多角的な評価から大規模言語モデルにおける事実知識の想起の要因分析

趙信 [1]　吉永直樹 [2]　大葉大輔 [2*]
[1] 東京大学大学院　[2] 東京大学 生産技術研究所
xzhao@tkl.iis.u-tokyo.ac.jp {ynaga,oba}@iis.u-tokyo.ac.jp

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．本研究では，多様なプロンプトを持つデータセット MyriadLAMA を構築し，それを用いた知識評価フレームワーク BELIEF を提案する．In-context learning (ICL) を用いて，LLM が有する知識を精度，一貫性，信頼性の観点から多角的な評価を可能にする．実験では，パラメタ数，事前学習コーパス，指示学習の有無などが異なる複数の LLM を対象に知識評価を行い，LLM における事実知識の想起において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，学習過程でテキストに含まれる関係知識を，暗黙のうちに獲得，保持することから，知識ベースとしての活用が期待されており，PLM が有する関係知識を評価する研究が行われるようになっている．PLM の保持する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 の出身国は [MASK] である）の空欄 ([MASK]) のエンティティを言語モデルで予測する LAMA probe [1] が用いられる．LAMA probe では，PLM の予測精度をもとに，モデルが有する知識の量を評価する．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予測精度が大きく変動することが報告されており [2, 3]，言語モデルの知識ベースとしての有用性をより適切に評価する手法を確立することが期待されている．

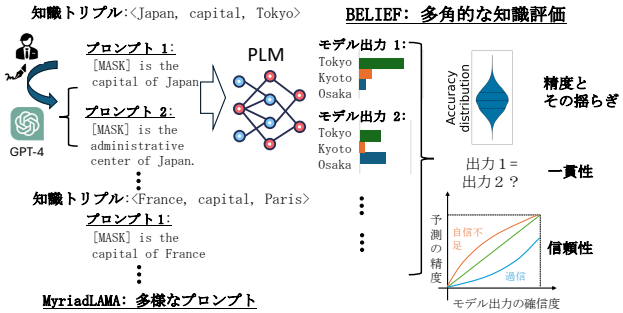

図 1　MyriadLAMA を用いた BELIEF 評価指標: BELIEF ベンチマークは，多様な事実プロンプトを活用して，LLM が持つ知識を精度，一貫性，信頼性の観点から評価する．

本研究では，単一プロンプトに依存した関係知識評価では PLM が有する知識を適切に評価することが困難であることを踏まえ，多様なプロンプトを用いた PLM の関係知識評価手法 BELIEF と，その評価のためのデータセット MyriadLAMA を構築する．MyriadLAMA は，関係知識のタイプごとに人手で用意した少数の構文的・意味的に多様なプロンプトをもとに，大規模言語モデル（LLM, 具体的には GPT-4）を用いて大量のプロンプトを自動生成することで，半自動で多様なプロンプトを構築する．BELIEF は，精度，一貫性，信頼性の多角的な観点から知識評価を行い，PLM の知識想起能力をより包括的に分析することができる．

本研究は LLM の評価に着目し，穴埋め文を LLM に適応できるように，in-context learning (ICL) を用いて評価を行う．実験では，Llama2 [4]，Llama3 [5]，および Phi3 [6] 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．評価結果に基づいて，LLM が事実知識の想起に対して，精度，一貫性，信頼性などの観点から事前学習における重要な要素を分析して，モデルサイズ，事前学習データの質と量，指示学習，使用する指示などが知識想起に与える影響とその傾向を明らかにした．

* 2024 年 4 月より株式会社 ELYZA に所属

2 MyriadLAMA の構築

本研究では，後述の多角的な関係知識評価を実現するため，既存の関係知識評価データセット LAMA-UHN [7] を拡張して MyriadLAMA を構築する．LAMA-UHN 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．各関係の種類に対して一つテンプレート表現（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．LAMA-UHN を使用した関係知識評価の手順は，まず関係テンプレートの [X] を対応する知識トリプルの主体で埋め，[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**）を生成し，言語モデルが知識トリプル中の対象を正しく予測できるかを評価する．

MyriadLAMA では，一つ知識に多様な言語表現（プロンプト）を提供するため，“関係” に対して多様なテンプレート表現を提供するのみならず，“主体” や “対象” などのエンティティの言語表現も拡張する．具体的には，表層的な言語表現の揺らぎを排して互いに独立な関係知識を考え，この関係知識に含まれるエンティティ（“主体” と “対象”）と関係の言語表現を言い換えにより多様化することで，多様なプロンプトを生成する．このとき，表層の揺らぎを縮退した知識トリプルを**固有トリプル**，各固有トリプルに含まれるエンティティと関係表現を具体化した知識トリプルを**派生トリプル**と呼び，区別することとする．詳しい拡張方法は § A を参照されたい. 拡張した結果，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 の関係知識に対して 6,492,800 のプロンプトを提供する．

3 BELIEF: 多角的な関係知識評価

本節では，LLM が有する知識をより適切に評価するため，多様なプロンプトを用いた関係知識評価手法 BELIEF を提案する．まず，マスク言語モデルを対象とするプロンプトを因果言語モデルである LLM の評価に適用できるように，in-context learning (ICL) に基づく事実知識評価手法を提案する．次に，BELIEF では，MyriadLAMA（2 節）の多様なプロンプトを使用することで，個別のプロンプトのバイアスの影響を低減した関係知識評価を行うとともに，言語モデルが想起する知識の量（精度）に加えて，モデルが提示する知識の一貫性と信頼性を考慮した評価方法を提案する．

3.1 In-context Learning による知識評価

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 [8]．事実知識評価のための 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．

1) タスク指示: LLM にマスク予測に関する指示を与え，対象のマスクされたプロンプトに対して単語で答えるようする．具体的に与えた指示は，‘‘`Predict the [MASK] in each sentence in one word.`’’ である．

2) 入出力事例: 知識評価における入出力事例選択の影響を評価するため，InstructGPT [9] の QA フォーマットに従い，4 種類の入出力事例を提案する．

zero-shot: 入出力事例を提供しない．
X-random: 全ての関係から X 個のプロンプトをサンプリングし，入出力事例とする．
X-relation: 同じ種類の関係で異なるテンプレートを持つプロンプトから X 個サンプリングする．
X-template: 同じ種類の関係・テンプレートのプロンプトから X 個サンプリングする．

LLM 出力の評価: LLM 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] の対応以外のテキスト（例えばプロンプトの一部）を生成する場合がある．表記の揺れ問題を解決するため，比較対象となる文字列を単語分割[1)] 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．

3.2 BELIEF の評価指標

BELIEF は，LLM 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．

1) ライブラリ nltk を使用する．

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 (Acc@1) 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 (Range) と標準偏差 (SD) で測定される．Acc@1 の計算は，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で，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．

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答え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．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，LLM の予測をど程度，信頼できるかを評価する指標である．BELIEF では，LLM が生成文字列への確信度とその精度の差として定義される**過信度**[2] を信頼性の評価に採用する．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，LLM は自身の予測の信頼性を正確的に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，過信度の値が正/負になると，モデルは自身の予測の信頼性を過大/小評価しているとみなせる．LLM が出力への確信度を計算するには，multinomial sampling[3] を用いて，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．

2） 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．

3） Multinomial sampling は，モデルが生成した全語彙の確率分布に基づいて次のトークンを確率的に選択する手法である．

LLMs	Acc@1	精度の揺らぎ↓		一貫性↑	過信度	1-word ratio
		Range	SD			
Llama2-7B 0-shot	.3385	.2602	.0299	.1269	-.1119	.4752
Llama2-7B 4-rand.	.4816	.2250	.0270	.2312	**-.0894**	.8247
Llama2-7B 4-rel.	.6286	.1221	.0150	.3753	-.1335	.9060
Llama2-7B 4-templ.	**.6616**	**.0294**	**.0036**	**.4163**	-.0933	**.9299**
Llama2-7B-IT 0-shot	.2925	.1980	.0253	.1151	.2605	.9069
Llama2-7B-IT 4-rand.	.4334	.1958	.0229	.2128	.2410	.9081
Llama2-7B-IT 4-rel.	.5576	.0791	.0092	.3341	**.1900**	.9314
Llama2-7B-IT 4-templ.	**.5896**	**.0439**	**.0050**	**.3687**	.2061	**.9380**

表 1 Llama2-7B と Llama2-7B-it における評価結果．

4 実験とその結果

4.1 実験設定

本研究は LLM が想起できる知識の違いを起こす原因の解明に着目し，3 種類の異なるパラメタ数の LLM（計 8 つ）に BELIEF を適用し，知識評価を行う．Llama2-7B/13B/70B, Llama3-8B は事前学習 LLM で，Llama2-7B-it, Llama3-8b-it, Phi3-mini(3.8B)/small(7B) は指示学習した LLM である．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 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．各モデルの事前学習データに関する情報は § B を参照されたい．

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計算コストが大きいため，全部のプロンプから 10,000 個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 $N = 50,000$ と設定する．

4.2 実験結果

LLM は ICL の指示に忠実か？ LLM が提案された ICL 設定の指示に従うかどうかを，平均精度 (Acc@1) と，指示通り 1 単語が生成されたプロンプトの割合（以下，1-word ratio），という 2 つの観点から評価する．表 1 は，Llama2-7B(-it) の BELIEF に基づく評価結果を示している．予想通り，指示学習を行った LLM は指示への忠実性が高いことが確認されたが，Few-shot 設定では，指示学習をしない LLM でも高い指示追従能力を示し，マスクプロンプトを用いた LLM の関係知識の評価を可能にする．

ICL は LLM の知識想起性能に影響を与えるか？さらに，文脈内に対象プロンプトと類似した例示

LLMs	Acc@1↑	精度の揺らぎ↓		一貫性↑	過信度
		Range	SD		
Llama2-7B	.6699	.0257	.0034	.4174	-.0933
Llama2-13B	.7080	.0235	.0031	.4326	**-.0662**
Llama2-70B	**.7784**	**.0190**	**.0024**	**.4449**	-.0690
Llama2-7B-IT	.6013	.0368	.0045	.3629	.2007
Llama2-13B-IT	.6482	.0301	.0038	.3656	.1708
Llama2-70B-IT	**.7232**	**.0258**	**.0031**	**.4226**	**.1026**
Llama3-8B	**.7316**	**.0194**	**.0025**	**.4060**	-.1119
Llama3-8B-it	.6563	.0252	.0032	.3752	**.0535**
Phi3-mini	.6106	.0314	.0039	**.3686**	**.0911**
Phi3-small	**.6668**	**.0306**	**.0039**	.3667	.1221

表 2 4-template ICL 設定での LLM に対する BELIEF の評価結果．計算コストのため，関係ごとに人手で作成した 5 つテンプレートのみを利用して評価をしている．

(4-template) が含まれる場合，全体的な指標の向上が確認された．表 1 に示されているように，zero-shot と few-shot ICL の Acc@1 の差を比較すると，少数の入出力事例を追加することで，LLM の事実知識を想起する能力が大幅に向上することがわかる．また，入出力事例の選択方法も結果に大きく影響する．3 種類の few-shot ICL 設定を比較した結果，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（精度の一貫性と変動の少なさ）が一貫して向上することが明らかになった. 4-template は，モデルが想起できる知識の上限をより正確に評価できる点を考慮し，以降の実験では 4-template のプロンプトを使用して評価を行う．

5 LLM 間の性能差に関する分析

何がモデルが有する知識の量に影響を与えるか？ 表 2 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．Llama3-8B は Llama2-13B より小型ながら優れた知識想起性能を示す．これは，Llama3 が Llama2 の 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，事前学習データ量の重要性を示唆する．さらに，Phi3 モデルは高品質な学習データを用いており，Phi3-small(7B) が Llama3-8B-IT の 1/3 程度の学習データ量にも関わらず，優れた知識想起性能を示していることから，学習データの質が性能改善に寄与することが分かる．最後に，指示学習を行った LLM が一貫して低い Acc@1 を示したことから，指示学習は LLM の指示への忠実性を向上させる一方で，知識想起能力を低下させる可能性があることが確認された．

何がモデルの頑健性に影響を与えるか？ 表 1 から，指示学習モデルは zero-shot や 4-random 設定で揺らぎの減少が見られる一方で，4-template 設定では性能が低下している．これにより，文脈よりも指示への依存が強まり，頑健性が低下する傾向が示され，指示学習によって指示への忠実性が向上する一方，多様な言語表現に対する意味理解が弱まる可能性が示唆される．また，モデルサイズが大きくなるほど一貫性が改善されることも示している．

何がモデルの信頼性に影響を与えるか？ 指示学習は LLM の出力信頼度を過度に高めることが明らかになった．表 1 から，指示学習されたモデルが出力に過信する傾向が示されている．（詳細は § C）事前学習では多様な言語データに基づいて信頼度が適切に調整される一方，指示学習は特定のタスクに特化するため，不確実性の表現が低下し，過信が増加する要因となる．また，大規模モデルでは信頼性が一貫して向上することも確認された．

6 関連研究

近年，LAMA Probe は PLM の知識評価手法として提案され [1]，多様な言語的側面での拡張が行われてきた [10, 2, 11, 12]．しかし，これらの研究は主にマスクモデルに限定されており，LLM に直接適用すると，モデルの知識理解力が過小評価される可能性がある．また，これらの手法には PLM の知識予測における信頼性評価が欠けること [13, 14] と表現の揺らぎに影響されるなどの問題がある [15, 16]．一方で，近年では因果モデルの知識想起能力を評価するための QA ベースのデータセットが開発されているが [17, 13, 18, 14]，これらのデータセットは質問形式に限定されており，信頼性のある頑健性評価を実施するのが困難である．

7 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ベンチマーク BELIEF 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 を用いて LLM の知識評価を行う．実験では，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，モデルの知識想起の精度，一貫性，過信度を分析した．結果として，モデルサイズ，学習データの質と量，指示学習，ICL が知識想起に与える影響とその傾向を明らかにした．

謝辞

本研究は，東京大学生産技術研究所特別研究経費，JSPS 科研費 JP21H03494 および JST, CREST, JPMJCR19A4 の支援を受けたものである．

参考文献

[1] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, **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. 2463–2473, Hong Kong, China, November 2019. Association for Computational Linguistics.

[2] Nora Kassner and Hinrich Schütze. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 7811–7818, Online, July 2020. Association for Computational Linguistics.

[3] Kanishka Misra, Allyson Ettinger, and Julia Rayz. Exploring BERT's sensitivity to lexical cues using tests from semantic priming. In Trevor Cohn, Yulan He, and Yang Liu, editors, **Findings of the Association for Computational Linguistics: EMNLP 2020**, pp. 4625–4635, Online, November 2020. Association for Computational Linguistics.

[4] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models, 2023.

[5] AI@Meta. Llama 3 model card. https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md, 2024.

[6] Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, et al. Phi-3 technical report: A highly capable language model locally on your phone, 2024.

[7] Fabio Petroni, Patrick Lewis, Aleksandra Piktus, Tim Rocktäschel, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. How context affects language models' factual predictions. **ArXiv**, Vol. abs/2005.04611, , 2020.

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, **Advances in Neural Information Processing Systems**, Vol. 33, pp. 1877–1901. Curran Associates, Inc., 2020.

[9] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll othersWainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, **Advances in Neural Information Processing Systems**, Vol. 35, pp. 27730–27744. Curran Associates, Inc., 2022.

[10] Lalchand Pandia and Allyson Ettinger. Sorting through the noise: Testing robustness of information processing in pre-trained language models. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 1583–1596, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[11] Amr Keleg and Walid Magdy. DLAMA: A framework for curating culturally diverse facts for probing the knowledge of pretrained language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Findings of the Association for Computational Linguistics: ACL 2023**, pp. 6245–6266, Toronto, Canada, July 2023. Association for Computational Linguistics.

[12] Xin Zhao, Naoki Yoshinaga, and Daisuke Oba. Tracing the roots of facts in multilingual language models: Independent, shared, and transferred knowledge. In Yvette Graham and Matthew Purver, editors, **Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 2088–2102, St. Julian's, Malta, March 2024. Association for Computational Linguistics.

[13] Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. 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. 9802–9822, Toronto, Canada, July 2023. Association for Computational Linguistics.

[14] Seiji Maekawa, Hayate Iso, Sairam Gurajada, and Nikita Bhutani. Retrieval helps or hurts? a deeper dive into the efficacy of retrieval augmentation to language models. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, **Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)**, pp. 5506–5521, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[15] Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. How can we know what language models know? **Transactions of the Association for Computational Linguistics**, Vol. 8, pp. 423–438, 2020.

[16] Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. Measuring and improving consistency in pretrained language models. **Transactions of the Association for Computational Linguistics**, Vol. 9, pp. 1012–1031, 2021.

[17] Jan-Christoph Kalo and Leandra Fichtel. Kamel: Knowledge analysis with multitoken en- tities in language models. In **Automated Knowledge Base Construction**, 2022.

[18] Jacek Wiland, Max Ploner, and Alan Akbik. BEAR: A unified framework for evaluating relational knowledge in causal and masked language models. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, **Findings of the Association for Computational Linguistics: NAACL 2024**, pp. 2393–2411, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[19] Hady Elsahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon Hare, Frederique Laforest, and Elena Simperl. T-REx: A large scale alignment of natural language with knowledge base triples. In **Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)**, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

LLMs	指示学習モデル	モデルサイズ	事前学習コーパス	
			トークン数	コーパスの種類
Llama2-7B	No	7B	2.0T tokens	公開されているオンラインデータの集合（個人情報を多く含むサイトは除外．事実知識の知識源は upsampling）
Llama2-13B	No	13B	2.0T tokens	
Llama2-70B	No	70B	2.0T tokens	
Llama2-7B-IT	Yes	13B	2.0T tokens	
Llama2-13B-IT	Yes	13B	2.0T tokens	
Llama2-70B-IT	Yes	70B	2.0T tokens	
Llama3-8B	No	8B	15T+ tokens	公開されているオンラインデータの集合（詳細不明，コードは Llama2 の 4 倍）
Llama3-8B-IT	No	8B	15T+ tokens	
Phi3-mini	Yes	3.8B	4.9T tokens	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
Phi3-small	Yes	7B	4.9T tokens	

表 3　本研究で扱う LLM の事前学習情報．

A　プロンプトの拡張方法

本節は，MyriadLAMA における関係知識の表現を拡張する方法の詳細を紹介する．表 4 に，LAMA-UHN と MyriadLAMA について，“主体-関係” ペア数，プロンプト数，関係テンプレート数と各種トリプル数をそれぞれ示す．

エンティティの拡張． LAMA-UHN に含まれる知識トリプルは Wikipedia に基づく知識ベース T-REx [19] のサブセットであり，その “主体-関係” ペアの対象の種類は限定されている．これに対して，MyriadLAMA では，“主体-関係” をキーとして T-REx の知識ベースを検索し，他の許容される “対象” を包含するように “対象” を拡張した．例えば，John Lennon が演奏できる楽器について，E_{guitar} のみを LAMA-UHN に含まれている場合，E_{piano} も含むように固有トリプルを拡張する．また，Wikidata[4] に含まれるエイリアスを使用して，主体および対象の表現も拡張した．例えば，E_{United Kingdom} は複数の表現 (United Kingdom, UK, Britain) として表せる．

関係テンプレートの言い換え． 既存研究では，関係テンプレートと意味的に等価な言い換え表現を人手 [15] や関係抽出手法を利用 [16] して少数生成しているが，本研究では，含意を伴う表現や異なる構文（例：陳述，質問-回答）を含む，より多様的な関係テンプレートを人手で生成した上で，さらに得られた言語表現の言い換えを LLM を用いて自動生成した．具体的な手順は以下の通りである．まず各 “関係” に対して 5 つの意味的・構文的に異なる関係テンプレートを作成し，次に GPT-4 API[5] を使用して各テンプレートの言い換えをそれぞれ 20 個生成した．全てのテンプレートの妥当性を人手で確認し，結果として，41 の関係に対して計 4100 件のテンプレートを作成した．

表 4　LAMA-UHN と MyriadLAMA の統計情報．

	LAMA-UHM	MyriadLAMA
関係テンプレート数	41	4100
固有トリプル数	27,106	34,048
派生トリプル数	27,106	21,140,500
主体-関係ペア数	24,643	24,643
プロンプト数	24,643	6,492,800

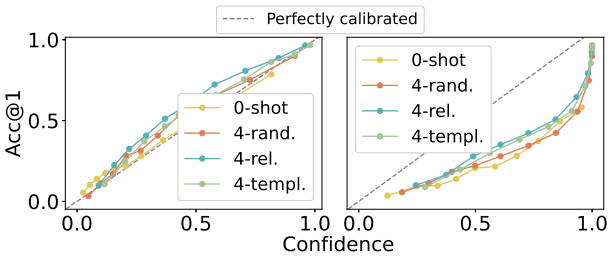

図 2　Llama2-7B と Llama2-7B-it に対する四種類の ICL 設定における精度と確信度のアライメント)．

B　モデルの事前学習情報

本研究で評価する LLM の事前学習情報およびモデルサイズと種類は表 3 で報告している．

C　Llama2-7b(-it) における過信度

図 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 Llama2-7B-it 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．

4）　https://www.wikidata.org/wiki/Wikidata:Data_access

5）　OpenAI: gpt-4-1106-preview