自己認知は LM as KB の信頼性を高めるか

井之上直也 [1,2]　原口大地 [1]　田中健史朗 [1]
白井清昭 [1]　Natthawut Kertkeidkachorn[1]
[1] 北陸先端科学技術大学院大学　[2] 理化学研究所
{naoya-i,s2110137,kenshiro.tanaka,kshirai,natt}@jaist.ac.jp

概要

Language **M**odels **as** **K**nowledge **B**ases (LM as KB) は，自然言語形式のプロンプトで知識の問い合わせを柔軟に行える一方で，信頼度の高い問い合わせ方法は未だ確立されていない．本稿では，LM as KB に自己認知機構を取り入れ，予測結果が不確実な場合に問い合わせを分解し，熟考的に検証する新しいアプローチ Back-off LMKB を検討する．評価実験では，GPT-4 及び GPT-3.5 に基づく Back-off LMKB を質問応答データセット StrategyQA [1] 上で検証し，自己認知機構の有効性，及び今後の課題を示す．

1　はじめに

LM as KB とは，自然言語形式で書かれた知識ベースを大規模言語モデル (LLM) に学習させることにより，自然言語形式のプロンプトで知識の問い合わせを柔軟に行うことを目指すパラダイムである [2].

LM as KB における未解決課題の一つは，知識の問い合わせ方法である．一般に，ある知識を表現する自然言語表現は複数考えられるが，LLM の出力は言語表現の微妙な変化に鋭敏であり，実際に言語表現によって同一の知識に対する問い合わせ結果が一貫しないことが示されている [3, 4]．例えば，文献 [4] では，*Anne Redpath* の死没地を問い合わせる際，プロンプト *Anne Redpath's life ended in* と *Anne Redpath's passed away in* とでは，LLM の返す結果がそれぞれ *London, Edinburgh* と異なることが指摘されており，LM as KB の信頼性を揺るがせている．

本研究では，こうした非一貫性は，予測の不確実性により捉えられるものと仮定し，LM as KB の信頼性を改善する方法を検討する．具体的には，図 1 に示すように，素の LM as KB (図 1, LMKB) の問い合わせ結果の不確実性をチェックする自己認知機構を取り入れる．結果に自信がない場合には，問い合わせを単純な命題に分解して論理的に再検証する

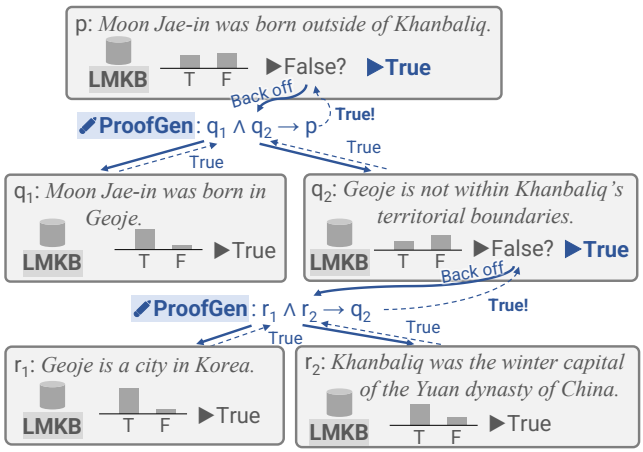

図 1　Back-off LMKB の概要．知識ベース LMKB から真偽が不確実な場合には解答を使用せず，間接証明生成モデル ProofGen を通して，別の知識を用いた論理的な真偽検証にバックオフする．

(図 1, ProofGen)．本研究の貢献は次の通りである:

- 問い合わせ結果の不確実性に応じて動的に問い合わせを分解する新たな枠組み Back-off LMKB を提案する（§3）．
- StrategyQA [1] から構築した真偽検証問題 457 問において，GPT-4 及び GPT-3.5 に基づく Back-off LMKB の有効性を定量的に示し（§4.2），エラー分析により今後の課題を明らかにする（§4.3）．

2　関連研究

LLM の生成結果の信頼性推定の研究は近年盛んに行われており [5, 6, 7, 8]，LLM にとって真偽の分からない知識を問われた場合には，不明と答える技術が成熟しつつあり，本研究ではこれを利用する．

質問応答や事実性判定の分野においては，複雑な質問を単純な質問に分解してから LLM に問い合わせ，その結果を統合することで，解答の精度が向上することが多数報告されている [9, 10, 11, 12]．しかしながら，いずれも予測の確実性を考慮した動的な分解は行っていない．

文献 [13] では，画像に対する質問応答 (VQA) において，確信度に応じて動的に質問を分解することの有効性が確認されている．しかしながら，VQA は入力画像に対する問い合わせ，LM as KB はパラメトリック知識に対する問い合わせであり，本研究とは実験設定が大きく異なる．また，§3 で示すような再帰的な真偽検証は検討されていない．

3 提案手法: Back-off LMKB

3.1 全体の流れ

自然言語形式の命題 p と知識ベース K が与えられたとき，p の真偽値が分かるならば True, False のいずれか，未知ならば Unknown と答えるモデルを構築する．形式的には次のような関数 f を設計する．

$$f(p) = \begin{cases} \text{True} & \text{if } K \rightsquigarrow p \\ \text{False} & \text{if } K \rightsquigarrow \neg p \text{ または } K \not\rightsquigarrow p \\ \text{Unknown} & \text{上記の両方が不確実} \end{cases} \quad (1)$$

ここで $\rightsquigarrow$ は意味的な含意を表す．

提案手法である Back-off LMKB の概要を図 1，アルゴリズムを Algorithm 1 に示す．引数について，p は検証対象の命題，d は再帰の上限，r は現在の再帰呼び出しの深さである．以下，解説する．

Step 1: K 上で事前訓練された LLM に基づく知識ベースモデル LMKB を構築し（§3.2），命題 p の真偽を検証する（2 行目）．ただし，LMKB は命題の真偽に確信が持てない場合に，Unknown を返すことができるとする．ここで True, False が決まる場合，もしくは再帰の上限に達した場合（$r = d$）には，真偽値を返して処理を終了する（3 行目）．

Step 2: §1 で述べたように，予測の不確実性は LLM の知識の非一貫性のサインであると考えられる．そこで，予測が不確実なとき，すなわち LMKB が Unknown を返した場合には，**間接証明**にバックオフし，p の真偽を論理的に検証する（4 行目）．

具体的には，まず p と論理的な関係にある命題の集合 $Q = \{q_1, q_2, ..., q_n\}$ を求めるモデル ProofGen を構築する（§3.3）．論理的な関係は命題論理式 ϕ で表され，例えば，十分条件の関係であれば，$(\phi, Q) = (q_1 \wedge q_2 \to p, \{q_1, q_2\})$ となる．

次に，$q_i \in Q$ の真偽値 y_{q_i} を検証し，仮定 $\Psi = \{q_i \mid q_i \in Q, y_{q_i} = \text{True}\} \cup \{\neg q_i \mid q_i \in Q, y_{q_i} = \text{False}\}$ を構築する．ここで，ϕ, Ψ より p または $\neg p$ が帰結できるならば（例えば $y_{q_1} = \text{True}, y_{q_2} = \text{True}$ の場合），対応する真偽値を返す．例えば，図 1 では，$\phi = q_1 \wedge q_2 \to p$ となるような命題 q_1, q_2 を求め，$q_1 = \text{True}, q_2 = \text{True}$ から p が True であることを帰結している．なお，$q_i \in Q$ の真偽検証では，q_i を入力として Step 1 から再帰的に処理を行う．

Algorithm 1 Back-off LMKB.

```
1:  function BackoffLMKB(p, d, r = 0)
2:      y ← LMKB(p) ∈ {True, False, Unknown}
3:      if y ∈ {True, False} or r = d then return y
4:      φ, Q ← ProofGen(p)          ▷ 間接証明にバックオフ
5:      for q_i ∈ Q do y_{q_i} ← BackoffLMKB(q_i, d, r + 1)
6:      Ψ ← {q_i | q_i ∈ Q, y_{q_i} = True}
7:      Ψ ← Ψ ∪ {¬q_i | q_i ∈ Q, y_{q_i} = False}
8:      if φ, Ψ ⊢ p then            ▷ p は真？
9:          return True
10:     else if φ, Ψ ⊢ ¬p then      ▷ p は偽？
11:         return False
12:     end if
13:     return Unknown              ▷ 証明失敗
14: end function
```

3.2 知識ベースモデル: LMKB

命題 p と知識ベース K が与えられたとき，p の真偽値が分かれば True, False のいずれかを，そうでなければ Unknown を返すモデルを構築する．

本研究では，K はウェブ上に存在するウェブページ集合とし，ウェブコーパスを用いて事前訓練された LLM を LMKB として用いる．具体的には，§A.1 のプロンプトの雛形を用いて言語モデルに p の真偽を問う．得られた応答の最初のトークンの生成確率分布から True と False の生成確率を取り出した上で正規化し，p の真偽値 Y の確率分布 $\pi(Y)$ とする．

予測の確信度については，LLM の生成確率に基づく確信度と正解率の間に高い相関があることが知られているため [7]，Y の正規化エントロピー $\text{NE}(Y) = -\sum_{y \in \{\text{True,False}\}} \pi(Y = y) \log \pi(Y = y) / \log 2 \in [0, 1]$ を計算し，$1 - NE(Y)$ を予測の確信度とする．最終的には，確信度が閾値 τ 以上ならば真偽値を返す．

$$\text{LMKB}(p) = \begin{cases} \underset{y \in \{\text{True,False}\}}{\text{argmax}} \ \pi(Y = y) & \text{if } 1 - \text{NE}(Y) \geq \tau \\ \text{Unknown} & \text{otherwise.} \end{cases} \quad (2)$$

3.3 間接証明生成モデル: ProofGen

命題 p が与えられたとき，p と論理的な関係 ϕ にある命題の集合 $Q = \{q_1, q_2, ..., q_n\}$ を求める．ϕ に様々なバリエーションを持たせることにより，

表 1　BoolQA の精度．下段は参考値．

モデル	精度
GPT-4 (Zero-shot)	74.8 (342/457)
GPT-3.5 (Zero-shot)	67.2 (307/457)
PaLM 540B (5-shot) [14]	73.9

様々な間接証明戦略を用いることができるが，今回は $\phi = q_1 \wedge q_2 \rightarrow p$ に固定し，$\phi, q_1, q_2 \vdash p$ より p を True と帰結する戦略のみを採用する．図 1 における *Moon Jae-in was born outside of Khanbaliq* の分解はその一例である．その他の戦略は §B を参照のこと．

p から Q を生成するために，§A.2 の雛形を用いてプロンプトを作成し，LLM に入力する．p が False である場合には，架空の事実を含む十分条件を生成する必要があるため，プロンプトでは，生成される命題 q_1, q_2 は偽の事実 (false facts) であることを許している．最後に，LLM の生成結果からパタンマッチにより q_1, q_2 を抽出し，ProofGen(p) の出力を $(\phi, Q) = (q_1 \wedge q_2 \rightarrow p, \{q_1, q_2\})$ とする．

4　評価実験

本節では，研究の大前提となる LMKB 単体の性能を確認し，LMKB の予測が不確実な場合の間接証明へのバックオフの効果を検証する．

4.1　実験設定

データセット　StrategyQA [1] (BigBench 版) [1)] の Validation Split 457 問を用いる．StrategyQA は，(a) 複数の知識を統合して答えるマルチホップ QA であり，(b) 解答に必要な知識が質問からは分からない質問を集めたデータセットであり，本研究の実験に適していると考えられる．GPT-4 により質問文を肯定形に変換して実験に利用する．

モデル　LMKB と ProofGen を実装するための LLM として，GPT-4 (gpt-4-1106-preview)，または GPT-3.5 (gpt-3.5-turbo-1106) を OpenAI API 経由で用いる[2)]．実験では，以下のモデルの性能を比較する．

- **BoolQA**: LMKB に真偽値を強制的に出力させる（提案手法 $\tau = 0$ と等価）．
- **BoolQA+Unk**: LMKB の検証結果をそのまま使う（提案手法 $d = 0$ と等価）．
- **Back-off LMKB**: 提案手法．再帰の深さの上限 $d = 1, d = 2$ それぞれを評価する．

1)　https://huggingface.co/datasets/tasksource/bigbench
2)　top_p=1.0, temperature=1.0 とした．

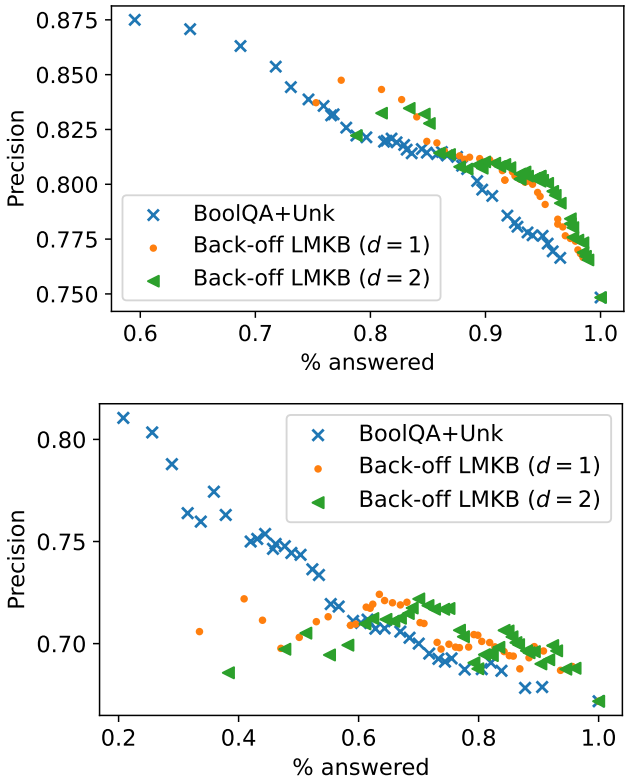

図 2　閾値 τ を変化させた場合の解答率と精度の変化（上: GPT-4, 下: GPT-3.5）．

評価指標　それぞれのモデルは「解答しない (Unknown)」という選択肢を取ることができ，解答対象の問題数が異なるため，性能を互いに比較することが難しい．そこで，Selective Classification タスク [15] のもとでモデルを評価する．すなわち，閾値 τ を変化させ，**解答率**と**精度**の変化をプロットする．本実験では τ を 0.025 きざみで 0.0-1.0 まで変化させ，解答率 = $\frac{\text{モデルが真偽値を返した問題数}}{\text{全問題数}}$，精度 = $\frac{\text{モデルが正しく真偽を予測した問題数}}{\text{モデルが真偽値を返した問題数}}$ とする．これにより，同一解答率の上での精度比較が可能になる．

4.2　実験結果

BoolQA の精度を表 1，BoolQA+Unk, Back-off LMKB の解答率と精度の関係を図 2 に示す．

LMKB は十分な精度を持ち，予測の不確実性を推定できるか？　表 1 より，LMKB は既存研究と比べても遜色ない精度で真偽検証ができることが確認できた．また，図 2 より，両言語モデルにおいて BoolQA+Unk の解答率と精度の間に負の相関が確認できた．これは，正規化エントロピーに基づく確信度を用いた LMKB が，予測の不確実性を適切に推定できていることを意味する．

表 2 LMKB が Unknown と答えた事例における，間接証明に基づく真偽検証の精度 (GPT-4)． % True は正解ラベルが True の問題の割合を示す．

(d, τ)	事例数	解答数	精度	% True
$(1, 0.25)$	43	23	71.4	62.8
$(2, 0.35)$	55	32	71.9	61.8

不確実な予測に対する間接証明へのバックオフは有効か？ 図 2 より，同一の解答率における BoolQA+Unk と提案手法の精度を比べると，GPT-4 ではすべての解答率において，GPT-3.5 では解答率 60%以上において，提案手法がより高い精度を得られた．これより，予測が不確実である場合の間接証明へのバックオフの有効性が確認できた．

さらに，図 2 より，$d = 1$ と $d = 2$ を比べると，$d = 2$ の精度が全体的に高い傾向にあることがわかる．これより，慎重な真偽検証を再帰的に行うことの有効性も確認できた．

図 3 に，BoolQA は False，BoolQA+Unk は Unknown と判断するところを，間接証明へのバックオフにより True と結論を上書きし，正解できた事例を示す．

間接証明による真偽判断はどの程度正確か？ GPT-4 の解答率 95%付近の実験結果について，間接証明に基づく真偽判断の性能を評価した．LMKB が Unknown と解答した事例における，間接証明に基づく真偽判断の精度を表 2 に示す．True を正解とする問題の割合はそれぞれ 62.8%, 61.8% 程度であり，Unknown を単純に True または False に倒す場合に比べて，10%以上高い精度で真偽判断を実現できていることになる．以上より，間接証明へのバックオフの効果を改めて確認できた．

4.3 間接証明のエラー分析

今後の課題を探るため，表 2 の $(d, \tau) = (2, 0.35)$ の結果において，間接証明が真偽判断を誤った事例 20 件を人手により分析した．

予測 = True，正解 = False: 9 件 6 件/9 件は，ProofGen は期待通りに動作していたが，LMKB の真偽判定誤りに起因するエラーが 3 件，LMKB の真偽判定は正しいが，データセットの質問がそもそも曖昧であり，True も正解といえるケースが 3 件あった．残りの 3 件/9 件は，ProofGen のエラーにより，生成された命題の集合が入力命題の十分条件でないという事例であった（図 4）．

予測 = Unknown，正解 = True: 11 件 9 件/11 件については，ProofGen にエラーはないが，5 件/9 件については，LMKB が少なくとも 1 つの命題を False または Unknown と判断したために，十分条件が成立していると判定されずに True と帰結しなかったケースであった (図 5)．残りの 4 件/9 件については，LMKB の真偽判定誤りにより True と帰結されなかったケースであった．残り 2 件/11 件は，前述同様，生成された命題の集合がそもそも十分条件になっていないものであった．

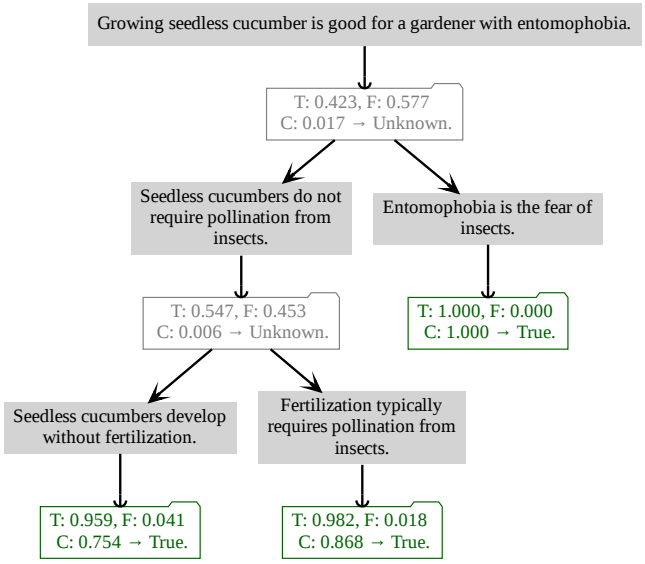

図 3 間接証明へのバックオフが有効な例．“T: ... F: ...” は LMKB の予測確率分布，“C:” は確信度，“→” は最終的な真偽検証結果を示す（式 2 による）．

5 おわりに

本稿では，LM as KB の信頼性向上を目指し，予測の不確実性を間接証明によるバックオフで補うアプローチを検討した．§4.3 のエラー分析により，大きく二つの課題が見えてきた．

第一に，LMKB の真偽検証の精度，及び予測の不確実性の推定精度の問題である．予測の不確実性の指標を，知識ベースにある命題とそうでない命題を明確に切り分けられる統制された実験条件下で研究していく必要がある．

第二に，間接証明戦略の不十分さである．現状の提案手法では，与えられた命題に対して一つの特定の形の証明を試みているが，これにより誤った Unknown の予測を生み出してしまった．一つの証明戦略について複数の命題集合を生成する，証明戦略を増やすなどの対策が必要である．また，最適な証明戦略の自動選択や，確実に真偽検証できる方向に探索的に証明を展開するなど，証明探索の研究なども今後の課題となる．

謝辞

本研究は JSPS 科研費 19K20332 の助成を受けたものです．所属研究室の坂井吉弘氏，趙羽風氏には論文に関して有益な助言をいただきました．両氏に感謝いたします．

参考文献

[1] Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. **Transactions of the Association for Computational Linguistics**, Vol. 9, pp. 346–361, 2021.

[2] 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.

[3] Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. How can we know when language models know? on the calibration of language models for question answering. **Transactions of the Association for Computational Linguistics**, Vol. 9, pp. 962–977, 2021.

[4] Lovisa Hagström, Denitsa Saynova, Tobias Norlund, Moa Johansson, and Richard Johansson. The effect of scaling, retrieval augmentation and form on the factual consistency of language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 5457–5476, Singapore, December 2023. Association for Computational Linguistics.

[5] Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. A survey of language model confidence estimation and calibration, 2023.

[6] 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.

[7] Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 5433–5442, Singapore, December 2023. Association for Computational Linguistics.

[8] Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 9004–9017, Singapore, December 2023. Association for Computational Linguistics.

[9] Pruthvi Patel, Swaroop Mishra, Mihir Parmar, and Chitta Baral. Is a question decomposition unit all we need? In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**, pp. 4553–4569, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[10] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 12076–12100, Singapore, December 2023. Association for Computational Linguistics.

[11] Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. Decomposed prompting: A modular approach for solving complex tasks. In **The Eleventh International Conference on Learning Representations**, 2023.

[12] Jiajie Zhang, Shulin Cao, Tingjian Zhang, Xin Lv, Juanzi Li, Lei Hou, Jiaxin Shi, and Qi Tian. Reasoning over hierarchical question decomposition tree for explainable question answering. 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. 14556–14570, Toronto, Canada, July 2023. Association for Computational Linguistics.

[13] Zaid Khan, Vijay Kumar BG, Samuel Schulter, Manmohan Chandraker, and Yun Fu. Exploring question decomposition for zero-shot vqa, 2023.

[14] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways, 2022.

[15] Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, **Advances in Neural Information Processing Systems**, Vol. 30. Curran Associates, Inc., 2017.

A　LLM に与えるプロンプト

A.1　LMKB

For the following question, provide your best guess. Give ONLY the guess. No other words or explanation. $\{p\}$. True or False?

A.2　ProofGen

To conclude a statement X, what premises are needed? Write two atomic premises P and Q.
- P and Q can be false facts. P and Q should contain all the world knowledge required to prove X. P and Q should be simple sentences consisting of only subject-verb-object.
- Write specific statements. Do not use pronouns. Use specific nouns. Give ONLY answer. No other words or explanations.
For example: P: <Premise 1> Q: <Premise 2>
X: $\{p\}$

B　その他の証明戦略の例

- Modus Tollens: $\phi = (p \to q_1)$ とし，$\phi, \neg q_1 \vdash \neg p$ より p を偽と帰結する．
- Disjunctive Syllogism: $\phi = (p \vee \neg p)$ とし，$\phi, \neg p \vdash p$ より p を真と帰結する．

C　間接証明のエラー

図 4，図 5 に，間接証明のエラーに起因する真偽検証誤りを示す．

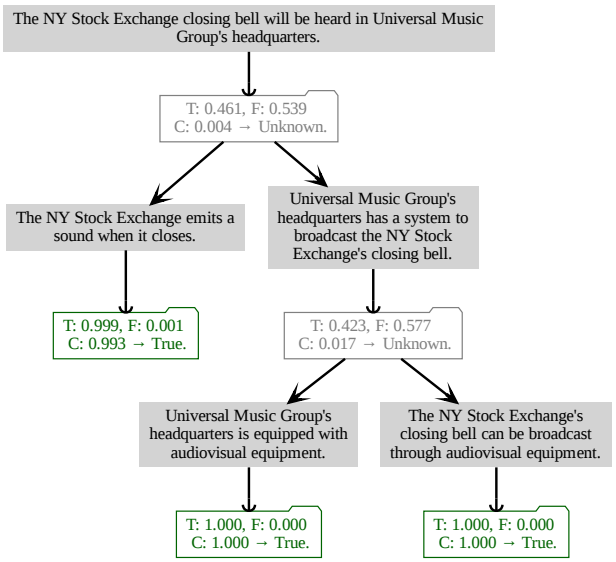

図 4　ProofGen のエラーの例（右下）．「UMG の本社にはオーディオ機器がある」「NYSE の closing bell はオーディオ機器で放送できる」だけでは「UMG の本社には NYSE の closing bell を放送するシステムがある」とは言い切れない．

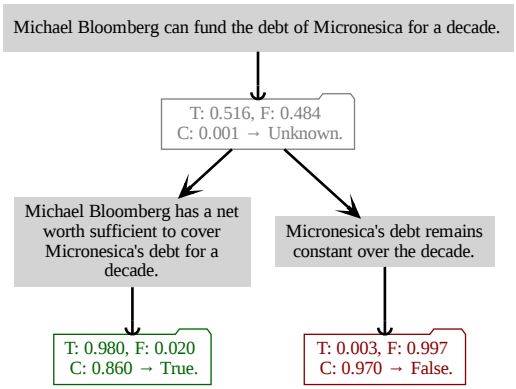

図 5　LMKB により生成された十分条件は正しいが LMKB が False を返したため，証明が失敗し，全体としては Unknown を返した．