時事情報に関する日本語 QA ベンチマーク『ニュース Q』

植木快 [1]　川畑輝 [1]　田口雄哉 [1]　新妻巧朗 [1]
浦川通 [1]　田森秀明 [1]　岡崎直観 [2]　乾健太郎 [3,4,5]
[1] 株式会社朝日新聞社　[2] 東京科学大学　[3]MBZUAI　[4] 東北大学　[5] 理化学研究所
{ueki-k1,kawabata-a,urakawa-t,niitsuma-t,taguchi-y2,tamori-h}@asahi.com
okazaki@comp.isct.ac.jp　kentaro.inui@mbzuai.ac.ae

概要

近年，大規模言語モデル（LLM）の性能は飛躍的に向上しているが，ハルシネーションと呼ばれる，誤情報を生成してしまう問題が知られており，生成主体の LLM が持つ知識の信頼性や事実性を測定することは重要である．特に，時事に関する誤情報は，ユーザーの実生活における意思決定や行動に直接的な影響を与えるリスクが大きい．そこで本研究では，様々なジャンルのニュースをもとに作成された，時事的知識を日本語で問うベンチマーク『ニュース Q』[1)] を提案する．主要な LLM のベンチマーク正答率を算出したところ，高品質な日本語コーパスを事前学習する重要性が示唆されるとともに，よりパラメータサイズの大きいモデルの方が時事的な知識をより正確に把握している傾向が確認された．

1　はじめに

近年，大規模言語モデル（LLM）の性能は飛躍的に向上しており，情報検索や意思決定支援などを通じて人々の生活に貢献している．しかしながら，ハルシネーション [1] と呼ばれる，誤情報を生成してしまう問題が知られており，そのリスクを見積もるためには LLM が保持している情報の事実性の程度を測定することが重要である．

誤情報の中でも時事に関するものには，事件・事故，災害，経済，選挙等に関する情報が含まれるため，ユーザーの実生活において，商品購入や旅行に関する経済的な意思決定のみならず，避難行動等の生命に関わる重大な意思決定に対しても，直接的な影響を与えるリスクが大きい．また LLM が保有する知識の中に誤った時事的知識が定着している場合，その知識に関連した誤情報の生成を助長するリスクがあるため，LLM が現実世界で起きた出来事について，いつ，どこで，どのように起きたのかを正確に把握できているか測定することは重要である．

言語モデルが持つ情報の事実性を測る先行研究には，英語 QA としては SimpleQA[2], CommonsenseQA[3], TriviaQA[4], RealtimeQA[5] などがある．一方，日本語 QA としては JAQKET[6], JtruthfulQA[7], JcommonsenseQA[8] などがあるが，一般常識や歴史・教養などに関する QA が主であり，時事情報に特化した日本語 QA は十分に整備されていない．

そこで本研究では，様々なジャンルのニュースをもとに作成された，時事的知識を日本語で問うベンチマーク『ニュース Q』を提案する．このベンチマークは朝日新聞デジタル公式アプリで提供されている時事クイズ『ニュース Q』をベースとしており[2)]，朝日新聞社の編集者が作問し，事実関係の確認を含む校正・校閲を経ているため，情報の正確性が高い．問題数は 2022 年度（穴埋め形式・4 択）が 434 問，2023 年度（QA 形式・3 択）が 335 問で，**計 769 問**である．LLM の内部知識をより多角的に分析できるようにするため，表 1 のように，各問題に対して，それがどこで起きた出来事についてなのか (場所)，時間的に変化する対象 (例：首相，人口，流行語) を考慮して解く必要があるか (時間依存)，どのジャンルに属する知識を問うているか (ジャンル) を表すラベルをアノテーションした．

主要なオープンモデルと商用モデルのベンチマーク正答率を算出したところ，様々なタスクで高性能な最先端の商用モデルであっても，海外の事象と比べて日本国内の事象の理解に乏しい傾向が確認され

1)　『ニュース Q』の利用については朝日新聞メディア研究開発センターの Web サイト (https://cl.asahi.com/api_data/newsq.html) を参照．

2)　https://digital.asahi.com/info/information/articles/SDI202304200001.html

表 1　ニュース Q の問題サンプル

問題と選択肢	ラベル
2022 年 4 月 1 日出題　（穴埋め形式・4 択） 過去最大の総額１０７兆５９６４億円を計上した２０２２年度予算が３月２２日、参院本会議で自民、公明両党と [?] などの賛成多数で可決、成立した。社会保障費や防衛費が過去最大規模となった。 1. 社会党　2. 国民民主党　3. 日本維新の会　4. 新党さきがけ	{ 場所：国内, 時間依存：あり, ジャンル：政治 }
2023 年 4 月 19 日出題　（QA 形式・3 択） 中国が台湾からの輸入を禁止したため、台湾から日本への輸出量が 8 倍以上に急増した食材は何でしょうか 1. バナナ　2. パイナップル　3. タピオカ	{ 場所：国内外, 時間依存：あり, ジャンル：経済 }
2023 年 9 月 26 日出題　（QA 形式・3 択） 消費者庁などは、故人との別れの際に棺（ひつぎ）の中に顔を入れないよう呼びかけています。なぜでしょうか 1. ドライアイスに触れ凍傷になる　2. 二酸化炭素中毒で死亡する恐れ　3. 小窓が閉じて頭をけがする危険	{ 場所：国内, 時間依存：なし, ジャンル：社会 }

表 2　アノテーション項目とラベル

場所	国内,　国内外,　海外
時間依存	あり,　なし
ジャンル	政治,　経済,　社会, 文化・芸能・スポーツ（文芸ス）, 医療・健康・教育（医健教）

表 3　アノテーションラベルの割合

場所	国内	国内外	海外		
	65.7	21.3	13.0		
時間依存	あり	なし			
	75.3	24.7			
ジャンル	政治	経済	社会	文芸ス	医健教
	10.9	27.2	26.0	23.8	12.1

た．また，日本国内の時事的知識に関する理解度の改善に日本語コーパスの継続事前学習が効果的であることが示唆された．全体としては，よりパラメータサイズの大きいモデルの方が時事的知識をより正確に把握している傾向が確認された．

2　アノテーション

本研究ではベンチマークとしての有用性を高めるため，問題ごとに表 2 の 3 項目について，社内で募った 3 名によるアノテーションを実施した．

いつ，どこで，どのようなことが起きたかに関する，モデルが持つ知識の正確性について分析するため，出来事が起きた場所を表す「場所」，時間的に変化する対象 (例：首相，人口，流行語) への理解が求められる問題を区別する「時間依存」，どのジャンルに属する知識を問うているかを表す「ジャンル」，というラベルを設けた．これらのラベルは 3 名のアノテーターが独立に付与し，少なくとも 2 名以上が一致したラベルを採用した．3 名とも不一致だった問題については，改めて最も適切なラベルについて議論し，合意したラベルを採用した．

アノテーションの結果，表 3 に示すように「国内」の問題が 65.7%を占め，また「時間依存あり」の問題が 75.3%と多いことが確認できた．また，どのジャンルの問題も 10.0%以上を占めており，幅広い時事的知識をカバーしていることが分かる．

3　実験

3.1　評価手法

表 4 のオープンモデル 6 種類 [9, 10, 11] および商用モデル 6 種類 [12, 13, 14] の合計 12 モデルに対して『ニュース Q』を用いたベンチマーク評価実験を行った[3)]．モデルは，パラメータサイズ，事前学習コーパス，およびモデル学習方針の差異が出るように選定した．評価用プロンプトには，出題日・問題文・選択肢・正答を 3 問分例示する 3-shot プロンプトを用いた[4)]．3-shot に用いるサンプルは，問題形式ごとに出題日が最も古い問題から 3 問とし，テストデータセットからは除外した．また，商用モデルの評価では，回答フォーマットを安定させるため system プロンプトに「選択肢番号のみを出力する」旨の指示を与えた[5)]．いずれのモデルの回答も，正規表現を用いて回答番号を機械的に抽出し，正誤判定した．回答番号が自動抽出できなかった問題は正答率の算出対象から除外した．

3)　いずれのモデルも temperature=0.0 を設定した．また，GPT-4o/4o-mini は Azure OpenAI Service，Gemini と Claude は Google Cloud Vertex AI を利用した．
4)　付録 A に使用した 3-shot プロンプトを示す．
5)　付録 B に使用した system プロンプトを示す．

表 4　ベンチマーク正答率

Human (計 100 問)	52.3
Human w/ Search (計 100 問)	**93.0**
Llama-3.1-8B-Instruct	46.1
Llama-3.1-70B-Instruct	64.7
Llama-3.1-Swallow-8B-Instruct-v0.1	57.0
Llama-3.1-Swallow-70B-Instruct-v0.1	**73.2**
gemma-2-2b-it	41.0
gemma-2-2b-jpn-it	39.3
GPT-4o	81.7
GPT-4o-mini	67.6
Gemini 1.5 Pro	78.8
Gemini 1.5 Flash	66.6
Claude-3.5-sonnet-v2	**83.2**
Claude-3.5-haiku	70.4

3.2　人間との比較

本ベンチマークの難易度を把握するとともに，モデルが正確な時事的知識を持っていれば正答可能な QA セットであることを確認するため，2022 年度版と 2023 年度版からそれぞれランダムに 50 問ずつ，計 100 問を抽出し，何も参照せずに回答する設定（Human），作問のもととなった新聞記事の参照と Google 検索を利用して回答する設定 (Human w/ Search) という 2 通りの設定で人間の正答率を測定した．各設定につきアノテーターとは別に社内で募った 3 名を被験者として平均を算出した．

4　結果と考察

本節では，各モデルのベンチマーク正答率を示し，結果から得られた知見を考察する．

4.1　ベンチマーク正答率

表 4 に，人間の正答率，オープンモデルの正答率，商用モデルの正答率を示す．

人間の正答率については，何も参照しないで回答した場合の正答率は 52.3%と，一定の難易度があることが確認された．また，記事参照と検索を用いる設定では，正答率は 93.0%を示し，適切な情報源にアクセスできれば容易に正答可能な QA セットであることも確認された．

続いて，オープンモデルの正答率を見ると，Llama-3.1-Swallow-70B-Instruct-v0.1 が最も高い正答率 73.2%を記録した．また日本語コーパスを追加で学習していない Llama-3.1-70B-Instruct の正答率 64.7%と比べると，8.5 ポイント向上していることから，日本語コーパスの学習が，より正確な時事的知識をもたらしていることが示唆される．同様の傾向が，日本語コーパスで追加事前学習を実施した Llama-3.1-Swallow-8B-Instruct-v0.1 と Llama-3.1-8B-Instruct の比較でも確認でき，こちらも日本語コーパスの追加事前学習後に正答率が 10.9 ポイント向上している．

モデルのパラメータサイズに着目すると，よりサイズが大きいほど正答率が高い傾向が見られ，Llama-3.1-70B-Instruct と Llama-3.1-8B-Instruct では 18.6 ポイント，Llama-3.1-Swallow-70B-Instruct-v0.1 と Llama-3.1-Swallow-8B-Instruct-v0.1 では 16.2 ポイント高い．また最もパラメータサイズの小さい gemma-2-2b シリーズの gemma-2-2b-jpn-it が最も低い正答率 39.3%を記録したこともその傾向に合致し，より大きなパラメータサイズのモデルの方が，より正確な時事的知識を保持していることが確認された．

商用モデルの正答率を見ると，Claude-3.5-sonnet-v2 が最も高い正答率 83.2%を記録したが，GPT-4o も 81.7%を記録しており，その差分は 1.5 ポイントと小さい．同シリーズにおけるハイエンドモデルと廉価版モデルを比較すると，GPT-4o と GPT-4o-mini で 14.1 ポイント，Gemini 1.5 Pro と Gemini 1.5 Flash で 12.2 ポイント，Claude-3.5-sonnet-v2 と Claude-3.5-haiku で 12.8 ポイントの差分があり，ハイエンドモデルの方がより正確な時事的知識を保有していることが示唆された[6)]．

4.2　場所別の比較

表 5 に，場所別に着目した結果を示す．すべてのモデルにおいて海外の問題の正答率が最も高く，国内あるいは国内外の正答率が海外と比べて低い傾向が確認された．これは，各モデルの事前学習データに日本国内で起きた出来事，あるいは日本と関係がある出来事に関する情報が十分に含まれていない可能性を示唆する．

一方，Llama-3.1-Swallow-70B-Instruct-v0.1 の正答率に着目すると，Llama-3.1-70B-Instruct における海外の正答率を維持したまま，国内および国内外の正答率が向上していることが分かる，これは，日本語コーパスの追加事前学習によって，海外の時事的知識を保持したまま，日本に関する時事的知識の正確性を高められる可能性を示している．

6)　ただし，本実験の商用モデルにおいては内部実装が未公開のため，言語モデル自体の能力が測定できているかは確認できていないことに注意されたい．

表 5　場所別・時間依存別・ジャンル別の正答率

	場所			時間依存		ジャンル				
	国内	国内外	海外	あり	なし	政治	経済	社会	文芸ス	医健教
Llama-3.1-8B-Instruct	42.3	47.5	57.1	44.0	52.6	54.3	41.8	51.0	39.2	51.6
Llama-3.1-70B-Instruct	58.1	60.6	**87.7**	62.1	72.6	64.2	64.4	67.5	56.4	**76.3**
Llama-3.1-Swallow-8B-Instruct-v0.1	54.6	54.5	65.6	53.8	66.3	61.7	54.8	59.8	53.0	59.1
Llama-3.1-Swallow-70B-Instruct-v0.1	**69.4**	**68.7**	**87.7**	**71.2**	**79.5**	**77.8**	**76.9**	**72.4**	**66.9**	75.3
gemma-2-2b-it	39.1	37.4	49.1	40.1	43.7	34.6	43.8	45.5	36.5	39.8
gemma-2-2b-jpn-it	36.9	37.4	47.9	38.4	42.1	35.8	44.2	41.5	35.4	34.4
GPT-4o	78.2	79.8	93.3	78.5	**91.1**	84.0	79.3	81.0	**84.0**	81.7
GPT-4o-mini	62.7	64.6	84.7	64.0	78.4	67.9	67.8	68.5	66.3	67.7
Gemini 1.5 Pro	73.5	**80.8**	93.9	75.9	87.4	81.5	77.4	80.0	76.2	81.7
Gemini 1.5 Flash	62.5	62.6	81.6	63.4	76.3	59.3	69.2	69.5	61.9	69.9
Claude-3.5-sonnet-v2	**79.2**	**80.8**	**96.9**	**81.2**	89.5	**87.7**	**83.7**	**86.5**	77.3	**82.8**
Claude-3.5-haiku	66.9	63.6	85.3	67.0	80.5	74.1	73.6	70.0	64.6	72.0

表 6　問題形式別（年度別）の正答率

年度	2022	2023
問題形式	穴埋め式	QA 式
選択肢数	4 択	3 択
Human (年度毎 50 問)	60.0	44.7
Human w/ Search (年度毎 50 問)	**90.7**	**95.3**
Llama-3.1-8B-Instruct	51.3	39.5
Llama-3.1-70B-Instruct	69.6	58.4
Llama-3.1-Swallow-8B-Instruct-v0.1	66.0	45.2
Llama-3.1-Swallow-70B-Instruct-v0.1	**80.7**	**63.6**
gemma-2-2b-it	46.6	33.7
gemma-2-2b-jpn-it	45.9	30.7
GPT-4o	90.3	70.5
GPT-4o-mini	76.8	55.7
Gemini 1.5 Pro	86.8	68.4
Gemini 1.5 Flash	76.1	54.2
Claude-3.5-sonnet-v2	**91.4**	**72.6**
Claude-3.5-haiku	79.1	59.0

4.3　時間依存別の比較

表 5 に，時間依存あり・なしに着目した結果を示す．時間依存がある問題はすべてのモデルで，時間依存がない問題より正答率が下がる傾向が確認された．時間依存がある問題は，時々刻々と変化する対象について常時正確に把握している必要性があり，難易度が高くなったと考えられる．

4.4　ジャンル別の比較

表 5 に，ジャンルに着目した結果を示す．オープンモデルでは Llama-3.1-Swallow-70B-Instruct-v0.1 が医療・健康・教育以外のジャンルで最も高い正答率を示したが，文化・芸能・スポーツ（文芸ス）の正答率が他ジャンルよりやや低い傾向が確認され，学習コーパス内に含まれる文芸スの時事的知識が相対的に少なかった可能性が示唆される．また，商用モデルでは Claude-3.5-sonnet-v2 が文芸ス以外のジャンルで最も高い正答率を示したが，こちらも文芸スの正答率が他ジャンルよりやや低い傾向が確認された．

4.5　問題形式別（年度別）の比較

表 6 に，問題形式に着目した結果を示す．QA 形式（2023 年度）は穴埋め式（2022 年度）より選択肢数が 1 つ少ないが，何も参照しない設定の人間とすべてのモデルの正答率において QA 形式（2023 年度）の方が正答率が低い傾向が確認された．これは問題形式の違いが主な要因と考えられる．穴埋め式は，文のマスクされた箇所に入る最も適切な単語や文節などを選択肢から選ぶ問題であるため，QA 形式より易しくなったと考えられる．

5　おわりに

本稿では，時事的知識の正確性を評価するための日本語 QA ベンチマーク『ニュース Q』を新たに提案し，その概要と主要な LLM に対する評価結果を報告した．実験の結果，モデル間の正答率に大きなばらつきが確認され，評価ベンチマークとして有用であることが示された．今後としては，(1) さらなる年度・期間の拡充による問題数・ジャンルの多様化，(2) 記事などの外部知識を参照した場合との比較試験，(3) 他の日本語 QA ベンチマークデータセットとの比較試験，などが挙げられる．本ベンチマークによって，今後，幅広い研究コミュニティで時事情報に関する LLM の性能評価や改善が一層促進されることを期待する．

謝辞

本研究のデータ収集にあたりご協力をいただいた朝デジ事業センターの方々に深く感謝いたします.

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, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. **ACM Transactions on Information Systems**, November 2024.

[2] Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. Measuring short-form factuality in large language models. arXiv:2411.04368, 2024.

[3] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4149–4158, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[4] Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In Regina Barzilay and Min-Yen Kan, editors, **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[5] Jungo Kasai, Keisuke Sakaguchi, yoichi takahashi, Ronan Le Bras, Akari Asai, Xinyan Velocity Yu, Dragomir Radev, Noah A. Smith, Yejin Choi, and Kentaro Inui. Realtime QA: What's the answer right now? In **Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track**, 2023.

[6] 鈴木正敏, 鈴木潤, 松田耕史, 西田京介, 井之上直也. JAQKET: クイズを題材にした日本語 QA データセットの構築. 言語処理学会第 26 回年次大会, 2020.

[7] 中村友亮, 河原大輔. 日本語 TruthfulQA の構築. 言語処理学会第 30 回年次大会, 2024.

[8] Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. JGLUE: Japanese general language understanding evaluation. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, **Proceedings of the Thirteenth Language Resources and Evaluation Conference**, pp. 2957–2966, Marseille, France, June 2022. European Language Resources Association.

[9] Llama Team. The Llama 3 Herd of Models. arXiv:2407.21783, 2024.

[10] Gemma Team. Gemma 2: Improving Open Language Models at a Practical Size. arXiv:2408.00118, 2024.

[11] 服部翔, 岡崎直観, 水木栄, 藤井一喜, 中村泰士, 大井聖也, 塩谷泰平, 齋藤幸史郎, Youmi Ma, 前田航希, 岡本拓己, 石田茂樹, 横田理央, 高村大也. Swallow コーパス v2: 教育的な日本語ウェブコーパスの構築. 言語処理学会第 31 回年次大会 (NLP2025), 2025.

[12] Anthropic. The Claude 3 Model Family: Opus, Sonnet, Haiku, 2024.

[13] OpenAI. GPT-4 Technical Report. arXiv:2303.08774, 2024.

[14] Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv:2403.05530, 2024.

A　3-shot プロンプト

2022 年度（穴埋め形式・4 択）

出題日:
{ サンプル 1 の出題日（yyyy/mm/dd）}
質問:
{ サンプル 1 の問題 }
選択肢:
1. { サンプル 1 の選択肢 1}
2. { サンプル 1 の選択肢 2}
3. { サンプル 1 の選択肢 3}
4. { サンプル 1 の選択肢 4}
答え: { サンプル 1 の正解番号 }

...（同じ形式でサンプル 3 まで）

出題日:
{ テストデータの出題日（yyyy/mm/dd）}
質問:
{ テストデータの問題 }
選択肢:
1. { テストデータの選択肢 1}
2. { テストデータの選択肢 2}
3. { テストデータの選択肢 3}
4. { テストデータの選択肢 4}
答え:

*2023 年度（QA 形式・3 択）は選択肢 3 まで

B　system プロンプト

2022 年度（穴埋め形式・4 択）

選択肢の中から最も適切な選択肢を選び、その番号を出力してください。出力は半角数字一文字（1〜4）だけにしてください。それ以外の文字は出力に含めないでください。

*2023 年度（QA 形式・3 択）は「半角数字一文字（1〜3）」