LLM の生成テキストの真偽検証のための日本語真偽判定データセットの構築

政野 美和 [1,2]　清丸 寛一 [2]　欅 惇志 [1]　堀尾 海斗 [3]
源 怜維 [2,3]　欅 リベカ [4]　中山 功太 [2]　橘 秀幸 [2]　河原 大輔 [2,3]
[1] 一橋大学 ソーシャル・データサイエンス学部
[2] 国立情報学研究所 大規模言語モデル研究開発センター
[3] 早稲田大学理工学術院　[4] 東京工科大学コンピュータサイエンス学部
5123053k@g.hit-u.ac.jp　{kiyomaru,nakayama,h_tachibana}@nii.ac.jp
a.keyaki@r.hit-u.ac.jp　keyakirbk@stf.teu.ac.jp
{kakakakakakaito,ray}@akane.waseda.jp　dkw@waseda.jp

概要

LLM の生成テキスト中の情報の真偽を検証するシステムの開発が進められている．真偽検証システムの真偽判定ステップでは，検索によって取得した根拠テキストによって生成テキストの情報が支持されるかを判定する．本研究では，厳密な真偽判定のみでは捉えきれない言説と根拠テキストの関係を扱うため，真偽判定ラベルセットを拡張して実用的な日本語真偽判定データセットの構築に取り組む．構築したデータセットに基づき，プロンプトベースの真偽判定手法の性能を評価した結果，推論の流れを指示するプロンプトを用いることで性能が改善することが明らかになった．

1　はじめに

大規模言語モデル（Large Language Model; LLM）の生成テキストにはハルシネーション（hallucination）と呼ばれる，もっともらしく見える誤情報が含まれることがある [1]．こうした背景のもと，真偽検証システム [2, 3, 4, 5] を LLM の生成テキストに対して応用する研究・開発が進んでいる [6]．我々の研究グループでも，LLM の日本語生成テキストを対象とした真偽検証システム [7, 8] を構築している．真偽検証システムは言説分解，根拠検索，真偽判定という三つのステップで構成される．言説分解では入力テキストを，一つの物事に関する性質や関係を表す独立した最小粒度の情報単位である言説に分解する．次に，根拠検索では分解された言説をクエリとして，根拠となる情報が格納された根拠データベースを検索し，各言説に対する根拠テキストを取得する．最後に，図 1 に示す真偽判定のステップでは，言説と取得した根拠テキストのペアに対して自然言語推論（Natural Language Inference; NLI）を行うことで，言説が根拠テキストによって支持（含意）されるかどうかを判定する．

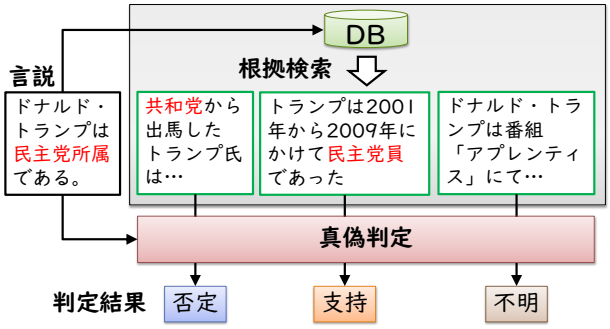

図 1　真偽判定の概要図

本論文ではこのうち，真偽判定に着目する．真偽判定は真偽検証システムの最終的な判定結果に直接影響するプロセスであり，システムの性能を実質的に担っている．真偽検証システムにおける高性能な真偽判定手法の構築には，言説・根拠テキスト・判定ラベルを含む評価用データセットが必須である．

本研究では我々が構築した言説分解データセット [8] に含まれる，生成テキストから分解された言説に対して真偽判定の評価用データセットを構築した．データセットの構築にあたり，言説と根拠テキストの関係を柔軟に評価するために，既存研究 [6] のラベル（支持，部分支持，否定，不明）に加えて，新たなラベル（推定支持，推定否定）を導入した．これらのラベルを用いた真偽判定アノテーション

のためにガイドラインを策定し，これに基づいて言説・根拠テキストペアに真偽判定ラベルを付与することでデータセットを構築した．

GPT-4o のプロンプトベースの真偽判定手法を構築し，作成したデータセットに基づいて性能を検証した．その結果，真偽判定の手順を明示的に記述した Chain-of-Thought プロンプトを使用することで，性能向上に効果があると確認された．

2 関連研究

真偽検証に関する研究は従来からフェイクニュースなどを対象として取り組まれていた [2, 3, 4]．その後，真偽検証のための代表的なデータセットである FEVER [9] やその後続のデータセット [10, 11, 12] が登場して，真偽検証システムの研究を牽引している．なお，これらのデータセットでは，真偽判定のラベルは，支持 (supported)，否定 (refuted)，不明 (not enough information) が採用されている．WiCE [5] は言説単位の真偽判定ラベルに加えて，トークン単位の真偽判定ラベルも付与しているため，より細かい粒度での真偽検証が可能となる．WiCE では Wikipedia の記事内から言説を抽出し，その文が引用しているウェブ記事を根拠テキストとして取得している．真偽判定ラベルには，支持 (supported)，部分支持 (partially-supported)，不支持 (not-supported) が用いられている．

LLM の生成テキストを対象とした真偽検証のためのフレームワークである Factcheck-Bench [6] は，前述の真偽検証システムより細かいステップから構成される真偽検証フレームワークの提案とベンチマークの構築を行っている．Factcheck-Bench で採用されている真偽判定ラベルは，支持 (support)，部分支持 (partially support)，否定 (refute)，無関係 (irrelevant) の 4 種類である．ただし，無関係とは，根拠テキストが言説に関連する内容を一切含まない場合を表すラベルである．

3 データセット構築

真偽判定の定量評価および分析のため，新たにデータセットを構築する．データセットの各事例は，言説，根拠テキスト，真偽判定ラベルの三つ組で構成される．

3.1 真偽判定ラベル

採用するラベルセットは以下の通りである：

- **完全支持**：言説全体を支持する直接的な言及が根拠テキストに含まれる
- **推定支持**：言説全体を支持すると推察される言及が根拠テキストに含まれる
- **部分支持**：言説の主要部を支持する言及が根拠テキストに含まれる
- **完全否定**：言説と矛盾する直接的な言及が根拠テキストに含まれる
- **推定否定**：言説と矛盾すると推察される言及が根拠テキストに含まれる
- **不明**：言説を支持する言及も言説と矛盾する言及も根拠テキストに含まれない

完全支持と完全否定はそれぞれ NLI における含意と矛盾に対応する．推定支持，部分支持，推定否定，不明は NLI では一括りに中立のラベルが与えられるが，本データセットでは柔軟な評価を目的として区別して扱う．例 (1) に推定支持の例を示す．

(1) **言説**：天平文化は聖武天皇の時代に最盛期を迎えました
根拠テキスト：天平文化は聖武天皇のころの文化です
真偽判定ラベル：推定支持

例 (1) の根拠テキストは言説を直接的に支持する内容ではない．しかし，語用論的観点から見ると，根拠テキストの「〜のころの文化」という表現は当該文化の中心的な時期を指すのが普通であり，成立期や衰退期を指して用いられるとは考えづらい．そのため，人間の自然な解釈として，聖武天皇の統治期が天平文化の最盛期と重なっていたと推論することが合理的である．このような事例を不明と区別してラベル付けすることで，論理的・形式的に厳密な真偽判定と，人間の自然な解釈に基づく真偽判定の双方に対する性能評価ができるようになる．

例 (2) には部分支持の例を示す．

(2) **言説**：ボブサップはアメリカ合衆国コロラド州出身です
根拠テキスト：ボブサップはアメリカ合衆国出身です
真偽判定ラベル：部分支持

このように言説全体を完全には支持できないが，言説が含む付属的な詳細を明確に否定する根拠も存在しない場合を部分支持として区別することで，情報

粒度の違いを考慮した真偽判定の評価が可能となる．なお，否定に関しては言説の一部が矛盾していれば言説全体も矛盾するため，部分否定のラベルは考えない．

3.2 アノテーション

我々が過去に構築した日本語言説分解データセット [8] を拡張する形でデータセットを構築した．日本語言説分解データセットは，日本語 QA データセット「AI 王 Version 2.0[1)]」(AIO) の問題，日本語対話データセット「LLM-jp Chatbot Arena Conversations[2)]」(CBA) のユーザ発話を入力とし，日本語 LLM「LLM-jp-3 13B Instruct[3)]」が出力した生成テキストに対して，人手で言説分解のアノテーションを行ったものである．

真偽判定のアノテーションを付与するにあたり，まず各言説に対する根拠テキストを取得した．根拠テキストの情報源として，LLM-jp コーパス v3[4)] を用いた．言説との関連性を保ちつつ多様な文書を取得するため，BM25 [13] に基づく全文検索の結果を Maximal Marginal Relevance によってリランキング [14] し，その上位 5 件を取得した．また，LLM への入力テキストは，生成テキストの根拠となる情報を含む場合がある（例：要約の対象テキスト）．このことを考慮し，検索によって取得した 5 件の文書に各言説に対応する LLM への入力テキストを加えた計 6 件を当該言説に対する根拠テキスト集合とした．

各言説と根拠テキストのペアについて，人手で真偽判定ラベルをアノテーションした．まず 2 名のアノテータが一次作業者として各ペアに対して独立に真偽判定ラベルを付与した．両者の判断が一致しない場合には，第三のアノテータがそれらを参照した上で最終的なラベルを決定した．アノテータが対象データを理解できない場合（例：理解に化学の専門知識が必要な場合など）には，「理解不能」のラベルを付与させた．理解不能のラベルが付与された事例は評価データセットから除外した．加えて，不注意によるアノテーションミスを防ぐため，支持または否定のラベルを付与する場合には，その根拠となる根拠テキスト中のスパンも合わせてアノテーションを行った．これは真偽判定の根拠まで含めて提示する真偽検証システムを構築した際の評価データとしても活用可能である．

表 1 データセットのラベル分布

	AIO		CBA	
ラベル	件数	割合	件数	割合
完全支持	3,983	0.292	1,798	0.124
推定支持	1,198	0.088	1,033	0.071
部分支持	555	0.041	436	0.030
完全否定	173	0.013	62	0.004
推定否定	70	0.005	64	0.004
不明	7,663	0.562	11,112	0.766

表 1 に，データセットの真偽判定ラベルの分布を示す（「理解不能」を除く）．また，2 名の一次作業者のアノテーションにおける Cohen の κ 係数は，AIO が 0.44，CBA が 0.34 であった．この結果は，真偽判定が言説と根拠テキストの関係を推論に基づいて解釈することがあり，判断が必ずしも一意に定まらない場合があるという，真偽判定タスクのアノテーションの難しさを反映している．

構築したデータセットをラベル分布が近くなるよう配慮した上で，開発データとテストデータに 1:1 の比率で分割した．その結果，AIO の開発データは 6,818 件，テストデータは 6,824 件，CBA の開発データは 7,256 件，テストデータは 7,249 件となった．

4 プロンプトによる真偽判定

本節では，構築した真偽判定データセットを用いて，プロンプトベースの手法の性能評価を行う．

4.1 実験設定

4 種類のプロンプトを設計し，完全支持，部分支持，推定支持，完全否定，推定否定，不明の 6 ラベルを付与する実験を行った．

まず，タスクの指示と言説とラベルの定義のみを含めたプロンプトを **base** プロンプトとして設定する．その他に，代表的なプロンプト手法である few-shot プロンプト，Chain-of-Thought (CoT) プロンプトの性能評価を行った．few-shot では base の内容に加え，言説，根拠テキスト，正解ラベルを組にした例を複数提示する．本研究では開発データから抽出した例を用いて，**6-shot** と **12-shot** の 2 種類を設定し，提示例の数と判定性能の関係を分析した．

1) https://sites.google.com/view/project-aio/dataset
2) https://huggingface.co/datasets/llm-jp/llm-jp-chatbot-arena-conversations
3) https://huggingface.co/llm-jp/llm-jp-3-13b-instruct
4) https://gitlab.llm-jp.nii.ac.jp/datasets/llm-jp-corpus-v3

表 2　プロンプトベースの真偽判定

プロンプト	正解率	適合率	再現率	F1
base	0.669	0.408	0.475	0.418
+ 6-shot	0.650	0.387	0.482	0.402
+ 12-shot	0.653	0.388	0.482	0.401
+ CoT 6-shot	**0.704**	**0.440**	**0.507**	**0.459**

表 3　各プロンプトのラベルごとの F1

ラベル	base	6-shot	12-shot	CoT
完全支持	0.754	0.760	0.762	**0.773**
推定支持	0.157	0.231	0.223	**0.278**
部分支持	0.202	0.191	0.201	**0.299**
完全否定	0.435	0.391	0.386	**0.481**
推定否定	**0.129**	0.068	0.071	0.077
不明	0.817	0.771	0.773	**0.818**

CoT では段階的に推論を行うため，base の内容に追加して，推論を行う手順と指示に従って段階的に推論した例を提示する．本研究では著者らが，6-shot で利用している例に対して推論過程を作成した．推論過程を指示するプロンプトを付録 A に示す．

真偽判定の性能評価には，モデルが付与したラベルをデータセットの正解ラベルと比較し，正解率，適合率，再現率，F1 それぞれのマクロ平均を用いた．実験に用いた LLM は gpt-4o[5]である．AIO のテストデータに対して各手法を 3 回ずつ実行し，その平均値を評価した[6]．

4.2　実験結果

表 2 に示す通り，4 つのプロンプトの中で CoT の判定性能が最も高い結果となった．表 3 に各手法のラベルごとの F1 を示す．CoT は 6-shot と比較して全てのラベルで改善傾向を示しており，特に部分支持と完全否定において大きな改善が見られた．図 2 は CoT による出力のうち，1 回目の試行の結果について混同行列を作成したものである．図 2 と 6-shot の混同行列との比較に基づく分析から，推定支持や完全支持を誤って部分支持と予測する傾向が弱まり，不明を誤って完全否定と予測するケースが減少しているため，2 つのラベルの性能が改善していることが分かった．

base に例を追加した few-shot (6-shot, 12-shot) は，

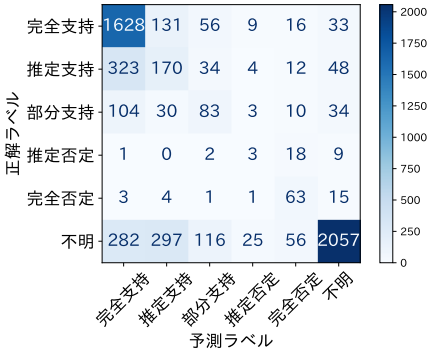

図 2　CoT プロンプトの混同行列

いずれも性能が低下する結果となった．詳細な分析は付録 B.1 に示す．

また，6-shot から 12-shot に増やしても性能は改善しなかった．few-shot で性能改善が限定的だった原因に関する詳細な分析は今後の課題である．

表 3 によると，最も性能がよい CoT でも，特に予測性能が低かったラベルは推定否定である．図 2 より，推定否定ラベルのエラーには，推定否定を完全否定と予測している誤りや，不明を推定否定と予測している誤りが多く存在することが分かる．

推定否定を完全否定と予測している事例には，前提条件が明示されていないにもかかわらず，同一対象に関する記述と仮定して判定しているケースが見られた．また，文の解釈の曖昧性を考慮せずに判定しているケースも確認された．不明を推定否定と予測している例としては，根拠テキストで言及されていない対象は存在しないものとして，短絡的な結論を述べているケースが見られた．それぞれの具体例を付録 B.2 に掲載する．

また，付録 C の否定系ラベルの再現率と，支持系ラベルの適合率の分析の結果，CoT が本真偽検証システムのハルシネーション検出という目的に照らしても有効な手法であることが示唆された．

5　おわりに

本研究では，LLM の生成テキストに対する真偽検証システムの構築のために，日本語真偽判定データセットを構築した．また，推論過程を追加したプロンプトにより，真偽判定性能が向上することが確認された．一方で，曖昧性を含む事例では誤りが多く，今後は，アノテーションガイドラインの改善やデータセットの拡充とともに，プロンプト設計の改善を通して，さらなる性能向上を目指す．

5）https://platform.openai.com/docs/models/gpt-4o
6）設計したプロンプトは LLM への入力テキストに対しては適切に機能しなかったため，以降の実験では除外した．

謝辞

本研究成果の一部は，データ活用社会創成プラットフォーム mdx を利用して得られたものである．また，本研究の一部は，JSPS 科研費（基盤研究 (B) (課題番号: 23H03686, 25K03178)，基盤研究 (C) (課題番号: 24K15066)，令和 7 年度次世代人工知能技術等研究開発拠点形成事業「生成 AI モデルの透明性・信頼性の確保に向けた研究開発拠点形成」，株式会社デンソーアイティーラボラトリとの共同研究の支援による．ここに記して謝意を表す．

参考文献

[1] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. **ACM Transactions on Information Systems**, Vol. 43, No. 2, pp. 1–55, 2025.

[2] Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. A Survey on Automated Fact-Checking. **Transactions of the Association for Computational Linguistics (TACL)**, Vol. 10, pp. 178–206, 2022.

[3] Xia Zeng, Amani S. Abumansour, and Arkaitz Zubiaga. Automated fact-checking: A survey. **Language and Linguistics Compass**, Vol. 15, , 2021.

[4] Neema Kotonya and Francesca Toni. Explainable Automated Fact-Checking: A Survey. In **Proc. of the COLING 2020**, 2020.

[5] Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. WiCE: Real-World Entailment for Claims in Wikipedia. In **Proc. of the EMNLP 2023**, 2023.

[6] Yuxia Wang, Revanth Gangi Reddy, Zain Muhammad Mujahid, Arnav Arora, Aleksandr Rubashevskii, Jiahui Geng, Osama Mohammed Afzal, Liangming Pan, Nadav Borenstein, Aditya Pillai, Isabelle Augenstein, Iryna Gurevych, and Preslav Nakov. Factcheck-Bench: Fine-Grained Evaluation Benchmark for Automatic Fact-checkers. In **Proc. of the Findings of the EMNLP 2024**, 2024.

[7] 政野美和, 欅リベカ, 欅惇志, 清丸寛一, 中山功太, 堀尾海斗, 源怜維, 橘秀幸, 河原大輔. LLM の生成テキストの真偽検証のための日本語言説分解データセットの構築. 第 265 回 自然言語処理研究発表会, 2025.

[8] 政野美和, 欅リベカ, 欅惇志, 清丸寛一, 中山功太, 堀尾海斗, 源怜維, 橘秀幸, 河原大輔. LLM の生成テキストの真偽検証のための日本語言説分解データセットの構築と評価. 言語処理学会第 32 回年次大会 (NLP2026), 2026.

[9] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In **Proc. of the NAACL 2018**, 2018.

[10] James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. The FEVER2.0 Shared Task. In **Proc. of the Second Workshop on Fact Extraction and VERification (FEVER)**, 2019.

[11] Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. The Fact Extraction and VERification Over Unstructured and Structured information (FEVEROUS) Shared Task. In **Proc. of the Fourth Workshop on Fact Extraction and VERification (FEVER)**, 2021.

[12] Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. The Fact Extraction and VERification Over Unstructured and Structured information (FEVEROUS) Shared Task. In **Proc. of the NeurIPS 2023**, 2023.

[13] Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. **Okapi at TREC-3**. British Library Research and Development Department, 1995.

[14] Jaime Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In **Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval**, pp. 335–336, 1998.

A CoT プロンプトの推論手順の指示

以下の手順で推論し、推論過程と解答を出力してください。
1. 言説の意味を解析し、主張内容を特定する
2. 関連テキストに言説と関連する内容が含まれるかを特定し、含まれる場合はその内容を抽出する。含まれない場合は推論を終了し、「不明」と判定して解答に進む。
3. 関連テキストから抽出した内容が言説に対してどのような立場（「完全否定」「推定否定」「完全支持」「推定支持」「部分支持」「不明」）であるかを判定する。
4. 「完全否定」「推定否定」「完全支持」「推定支持」「部分支持」「不明」のいずれに該当するか判定する

B 結果の分析

B.1 base と few-shot のラベル別分析

表 2 に示す通り， few-shot は base と比較して性能改善が見られなかった．ラベルごとの予測性能を見ると，few-shot では，base において目立っていた推定支持を部分支持と予測する誤りが減少し，推定支持の性能が改善していた．一方で，推定否定，完全否定，不明については base, few-shot の両者において，推定否定を完全否定と予測する誤りや，不明を推定否定，完全否定と予測する傾向が見られ，few-shot ではこれらの誤りが増加していた．

B.2 推定否定ラベルのエラー分析

推定否定を完全否定と予測している事例の分析から，以下の２つのようなケースが確認された．

1 つ目は，言説と根拠テキストの前提条件や指示対象の曖昧性を考慮せず，同一の対象について述べていると仮定して判定してしまうケースである．例えば，言説が「FIFA ワールドカップ」について述べている一方，根拠テキストには「ワールドカップ」という用語のみが用いられている場合，アノテータは「ワールドカップ」が「FIFA ワールドカップ」ではなく，他競技の大会である可能性を考慮して推定否定と判定している．一方で，プロンプトによる LLM の判定は主語の曖昧性を検討せず，同一対象への言及と解釈して完全否定と判定していた．

表 4 否定系ラベルの再現率と支持系ラベルの適合率

	否定系の再現率	支持系の適合率
base	0.3898	0.3657
+ CoT 6-shot	0.4075	0.4162

２つ目は，文の解釈の曖昧性を考慮せずに判定してしまうケースである．例えば，言説中の「火星には現在 4 つの衛星が確認されています」という記述に対し，「フォボスとデイモスは、火星の 2 つの衛星の名前にもなっています」と述べる根拠テキストが与えられた場合である．根拠テキストの表現では，火星の衛星の数は 2 つなのか，2 つ以上なのかは曖昧だが，LLM は火星の衛星は 2 つであると解釈し，完全否定と判定している．

また，不明を推定否定と誤判定した事例として，根拠テキストで言及されていない対象を存在しないものとみなし，短絡的な結論を述べるケースが見られた．例えば，「ドストエフスキーは『父と子』という作品を残した」という言説に対し，ドストエフスキーの作品をいくつか列挙した根拠テキストが与えられたときに，根拠テキスト中に『父と子』という作品名が登場しないことのみを理由に，推定否定と判定している例が確認された．

C ハルシネーション検出を考慮した評価

本論文では，手法間の性能差を主に F1 に基づいて議論してきた。しかし，本研究で構築している真偽検証システムの目的がハルシネーション検出であることを踏まえると，否定系ラベル（完全否定・推定否定）の再現率が重要な評価指標となる．また，否定と判定すべき事実を支持と誤判定することは，ハルシネーションを見逃すことになり，重大な問題である．そのため，支持系ラベル（完全支持・推定支持・部分支持）については適合率が重要である．

以上の考えに基づき， F1 に基づく性能が最も高かった CoT について，否定系ラベルの再現率と支持系ラベルの適合率をそれぞれ算出し，その算術平均を評価した．base についても同様の評価を行い，結果を比較した．表 4 の結果を見ると， CoT はいずれの指標においても base を上回っており，ハルシネーション検出という本システムの目的に照らして，有効な手法の一つであることが示唆される．しかし，その改善幅は限定的であるため，今後はさらなる真偽判定手法の性能改善に取り組む必要がある．