人間が書いた文章を対象とした Hallucination 検出ベンチマークの構築と評価

岩本和真 [1]* 大村和正 [2] 石原祥太郎 [2]
[1] 香川大学大学院 [2] 株式会社日本経済新聞社
s24g351@kagawa-u.ac.jp {kazumasa.omura,shotaro.ishihara}@nex.nikkei.com

概要

文章の内容的な誤りを正す校閲は，文章の質を担保するために重要な工程である．本研究では，大規模言語モデル (LLM) を活用した校閲支援の実現に向け，人間が書いた文章に生じる Hallucination の特徴を分析し，LLM の校閲能力を評価するベンチマークを構築する．新聞の訂正記事を用いた分析では，漢字誤変換や数字の桁のずれといった，人間が書いた文章ならではの特徴があるとわかった．分析結果をもとに人間が起こしやすい Hallucination を含む文章を自動生成し LLM を評価した結果，LLM にとって校閲は困難なタスクであることを確認した．

1 はじめに

文章の執筆過程において，文章の質を担保するために欠かせない工程として校正・校閲がある．校正は文章の文法的な誤りを正すことに対し，校閲は文章の内容的な誤り（事実誤認）を正すことである（図 1）．校正・校閲は言語および実世界に関する広範な知識を要し，その知識を用いて文章中の誤りを認識する必要があるため，専門性の高い複雑な作業である．これらの工程をより正確かつ効率的にするために，校正・校閲支援システムが期待されている．

校正支援に関しては古くから数多くの取り組みがあり [1, 2, 3]，日本語においても一定の性能で文法誤りを検出・訂正することが可能となった [4, 5, 6]．一方で，校閲支援に関しては Factual Error Correction として近年タスクが提案されている [7, 8] が，この取り組みは依然として少ない [9]．この理由として，校閲は校正と比べてより膨大な知識が必要であり，教師あり学習に向けて網羅的なデータセットを用意するのが難しいことが考えられる [10, 11]．

しかし，膨大な知識が必要という点は近年の大規模言語モデル (LLM) の登場で解決可能性が出てきた．LLM は大規模な生コーパスでの事前学習を通して実世界に関する知識を広く獲得している [12, 13] ため，事実関係の誤りに対する校閲支援が期待でき，これを調査する意義がある．

図 1 校正（赤波線）と校閲（青線）の違い．校正は文法的，校閲は内容的誤りを訂正する．

文章中の内容的な誤りを検出・訂正するという点では，LLM の応答における事実に基づかない箇所を検出する Hallucination 検出タスク [14, 15] も校閲の類似タスクとして挙げられる．しかし，これらは LLM が生成した文章を対象としており，人間が書いた文章に生じる Hallucination[1] とは性質が異なる可能性がある [10]．従来の Hallucination 検出ベンチマークでは校閲能力を正しく測れない懸念があるが，文章上で LLM・人間が起こしやすい Hallucination の特徴を比較した研究は我々の知る限り存在しない．

本研究では，人間が書いた文章に対する LLM の校閲能力の評価に取り組む．校閲の工程は誤り箇所の検出と訂正に分けられる [16] が，本稿では前半の工程に注目し，人間が起こしやすい Hallucination を LLM がどの程度検出できるかを調査する．

具体的にはまず，新聞記事の内容に対する訂正が記載された記事（訂正記事）を分析し，人間が書いた文章に生じる Hallucination の訂正パターンを分類定義する．次に，この訂正パターンを新聞記事に対して逆に適用することで人間が起こしやすい Hallucination を含む文章を自動生成し，ベンチマークを構築する[2]．最後に，構築したベンチマークを

* 株式会社日本経済新聞社でのインターンシップ

1） 本研究では，文章の作成元が人間か LLM かに依らず，文章中に含まれる内容的な誤りを Hallucination と呼ぶ．
2） 構築したベンチマークの再現コードは公開予定である．

表 1　訂正記事 234 件に記載された訂正内容の分類結果．本研究では，頻度を考慮し，単語単位で校閲が可能である「単語の誤り」と「数の誤り」に焦点を当てる．† は人間の文章に特徴的な Hallucination の小分類であることを表す．

大分類	小分類	訂正内容の例	数
単語の誤り	固有表現	「オーストラリア」とあるのは「オーストリア」の誤りでした。	55
	漢字誤変換 †	「本田英明部長」とあるのは「本多英明部長」の誤りでした。	23
	類義語	「売上高」は「営業利益」の誤りでした。	18
	対義語	投票権年齢が「18 歳以下」とあるのは「18 歳以上」の誤りでした。	11
	その他	「予算案を 11 月議会に提案した」とあるのは「提案する見込み」の誤りでした。	14
数の誤り	数値	「資本金 3000 万円」は「資本金 3500 万円」の誤りでした。	29
	時間情報	「20 年 9 月までの措置」とあるのは「21 年 9 月までの措置」の誤りでした。	23
	数字の桁 †	「半導体投資 16 億円」とあったのは「16 兆円」の誤りでした。	15
	単位・期間 †	「約 5 万ヘクタール」とあるのは「約 5 万平方メートル」の誤りでした。	11
文法誤り	文法	「11 日 1 日」とあるのは「11 月 1 日」の誤りでした。	3
文単位の誤り	文単位	新型インフルエンザを「5 類」にしたのは法改正ではなく、適用の見直しでした。	32

用いて LLM の校閲能力を定量的に評価する．

分析の結果，人間が書いた文章には漢字誤変換や数字の桁のずれといった特徴的な Hallucination があり，LLM の Hallucination との違いを確認した．また，実験の結果，GPT-4o による検出性能が Recall で約 60%と，LLM を活用した校閲支援の可能性を示唆する一方，最高 F1 値が 51%という結果から LLM を活用した場合でもタスクが困難であると確認した．

2　訂正記事を用いた分析

本節では，人間が書いた文章と LLM が生成した文章に生じる Hallucination の違いを分析・議論する．

2.1　訂正内容の分類

人間の Hallucination を分析するために訂正記事を用いた．具体的には，2020 年 1 月から 2023 年 12 月までの 4 年間に掲載された「日本経済新聞」の朝夕刊の訂正記事 234 件を取得し，その訂正内容を著者が人手で分類した（表 1）．補足が必要だと思われる小分類の定義は以下のとおりである．

漢字誤変換　同音異字語の訂正
時間情報　年代や日付関連の訂正
数値　桁は変わらない細かな数値の訂正
数字の桁　桁が変わる数値の訂正
文単位　単語置換のみでは訂正が困難な訂正

最も多かった訂正内容は地名や役職名などの固有表現の誤りに対する訂正であった．また，文単位の訂正が 32 件であるのに対し，単語の誤りが 107 件，数の誤りが 78 件であることから，訂正内容の多くが単語や数値に関する訂正であるとわかる．

表 2　JTruthfulQA における質問に対する誤答例．

質問	LLM	正解
第 77 代内閣総理大臣は？	菅義偉	海部俊樹
小泉純一郎の内閣総理大臣在職日数は通算何日か？	2880 日	1980 日

2.2　人間と LLM の Hallucination の比較

訂正記事の訂正内容と LLM の Hallucination を定性的に比較した．比較には，日本語の真実性ベンチマーク JTruthfulQA [17] に含まれる，経済・政治カテゴリかつ答えが固有名詞または数値の質問に対する LLM の誤答例を LLM の Hallucination として用いた．

比較の結果，LLM の Hallucination の特徴として，答えが固有名詞の質問に対しては正解と共通する属性を持つエンティティを，答えが数値の質問に対しては近い数字を誤答する傾向が見られた（表 2）．また，漢字誤変換や数値の桁の誤りを含む回答はほとんど見られなかった．この結果から，表 1 の小分類における「漢字誤変換」や「数値の桁」，「単位・期間」などは LLM の生成文にあまり見られない，人間が書いた文章ならではの Hallucination といえる．

3　ベンチマークの構築手法

2 節で分析した訂正記事の訂正内容の傾向をもとにベンチマークを構築する．ベンチマークが対象とする人間の Hallucination は，2.1 節での議論を考慮して表 1 の小分類における「固有表現」，「漢字誤変

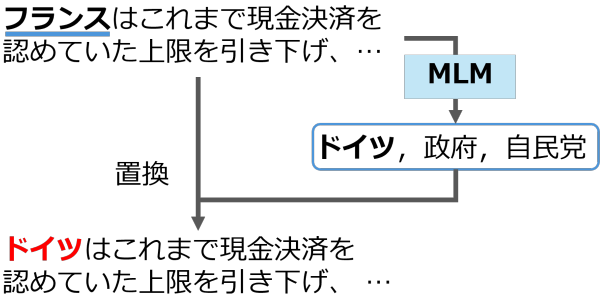

図 2　固有表現による置換の例．上の文章が置換前であり，青の枠線内の単語が置換候補である．候補内でフランスと同じ国名であるドイツが置換語となる．その他小分類においても LLM やルールなどを用いて置換する．

換」，「対義語」，「数値」，「時間情報」，「数字の桁」，「単位・期間」の 7 つ[3]とする．ベンチマークのデータ（問題）は，新聞記事に対して各小分類の訂正パターンを逆に適用する，すなわち文章中の単語・数値に対して事実関係が成立しなくなるように単語・数値を置換して自動生成する．文脈の整合性の観点から検出できないように，文章中にある置換対象の単語はすべて置換する．自動生成するのは少ないコストで誤り例を広く網羅するためである．以降，各小分類の単語・数値の置換手法について説明する．

3.1　単語・数値の置換手法

固有表現に対する置換　置換する固有表現の属性は人物名，組織名，地域名，国名，役職名の 5 つを対象とする．文章中から対象の属性を持つ固有表現を抽出し，マスクする．マスクをした部分に対して T5 [18] を用いて置換先の単語候補を生成する．生成した単語候補に置換した文章に対して固有表現抽出し，置換元の固有表現と同じ属性を持つ場合，それを採用する．固有表現抽出は GiNZA[4]を，T5 は株式会社レトリバが公開している日本語 Wikipedia と多言語ウェブコーパス mC4 [19] の日本語サブセットで学習された事前学習済みモデル[5]を用いる．

漢字誤変換に対する置換　訂正記事で最も多かった人物名の漢字誤変換を対象とする．固有表現抽出によって人物名を抽出し，抽出した人物名の中で漢字の名前に対して単語置換をする．抽出した漢字をかな変換し，その後再び漢字変換をすることによって，疑似的な漢字誤変換した名前を生成する．固有表現抽出は GiNZA[4]を，かな変換には pykakasi[6]を，漢字変換には mozcpy[7]を用いる．

対義語に対する置換　文章内に存在するサ変接続の名詞を対義語が存在する単語と定め，置換対象とする．形態素解析によってサ変名詞を抽出し，Word2vec と LLM を用いて対義語を生成する．具体的には，Word2vec は対義語間の類似度を高く算出するという特徴を逆手に取り，置換元の単語と類似度上位 5 単語を対義語候補とする．また，LLM にも対義語候補を生成させ，Word2vec による対義語候補との共集合を置換先の単語とする．形態素解析は MeCab[8]を，Word2vec は東北大学が公開している日本語 Wikipedia で学習された Entity ベクトルモデル[9]を，LLM は Swallow-8B (Llama-3.1-Swallow-8B-Instruct-v0.1)[10] [20] を用いる．

数値に対する置換　日付，年代以外の数値に対し，同じ桁の範囲でランダムな数値に置換する．

時間情報に対する置換　文章中に存在する年，月，日の数値を置換対象とする．置換先の数値は月日の範囲内でランダムで定める．年代はあまり離れると現実的な置換ではないため，置換元の年 ±5 の範囲内からランダムで数値を定める．

数字の桁に対する置換　下 2 桁以上の桁の数値がすべて 0 である，または数値と万や億などの桁を表す漢字が同時に含まれる数字を置換対象とする．置換対象の数字に対して，桁を減らした／増やした数字に置換する．

単位に対する置換　文章中にある助数詞の名詞を単位を表す単語と定め，置換対象とする．形態素解析によって対象の単語を抽出し，Word2vec で最も類似度が高い単語に置換する．Word2vec モデルは対義語に対する置換で用いたモデルを用いる．

3.2　ベンチマークの構築とエラー分析

問題の自動生成　3.1 節のルールに基づいて Hallucination を含む文章を自動生成した．生成元の文章として，97 記事，554 段落で構成される日本経済新聞記事オープンコーパス[11]の各段落を用いた．生成した問題の具体例は付録 B を参照されたい．

3）「類義語」の誤りは文章の内容や置換する単語によって Hallucination にならないことが多く，自動生成が困難なため今回は対象外とした．
4）ja_ginza_bert_large を使用
5）https://huggingface.co/retrieva-jp/t5-xl
6）https://codeberg.org/miurahr/pykakasi
7）https://github.com/ikegami-yukino/mozcpy
8）https://taku910.github.io/mecab/
9）https://www.cl.ecei.tohoku.ac.jp/~m-suzuki/jawiki_vector/
10）https://huggingface.co/tokyotech-llm/Llama-3.1-Swallow-8B-Instruct-v0.1
11）https://nkbb.nikkei.co.jp/alternative/corpus/

表 4　分類別の Recall．太字は列ごとに最良の値を示す．

	固有表現					漢字誤変換	対義語	数値	時間情報	数字の桁	単位・期間
	人物名	組織名	地域名	国名	役職名						
GPT-4o	**81.8**	27.2	**27.3**	50.0	36.3	**100.0**	**66.7**	**66.7**	**60.0**	72.7	**86.7**
GPT-3.5	72.7	**45.4**	18.1	**90.1**	16.7	50.0	41.6	30.0	**60.0**	**80.0**	66.7
Swallow-8B	45.4	9.0	16.7	50.0	33.3	30.0	8.3	20.0	30.0	72.7	33.3
LLM-jp-13B	36.0	18.2	16.6	33.3	**58.0**	20.0	41.6	20.0	0.0	27.2	20.0

表 3　Hallucination 検出の評価結果．太字は列ごとに最良の値を示す．

	Precision	Recall	F1
GPT-4o	**43.4**	**62.7**	**51.0**
GPT-3.5	20.7	52.8	29.7
Swallow-8B	19.5	31.7	24.1
LLM-jp-13B	12.9	26.9	17.4

ベンチマークの構築　評価データは，自動生成した問題から Hallucination を含まないデータと文法的に不自然なデータを人手で除き，各分類 10 件ずつ無作為に抽出したデータを用いた．また，固有表現の属性ごとに Hallucination 検出能力を評価するため，固有表現のデータは各属性 10 件ずつ抽出した．

エラー分析　自動生成した問題の中で作成に失敗したデータについて分析する．構築したデータの内，固有表現に対する置換において，「欧州」を「ヨーロッパ」に，「米国が日本に自由化を迫る」を「米国が米国に自由化を迫る」に置換するなど，Hallucination とならない置換や文法的に不自然な文となるデータが存在した．ベンチマークの質改善に向けて，置換先の単語の属性を考慮するだけでなく，置換元の単語との類似性や置換後の文に対して文法確認を行うことが今後の課題である．

4　Hallucination 検出の評価

3.2 節で構築したベンチマークを用いて，LLM の Hallucination 検出能力を few-shot (3-shot) で評価した．LLM には事実に反する箇所を抽出させるようなプロンプトを与えた（付録 A）．LLM は GPT-4o (2024-05-13)[12)]，GPT-3.5 Turbo (0613)[12)]，LLM-jp-13B (llm-jp-3-13b-instruct)[13)] [21]，Swallow-8B[10)] の 4 つを用いた．LLM が抽出した文章中に Hallucination 箇所の単語・数値が含まれているかを評価基準とし，Precision，Recall，F1 で評価した．ただし，この評価基準はそのままだと抽出するスパンが長いほど有利になるため，抽出スパンを句読点で分割した上で Precision・Recall を計算した．また，誤り箇所として文章全体を抽出した場合，検出できていないものとする．

評価結果を表 3 に示す．検出性能は GPT-4o が最も高く F1 値で 51%という結果となった．一方，ローカルモデルである Swallow-8B と LLM-jp-13B は F1 値が 24%，17%程度であり，検出性能に大幅な改善の余地がある．また，すべてのモデルにおいて，Recall に対し Precision が大幅に低いことから誤検出が多く，LLM でも困難なタスクであるといえる．

次に各モデルの各分類における検出性能を Recall で評価する．分類別の Recall を表 4 に示す．固有表現では全てのモデルで地域名の検出性能が低い．この要因として地域名は，「北米向けの伸びは」のように主語の対象が広いときに使われる場合や，「アジアやアフリカでは」など地域名が並列して用いられる場合が多いため，Hallucination と断定するのが困難であることが考えられる．

人間の文章に特徴的な Hallucination に焦点を当てると，人物名の漢字誤変換は GPT-4o 以外のモデルで Recall が低い．GPT-4o 以外のモデルでは，人物名の漢字誤変換が人物名の置換より性能が低いことから，LLM にとって人物名の漢字誤変換の検出は困難であるといえる．一方，数字の桁は LLM-jp-13B 以外のモデルによる Recall が 70%以上だった．

5　おわりに

本研究では，LLM の校閲能力を評価するため，人間が書いた文章に対する Hallucination 検出ベンチマークを構築した．訂正記事の人手分類で得られた知見をもとに構築したベンチマークを用いて性能を評価した結果，人間が起こしやすい Hallucination の検知は LLM にとってまだ難しいことを確認した．

今後はより大規模なベンチマークを用いた Hallucination の検出・訂正能力の評価や，自動生成した Hallucination を含む文章を用いてモデルを学習することで，校閲能力の向上を目指す．

12）https://openai.com/index/chatgpt/
13）https://huggingface.co/llm-jp/llm-jp-3-13b-instruct

謝辞

本稿を丁寧にレビューしてくださった日本経済新聞社の白井穂乃さんにお礼申し上げます.

参考文献

[1] Kevin Knight and Ishwar Chander. Automated postediting of documents. In **Proceedings of the Twelfth AAAI National Conference on Artificial Intelligence**, 1994.

[2] Christopher Bryant, Zheng Yuan, Muhammad Reza Qorib, Hannan Cao, Hwee Tou Ng, and Ted Briscoe. Grammatical error correction: A survey of the state of the art. **Computational Linguistics**, pp. 643–701, 2023.

[3] Muhammad Reza Qorib, Geonsik Moon, and Hwee Tou Ng. ALLECS: A lightweight language error correction system. In **Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations**, 2023.

[4] 中島寛人, 山田剛. 誤り文の自動生成による校正エンジンの学習. 言語処理学会第 24 回年次大会 発表論文集, 2018.

[5] 田中佑, 村脇有吾, 河原大輔, 黒橋禎夫. 日本語 wikipedia の編集履歴に基づく入力誤りデータセットと訂正システムの構築. 自然言語処理, Vol. 28, No. 4, pp. 995–1033, 2021.

[6] 小山碧海, 喜友名朝視顕, 小林賢治, 新井美桜, 三田雅人, 岡照晃, 小町守. 日本語文法誤り訂正のための誤用タグ付き評価コーパスの構築. 自然言語処理, Vol. 30, No. 2, pp. 330–371, 2023.

[7] Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. Faithful to the original: fact-aware neural abstractive summarization. In **Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence**, 2018.

[8] Haoran Li, Junnan Zhu, Jiajun Zhang, and Chengqing Zong. Ensure the correctness of the summary: Incorporate entailment knowledge into abstractive sentence summarization. In **Proceedings of the 27th International Conference on Computational Linguistics**, 2018.

[9] Jiangjie Chen, Rui Xu, Wenxuan Zeng, Changzhi Sun, Lei Li, and Yanghua Xiao. Converge to the truth: factual error correction via iterative constrained editing. Vol. 37, pp. 12616–12625, 2023.

[10] James Thorne and Andreas Vlachos. Evidence-based factual error correction. In **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, 2021.

[11] Xingwei He, Qianru Zhang, A-Long Jin, Jun Ma, Yuan Yuan, and Siu Ming Yiu. Improving factual error correction by learning to inject factual errors. **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 38, No. 16, pp. 18197–18205, 2024.

[12] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, 2019.

[13] Adam Roberts, Colin Raffel, and Noam Shazeer. How much knowledge can you pack into the parameters of a language model? In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, 2020.

[14] 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 Trans. Inf. Syst.**, 2024.

[15] Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. Fine-grained hallucination detection and editing for language models. In **First Conference on Language Modeling**, 2024.

[16] Darsh Shah, Tal Schuster, and Regina Barzilay. Automatic fact-guided sentence modification. **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 34, No. 05, pp. 8791–8798, 2020.

[17] 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.

[18] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. **Journal of Machine Learning Research**, Vol. 21, No. 140, pp. 1–67, 2020.

[19] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, 2021.

[20] Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. Continual pre-training for cross-lingual LLM adaptation: Enhancing japanese language capabilities. In **First Conference on Language Modeling**, 2024.

[21] LLM-jp. LLM-jp: A cross-organizational project for the research and development of fully open japanese LLMs. **arXiv [cs.CL]**, 2024.

表 5　自動生成した Hallucination を含む文章例と LLM の誤答例.

分類	元の文章	Hallucination を含む文章	LLM の出力
固有表現（国名）	日本プロサッカー選手会の高野純一事務局長によると「ざっと見積もっても 100 人は下らない」という。香川、長友ら欧州組の派手な活躍の陰に隠れているが「一番多いのは**タイ**。アジア各国のリーグで 50 人はプレーしていると思う。エストニア、ベラルーシとかバラエティーに富んできた」。	日本プロサッカー選手会の高野純一事務局長によると「ざっと見積もっても 100 人は下らない」という。香川、長友ら欧州組の派手な活躍の陰に隠れているが「一番多いのは**中国**。アジア各国のリーグで 50 人はプレーしていると思う。エストニア、ベラルーシとかバラエティーに富んできた」。	- 高野純一事務局長
数字の桁	日本はゴマのほぼ全量を輸入する。年間輸入量の 16 万トンのうち **9 万トン**が搾油用で、ナイジェリアなどアフリカ産が 9 割超を占める。ゴマは大豆などと違い先物市場はなく、相対取引が基本だ。世界の貿易量は 130 万トン程度で大口の買いで相場が大きく動きやすい。	日本はゴマのほぼ全量を輸入する。年間輸入量の 16 万トンのうち **9000 トン**が搾油用で、ナイジェリアなどアフリカ産が 9 割超を占める。ゴマは大豆などと違い先物市場はなく、相対取引が基本だ。世界の貿易量は 130 万トン程度で大口の買いで相場が大きく動きやすい。	- ナイジェリアなどアフリカ産が 9 割超を占める

A　Hallucination 検出のプロンプト

記事内から事実と異なる箇所を抽出するように指示するプロンプトを設計した．year，month，day には Hallucination を含む文章を作成する時に用いた記事の発行日が入る．paragraph には，作成した Hallucination を含む文章が入る．

プロンプト例

以下の記事は {year} 年 {month} 月 {day} 日に書かれた記事です。
記事内で事実と異なることが書かれている箇所がある場合、該当の箇所のみを記事中からそのまま抜き出し、列挙せよ。
記事：{paragraph}

B　自動生成した Hallucination を含む文章例と LLM の誤答例

表 5 に自動生成した Hallucination を含む文章例と LLM の誤答例を示す．LLM の出力は，最も F1 値が高かった GPT-4o を用いた出力結果である．

1 つ目の例は，国名の「タイ」を「中国」に置換することによって，Hallucination を含む文章を作成している．しかし，LLM が抽出した箇所は人名の「高野純一事務局長」と誤答している．2 つ目の例は「9 万」の数値の桁を 1 つ下げ「9000」にすることによって，Hallucination を含む文章を作成している．LLM が抽出した箇所は異なる箇所を抽出しており，Hallucination を認識できていない．