大規模言語モデルが有する事実知識の多角的な評価

趙 信[1,a)]　吉永 直樹[2,b)]

概要：大規模言語モデル（LLM）は膨大なテキストから学習され，実世界の知識を内包する一方で，幻覚など知識の運用に問題があることが知られている．そのため，LLM の知識評価を行うことが重要であるが，LAMA probe など既存の言語モデルの知識評価手法は，マスク言語モデルを想定しており，因果言語モデルである LLM の知識評価にそのままでは適用できない．本論文では，多様なプロンプトを用いた知識評価フレームワーク BELIEF を拡張し，in-context learning（ICL）で LLM が有する知識を多角的に評価した．具体的には，マスクを予測するタスクにおいて zero-shot および few-shot の ICL で精度・一貫性・信頼性の観点で LLM の知識評価を行う．実験では，Llama-3 70B を含む複数の LLM の知識評価を行い，BERT と比べてより多くの知識を内包することを示すとともに，事前学習における事実知識の獲得に重要な要素を明らかにする．

1. はじめに

大規模テキストから学習した事前学習済み言語モデル（PLM）は，学習データに含まれる事実知識を暗黙のうちに獲得，保持することから，質問応答（QA）[1]，研究支援 [2]，知識推論 [3] など，事実知識を必要とするタスクで成功を収めている．PLM にどれだけの知識が含まれているか，また知識が事前学習を通してどのように獲得されるかを理解することは，知識学習の観点で PLM を改善するために重要である．

言語モデルの保持する関係知識を評価する手法としては，関係知識を表現する穴埋め文（プロンプト，例: “John Lennon was born in [MASK]”）の空欄 [MASK] のエンティティを言語モデルで予測する LAMA probe [4] が用いられる．[MASK] トークンの予測精度を測定することにより，PLM を持つ知識を定量的に評価できる．一方で，単一のプロンプトのみで関係知識の有無を評価すると，その結果はプロンプトの言語表現の些細な違いに強い影響を受けてしまう [5], [6], [7]．そこで，正確に PLM が有する知識の量を測定するため，同じ事実に対して複数のプロンプトを使用する評価手法が提案されている [8], [9], [10]．さらに我々は，より多様な知識プロービングデータセット MyriadLAMA と知識評価フレームワーク BELIEF[11] を提案した．BELIEF は，精度，一貫性，信頼性を含む多様な観点から PLM の知識評価を行い，PLM の知識理解能力をより包括的に分析することができる．

穴埋め文による評価は，プロンプト内でマスクされたトークンを予測するもので，マスク言語モデル（MLM）が内包する事実知識の有無を調べるために設計されている．しかし，GPT-4 [12] や Llama2 [13] のような大規模言語モデル（LLM）は因果言語モデル（CLM）であり，単純に穴埋め文のみを入力して [MASK] を予測することはできない．CLM は所与のトークン列の後続トークンを予測するタスクで学習されるため，[MASK] 後方の言語情報を利用できないためである（例:`[MASK]` は日本の首都である）．既存研究 [14], [15] では，CLM の知識を評価するために QA スタイルのプロンプトを用いているが，これはプロンプトの言語表現の多様性に乏しく，言語の多様性が知識評価に与える影響を測定するのが難しい．また，プロンプトセットの違いから，MLM と CLM を正確かつ公平に比較することも困難になっている．

本研究では，in-context learning（ICL）を用いて，BELIEF を CLM に適用できるようにした BELIEF-ICL（§3）を提案する．具体的に，複数のタスク指示と入出力事例の種類を組み合わせて，知識評価のための ICL 設定を検討する．また，BELIEF の評価指標（精度，その揺らぎ，一貫性，信頼性）を CLM に適応した計算方法を提案する．

実験では，Llama2 [13]，Llama3 [16]，および Phi3 [17] など，複数の LLM を対象として，異なる ICL 設定で知識評価を行った（§4）．評価結果に基づいて，LLM が事実知識をどのように学習し，モデル内で表現するかについての理解を深めるための詳細な分析を行った (§5,§6,§7)．

[1] 東京大学大学院 情報理工学系研究科
[2] 東京大学 生産技術研究所
[a)] xzhao@tkl.iis.u-tokyo.ac.jp
[b)] ynaga@iis.u-tokyo.ac.jp

2. 予備知識：BELIEF

本節では，多様なプロンプトを用いて言語モデルが持つ知識を多角的に評価するフレームワークである BELIEF [11] を紹介する．BELIEF は，MyriadLAMA [11] など各事実に対して複数のプロンプトを提供する知識プロンプトデータセットを用いて，精度とその揺らぎ，一貫性，信頼性の観点から PLM の事実知識理解能力の評価を行う．以下では，本研究でも用いる MyriadLAMA データセットを紹介したのち，本研究で拡張する BELIEF について説明する．

2.1 MyriadLAMA

MyriadLAMA は，Wikipedia から抽出された事物間の関係知識に対応する複数のプロンプトで構成され，各関係は**知識トリプル**〈主体, 関係, 対象〉(例: 〈東京, 首都, 日本〉) の形で表現される．MyriadLAMA は，各関係に対して様々なテンプレート表現（以下，**関係テンプレート**，例：[X] は [Y] の首都である）が提供されている．MyriadLAMA を使用した事実知識評価の基本的な手順は，まず対象の知識トリプルで対応する関係テンプレートを埋め，[Y] を [MASK] トークンに置き換えて**マスクプロンプト** (以下，**プロンプト**) を生成する．次に，評価対象の言語モデルにプロンプトを入力して対象を正しく推測できるかを確認する．MyriadLAMA では，各事実に対応するプロンプトを多様化するために，主体・対象の言語表現を拡張し，各関係に対して多数の関係テンプレートを作成している．具体的には，主体・対象の言語表現はエイリアスを参照することで拡張している．関係テンプレートは、各関係に対して 5 つのテンプレートを手作業で作成し，それぞれを GPT-4 を使用して 19 回言い換えて，結果として関係ごとに 100 のテンプレートを生成することで多様化している．MyriadLAMA は 41 の関係に対する 4,100 の関係テンプレートをもとに，これと多様なエンティティの言語表現を組み合わせることで，24,643 の事実に対して 6,492,800 のプロンプトを提供する．

2.2 BELIEF における多様な評価指標

BELIEF は，MLM が知識理解における精度，一貫性，信頼性を，以下の評価指標により評価する．

精度とその揺らぎ．BELIEF は，正確な精度評価を行うため，複数のプロンプトセット（N 個）に基づいて複数の精度値を算出し，分布の観点からモデル性能（平均精度など）を評価する．各セットは，各事実に一つだけのプロンプトをサンプリングする形で作られ，それぞれのセットには事実の数（MyriadLAMA であれば 24,643）と同じ数のプロンプトが含まれる．これら N 個のプロンプトセットから算出された N 個の精度値は，平均精度と精度の揺らぎを計算するために用いられ，精度の揺らぎは範囲と標準偏差で測定される．

一貫性．BELIEF は，各事実に対し，多様なプロンプトを PLM に入力して予測されたトークンの一致率によって言語モデルが有する知識の一貫性を評価する．一貫性と先述の精度の揺らぎは，PLM による知識理解の頑健性を評価するための指標である．

過信度（信頼性）．過信度は，PLM の予測をどれ程度，信頼できるかを評価する指標である．BELIEF では，PLM が予測した top-1 トークンの確率（以降，確信度）とその精度の差として定義される**過信度**[*1]を信頼性の評価に採用する．具体的には，モデルが [MASK] トークンに対して予測した “対象” の出力確率の最大値と実際の精度との乖離を測定し，言語モデルが自身の予測をどれ程過信しているかを評価する．過信度の値がゼロに近ければ近いほど，つまり言語モデルの確信度と精度が近いほど，言語モデルは自身の予測の信頼性を過大評価しているとみなせ，予測の確信度を信頼することができる．また，過信度の値が負になると，モデルは自身の予測の信頼性を過小評価しているとみなせる．

3. BELIEF-ICL

本節では，CLM として学習された LLM に適用できるように，in-context learning (ICL) を用いて BELIEF を拡張した BELIEF-ICL を提案する．

3.1 事実知識評価ための In-context Learning

LLM は，In-context learning により複雑なタスクを推論のみで解くことができる [18]．事実知識評価ための ICL を設計する上では，MyriadLama が提供する評価対象の**プロンプト**に加えて，**タスク指示**，**入出力事例**を考える必要がある．以下で，それぞれについて具体的に述べる．

3.2 タスク指示

本研究では，二種類の ICL タスクを提案する．

QA-style ICL. MyriadLAMA で利用可能な QA スタイルの関係テンプレート[*2]を使用すれば，質問-回答形式のプロンプトを作成できる．各 QA スタイルのプロンプトは，主語と関係が質問を形成し，目的語を回答する形式（例：“`Who developed [X]? [Y].`”）に従う．次に，InstructGPT [19] で使用されている形式に従い，ランダムな QA ペアからな

[*1] 過信度の計算は具体的に，プロンプトを確信度降順でソートし，これを複数のビンに分割しする．次に，各ビンに対して，精度の平均および確信度の平均をそれぞれ求めて，これらの差分を全てのビンに渡って平均することで言語モデルが “対象” を予測する際の過信度を評価する．

[*2] MyriadLAMA は，各関係に対して 20 個の QA スタイルのテンプレートを提供しており，これらは予測対象のマスクトークンがテンプレート末尾に来ることから CLM の評価に直接使用することができる．

る few-shot プロンプトを使用する．MyriadLAMA では，全ての関係知識の対象が一トークンで答えられるように設計されていることから，“Answer each question in one word.” という指示を先頭に付け加える．

マスク予測（MP-style）ICL. 次に，全てのテンプレートを活用可能なマスク予測（MP-style）ICL 設定を説明する．タスク指示は，‘‘Predict the [MASK] in each sentence in one word.’’ という形で表されている．few-shot 例や質問のプロンプトでは，BELIEF と同じ規則に従い，関係テンプレート内の対象代用語 [Y] を ‘‘[MASK]’’ に置換する．

3.3 入出力事例

前節で説明した指示のうち，特に MP-style ICL については，CLM と MLM というタスクの齟齬から，言語モデルが有する知識が過小評価される可能性がある．この点は少数の入出力事例を与えることで緩和できるため，以下の 4 種類の入出力事例を比較する．

zero-shot: 入出力事例を提供しない．

X-random: 全ての関係から X 個の事実をサンプリングし，入出力事例とする．

X-relation: 同じ関係から X 個の事実をサンプリングするが，テンプレートはその関係が持つ 100 のテンプレートからランダムサンプリングする．

X-template: 同じ関係と同じテンプレートから X 個の事実をサンプリングし，few-shot 学習例とする．

Few-shot の設定では，評価対象の関係知識以外の知識からサンプルされる．例えば，MP-style 1-relation のプロンプトは以下の通りである．

```
Predict the [MASK] in each sentence in one word.
Q: [MASK] consists of pharmacy.
A: biology.
Q: [MASK] consists of environmental factors.
A:
```

その他のプロンプトの例は §A.1 を参照されたい．

3.4 評価指標と評価方法

MLM では，単一のマストークンをプロンプトとして設定することで，回答候補を事前に定義することが可能である．しかし，CLM はトークン数に制限なく回答を生成するため，正解とモデルの生成文字列間の照合を行う際に問題が生じる．具体的には，正解の対象の言語表現に関する冠詞の有無や単数・複数形などの表記揺れが生じたり，[MASK] 以外に対応する表現以外のテキスト（例えばプロンプトの一部）を追加で生成する場合がある．例えば，“John Lennon can play [MASK]” のプロンプトに対して，モデルが “guitars” と “a guitar” を生成する場合，両方とも正解とみなせる． そこで本研究は，この問題を解決するために BELIEF の定義をもとに，CLM に適用できる精度，精度の揺らぎ，一貫性，過信度の評価手法を提案する．

以上の指標の計算において必要となる基本操作は，二つの文字列（精度を計算する際は生成テキストと正解となる対象の言語表現，一貫性や過信度を計算する際は二つの生成テキスト）の間の照合である．ここでは，まずその方法を定義する．比較対象となる文字列をトークン化し，見出し語化することで正規化する．例えば,“a guitar” と “guitars” はそれぞれ “a guitar” と “guitar” に正規化される．次に，一つの正規化された文字列がもう一つの文字列に含まれる場合 (部分マッチング)，二つの文字列はマッチングしていると見なす．上の例の場合は，“guitar” が “a guitar” に含まれているため，二つ文字列はマッチングするとみなす．

精度とその揺らぎ． 精度は，モデルから貪欲法によるデコード戦略を用いて生成された文字列と正解の文字列との照合に基づいて計算される．精度計算において注意すべき点は，マッチングの判断が一方向，すなわち正解が生成文字列に含まれるかどうかのみを考慮する点である．つまり，対象の言語表現が正規化された生成文字列に含まれる場合，正解が出力されたとみなす．単方向のマッチングを採用する理由は，モデルが無関係な単語を生成することで誤った判断が行われる可能性があるためである．例えば，首都に関する事実知識において，正解が “Tokyo city” である場合，生成された文字列が “city” であると，逆方向にはマッチングしているが，正しい答えとするべきではない．

具体的な計算について，BELIEF-ICL は，BELIEF と同じ手法を用い，最初に N 個のプロンプトセットを収集し，各プロンプトセット内の正解数に基づき，精度（Acc@1）を計算する．その後，N 個の精度値を用いて平均精度と精度の揺らぎを求める．BELIEF の定義に従い，Acc@1 の揺らぎを範囲と標準偏差で測定する．

一貫性． 二つのプロンプトの生成文字列の一貫性を評価する際は，両方向からのマッチング関係を検証する．一貫性の計算は BELIEF で提供されている定義に従う．

過信度（信頼性）． 過信度の計算には，モデル生成の確信度（確率）を取得する必要がある．BELIEF では，単一の出力トークンを仮定し，その確率を確信度として使用するが，CLM では複数トークンを生成する場合があり，この計算方法がそのまま適用できない．したがって，その出力の確信度を計算するためには，その対象の言語表現と一致できる文字列を列挙し，それらの確率を計算して合計する必要があるが，これは計算上現実的ではない．この問題を解決するために，本研究では，まず，multinomial sampling[*3] を

[*3] Multinomial sampling は，モデルによる全語彙に対する確率分布に基づいて次のトークンをサンプリングするデコード戦略である．

LLMs	モデルサイズ	事前学習データ サイズ	リソース
Llama2-7B	7B	2.0T	公開されているオンラインデータの集合
Llama2-13B	13B	2.0T	（個人情報を多く含むサイトは除外.
Llama2-70B	70B	2.0T	事実知識の知識源は upsampling）
Llama3-8B	8B	15T+	公開されているオンラインデータの集合
Llama3-70B	70B	15T+	（詳細不明，コードは Llama2 の 4 倍）
Phi3-mini	3.8B	3.3T	教育用データやコードを含む高品質文書，教科書的な生成テキスト，高品質チャット

表 1 本研究で扱う LLM の事前学習情報.

用いて，各プロンプトに対して 100 回の文字列生成を行う．次に，プロンプトが貪欲デコード戦略によって生成した文字列と，M 回のサンプリング（以降の実験では $M = 100$）によって得られた回答との間でマッチング率を測定する．このマッチング率は，プロンプトの生成に対する確信度として使用される．この計算方法を MLM に適用する場合，予測トークンの確率分布を近似するものとなることに注意されたい（言い換えると，MLM に対して BELIEF を適用して計算した確信度と，BELIEF-ICL における確信度は比較可能である）.

4. 実験設定

本研究では，3 種類の LLM の異なるパラメタ数のモデル（計 6 つ）に BELIEF-ICL を適用し，知識評価を行う．具体的に，Llama2-7B,[*4] Llama2-13B,[*5] Llama2-70B,[*6] Llama3-8B,[*7] Llama3-70B,[*8] Phi3-mini[*9]である．

これらのモデルは全てデコーダベースの Transformer アーキテクチャーを使用している．本研究で評価する LLM の事前学習情報は表 1 で示す．同じ種類でサイズが異なるモデルは，パラメータ数のみに異なる．. Llama3 は，公開されているオンラインデータで，Llama2 の 7 倍に相当する 15 兆トークン以上から学習されている．Phi3-mini は，パラメータが 3.8B と比較的小さく，事前学習データセットは 3.3 兆トークンと少なめである．ただし，このデータセットは高品質で教科書的な素材で構成されている．その他トークナイザや活性化関数などの他の要素にも若干の違いはあるが，本研究は事前学習コーパスとパラメータサイズの違いに焦点を当てて分析を行う.

我々は評価設定について，Llama2-7B, Llama3-8B, Phi3-mini を含む 8B 以下の LLM で，MyriadLAMA にある全ての関係テンプレートを用いて評価を行った．13B 以上のパラメタ数の LLM（LLama2-13B，LLama2-70B，LLama3-70B）に対しては，計算コストのため，各関係に手作業で作成した五個の関係テンプレートのみで評価を行った．なお，

[*4] https://huggingface.co/meta-llama/Llama-2-7B
[*5] https://huggingface.co/meta-llama/Llama-2-13B
[*6] https://huggingface.co/meta-llama/Llama-2-70B
[*7] https://huggingface.co/meta-llama/Meta-Llama-3-8B
[*8] https://huggingface.co/meta-llama/Meta-Llama-3-70B
[*9] https://huggingface.co/microsoft/Phi-3-mini-4k-instruct

ICL 設定	平均精度 (Acc@1) ↑		一語生成率 ↑	
	QA	MP	QA	MP
zero-shot	**.5066**	.4534	**.5285**	.4802
4-random	.5429	**.5591**	.7996	**.8058**
4-relation	.6582	**.6649**	.9187	**.9246**
4-template	.6687	**.6765**	.9216	**.9266**

表 2 Llama2-7B における指示忠実性の分析.

MyriadLAMA が含む QA スタイルのテンプレートの数は各関係に対して 20 であり，QA-style ICL 設定の実験に利用できるプロンプトの数は MP-style ICL 設定の 5 分の 1 に相当する．過信度を計算する際，各プロンプトから 100 回文字列をサンプリングすることはかなりの計算コストがかかるため，本研究では MyriadLAMA が有する 6,492,800 のプロンプトから 10,000 プロンプトをランダムにサンプリングして計算に利用した．最後に，精度の揺らぎを正確に捉えるために BELIEF [11] と同様に $N = 50,000$ と設定する．また，公平な比較を保証するために異なる LLM のプロンプトサンプリングでは，一貫したシードを採用した.

以下の節では，本節で説明した設定に沿って，BELIEF-ICL の評価結果を用いて様々な LLM を分析し，LLM の事実知識理解能力についての理解を深める.

5. ICL 設定が知識評価に与える影響

本設では，Llama2-7B を用いて，BELIEF-ICL を LLM の評価に用いる際の ICL 設定の影響を確認し，エンコーダベースの PLM である BERT との知識評価の比較を行う.

5.1 LLM は ICL の指示に忠実か？

我々はまず，LLM が提案された ICL 設定の指示に従うかどうかを，生成された事実知識の精度と，指示通り 1 単語が生成されたプロンプトの割合（以下，一語生成率），という 2 つの観点から評価する．ここでは，QA-style ICL 設定と MP-style ICL を公平に比較するため，Llama2-7B 上で，各関係に対して 20 個の QA 型テンプレートを用いて，両設定で共通のテンプレートを評価に用いた評価を行った.

表 2 に，Llama2-7B の ICL の指示への忠実度を，平均精度 (Acc@1) と一語生成率を用いて個別に評価した結果を示す．zero-shot 設定の下では，QA-style の指示が知識の精度と一語生成率の両方でより優れた性能を持つことが明らかになった．これは，QA-style の指示の方が，CLM が解くタスクと親和性が高いことに起因すると考えられる．しかし，この優劣は few-shot 設定では逆転しており，入出力事例により，MP-style でも QA-style と遜色なく指示に忠実な出力が可能なことが明らかとなった．Few-shot 設定では，一語生成率が大きく改善しており，評価対象のプロ

	PLMs	平均精度 (Acc@1) ↑	精度の揺らぎ ↓ 範囲	精度の揺らぎ ↓ 標準偏差	一貫性 ↑	過信度
BERT	$\text{BERT}_{\text{base}}$ *11	.1095	.1534	.0217	.1682	.2154
	$\text{BERT}_{\text{large}}$ *12	.1102	.1574	.0220	.1713	.2052
	BERT_{wwm} *13	.1364	.1517	.0208	.1524	.1000
Llama2-7B	zero-shot	.3385	.2602	.0299	.1269	**-.0713**
	4-random	.4816	.2250	.0270	.2312	-.0894
	4-relation	.6286	.1221	.0150	.3753	-.1335
	4-template	**.6616**	**.0294**	**.0036**	**.4163**	-.0933

表 3 BERT と Llama2-7B における評価結果.

ンプトに近い入出力事例*10で，モデルの指示への忠実性が著しく向上することが分かった．この点を踏まえて，以降の実験では MP-style のプロンプトを用いて LLM の知識評価を行う．

5.2 ICL は LLM の知識検索性能に影響を与えるか？

本節では，BELIEF-ICL において，MP-style ICL プロンプトで入出力事例として 4 事例（$X = 4$）を用いて Llama2-7B の知識評価を行い，BELIEF [11] で報告されている 3 つの BERT ベースの PLM の評価結果と比較する．

表 3 に示すように，zero-shot と few-shot ICL の Acc@1 の差を見ると，少数の入出力事例の追加により LLM の事実知識の検索能力が大幅に改善することを確認できる．また，入出力事例の選択方法の影響も大きい．表 3 の 3 つの few-shot ICL 設定を比較すると，より対象プロンプトとの関連性が高い入出力事例を使用することで，精度（Acc@1）と頑健性（精度の揺れと一貫性）が一貫して向上することが観察された．

さらに，Llama2-7B は，BERT より優れた知識検索能力を示した．表 3 に示されているように，Llama2-7B は zero-shot ICL でも 3 つの BERT モデルを大幅に上回る予測精度と絶対値の小さい過信度が得られている．また，4-relation や 4-template の few-shot 設定で，Llama2-7B は精度の揺らぎや一貫性の観点で BERT モデルより優れた性能を示している．図 1, 2 に，過信度の計算で用いた確信度でグループ化したプロンプトに対する精度を示す．Llama2-7B が BERT モデルと比較して，モデルの確信度の値に寄らず，0 に近い過信度（=確信度-Acc@1）を示すことが分かる．

6. LLM に共通する知識評価結果の分析

本節では，多様な LLM を評価対象に含めて §5.2 で示された LLM の性能について，精度，一貫性，信頼性の 3 つの側面から深く分析する．本節では特に，PLM の共通点に焦点を当てる．モデル間の違い，例えば，CLM と MLM の違いや異なる LLM 間の違いについては，§7 で議論する．

*10 例えば，X-random は他の関係を事例として引用することがある一方，X-relation では同じ関係の知識だけを事例にしており，プロンプトで回答する知識に近い．さらに，X-template は同じ関係の同じ関係テンプレートに基づく事例を用いており，X-relation よりもさらに評価プロンプトに近い事例を与えている．

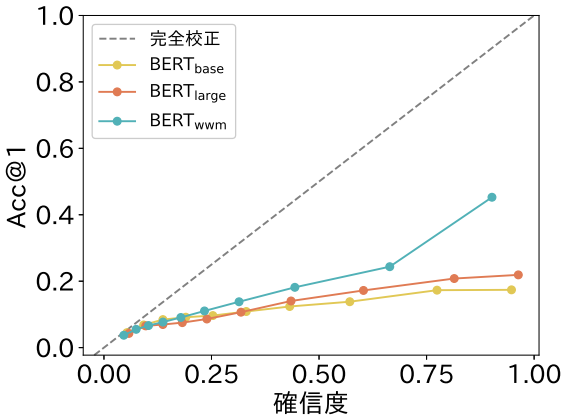

図 1 3 つの BERT モデルにおける精度と確信度のアライメント.

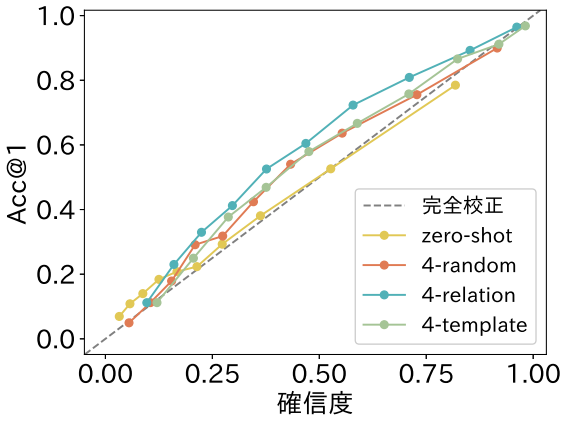

図 2 Llama2-7B の四種類の ICL 設定における精度と確信度のアライメント.

6.1 プロービングによる知識検索の上限

表 3 の一貫性の評価から，LLM では用いるプロンプトによって引き出せる知識の量に大きなばらつきがあることが示唆される．そこで，本節ではまず多様なプロンプトを用いることで，モデルが示す知識検索能力がどのように変化するかを調査する．

まず，各モデルについて，各関係テンプレートが正確に捉えた事実を個別に収集し，その合計数を関係知識数と関係テンプレート数の積で割った平均値を計算する．次に，各関係ごとに精度が最大となる関係テンプレートを一つ選んで知識評価に用いた時の精度を関係ごとに単一テンプレートを選ぶ場合を知識検索性能の最大値として計算する．最後に，LLM に含まれる事実の上限を測定するために，関係知識ごとに正解が得られる関係テンプレートを選んだ場合の精度を上限として計算する．以上の zero-shot ICL 設定での LLM の評価結果を，BERT モデルの測定結果と共

PLMs	平均値	最大値	上限
$BERT_{base}$	.1441	.4248	.6209
$BERT_{large}$	.1554	.4451	.6556
$BERT_{wwm}$	.1884	.4501	.6636
Llama2-7B	.3629	.6577	.8153
Llama3-8B	.3712	**.7099**	**.8756**
Phi3-mini	**.3720**	.6351	.8383

表 4 Zero-shot 知識評価の精度．平均値は全テンプレートを用いたときの精度の平均，最大値は関係ごとに最高精度の関係テンプレートを用いたときの精度，上限は関係知識ごとに最高精度のテンプレートを選んだときの精度である．

に表 4 に示す．

この結果から，全テンプレートに対する平均的な知識検索の精度に対し，関係ごとに最高精度の関係テンプレートを選んで用いた場合の精度には約 30 %の増加が見られる．これは，関係テンプレートの言語表現の多様性に対する言語モデルの脆弱性，言い換えると適切な関係テンプレートを選択する重要性を示している．さらに，最大値と上限との間に約 20 %の差が存在することは，関係知識ごとに最適となる関係テンプレートを異なることを示しており，全ての関係知識に対して普遍的に有効な関係テンプレートが存在しないことを示唆する．上限によって示される精度は平均値より 50%近く高い値となっており，多様なプロンプトを関係知識ごとにうまく組み合わせることで，モデルの有する知識を最大限に引き出すことが可能となることを示唆する．

6.2 プロービングによる知識検索の脆弱性の原因

表 3 に示された 3 種の BERT モデルと zero-shot 設定の Llama2-7B の大きな精度揺らぎと低い一貫性は，プロンプトの多様性に対する知識評価の脆弱性を示唆している．本節では，主体の言語表現と関係テンプレートの言語表現の多様性が，精度の揺らぎと一貫性の測定にどのように影響するかを分析する．

ここでは，BERT モデルを使用して分析を行う．具体的には，まず MyriadLAMA の全てのプロンプトから，以下の 2 つの方法で全プロンプトをプロンプトサブセットに分割し，それぞれ，各サブセットごとに評価を行ってその平均を計算する．

(1) 主体の言語表現の多様性が知識検索の脆弱性に与える影響を測るため，各サブセットを，主体のみ多様で各関係は一つの関係テンプレートしか含まないよう構成する．
(2) 関係テンプレートの言語表現の多様性が知識検索の脆弱性に与える影響を測るために，各サブセットを，関係テンプレートのみ多様で各知識は一つの主体表現しか含まないよう構成する．

PLMs	一貫性 ↑		Acc@1 範囲 (最小/大値)	
	主体	関係	主体	関係
$BERT_{base}$	.5745	.1504	.0673/.1441	.0000/.3534
$BERT_{large}$	.5497	.1548	.0714/.1554	.0007/.3728
$BERT_{wwm}$	.5005	.1057	.0831/.1884	.0015/.3695

表 5 言語モデルの知識検索の頑健性に対する主体と関係テンプレートに対する言語表現の多様性の影響．表中で，主体，関係はそれぞれ，主体のみ多様化したプロンプトサブセット，関係のみ多様化したプロンプトサブセットでの結果を示す．

トークン数	Llama2-7B	Llama3-8B	Phi3-mini
1	-.1536	-.1030	.0322
2	-.1566	-.0906	.0502
3	-.1010	-.0297	.0948
4	-.1772	-.0546	.0838
5	-.0491	.0573	.1721
1 to 5	-.1479	-.0985	.0471

表 6 異なるトークン数のプロンプトセットに対する過信度．

表 5 にこの二種類のプロンプトサブセットを用いたときの頑健性の評価結果を示す．表から，関係テンプレート表現の多様性が主体表現の多様性と比べて精度の範囲（最小/最大）および一貫性に大きな影響を与えることが分かる．

6.3 モデル出力の信頼性に影響を与える要因

表 3 は，BERT モデルと Llama2-7B の間で過信度に大きな違いがあることを示しており，BERT モデルは自身の結果を過信している一方で，Llama2-7B は過信していないことが分かる．本節では，その違いをもたらす理由について探究し，信頼性に影響を与える可能性のあるいくつかの要因を調査する．

生成トークン数．まず，生成されたトークンの数が過信度にどのように影響するかを調査する．MLM と CLM の過信度計算における違いとして，BELIEF は MLM を単一マスクのプロンプトで評価する一方，CLM は冗長に生成した出力を用いて複数トークンから構成されうる対象の確信度を計算する．CLM と MLM を公平に比較するために，CLM の生成トークン数ごとに過信度を計算する．まず，各 CLM ごとに MyriadLAMA のプロンプトセットを生成したトークン数に基づいてグループに分ける．各グループについて，生成文字列全体の確率を計算し，1 から 5 トークンのプロンプトセットに対してそれぞれ過信度を計算する．Llama2-7B の 4-template ICL 設定において，生成された文字列のトークン数が 5 トークン以内のプロンプトは，全体の 99.06 %をカバーしている．

表 6 に Llama2-7B，Llama3-8B，および Phi3-mini の各グループの過信度を示す．結果として，生成されるトーク

PLMs	平均精度 (Acc@1) ↑	精度の揺らぎ ↓		一貫性 ↑	過信度
		範囲	標準偏差		
Llama2-7B	.6699	.0257	.0034	.4174	-.0933
Llama2-13B	.7080	.0235	.0031	.4326	**-.0662**
Llama2-70B	**.7784**	**.0190**	**.0024**	.4449	-.0690
Llama3-8B	.7316	.0194	.0025	.4060	-.1119
Llama3-70B	**.8211**	**.0139**	**.0017**	**.4636**	-.0812
Phi3-mini	.6107	.0295	.0039	.3684	.090

表 7 4-template ICL 設定での LLM に対する BELIEF-ICL の評価結果．全てのモデルで，人手で作成したテンプレート 5 つのみを利用して評価をしている．

ンの数が増えるにつれてモデルの過信度が上がることが分かる．しかし，LLM が単一トークンを生成したプロンプトの結果（表 6 のトークン数 1 の行）を BERT モデルの結果（表 3）と比較すると，過信度に大きな差が見られ，全体的な傾向として CLM である LLM が小規模な MLM よりより 0 に近い確信度で予測を行っていることが分かる．

モデルサイズ．表 7 に異なるサイズのモデルを評価結果[*14]を示す．表 7 の過信度の値と，表 3 における $\text{BERT}_{\text{base}}$ から $\text{BERT}_{\text{large}}$ への過信度の改善と同様に，同じ種類のモデルでは，Llama2-13B から Llama2-70B で悪化している例を除き，モデルサイズが大きくなるにつれて，過信度に改善が見られる（0 に近づいている）．結論として，モデルサイズを大きくすることで，精度と確信度のキャリブレーションが改善し，モデルが生成結果とともに出力する確信度（確率）がより信頼できるものとなることが示唆される．

関係テンプレート．最後に，LLM が全ての関係テンプレートで一貫した過信度を示すかどうかを検討する．ために，我々は 1,000 以上の単一トークンの回答を持つテンプレートのみを使用する．各テンプレートに対して，単一トークンを生成したプロンプトの精度と生成トークンの確率を用いて過信度を計算する．具体的に，zero-shot 設定で Llama2-7B を用いた関係 P30[*15]のプロービング結果を調査したところ，81 のテンプレートが条件を満たしていることが分かった．これらのテンプレートの平均過信度は-0.2741 であり，標準偏差は 0.1119 である．最大の過信度は 0.0857，最小の過信度は-0.4558 であり，テンプレート間で過信度に大きな差異があることが分かった．これは，多様なテンプレートが PLM の信頼性の総体的な評価を提供する一方で，特定のテンプレートを用いて知識を探索する場合には，個別にモデルの信頼性を分析する必要があることを示唆している．

[*14] 4 節で述べたように，13B 以上のモデルでは，評価コストから MP スタイルの 4 テンプレート ICL 設定で人手で作成したテンプレート 5 つのみを使用して評価しているが，本節では，比較のため，8B 以下のモデルもこの人手で作成したテンプレートのみを利用して評価した結果をもとに議論する．

[*15] https://www.wikidata.org/wiki/Property:P30

PLMs	平均精度 (Acc@1) ↑	精度の揺らぎ ↓		一貫性 ↑	過信度	一語生成率 ↑
		範囲	標準偏差			
Llama2-7B	.3385	.2602	.0299	.1269	**-.0731**	.4752
Llama3-8B	.3427	.2864	.0350	.0240	-.1119	.1572
Phi3-mini	**.3496**	**.2538**	**.0292**	**.1464**	.1753	**.8736**

表 8 Evaluation result of BELIEF-ICL on LLMs with non-context.

7. LLM 間の知識評価の性能差に関する分析

本節では，異なる LLM の性能を比較し，事前学習中の事実知識学習に影響を与える要因を探究する．

7.1 何がモデルが有する知識の量に影響を与えるか？

モデルサイズ．表 7 に示したように，より大規模な LLM は小規模な LLM よりも高い精度を達成している．例えば，Llama2 の 70B モデルの平均精度は 7B モデルを 10 %上回っている．さらに，大規模な LLM は，基本的に異なるプロンプトに対する一貫性と過信度も優れている．

事前学習コーパス．表 7 から Llama3-8B は Llama2-13B より小さいにも関わらず、Llama3-8B は優れた知識検索性能を持つことが分かる。これは、Llama3 が Llama2 より 7 倍の事前学習コーパスを使用しているためと考えられる。さらに、Llama3-70B が Llama2-70B を上回るという結果はこれを裏付けている．これにより，事前学習データの量は知識取得において重要であることを確認した．

一方で，表 8 に示す全量の MyriadLAMA を使用した zero-shot 設定での評価において，Phi3-mini は Llama2-7B と Llama3-8B より優れた知識検索性能と，指示への高い忠実性（一語生成率）を達成している．Phi3-mini が Llama2-7B や Llama3-8B の半分程度のモデルサイズ（3.8B）であることと，モデルサイズが知識検索性能を改善することを考えると，興味深い結果である．Phi3 モデルは高品質な学習データを用いており，学習データの質が性能改善に寄与している可能性がある．

7.2 異なる PLM が好む関係テンプレートは異なるか？

次に，様々な PLM が異なる関係に対する知識検索で同じ関係テンプレートで相対的に高い（あるいは低い）知識検索性能を示すかどうかを調査する．各関係において 100 のテンプレートにわたるテンプレート順位の相関係数を計算し、その平均を取ることで，モデルが好む関係テンプレートの一致度を定量化する．

図 3 に，zero-shot ICL 設定で求めた関係テンプレート順位に対してモデル間で計算した Kendall の順位相関係数を示す．図から分かるように，モデル間でテンプレート順位に顕著な違いがあることが分かった．例えば，“A demarcation exists linking [Y] to [X].” というテンプレー

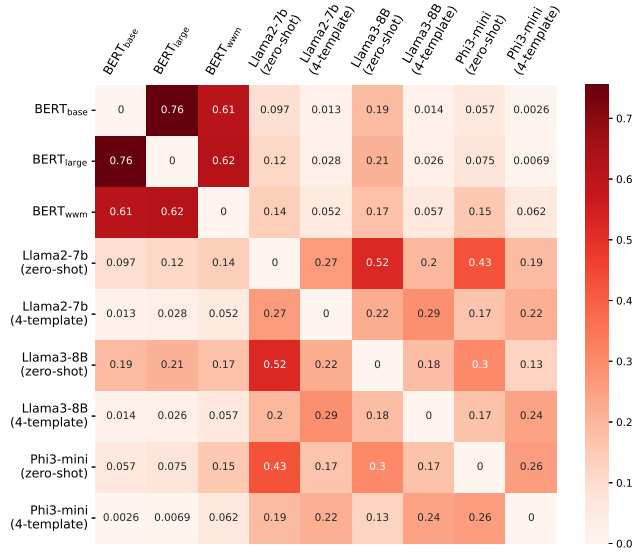

図 3 関係テンプレートに対する精度に関する PLM 間の順位相関: 各セル中の値は Kendall の順位相関係数.

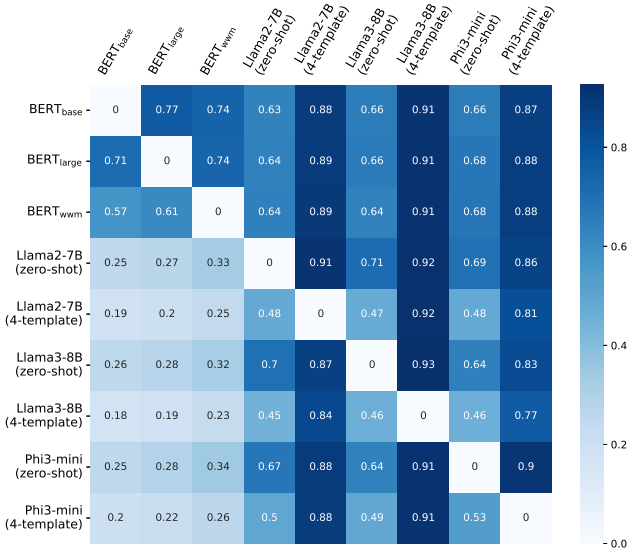

図 4 異なる PLM および ICL 設定においてモデルが正しく予測した関係知識の一致率. 各行は, 左端のモデルが正しく予測した関係知識に対し, 上部に記載されたモデルが正しく予測した関係知識の割合を示す.

トは, Phi3-mini において P47 関係[*16]で最も高い精度を持つが, Llama2-7B ではこのテンプレートの順位が 100 中 72 位に低下している. さらに, 図 3 から, MLM と CLM の間でモデルが好む関係テンプレートの一致度は, 同じタイプの PLM 間の一致度を下回ることが分かる.

7.3 異なる PLM が内包する知識はどの程度異なるか？

最後に, 各モデルが内包する知識集合の違いを評価する. 具体的に, 我々は関係テンプレートごと正しく予測された関係知識を収集し, その関係知識の集合をモデルが正しく予測した知識とする. ここでは, 各モデルが有する関係知識のうち, 他のモデルが有していた関係知識の割合を前者のモデルに対する後者のモデルの一致率（非対称）として計算する. 図 4 に結果を示す. BERT モデル間の平均一致率は 69.07 %, Llama2-7B, Llama3-8B, と Phi3-mini の zero-shot ICL 設定の平均一致率は 67.58 %となっている. それに比べて, CLM と MLM 間の不一致率は低くなり, 47.10 %しかない. また, 複数 LLM の zero-shot と 4-template の一致率から, 入出力事例を取り入れることでモデルが引き出せるできる知識は包括的に向上するが, zero-shot 設定しか引き出せない知識がおよそ 10%あることが分かった.

8. おわりに

本研究では, 多様なプロンプトに基いて言語モデルの多角的な知識評価を行うフレームワークである BELIEF を, 因果言語モデルに基づく LLM に拡張した BELIEF-ICL を提案した. 提案手法は, 簡単な指示と少数事例に基づく in-context learning (ICL) を用いて LLM の知識評価を行う. 実験では, Llama3-70B を含む様々な LLM に対し, ICL に基づく知識評価の有効性と, ICL 設定が LLM の知識評価に与える影響を調査し, モデルに依存しない知識検索の精度, 一貫性, 過信度の傾向について分析を行った. さらに, モデルサイズや学習データの異なる LLM 間での比較を行い, モデルが有する知識の量に影響する要因を調査し, 異なる LLM 間での知識検索の結果の違いについてテンプレートへの依存性と検索可能な知識の差異の観点から分析を行った.

謝辞 本研究は JSPS 科研費 JP21H03494, および JST, CREST, JPMJCR19A4 の支援を受けたものである.

[*16] https://www.wikidata.org/wiki/Property:P47

参考文献

[1] Kamalloo, E., Dziri, N., Clarke, C. and Rafiei, D.: Evaluating Open-Domain Question Answering in the Era of Large Language Models, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Rogers, A., Boyd-Graber, J. and Okazaki, N., eds.), Toronto, Canada, Association for Computational Linguistics, pp. 5591–5606 (online), DOI: 10.18653/v1/2023.acl-long.307 (2023).

[2] Fang, Y., Liang, X., Zhang, N., Liu, K., Huang, R., Chen, Z., Fan, X. and Chen, H.: Mol-Instructions: A Large-Scale Biomolecular Instruction Dataset for Large Language Models, *ICLR*, OpenReview.net, (online), available from ⟨https://openreview.net/pdf?id=Tlsdsb6l9n⟩ (2024).

[3] Kim, J., Kwon, Y., Jo, Y. and Choi, E.: KG-GPT: A General Framework for Reasoning on Knowledge Graphs Using Large Language Models, *Findings of the Association for Computational Linguistics: EMNLP 2023* (Bouamor, H., Pino, J. and Bali, K., eds.), Singapore, Association for Computational Linguistics, pp. 9410–9421 (online), DOI: 10.18653/v1/2023.findings-

emnlp.631 (2023).
[4] Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y. and Miller, A.: Language Models as Knowledge Bases?, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (Inui, K., Jiang, J., Ng, V. and Wan, X., eds.), Hong Kong, China, Association for Computational Linguistics, pp. 2463–2473 (online), DOI: 10.18653/v1/D19-1250 (2019).
[5] Kassner, N. and Schütze, H.: Negated and Misprimed Probes for Pretrained Language Models: Birds Can Talk, But Cannot Fly, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Jurafsky, D., Chai, J., Schluter, N. and Tetreault, J., eds.), Online, Association for Computational Linguistics, pp. 7811–7818 (online), DOI: 10.18653/v1/2020.acl-main.698 (2020).
[6] Misra, K., Ettinger, A. and Rayz, J.: Exploring BERT's Sensitivity to Lexical Cues using Tests from Semantic Priming, *Findings of the Association for Computational Linguistics: EMNLP 2020* (Cohn, T., He, Y. and Liu, Y., eds.), Online, Association for Computational Linguistics, pp. 4625–4635 (online), DOI: 10.18653/v1/2020.findings-emnlp.415 (2020).
[7] Ravichander, A., Hovy, E., Suleman, K., Trischler, A. and Cheung, J. C. K.: On the Systematicity of Probing Contextualized Word Representations: The Case of Hypernymy in BERT, *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics* (Gurevych, I., Apidianaki, M. and Faruqui, M., eds.), Barcelona, Spain (Online), Association for Computational Linguistics, pp. 88–102 (online), available from ⟨https://aclanthology.org/2020.starsem-1.10⟩ (2020).
[8] Jiang, Z., Xu, F. F., Araki, J. and Neubig, G.: How Can We Know What Language Models Know?, *Transactions of the Association for Computational Linguistics*, Vol. 8, pp. 423–438 (online), DOI: 10.1162/tacl_a_00324 (2020).
[9] Elazar, Y., Kassner, N., Ravfogel, S., Ravichander, A., Hovy, E., Schütze, H. and Goldberg, Y.: Measuring and Improving Consistency in Pretrained Language Models, *Transactions of the Association for Computational Linguistics*, Vol. 9, pp. 1012–1031 (online), DOI: 10.1162/tacl_a_00410 (2021).
[10] Newman, B., Choubey, P. K. and Rajani, N.: P-Adapters: Robustly Extracting Factual Information from Language Models with Diverse Prompts, *International Conference on Learning Representations*, (online), available from ⟨https://openreview.net/forum?id=DhzIU48OcZh⟩ (2022).
[11] 趙　信，吉永直樹，大葉大輔：多様なプロンプトを用いた言語モデルの多角的な知識評価，情報処理学会 研究報告自然言語処理（NL）（2024).
[12] OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D. et al.: GPT-4 Technical Report (2024).
[13] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S. et al.: Llama 2: Open Foundation and Fine-Tuned Chat Models (2023).
[14] Lin, S., Hilton, J. and Evans, O.: TruthfulQA: Measuring How Models Mimic Human Falsehoods, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Muresan, S., Nakov, P. and Villavicencio, A., eds.), Dublin, Ireland, Association for Computational Linguistics, pp. 3214–3252 (online), DOI: 10.18653/v1/2022.acl-long.229 (2022).
[15] Mallen, A., Asai, A., Zhong, V., Das, R., Khashabi, D. and Hajishirzi, H.: When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Rogers, A., Boyd-Graber, J. and Okazaki, N., eds.), Toronto, Canada, Association for Computational Linguistics, pp. 9802–9822 (online), DOI: 10.18653/v1/2023.acl-long.546 (2023).
[16] AI@Meta: Llama 3 Model Card, `https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md` (2024).
[17] Abdin, M., Jacobs, S. A., Awan, A. A., Aneja, J., Awadallah, A., Awadalla, H., Bach, N., Bahree, A., Bakhtiari, A. et al.: Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone (2024).
[18] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I. and Amodei, D.: Language Models are Few-Shot Learners, *Advances in Neural Information Processing Systems* (Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. and Lin, H., eds.), Vol. 33, Curran Associates, Inc., pp. 1877–1901 (2020).
[19] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J. and Lowe, R.: Training language models to follow instructions with human feedback, *Advances in Neural Information Processing Systems* (Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K. and Oh, A., eds.), Vol. 35, Curran Associates, Inc., pp. 27730–27744 (online), available from ⟨`https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf`⟩ (2022).

付　　録

A.1 In-context learning プロンプトの例

本節では，本研究で導入した 8 つの ICL 設定のプロンプトの例を示す．この 8 つのパターンは，2 種類の指示形式（QA-style 指示と MP-style 指示）と 4 種類の入出力事例を組み合わせたものである．

A.1.1 MP-style/zero-shot

```
Predict the [MASK] in each sentence in one word.
Q: [MASK] consists of LAUPT.
A:
```

A.1.2 MP-style/4-random

```
Predict the [MASK] in each sentence in one word.
Q: [MASK] is the administrative center of Jiangsu.
A: Nanjing.
Q: Mar del Plata and [MASK] are sister cities that have been developing together.
A: Havana.
Q: Malawi has established diplomatic ties with [MASK].
A: Australia.
Q: Which country is House of Representatives located? [MASK].
A: Libya.
Q: [MASK] consists of LAUPT.
A:
```

A.1.3 MP-style/4-relation

```
Predict the [MASK] in each sentence in one word.
Q: What is the overarching group for Panzer Division Kempf? [MASK].
A: Wehrmacht.
Q: To whom does Mount Bulusan relate? [MASK].
A: Luzon.
Q: Who is responsible for Army National Guard? [MASK].
A: National Guard.
Q: What group is pharmacy a part of? [MASK].
A: biology.
Q: [MASK] consists of environmental factors.
A:
```

A.1.4 MP-style/4-template

```
Predict the [MASK] in each sentence in one word.
Q: [MASK] consists of Panzer Division Kempf.
A: Wehrmacht.
Q: [MASK] consists of Mount Bulusan.
A: Luzon.
Q: [MASK] consists of Army National Guard.
A: National Guard.
Q: [MASK] consists of pharmacy.
A: biology.
Q: [MASK] consists of environmental factors.
A:
```

A.1.5 QA-style prompts

QA-style のプロンプトでは，指示を “Answer each question in one word.” に置き換える．その他の設定は MP-style のプロンプトと同じである．以下に，QA-style/4-template プロンプトの例を示す．

```
Answer each question in one word.
Q: Which entity does Panzer Division Kempf belong to?
A: Wehrmacht.
Q: Which entity does Mount Bulusan belong to?
A: Luzon.
Q: Which entity does Army National Guard belong to?
A: National Guard.
Q: Which entity does pharmacy belong to?
A: biology.
Q: Which entity does environmental factors belong to?
A:
```

A.2 LLM における評価結果のまとめ

表 A·1 と表 A·2，MP-style の指示テンプレートを用いた場合の LLM の評価結果を示す．

PLMs		平均精度 (Acc@1) ↑	精度の揺らぎ ↓		一貫性 ↑	過信度
			範囲	標準偏差		
Llama2-7B	zero-shot	.3385	.2602	.0299	.1269	**-.1119**
	4-random	.4816	.2250	.0270	.2312	-.0894
	4-relation	.6286	.1221	.0150	.3753	-.1335
	4-template	**.6616**	**.0294**	**.0036**	**.4163**	-.0933
Llama3-8B	zero-shot	.3427	.2864	.0350	.0240	-.1119
	4-random	.5205	.2033	.0273	.2156	-.0789
	4-relation	.6871	.1236	.0156	.3659	-.0783
	4-template	**.7268**	**.0220**	**.0026**	**.4015**	**-.0582**
Phi3-mini	zero-shot	.3496	.2538	.0292	.1464	.1752
	4-random	.4191	.2223	.0270	.1648	.1182
	4-relation	.5411	.1636	.0205	.2472	.1062
	4-template	**.6066**	**.0401**	**.0047**	**.3611**	**.0888**

表 A·1 8B 以下のモデルに対して四種類の MP-style ICL 設定で MyriadLAMA の全データを評価した結果.

PLMs		平均精度 (Acc@1) ↑	精度の揺らぎ ↓		一貫性 ↑	過信度
			範囲	標準偏差		
zero-shot	Phi3-mini	.4245	**.1961**	**.0245**	.2065	.1604
	Llama2-7B	.4311	.2014	.0249	.1932	**-.0922**
	Llama3-8B	.4224	.2820	.0353	.1269	-.1438
	Llama2-13B	.4785	.2131	.0260	.1437	-.1673
	Llama2-70B	.5675	.2126	.0280	.0359	-.0988
	Llama3-70B	**.5974**	.2137	.0278	**.2290**	-.1438
4-template	Phi3-mini	.6107	.0295	.0039	.3684	.0909
	Llama2-7B	.6699	.0257	.0034	.4174	-.0933
	Llama3-8B	.7316	.0194	.0025	.4060	-.1119
	Llama2-13B	.7080	.0235	.0031	.4326	**-.0662**
	Llama2-70B	.7784	.0190	.0024	.4448	-.0690
	Llama3-70B	**.8211**	**.0139**	**.0017**	**.4636**	-.0812

表 A·2 全ての LLM（6 つ）に対して，zero-shot と 4-template の ICL 設定で，人手で作成したテンプレートに基づく部分データの評価結果.