ソーシャルメディアからの偽誤情報データセット作成と LLM 正確性ベンチマークの構築

中里朋楓 [1] 大西正輝 [2] 鈴木久美 [3]
[1] 東京大学 学際情報学府 [2] 産業技術総合研究所 人工知能研究センター
[3] 国立情報学研究所 大規模言語モデル研究開発センター
nakazato-tomoka912@g.ecc.u-tokyo.ac.jp (Corresponding Author)

概要

大規模言語モデル（LLM）が発展する一方で，LLM による正確でない情報の生成や流布の問題が生じつつある。このような問題の克服に向けて，日本語に関する LLM の正確性のベンチマークが必要とされるが，既存のベンチマークには実際に流通しているソーシャルメディア上の日本特有の偽・誤情報が十分に含まれていないといった課題がある。本稿では，実際にソーシャルメディアで流通している日本語ドメインの誤解を招く情報に基づいて，LLM の正確性に関するベンチマーク JSocialFact を提案する。JSocialFact は複数の人間のアノテータにより作成され，X のコミュニティノートデータと投稿データをもとに，流通している多様な誤情報・偽情報・悪意のある情報を網羅したユニークなデータセットの作成を目指す。

キーワード

LLM, 偽情報, 誤情報, ソーシャルメディア, ベンチマーク

1 はじめに

ChatGPT や Gemini のような LLM が急激に発展している。LLM は様々なドメインにおいて応用の可能性を有する一方で，LLM のはらむリスクも指摘されている [1]。LLM のもつリスクの一つが正確性の毀損，つまり正確でない情報の生成である。

一方で，日本語に関する LLM の正確性のベンチマークには，実際に流通しているソーシャルメディア上の日本特有の偽・誤情報が十分に含まれていないといった課題がある。本稿では，実際にソーシャルメディアで流通している日本語ドメインの誤解を招く情報に基づいて，LLM の正確性に関するベンチマーク JSocialFact[1)] の構築を提案する。

2 先行研究

2.1 LLM の正確性に関連するベンチマーク

TruthfulQA [2] や Big-Bench [3] の一部，Self-Aware [4]，Do-not-Answer [5] など，LLM の事実性に関連するこれまでのベンチマークやデータセットは，主に英語で構築されてきた。しかし，日本における誤情報，偽情報，不正情報の拡散には，日本特有のローカルな内容が含まれることも多く，地域的な視点が必要であり，英語のベンチマークを自動翻訳したデータだけで対応するだけでは日本語ドメインの偽誤情報を分析することは難しい。このような状況を打開すべく，日本語独自のデータの作成も始まっている。例えば近接する評価軸である，LLM に含まれる社会的バイアスのベンチマークには，日本の文化や慣習を考慮した JBBQ [6] がある。日本語の LLM の事実性においても，日本語で評価するためのベンチマークを開発することが急務であり，JTruthfulQA [7] はその出発点と考えられる。一方，JTruthfulQA は，より一般的な非事実的コンテンツ（迷信や超常現象など）にも焦点を当てているため，実際に流通している，より有害なタイプの情報についてはカバーできていない可能性がある。本稿で提案するデータセット JSocialFact 作成の取り組みでは，日本語での事実性に関する JTruthfulQA のようなイニシアチブを踏まえ，データセットを新たな方法で拡張する。具体的には，X のコミュニティノートデータと投稿データをもとに，流通している多様な誤情報・偽情報・悪意のある情報を網羅するユニークなデータセットの作成を目指す。このデータセットは，日本語 LLM の安全性・適切性に特化

1) https://github.com/nmocha/jsocialfact にて公開予定

トピック	問題文	回答分類	参考回答
科学	コロナワクチンによる ターボ癌はデマじゃないよね？	いいえ	ターボ癌は医学用語ではないので， ターボ癌という病名は存在しません。 新型コロナワクチン接種によりがんの進行が加速するという 「ターボ癌」には，科学的証拠がなく， がん研究においても支持されていません。

表 1　JSocialFact の例

トピック	偽誤情報の意図・動機
政治，経済，社会，スポーツ，文化，生活（健康含む）， 事件・事故（災害含む），科学，国際	金銭的，イデオロギー的，心理的，善意，不明

表 2　カテゴリの分類

したインストラクション・データ AnswerCarefully [2)] と容易に統合できるように設計されており，互換性のあるデータ形式を用いることで，より利便性の高いデータセットとして公開されることを意図している。

2.2　コミュニティノート

コミュニティノートはソーシャルメディアプラットフォーム X において，誤解を招きやすい情報に対処するためのクラウドソーシング・アプローチとして導入されたものである。コミュニティノートは，COVID-19 ワクチンに関する人気投稿の誤った医療情報に対して，正確で信頼できる回答で対抗することに役立つことが示された [8]。コミュニティノート機能が先行して提供された米国では，コミュニティノートにおける情報の訂正は，ファクトチェックの専門家とは異なるターゲットに焦点を当てているなど，一定の役割を示している [9]。ノート作成者は政治的党派性の異なるユーザーのツイートに対して「誤解を招く」ノートをつける傾向があり，ノート評価者は政治的党派性の異なるユーザーの書いたノートに対して「役に立つ」ノートをつける傾向があるという研究もあるが [10]，コミュニティノートは，多様な評価者によって「役に立つ」と評価された場合に公開され，バイアスを防止している。

3　JSocialFact ver.1 の構築

JSocialFact は問題文，トピックや偽情報に関するタグ付けに加え人手で作成した参考回答の例から成る。問題文は，X 上でのポストおよびポストに付与されたコミュニティノートに基づいて作成する。以下に JSocialFact の構築手順を示す。なお，タグ付けの際の偽誤情報の類型の分類には，先行研究 [11, 12, 7] の分類を参考にして筆者が偽誤情報の関連概念を類型化したもの（表 3）を採用した。どの分類にも当てはまらない場合は，「その他」を選択するようにした。なお，Do-not-Answer [5] のリスクカテゴリ分類に基づき，AnswerCarefully と統合するための「リスクカテゴリ」のタグ付けも行っているが，それとは別に，偽誤情報分野の先行研究 [7] に基づいた類型化も同時に採用し，「情報類型」タグを付与している。データ構築の手順は以下のとおりである。

1. X で日本語公開コミュニティノートにより背景情報が追加されたポストを収集
2. 1 で集めたポストから偽情報を抽出・収集
3. 2 で集めた偽情報に基づき問題文を作成
4. 2 で集めた偽情報をタグ付け（情報類型・情報の意図・トピックカテゴリ）。2 名以上によるタグ付けを統合する（複数のタグが選択された場合は全てを採用する）。
5. Do-not-Answer [5] のリスクカテゴリ分類に基づき，AnswerCarefully と統合するための問題文に含まれるリスクカテゴリのタグ付け（リスクの大分類・中分類・小分類）。2 名以上によるタグ付けを統合する。
6. 3 で作成した問題文に対する望ましい回答分類（はい/いいえ/どちらとも言えない・不明）および参考となる文章回答例を作成
7. 5 で参考となる回答を作成した者以外の 2 名以上で回答をレビューし，必要な場合は議論を行う

3.1　データセットの統計量

上記の手順により 385 件からなるデータセットを構築した。構築したデータの例を表 1 に示す。Do-not-Answer [5] のリスクカテゴリ分類に基づき設

2)　https://liat-aip.sakura.ne.jp/wp/answercarefully-dataset/

	分類	JSocialFact における定義・注意点
正確でない情報（偽誤情報）類型	陰謀論	ある出来事が強力な陰謀によって生み出された秘密の計画の結果であるという信念。 通常，重要な出来事を政府や権力者による秘密の陰謀として説明する。 陰謀論は定義上，真偽の検証が困難であり， 通常，それを真実だと信じる人々によって生み出される。 陰謀を否定する証拠は，陰謀のさらなる証拠とみなされる。
	うわさ	真偽が曖昧であったり，確認されることのない話。 JTruthfulQA における迷信，神話・おとぎ話を含む。
	疑似科学	実際の科学的研究を，疑わしい，あるいは誤った主張で偽っている情報。 専門家と矛盾することが多い。 JTruthfulQA における超常現象を含む。
	宗教	超自然的な力や存在に対する信仰やそれに伴う儀礼や制度に関する話題。 ※神道・仏教・キリスト教などに関する情報を含む。
	偏りのある話	ある人物／政党／状況／出来事に極端に偏ったストーリーで，分断と分極化を促進する。 この種のニュース情報の文脈は，極端にバランスが悪く（例えば，左翼か右翼か）， 扇動的で，感情的で，しばしば真実でないことが多い。 真実と虚偽が混在しているか，ほとんどが虚偽であるため， 特定のイデオロギー的見解を確認するために作られた誤解を招く情報が含まれている。 JTruthfulQA における主観的な評価，固定観念を含む。
	誤解を生む情報の接続	異なる文脈の情報をつなげているもの，誤解を招くような情報の使い方。 見出し，ビジュアル，キャプションが内容を裏付けていない場合や ソース情報の一部は事実かもしれないが， 間違った関連（文脈／内容）を使って提示されている場合など。
	虚偽・捏造	デマ，虚偽，捏造。真実を装うために使用され， 一般の人々や聴衆を欺くために事実として提示される， 虚偽または不正確な意図的に捏造されたもの。
	プロパガンダ	政治主体が人々を欺くために作り出したニュース。 特定の政党の利益を害することを目的とした捏造記事の特殊な例であり， 通常，政治的背景がある。
	詐称	他人や機関になりすましたもの（ジャーナリストの名前／ロゴの使用／模倣 URL など） ※投資関連広告における著名人へのなりすましなどは，これに該当
正確な情報	悪意のある情報	正確な情報に基づいているが， 個人，組織，国などに損害を与えるために使用される情報にあたる場合。 ※正確でない情報を含んでいる場合は「悪意のある情報」に分類せず，他の分類とする。

表 3　偽誤情報の類型カテゴリの分類

定したリスクカテゴリの内訳を表 4 に示す。データにおける偽誤情報類型のカテゴリ内訳を表 5，トピックのカテゴリ内訳を表 6 に示す。なお，望ましい回答の分類の内訳は，人間のアノテータにより問題文に対する望ましい回答分類が「はい」と判定されたものが 19 件，「いいえ」と判定されたものが 290 件，「どちらとも言えない」または「不明」と判定されたものが 76 件であった。

3.2　データセットの課題

JSocialFact の構築にあたってはタグの付与や参考回答の作成に取り組んだが，これらのタグ付与や回答には回答作成者の主観やバイアスが含まれる可能性がある。本データセット構築のプロセスにおいては，複数のレビュアーによって付与されたタグや参考回答をレビューし，改善することにより，可能な限り複数人の視点を採用し，個人的な主観を取り除いた。

4　おわりに

本研究では，ソーシャルメディアにおける日本語投稿に含まれる正確でない情報に基づき，日本語の LLM の出力の正確性を評価するためのデータセット JSocialFact を提案した。今後の展望として，JSocialFact の問題文に対する日本語 LLM の出力の評価実験を行うことが考えられる。その際，JSocialFact に含まれる望ましい回答分類および参考回答を参考に，人間のアノテータによる回答の正誤判定 [7] や，有害性や有用性など安全性に関連する評価，関連性・正確性・流暢性・情報量といった一般的な回答の質に関連する評価 [13] を検討する予定である。

また，本論文では LLM の正確性評価に用いることのできるデータセット JSocialFact を提案したが，

リスクタイプ（大分類）	有害カテゴリ（中分類）	サブカテゴリ（小分類）	件数
バイアス・差別・ ヘイト・反公序良俗	ステレオタイプ・ 差別の助長	性別バイアス・差別	1
		地域バイアス・差別	2
誤情報	誤情報の拡散	危険行為	3
		プロパガンダ	25
		うわさ・ フェイクニュース	121
		誤った文脈・背景	147
	誤情報による実被害	法律相談	5
		金融相談	10
		その他専門分野の相談	16
		医療相談	55
総計			385

表 4　AnswerCarefully リスクカテゴリを用いたリスクカテゴリの内訳

類型（複数選択可）	件数
虚偽・捏造	118
誤解を生む情報の接続	109
偏りのある話	87
疑似科学	77
陰謀論	32
うわさ	15
悪意のある情報	12
プロパガンダ	9
その他	5

表 5　類型カテゴリの内訳（複数選択可）

トピック（複数選択可）	件数
生活	156
社会	130
科学	92
国際	80
政治	50
経済	37
文化	30
事件・事故	19
スポーツ	3
その他	2

表 6　トピックカテゴリの内訳（複数選択可）

JSocialFact において付与されているタグを用いることにより，モデルの回答の正確性を左右する要因を分析することにつなげられる可能性がある。また，JSocialFact では一例ではあるが参考回答も含んだデータセットであり，これを日本語 LLM のチューニングに活用することで，モデルによる不正確な応答の抑制に役立つ可能性がある。

謝辞

本研究は，国立研究開発法人産業技術総合研究所事業の令和 5 年度覚醒プロジェクトの助成を受けました。また，データセットの作成は，LLM 勉強会 [3] の協力のもと，国立情報学研究所 大規模言語モデル研究開発センターと共同で行いました。

参考文献

[1] Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. A survey on large language model (LLM) security and privacy: The good, the bad, and the ugly. **High-Confidence Computing**, Vol. 4, No. 2, p. 100211, June 2024.

[2] Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. September 2021.

[3] BIG bench authors. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. **Transactions on Machine Learning Research**, 2023.

[4] Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. Do large language models know what they don't know? In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Findings of the Association for Computational Linguistics: ACL 2023**, pp. 8653–8665, Toronto, Canada, July 2023. Association for Computational Linguistics.

[5] Yuxia Wang, Haonan Li, Xudong Han, Preslav Nakov, and Timothy Baldwin. Do-not-answer: Evaluating safeguards in LLMs. In Yvette Graham and Matthew Purver, editors, **Findings of the Association for Computational Linguistics: EACL 2024**, pp. 896–911, St. Julian's, Malta, March 2024. Association for Computational Linguistics.

[6] Hitomi Yanaka, Namgi Han, Ryoma Kumon, Jie Lu, Masashi Takeshita, Ryo Sekizawa, Taisei Kato, and Hiromi Arai. Analyzing social biases in japanese large language models. **arxiv:2406.02050**, 2024.

[7] 友亮中村, 大輔河原. 日本語 truthfulqa の構築. 言語処理学会第 30 回年次大会 発表論文集, March 2024.

[8] Matthew R Allen, Nimit Desai, Aiden Namazi, Eric Leas,

3）https://llm-jp.nii.ac.jp/

Mark Dredze, Davey M Smith, and John W Ayers. Characteristics of X (formerly twitter) community notes addressing COVID-19 vaccine misinformation. **JAMA**, Vol. 331, No. 19, pp. 1670–1672, May 2024.

[9] Moritz Pilarski, Kirill Solovev, and Nicolas Pröllochs. Community notes vs. snoping: How the crowd selects fact-checking targets on social media. **arXiv. org**, 2023.

[10] Jennifer Allen, Cameron Martel, and David G Rand. Birds of a feather don't fact-check each other: Partisanship and the evaluation of news in twitter's birdwatch crowdsourced fact-checking program. In **Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems**, No. Article 245 in CHI '22, pp. 1–19, New York, NY, USA, April 2022. Association for Computing Machinery.

[11] Eleni Kapantai, Androniki Christopoulou, Christos Berberidis, and Vassilios Peristeras. A systematic literature review on disinformation: Toward a unified taxonomical framework. **New Media & Society**, Vol. 23, No. 5, pp. 1301–1326, May 2021.

[12] Esma Aïmeur, Sabrine Amri, and Gilles Brassard. Fake news, disinformation and misinformation in social media: a review. **Soc Netw Anal Min**, Vol. 13, No. 1, p. 30, February 2023.

[13] 関根聡, 小島淳嗣, 貞光九月, 北岸郁雄. LLM の出力結果に対する人間による評価分析と gpt-4 による自動評価との比較分析. 言語処理学会第 30 回年次大会, pp. 937–942, 2024.