbf{p}}_{\mathrm{cf}}\right)^{-1}, \quad \mathbf{b} = \mathbf{0} \quad (3)$$

本研究では，Zhao らの手法を「一括補正」と呼ぶ．すなわち，式 (2) をテストセット内のすべての問題 D に対して適用し，バイアスを補正する．

3.2.2 選択的補正 (提案手法)

本研究では，ICL 依存問題に該当すると推定される問題に対してのみキャリブレーションを適用し，他の問題では補正を行わない「選択的補正」を提案する．具体的には，3.1 節で定義した ICL 依存問題の集合 D_{ICL} に含まれる問題にのみ式 (2) を適用し，その他の問題はモデルの出力確率 $\hat{\mathbf{p}}$ をそのまま使用する．

4 実験

4.1 モデル

本研究では，Llama-3.2-1B[1] と Llama-3.1-8B[2] を用いる．事前実験では，Llama-3.2-1B による回答はプロンプトで指示した出力フォーマットに従わないことが多かった．このため，本研究では回答を直接文字列解析する方法ではなく，次トークン出力確率を比較することで正答を推定している．具体的には，各推論で top-k=10 の候補トークンを取得し，その確率分布から選択肢ごとのスコアを計算して最も確率が高いものを回答と判定する．

4.2 実験設定

医療系の QA タスクとして，MIRAGE ベンチマークのうち MMLU-Med・MedQA-US・PubMedQA*・BioASQ-Y/N の 4 タスクを採用し，それぞれから 100 件ずつをテストセットとして抽出する．MMLU-Med と MedQA-US は 4 択問題，PubMedQA*と BioASQ-Y/N は Yes/No 問題[3]である．残りの問題を ICL 例示用のサンプル源として利用し，異なる例示を組み合わせることで精度のばらつきやバイアスの動向を観測する．ランダム選択による結果への影響を軽減するために，テストセットの作成と評価を 3 回繰り返す．

また本研究では，Zhao らのキャリブレーション実装を参考に，コンテンツフリー入力として `N/A`，`[MASK]`，空文字列の 3 つを用意し，それぞれの出力確率を取得・平均して安定した補正パラメータを推定する．本実装を一括補正および選択的補正の両方で適用し，RAG の有無とも組み合わせて比較する．各タスクには，タスクと無関係な医療文献集合 (Corpus

1) https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct
2) https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
3) PubMedQA*は Maybe も選択肢に含む．

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

A) と該当タスクに有益な情報を含む医療特化文献 (Corpus B) を用意し, BM25 を用いて 1 件のスニペットを抽出してプロンプトに付与することで, RAG が有益な情報をもたらすか否かを検証する. ICL 例示の Shot 数は 1・4・8 の 3 種類を採用し, ランダムに 10 個の例示を作成する. なお, Zero-shot は例示が存在せず, 例示の違いによる回答の変動が確認できないため, 実験では対象外とした. 本研究で使用したプロンプトを付録 B に示す.

4.3 評価指標

テストセット 100 件それぞれの正答率を計測し, 3 回の実験結果から標準偏差を算出して正答率のばらつきを評価する.

また, ICL 依存問題の数をバイアス指標とし, RAG 導入やキャリブレーション適用によってそれらがどの程度減少するか, すなわちバイアスがどれほど抑制されるかを確認する. さらに, 選択的補正により, 全体の正答率や ICL 依存問題数にどのような影響が及ぶかを分析する.

5 実験結果と考察

5.1 RAG の導入による性能比較

表 1 は, RAG を導入しない (No RAG) 場合と, コーパス A または B を利用する場合について, MMLU-Med, MedQA, PubMedQA*, BioASQ-Y/N の 4 タスク・2 モデル (1B,8B)・few-shot 数 (1,4,8) の組み合わせ毎の正答率 (%±標準偏差) を示している.

まず, MMLU-Med と MedQA において, RAG を利用しても正答率の向上が限定的であり, No RAG よりも低下する例が確認できる. たとえば MedQA において, 1B モデルでは RAG(A) が No RAG に比べて+3〜4 ポイント程度改善する条件があるが, 8B モデルでは RAG(B) の導入でも No RAG に比べて+0.7 ポイントの上昇にとどまり, ほかの条件では正答率が低下している. MMLU-Med と MedQA は幅広い医療領域の質問を含むため, 外部知識から適切な情報を取得できないことが多いためと考えられる.

一方, PubMedQA*と BioASQ-Y/N では, 外部知識として PubMed を参照できる RAG(B) の使用にて顕著な向上が確認できる. たとえば, BioASQ-Y/N の 8B モデル (Shots=4,8, RAG(B)) の場合に, 精度が 85%前後に達し, No RAG や RAG(A) に比べて+8〜10 ポイントほど上昇している. PubMedQA*においても, 同様の条件下で 70%台後半の精度が確認でき, いずれも医療論文データベースである PubMed との親和性が高いことが寄与していると考えられる.

5.2 ICL とバイアスの分析

表 1 の括弧内は, 式 (1) にて定義される ICL 依存問題の数を示している. 結果を見ると, Yes/No 形式の PubMedQA*や BioASQ-Y/N では, 1B モデルで依存問題が比較的多いものの, 8B になると依存問題数が大幅に減少し, 正答率が大きく向上している. 一方, MMLU-Med や MedQA のような形式のタスクでは, Shot 数の増加が必ずしも正答率向上に結びつかず, 依存問題数も増加するケースが見られる.

これらの傾向から, Yes/No タスクではモデル規模が拡大するほど少数の例示をより適切に考慮できるようになるため, 依存問題の減少と精度向上が同時に起こりやすいと考えられる. 対照的に, MMLU-Med や MedQA のような形式のタスクでは, ICL 例示の内容や順序がモデルによる誤答と深く関係し, その影響が選択肢の数や配置によるバイアスより大きいと考えられ, このことが Shot 数の増加が正答率の低下を招く要因となっている可能性がある.

5.3 キャリブレーションの効果

まず, キャリブレーションによって No RAG の精度が向上する条件では, RAG(A) や RAG(B) を導入している場合でも同様に精度が上昇する傾向がある. たとえば MMLU-Med の 1B モデル (Shots=4,8) では, No RAG における一括補正が約+1〜2 ポイント精度を改善しているが, RAG(A) や RAG(B) でも+1〜2 ポイント程度向上し, キャリブレーションの汎用的な効果が確認できる.

しかし, RAG(B) のようにタスクに関連する知識を十分に参照できる条件では, キャリブレーションが逆効果となる場合もある. たとえば PubMedQA* の 1B モデル (Shots=4, RAG(B)) では, キャリブレーションなしで 56.0% に達している一方, 一括補正後は 48.7% に低下している. これは, RAG(B) がタスク関連知識を十分に活用しているため, 追加のバイアス補正がノイズとなりうることを示唆している.

一方, 平均精度がわずかに下がっても標準偏差が減少し, 回答の安定性が高まるケースもある. たとえば MedQA の 1B モデル (Shots=1, RAG(B)) では, キャリブレーション前が 35.8 ± 5.5%に対し, 一括補正後

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

表 1 Shot 数別・RAG(スニペット=1) 有無別の正答率 (%±標準偏差) と ICL 依存問題数 (括弧内).
MMLU-Med・MedQA は (A)=StatPearls, (B)=Textbooks, BioASQ-Y/N・PubMedQA*は (A)=StatPearls, (B)=PubMed

		Shots=1			Shots=4			Shots=8		
Model	**Calibration**	**No RAG**	**RAG(A)**	**RAG(B)**	**No RAG**	**RAG(A)**	**RAG(B)**	**No RAG**	**RAG(A)**	**RAG(B)**
MMLU-Med										
1B	No calibration	35.3 ± 5.4 (76)	41.8 ± 6.6 (55)	**43.8 ± 5.8** (51)	43.7 ± 5.0 (41)	44.5 ± 3.7 (36)	43.5 ± 4.3 (39)	42.7 ± 4.5 (38)	42.8 ± 3.7 (37)	41.9 ± 4.5 (39)
	一括補正 [7]	**38.5 ± 5.3**(64)	42.7 ± 5.2(52)	42.3 ± 6.0(57)	**45.4 ± 3.6**(33)	**46.2 ± 3.5**(27)	45.6 ± 4.3(30)	**43.1 ± 3.7**(39)	**44.1 ± 3.8**(35)	**43.0 ± 3.8**(37)
	選択的補正	38.3 ± 4.1(57)	**42.8 ± 4.9**(40)	43.4 ± 4.8(40)	45.0 ± 3.0(21)	46.0 ± 2.8(19)	**45.6 ± 3.3**(18)	43.6 ± 3.3(25)	43.5 ± 3.3(24)	42.7 ± 3.9 (25)
8B	No calibration	41.7 ± 8.1 (63)	44.5 ± 9.3 (53)	43.0 ± 11.6 (58)	26.2 ± 7.3 (69)	27.3 ± 9.3 (57)	31.5 ± 10.5 (59)	30.6 ± 6.3 (49)	28.1 ± 9.7 (50)	34.2 ± 11.1 (50)
	一括補正 [7]	**47.0 ± 7.3**(69)	**50.4 ± 8.1**(59)	**49.6 ± 10.1**(67)	**32.6 ± 8.4**(76)	**33.5 ± 9.9**(72)	**37.6 ± 11.6**(75)	**35.9 ± 7.3**(65)	**32.7 ± 9.8**(64)	**39.6 ± 11.1**(62)
	選択的補正	45.6 ± 7.8(57)	47.9 ± 8.7(46)	47.1 ± 10.2(53)	30.1 ± 7.5(49)	30.2 ± 8.9(49)	34.2 ± 10.5(55)	33.9 ± 6.4(42)	30.0 ± 9.2(41)	36.8 ± 10.7 (43)
MedQA										
1B	No calibration	**32.8 ± 5.8** (76)	**36.2 ± 5.6** (58)	**35.8 ± 5.5** (55)	**35.5 ± 4.7** (46)	**36.8 ± 3.9** (45)	**35.7 ± 3.9** (43)	**35.6 ± 4.0** (36)	**36.6 ± 3.3** (38)	**35.9 ± 3.0** (35)
	一括補正 [7]	32.7 ± 4.2(67)	34.2 ± 3.7(57)	33.2 ± 4.4(54)	32.5 ± 5.0(40)	34.4 ± 5.2(44)	33.2 ± 4.6(41)	33.3 ± 4.6(43)	34.6 ± 4.1(36)	33.6 ± 3.6(35)
	選択的補正	32.7 ± 4.5(56)	35.5 ± 4.1(42)	35.3 ± 4.9(35)	34.2 ± 4.6(28)	36.1 ± 4.5(29)	34.8 ± 4.6(28)	34.5 ± 3.7(30)	36.0 ± 3.4(24)	35.8 ± 2.9(22)
8B	No calibration	47.4 ± 9.2 (53)	46.0 ± 6.5 (41)	48.1 ± 6.0 (47)	32.2 ± 9.7 (64)	27.0 ± 6.4 (55)	30.1 ± 8.3 (56)	33.2 ± 7.9 (52)	25.9 ± 6.2 (46)	28.6 ± 7.0 (50)
	一括補正 [7]	**48.4 ± 8.6**(61)	**47.4 ± 6.3**(58)	**48.1 ± 5.8**(57)	**34.3 ± 8.3**(80)	**29.9 ± 6.3**(81)	**32.4 ± 6.5**(76)	**34.5 ± 7.1**(63)	**28.5 ± 6.1**(62)	**30.5 ± 5.3**(57)
	選択的補正	48.0 ± 8.9(49)	45.8 ± 6.9(39)	47.9 ± 5.7(43)	33.6 ± 8.9(59)	27.5 ± 7.0(51)	30.8 ± 7.6(50)	33.7 ± 7.6(43)	26.8 ± 6.2(37)	29.4 ± 6.2(38)
PubMedQA*										
1B	No calibration	44.3 ± 10.9 (98)	47.2 ± 9.6 (87)	50.1 ± 8.1 (80)	46.6 ± 7.9 (74)	**48.5 ± 7.6** (57)	**56.0 ± 5.5** (55)	**49.0 ± 8.3** (56)	**48.5 ± 7.0** (47)	**53.5 ± 6.3** (53)
	一括補正 [7]	46.8 ± 7.3(77)	47.0 ± 7.2(73)	**55.7 ± 7.0**(57)	**47.7 ± 8.2**(66)	45.4 ± 6.8(60)	48.7 ± 6.7(58)	45.0 ± 8.0(48)	41.5 ± 5.6(51)	43.7 ± 7.4(61)
	選択的補正	**46.8 ± 7.3**(76)	**47.4 ± 6.9**(66)	55.4 ± 7.4(52)	46.8 ± 8.5(61)	46.5 ± 6.7(42)	51.7 ± 5.8(43)	46.0 ± 7.1(38)	44.9 ± 5.7(34)	48.3 ± 6.1(41)
8B	No calibration	**53.0 ± 7.0** (43)	**53.8 ± 5.8** (36)	**76.1 ± 4.5** (11)	**55.5 ± 4.1** (22)	**55.6 ± 3.8** (24)	**77.5 ± 4.0** (8)	**55.2 ± 4.3** (22)	**55.6 ± 4.2** (23)	**78.2 ± 3.2**(7)
	一括補正 [7]	49.6 ± 5.7(39)	48.9 ± 5.7(50)	74.3 ± 4.8(12)	51.8 ± 5.4(29)	52.2 ± 3.5(22)	76.3 ± 4.4(7)	49.0 ± 5.6(26)	50.2 ± 4.7(21)	77.0 ± 3.6(5)
	選択的補正	52.0 ± 4.1(25)	52.6 ± 3.4(27)	75.5 ± 4.2(8)	55.1 ± 4.0(8)	54.4 ± 3.4(10)	76.9 ± 4.3(4)	54.2 ± 3.2(5)	54.9 ± 4.1(4)	77.9 ± 3.3(2)
BioASQ-Y/N										
1B	No calibration	52.4 ± 14.1 (66)	57.7 ± 10.8 (53)	59.7 ± 10.1 (56)	**54.3 ± 10.9** (77)	**57.2 ± 8.9** (66)	**63.1 ± 7.9** (46)	**54.3 ± 10.6** (84)	**53.4 ± 10.4** (66)	**57.0 ± 11.4** (53)
	一括補正 [7]	54.0 ± 14.9(88)	53.0 ± 13.7(84)	57.9 ± 14.2(72)	50.2 ± 8.4(60)	52.5 ± 8.1(43)	60.0 ± 8.3(38)	52.3 ± 8.8(54)	48.8 ± 7.7(32)	52.9 ± 7.7(37)
	選択的補正	**56.3 ± 13.2**(58)	**58.1 ± 10.0**(47)	**62.1 ± 9.5**(41)	51.4 ± 7.8(50)	54.2 ± 7.0(32)	62.6 ± 7.3(23)	53.0 ± 8.7(51)	49.8 ± 7.8(26)	54.1 ± 7.6(29)
8B	No calibration	75.3 ± 4.0 (19)	76.3 ± 3.2 (19)	**85.5 ± 2.9** (6)	78.8 ± 4.0 (13)	77.7 ± 3.5 (12)	**85.5 ± 2.7** (4)	78.6 ± 3.3 (10)	78.7 ± 3.6 (12)	**85.9 ± 3.0** (5)
	一括補正 [7]	72.2 ± 6.5(25)	72.0 ± 7.1(34)	83.5 ± 4.2(10)	76.2 ± 3.6(13)	74.8 ± 3.3(16)	84.5 ± 3.4(5)	75.6 ± 5.8(15)	74.1 ± 4.1(17)	84.8 ± 2.9(4)
	選択的補正	**75.9 ± 3.2**(11)	**78.0 ± 4.9**(13)	85.1 ± 2.9(3)	**80.5 ± 2.1**(1)	**79.0 ± 3.0**(2)	84.5 ± 2.2(1)	**80.1 ± 2.4**(1)	**79.5 ± 2.3**(2)	85.3 ± 2.5(0)

は 33.2 ± 4.4%, 選択補正後は 35.3 ± 4.9%と, 平均正答率はわずかに低下しているが標準偏差は小さくなっており, ばらつきを抑える効果が認められる.

次に，ICL 依存問題数（表 1 の括弧内）に着目すると，一括補正が精度を向上させながらも ICL 依存問題数を増やしてしまう例があるのに対し，選択的補正では平均精度の改善幅が小さいまたは改善がみられない場合でも ICL 依存問題数を大きく減少させる傾向が確認される. たとえば，MedQA の 1B モデルにおいて，No calibration の正答率に対し，選択的補正後は平均正答率がほぼ横ばいまたは微減であるにもかかわらず，ICL 依存問題数は約 10〜20 件減少している. これは，選択的補正が ICL 例示によるバイアスの影響を集中的かつ効果的に抑制しているためだと考えられる.

さらに, ICL 依存問題数が特に多く, No RAG や有益な情報が得にくい RAG(A) のような条件では, 本研究にて提案する選択的補正によって従来の一括補正を上回る結果が得られることがある. たとえば PubMedQA*の 1B(Shots=1, RAG(A)) では, No RAG が 47.2%に対し, 一括補正後は 47.0%, 選択的補正後は 47.4%と, わずかながら提案手法による改善が確認できる. また BioASQ-Y/N でも, 1B（Shots=1）において, 従来手法の 54.0%に対して提案手法が 56.3%と 2 ポイント以上向上している. さらに 8B（Shots=1）では, ベースラインの 75.3%に対し, 従来手法では 72.2%と精度が下がるが, 提案手法では 75.9%と改善が見られる. Shots=4 や 8 の条件でも同様の傾向が確認できる. これらは特定条件下において曖昧性の高い問題に絞って補正を施すメリットを示唆する.

以上の結果から，多くの場合キャリブレーションはプラスに作用するものの，外部知識を活用できる RAG(B) のように既に高精度を実現している条件では，必ずしも有効とは限らないことがわかる. また，少数の特定問題だけを補正する選択的補正は，一括補正のような性能低下リスクを抑えつつ，曖昧性の高い問題での精度をより改善するだけでなく，ICL 依存問題数の大幅な削減にも寄与する. これらの知見は，キャリブレーション手法をさらに洗練し，RAG の活用条件やタスク特性に応じて適切に使い分けることで，LLM のバイアス抑制と安定的な性能向上を両立する余地があることを示している.

6 おわりに

本研究では，ICL 例示によるバイアスを抑制するためのキャリブレーション手法を，外部知識参照を行う RAG と組み合わせた際の効果を医療系タスクを用いて検証し，ICL 例示の違いで回答が大きく変動する「ICL 依存問題」に焦点を当てた. 特に，それらにのみバイアス補正を施す「選択的補正」を提案し，従来の一括補正よりも ICL 依存問題数を大きく削減できるだけでなく，特定条件下では, 一括補正を上回る結果が得られることを示した.

今後は, 本論文にて用いた手法以外のキャリブレーション手法の検討や, ICL 例示自体の設計やタスク形式の特性を詳細に分析したいと考える.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

参考文献

[1] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. A survey on in-context learning. In **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 1107–1128, 2024.

[2] Yu Fei, Yifan Hou, Zeming Chen, and Antoine Bosselut. Mitigating label biases for in-context learning. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 14014–14031, 2023.

[3] Zhixiong Han, Yaru Hao, Li Dong, Yutao Sun, and Furu Wei. Prototypical calibration for few-shot learning of language models. In **The Eleventh International Conference on Learning Representations**, 2023.

[4] Han Zhou, Xingchen Wan, Lev Proleev, Diana Mincu, Jilin Chen, Katherine A Heller, and Subhrajit Roy. Batch calibration: Rethinking calibration for in-context learning and prompt engineering. In **The Twelfth International Conference on Learning Representations**, 2024.

[5] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In **Proceedings of the 34th International Conference on Neural Information Processing Systems**, NIPS '20, 2020.

[6] Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. Benchmarking retrieval-augmented generation for medicine. In **Findings of the Association for Computational Linguistics: ACL 2024**, pp. 6233–6251, 2024.

[7] Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models, 2021.

[8] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. **Applied Sciences**, Vol. 11, No. 14, 2021.

[9] Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. **Foundations and Trends in Information Retrieval**, Vol. 3, pp. 333–389, 01 2009.

[10] Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. Pyserini: A python toolkit for reproducible information retrieval research with sparse and dense representations. In **Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval**, SIGIR '21, p. 2356–2362, 2021.

[11] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In **International Conference on Learning Representations**, 2021.

[12] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. PubMedQA: A dataset for biomedical research question answering. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 2567–2577, 2019.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

A　付録 A: MedRAG と MIRAGE の詳細

A.1　コーパス

PubMed　PubMed[4]は最も広く利用されている生物医学文献リソースであり，3,600 万を超える論文を含むと報告されている．本研究で参照している MEDRAG では，タイトルと抄録が有効な約 2,390 万件の PubMed 記事を用いている．

StatPearls　StatPearls[5]は，UpToDate[6]に類似した，臨床現場で参照される支援ツールである．MEDRAG では，NCBI Bookshelf[7]から取得可能な 9,330 の StatPearls 記事をパラグラフ単位でスニペット化し，関連する見出し情報をタイトルとして付与している．

Textbooks　Textbooks[8][8] は，USMLE(米国医師免許試験) の受験者に広く使われている 18 種類の医学教科書を集約したコレクションである．MEDRAG では，各教科書を 1,000 文字以下のチャンクに分割してスニペット化しており，LangChain[9]の RecursiveCharacterTextSplitter を用いて前処理を行っている．

A.2　リトリーバー

BM25　BM25[9] は典型的な Bag-of-Words ベースのリトリーバーであり，TF-IDF に基づく語彙マッチングを行う．MEDRAG では，Pyserini[10][10] を用いて BM25 のインデックスを作成し，すべてのスニペットを検索対象としている．

A.3　各タスクの概要

- **MMLU-Med**：MMLU (Massive Multitask Language Understanding)[11] [11] からバイオ医学領域に関連する 6 つのタスクを抜粋したもの．計 1089 問．
- **MedQA-US**：米国医師免許試験 (USMLE) の英語版試験を元に作成された 4 択問題．計 1273 問．
- **PubMedQA***：PubMedQA[12] [12] から文脈情報を除去した修正版．科学文献を基に Yes/No/Maybe 形式で回答する問題を含む．計 500 問．
- **BioASQ-Y/N**：BioASQ[13]のうち Yes/No 形式の問題を抽出したサブセット．最新版 (2019-2023)5 年間のタスクから計 618 問．

B　付録 B: 使用したプロンプト例 (One-shot の場合)

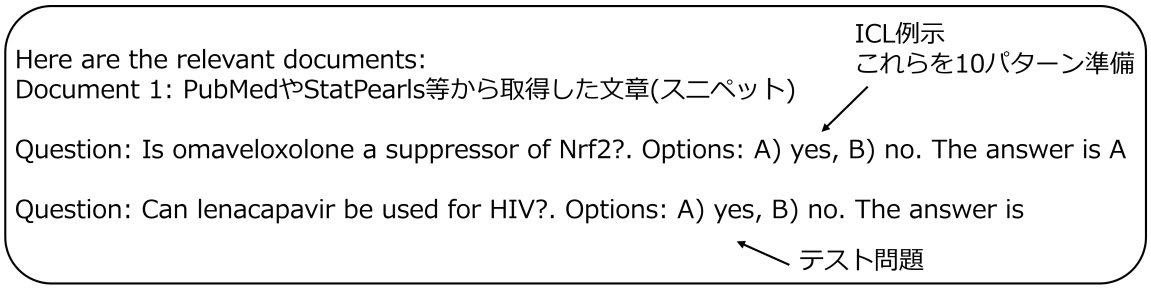

図 1　One-shot プロンプトの例

4）https://pubmed.ncbi.nlm.nih.gov/
5）https://www.statpearls.com/
6）https://www.uptodate.com/
7）https://www.ncbi.nlm.nih.gov/books/NBK430685/
8）https://github.com/jind11/MedQA
9）https://www.langchain.com/
10）https://github.com/castorini/pyserini
11）https://github.com/hendrycks/test
12）https://pubmedqa.github.io/
13）http://bioasq.org/

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).