

# JHARS: RAG 設定における日本語 Hallucination 評価ベンチマークの構築と分析

亀井遼平<sup>1\*</sup> 坂田将樹<sup>1\*</sup>

邊土名朝飛<sup>3,2</sup> 栗原健太郎<sup>2,3</sup> 乾健太郎<sup>5,1,4</sup>

<sup>1</sup> 東北大学 <sup>2</sup> 株式会社 AI Shift

<sup>3</sup> サイバーエージェント <sup>4</sup> 理化学研究所 <sup>5</sup> MBZUAI

{ryohei.kamei.s4, sakata.masaki.s5}@dc.tohoku.ac.jp

{hentona\_asahi, kurihara\_kentaro}@cyberagent.co.jp kentaro.inui@mbzuai.ac.ae

## 概要

大規模言語モデルの hallucination (与えられた情報源に存在しない内容を生成する現象) は、実用上での重要な課題となっている。本研究では、日本語における hallucination 評価のための包括的なベンチマーク **JHARS** (Japanese Hallucination Assessment in RAG Settings) を構築し、最新の GPT-4o を含む 3 つのモデルを対象に分析を行った。その結果、hallucination 発生率は低い一方、事実確認が必要な重大な hallucination が検出された。また、自動検出における高い適合率と再現率の両立は困難であるものの、重大な hallucination に関しては高い再現率で検出可能であることが示された。これは、LLM 自身による出力の検証が、ユーザへの事実確認支援として機能する可能性を示唆している。

 [github.com/cl-tohoku/JHARS](https://github.com/cl-tohoku/JHARS)

## 1 はじめに

大規模言語モデル (LLM) は幅広い分野での活用が期待され、様々な業務プロセスの自動化や効率化に向けた検討が進められている。しかし、実用環境での LLM の適用における重要な課題として、hallucination の問題が指摘されている。一般に hallucination とは、与えられた情報源に存在しない内容を LLM が生成する現象を指す [1, 2]。特に、検索拡張生成 (Retrieval-Augmented Generation: RAG) の設定において、この問題は実用上の重要な課題となっている。図 1 に示すように、LLM の生成する応答は一見適切に見えるが、参考文献に記載のない情報が含まれることがある。それらの情報に対する事

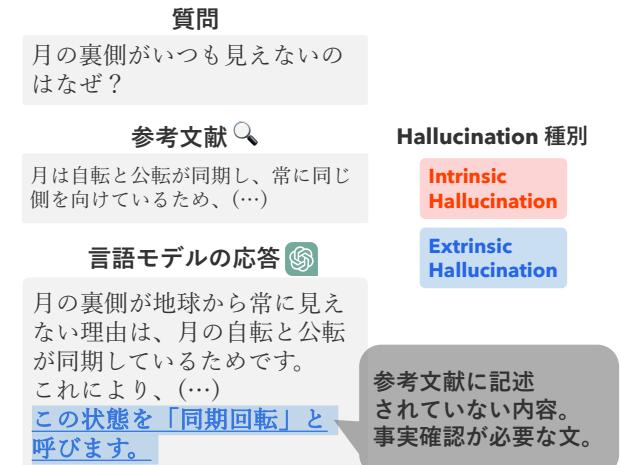


図 1 収集された hallucination の例。

実確認や修正作業は、実務者のコスト増大を招く。

この問題に対処するための第一歩として、hallucination を定量的に評価するための包括的なベンチマークの整備が求められている。これまでにも、hallucination 評価のためのベンチマーク作成は行われてきた [3, 4, 5, 6, 7]。しかし、既存のベンチマークの多くは英語や中国語を対象としており、日本語における hallucination の評価データセットは十分に整備されていない。そのため、日本語に対応した LLM における hallucination の発生頻度や、その検出の困難さについては十分な知見が得られていない。この状況は、説明責任が求められる実務タスクへの LLM の導入を妨げる要因となっている。

そこで本研究では、RAG 設定における日本語 LLM の応答 450 件にアノテーションを行い、日本語 hallucination 評価ベンチマーク **JHARS** (Japanese Hallucination Assessment in RAG Settings) を構築した。このベンチマークを用いた分析の結果、LLM の応答文に対する hallucination 発生率は比較的低いものの、

\* Equal contribution

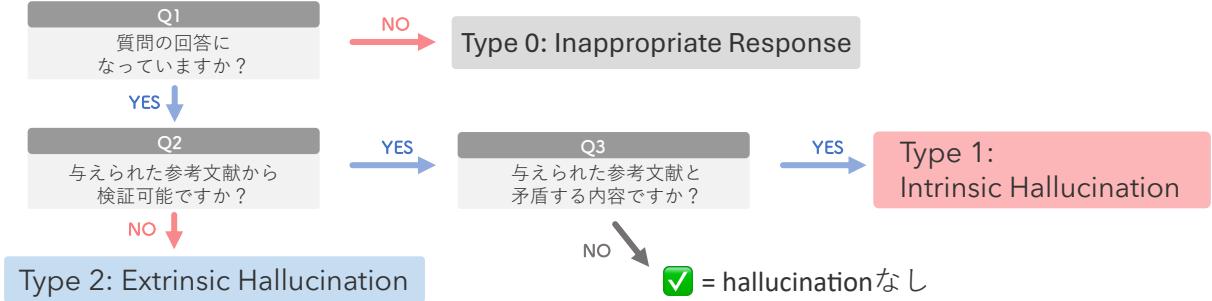


図 2 データセット構築時に使用したフローチャート。

事実確認が必要な重大な hallucination が検出された。また、LLM の応答文に含まれる hallucination の自動検出において、高い適合率と再現率の両立は困難であることが明らかになった。一方で、事実確認が必要な重大な hallucination については高い再現率で検出可能であることが示された。この結果は、LLM 自身による生成結果の検証が、事実確認を要する箇所の特定と警告に有効である可能性を示唆している。

## 2 JHARS データセット

本研究の目的は、RAG 設定において日本語 LLM が生成する hallucination の種類と程度を調査することである。そのため、先行研究 [5, 6] に倣って RAG 設定での日本語 hallucination 評価ベンチマーク JHARS を作成した。

**hallucination の定義** 一般に、自然言語生成タスクにおける hallucination とは、生成された内容が、入力として与えられたテキストや参考文献に対して意味をなさない、または忠実でないことを指す [2]。この hallucination は Intrinsic Hallucination と Extrinsic Hallucination の 2 つの主要なタイプに分類することができる [1, 2, 8, 9]。Intrinsic Hallucination とは、LLM の出力が入力されたテキストや参考文献と矛盾していることを指す。Extrinsic Hallucination とは、入力されたテキストや参考文献から事実正誤性を検証できない LLM の出力を指す<sup>1</sup>。本研究でもこちらの定義を採用してアノテーションを実施した<sup>2</sup>。

### 2.1 データセットの構築

**応答生成のベースデータセット** 我々は Wikipedia の記事検索を含む日本語の質問応答データセットである wikipedia-human-retrieval-ja[10]

表 1 アノテーション時のアノテータのラベル一致率

	文数	割合
3 人一致	2130	90.37%
2 人一致	218	9.25%
一致なし	9	0.38%

を用いて応答の生成を行った。これは応答に必要な参考文献が与えられており、長文形式での応答が必要な点で本研究に適している。

**LLM を用いた応答生成** hallucination 評価のためのデータセットの開発が活発になりつつある一方、それらの多くの場合で、特定の種類の hallucination を人為的に生成し、収集するためのテクニックが採用されている [3, 11, 12]。具体的には、hallucination が起こるようなプロンプトで指示したり、単に出力に矛盾を挿入したりするといったテクニックである。これは hallucination が発生する数を増やすのに有効であるものの、自然に発生した hallucination の分布と大きく異なることがある点が指摘されている [5]。我々は自然に発生する hallucination を評価するため、hallucination を人為的に発生させるテクニックは使用せず、むしろ参考文献以外の情報は使用しないように指示を与えた。

また、LLM の応答に hallucination が含まれる要因として、参考文献の品質と LLM の性能という 2 つの可能性が考えられる。本研究では、LLM の性能に起因する hallucination を評価するため、以下の 2 点を実施した。(i) 質問応答に必要な参考文献のみを使用した。(ii) 応答を生成する前に、GPT-4o[13] を用いて参考文献が十分な情報を含んでいるかのチェックを行った。以上の 2 点により、RAG の設定において応答生成のための参考文献が正しく検索され、無関係なテキストは入っていないという前提で、LLM の応答に hallucination のラベルを付与する。

1 Intrinsic Hallucination と Extrinsic Hallucination の具体例については表 4 を参照されたい。

2 なお本研究では常識的に正しい箇所は hallucination に含まないとした。(例:「岸田総理」 = 「岸田文雄総理」であることは常識的に正しい。)

表2 文単位の hallucination の統計情報. None は hallucination 無しを示す. 括弧内は割合を表す.

モデル名	None	Intrinsic	Extrinsic	null(3人のラベルが全て違う)	合計
GPT-4o-mini	959 (97.0%)	1 (0.1%)	24 (2.4%)	5 (0.5%)	989
GPT-4o	866 (97.9%)	0 (0.0%)	17 (1.9%)	2 (0.2%)	885
Llama-3.1-Swallow-8B-Instruct-v0.1	474 (98.3%)	0 (0.0%)	6 (1.2%)	2 (0.4%)	482

応答生成モデルは、GPT-4o, GPT-4o-mini<sup>3</sup>[14]に加え、Llama 3.1 の日本語機能を強化したものである Llama-3.1-Swallow-8B-Instruct-v0.1[15, 16]を使用した。GPT-4o, GPT-4o-mini は比較的最近にリリースされ広く社会で用いられているモデルである。Llama-3.1-Swallow-8B-Instruct-v0.1 はパラメータ数が小さい日本語 LLM がどの程度 hallucination を含む応答を生成するかを調査するために採用した。これらの 3 つのモデルを用いて、各モデル 150 件の質問文に対して応答（計 450 応答）を生成させた。

**アノテーション** 本研究のアノテーションのフローチャートを図2に示した。本研究では1文単位でアノテーションを付与した。hallucination に関するアノテーションを付与している先行研究[6, 17]に倣い、“はいいいえ”で答えられる質問に複数回答えることで自動的にラベルが決まるようにした。

我々は、アノテータ間の合意を取ることを目的として、各文に3人のアノテータを割り当てた。なお、全体のアノテーションを実行する前に、アノテータに対し、ガイドラインを読んだ上で 10 件のテストアノテーションを実施した。その後、不明点や質問等に回答するための説明会を実施した。本研究におけるアノテータ間のラベルの一致率を表1に示した。表1より、全員不一致の割合は 0.38% と低く、妥当性の高いデータセットであると考えられる。

## 2.2 データセットの分析

定量的な分析として、本データセットのラベルの統計情報を表2に示した。Intrinsic Hallucination と Extrinsic Hallucination の発生件数は英語 hallucination 評価ベンチマークを構築している先行研究[5, 6]と比較して少なかった。Intrinsic Hallucination の割合は特に少なく、GPT-4o-mini における1件のみであった。これらの要因として、RAG による応答生成のための参考文献が正しく検索された前提であるということと、LLM の性能が以前より向上していることが考えられる。また、応答単位の hallucination の発生件数を表3に示した。表2, 3 より、応答単位、文単位

表3 応答単位の hallucination の発生件数（割合）

モデル名	件数(割合)
gpt-4o-mini	16 (10.7%)
gpt-4o	11 (7.3%)
Llama-3.1-Swallow-8B-Instruct-v0.1	5 (3.3%)

のいずれにおいても GPT-4o-mini の hallucination の割合が 3 つのモデルで最も高く、Llama-3.1-Swallow-8B-Instruct-v0.1 の割合が最も低かった。この原因として、Llama-3.1-Swallow-8B-Instruct-v0.1 の出力の文長が他のモデルに比べて短いことが考えられる。

定性的な分析として、hallucination の発生事例を表4に示した。Intrinsic Hallucination の事例では、“14,534 人”の箇所は誤りであり、正しくは“約 1.6 万人”である。Extrinsic Hallucination の事例では、“この状態を「同期回転」と呼びます。”という箇所は参考文献には書いておらず、事実確認が必要である。Mishra ら[6]によると、Extrinsic Hallucination は、事実確認が必要となるものと、主観的な感想（例：“おもしろい”）に大別できる。本データセットにおいて、上記の違いについて著者らの目視で確認したところ、全応答中、平均約 1.6% が事実確認が必要となる事例となっていた。RAG 設定において参考文献が正しく検索されている前提を踏まえると、上記の hallucination 事例は文書検索処理の問題ではなく、LLM の生成特性に起因すると考えられる。このような応答はユーザに誤った知識を与える可能性があり、今後無くしていくべきである。

## 3 実験：hallucination の自動検出

§2.2 の分析により、最近の日本語 LLM であっても hallucination が一定数発生することが明らかになった。この結果は、LLM の実用において hallucination の検出が重要な課題となることを意味している。そこで本節では、「現状の日本語 LLM が起こす hallucination は自動検出可能か」という問い合わせを検証する。実験の結果、日本語 LLM の生成結果に含まれる hallucination の自動検出において、高い適合率と再現率の両立は困難であることが明らかになった。一方、事実確認が必要な重大な hallucination に

<sup>3</sup> GPT モデルは 2024 年 10 月時点でのモデルを使用。

表4 Intrinsic Hallucination（赤色塗りつぶし）と Extrinsic Hallucination（青色塗りつぶし）の事例。(...)は省略を表す。

応答モデル名: GPT-4o-mini

質問文: 四万十町はどの県ですか

参考文献: 高知県西南部に位置し、高岡郡に属する約人口 1.6 万人（2021 年 8 月 31 日現在 [1]）の町。四万十川の中流に位置し、東は太平洋（土佐湾）に面し、(…)

応答文: 四万十町は高知県に位置する町です。(…). 町の面積は 642.30 平方キロメートルで、

2021 年 8 月 31 日現在の総人口は **14,534 人** です。また、(…)

応答モデル名: GPT-4o

質問文: 月の裏側がいつも見えないのはなぜ？

参考文献: 月は自転と公転が同期し、常に同じ側を向いているため、(…)

応答文: 月の裏側が地球から常に見えない理由は、月の自転と公転が同期しているためです。これにより、月は常に同じ側を地球に向けています。この状態を「同期回転」と呼びます。しかし (…)

表5 hallucination 検出の F1 スコア (%). Intrinsic Hallucination の事例は 1 件のみであることに注意。

Detector	Generator: GPT-4o	Generator: GPT-4o-mini	Generator: Llama-3.1-Swallow-8B-Instruct-v0.1			
Detector	Intrinsic	Extrinsic	Intrinsic	Extrinsic	Intrinsic	Extrinsic
GPT-4o (zero shot)	-	<b>14.46</b>	0	15.75	-	10.81
GPT-4o w/ 5-shot	-	14.43	0	<b>17.07</b>	-	<b>14.63</b>
GPT-4o w/ 10-shot	-	10.08	0	16	-	10.53
GPT-4o w/ 30-shot	-	9.76	0	13.33	-	8.51

については高い再現率での検出が可能であった。これは、LLM の生成結果を LLM 自身でチェックし、hallucination 発生の可能性がある箇所を警告するアプローチが実用的である可能性を示唆している。

### 3.1 実験設定

検出モデルとして、HELM [18] 及び Nejumi LLM リーダーボード [19] で高い性能を示している GPT-4o を採用した。検出モデルは、入力された質問と参考文献に基づき、各応答文に対して “hallucination なし”, “Intrinsic Hallucination”, “Extrinsic Hallucination”, の 3 ラベルのいずれかを付与する。JHARS の応答数は 450 件と限られているため追加学習は適切でないと判断し、文脈内学習を採用した。プロンプトには、Ji ら [7] の英語プロンプトを日本語に翻訳したものを使用した。評価には、JHARS を 8:2 の比率で分割し、8 割を評価データ、残り 2 割を少数事例学習用の例示データとして使用した。評価指標には、各クラスの適合率と再現率から算出した F1 スコアを採用し、hallucination の検出性能を測定した。

### 3.2 実験結果

GPT-4o を用いた hallucination 検出の F1 スコアを表5 に示した<sup>4</sup>。hallucination 検出において、再現率は高い値を示したもののが適合率が低く、結果として

<sup>4</sup>適合率と再現率の結果は Appendix 表7 に記載。

F1 スコアは 15% 程度に留まった。これは、偽陽性が多く発生する傾向にあることを意味しており、高い適合率と再現率を両立した hallucination 検出は困難であることを示している。また、少数事例学習における事例数の増加が必ずしも F1 スコアの向上につながらないことも判明した。特に 30 事例を使用した場合において、最も低い F1 スコアを記録した。これについては、入力長の増加が LLM の性能低下を招くことが指摘されており [20]、本実験においても同様の現象が発生したと考えられる。

一方、§2.2 で言及した事実確認が必要な Extrinsic Hallucination についての検出性能を分析したところ、7 件の事例全てを正しく検出できていることを確認した。この結果は、LLM 自身による生成結果のチェックが、hallucination 発生の可能性がある箇所をユーザに警告し、事実確認を促す有効な手段となり得ることを示唆している。

## 4 おわりに

本研究では、日本語における hallucination 評価のための包括的なベンチマーク JHARS を構築し、3 つの LLM の応答を分析した。今後、本研究で用いたフローを活用してデータの拡充を目指す。加えて、本研究では文単位でアノテーションを付与したが、スパン単位でアノテーションを付与し、よりきめ細かい hallucination 評価データセットの構築を目指す。

## 謝辞

本研究は株式会社 AI Shift と東北大学の共同研究により実施した。

## 参考文献

- [1] Zwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, Vol. 55, No. 12, March 2023.
- [2] 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.
- [3] Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. HaluEval: A large-scale hallucination evaluation benchmark for 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. 6449–6464. Association for Computational Linguistics, December 2023.
- [4] Xun Liang, Shichao Song, Simin Niu, Zhiyu Li, Feiyu Xiong, Bo Tang, Yezhaohui Wang, Dawei He, Cheng Peng, Zhonghao Wang, and Haiying Deng. UHGEval: Benchmarking the hallucination of Chinese large language models via unconstrained generation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5266–5293. Association for Computational Linguistics, August 2024.
- [5] Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10862–10878. Association for Computational Linguistics, August 2024.
- [6] 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.
- [7] Zwei Ji, Yuzhe Gu, Wenwei Zhang, Chengqi Lyu, Dahua Lin, and Kai Chen. ANAH: Analytical annotation of hallucinations in large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, August 2024.
- [8] Yichong Huang, Xiachong Feng, Xiaocheng Feng, and Bing Qin. The factual inconsistency problem in abstractive text summarization: A survey, 2023.
- [9] Wei Li, Wenhao Wu, Moye Chen, Jiachen Liu, Xinyan Xiao, and Hua Wu. Faithfulness in natural language generation: A systematic survey of analysis, evaluation and optimization methods, 2022.
- [10] Inc. Baobab. wikipedia-human-retrieval-ja, 2024. <https://huggingface.co/datasets/baobab-trees/wikipedia-human-retrieval-ja> [Accessed : 10/2024].
- [11] Tianyu Liu, Yizhe Zhang, Chris Brockett, Yi Mao, Zhifang Sui, Weizhu Chen, and Bill Dolan. A token-level reference-free hallucination detection benchmark for free-form text generation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6723–6737. Association for Computational Linguistics, May 2022.
- [12] Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7052–7063. Association for Computational Linguistics, November 2021.
- [13] OpenAI. GPT-4o, 2024. <https://platform.openai.com/docs/models#gpt-4o> [Accessed : 10/2024].
- [14] OpenAI. GPT-4o mini, 2024. <https://platform.openai.com/docs/models#gpt-4o-mini> [Accessed : 10/2024].
- [15] 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 *Proceedings of the First Conference on Language Modeling*, COLM, p. (to appear), October 2024.
- [16] Naoaki Okazaki, Kakeru Hattori, Hirai Shota, Hiroki Iida, Masanari Ohi, Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Rio Yokota, and Sakae Mizuki. Building a large japanese web corpus for large language models. In *Proceedings of the First Conference on Language Modeling*, COLM, p. (to appear), October 2024.
- [17] Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Osmar Zaiane, Mo Yu, Edoardo M. Ponti, and Siva Reddy. FaithDial: A faithful benchmark for information-seeking dialogue. *Transactions of the Association for Computational Linguistics*, Vol. 10, pp. 1473–1490, 2022.
- [18] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. Featured Certification, Expert Certification.
- [19] Kei Kamata, Taichi Ibi, Yuya Yamamoto, Kazuki Kurosawa, Ryo Kanazawa, and Akira Shibata. Nejumi LLM リーダーボード 3, 2024.
- [20] Mosh Levy, Alon Jacoby, and Yoav Goldberg. Same task, more tokens: the impact of input length on the reasoning performance of large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15339–15353. Association for Computational Linguistics, August 2024.

## A データセット作成の参考情報

アノテーション時のインターフェースを図 3 に示した。なお、アノテーションで使用したガイドラインは以下のリンクから確認できる。

[https://docs.google.com/document/d/1T1iRGCDm384DTh7Hs11T4YRCpFaTDdK3-e\\_gLSiiV\\_c/edit?usp=sharing](https://docs.google.com/document/d/1T1iRGCDm384DTh7Hs11T4YRCpFaTDdK3-e_gLSiiV_c/edit?usp=sharing)

質問文  
1: ガンバ大阪の「ガンバ」はどういう意味？  
参考文献  
1: クラブ名の「ガンバ (GAMBA)」は、イタリア語で「脚」を意味し「脚」によってシンプルで強いチームを目指すという意味があります。また、日本語の「頑張る」にも通じ[1]。また、日本語の「頑張る」にも通じ[1]。チーム一丸となって勝利を目指して「頑張る」ことや、大阪をホームタウンとし関西のファンの声援を受けて「頑張る」こと、そして日本一、世界一のクラブ組織を目指して「頑張る」サッカークラブであるという意味が込められています。

回答全文  
Q1: 答が正しいかどうかにかかわらず、回答として適切な文章になっていますか？  
● はい  
○ いいえ

回答文1  
Q2: 矛盾があるかどうかにかかわらず、与えられた参考文献から回答文の矛盾をチェックできる文章ですか？  
(矛盾が断定できるもの・できないものの両方が含まれている場合は「わからない」を選択)  
● はい

図 3 アノテーション時のインターフェース。

## B Extrinsic Hallucination の事例

本文に記載しなかった Extrinsic Hallucination の事例をいくつか表 6 に示した。

表 6 Extrinsic Hallucination (青色塗りつぶし) の事例。 (...) は省略を表す。

応答モデル名: GPT-4o

質問文: ハイダイナミックレンジとはどのような機能ですか？

参考文献: ハイダイナミックレンジ(英語: High dynamic range, HDR)は、通常よりも高いダイナミックレンジのことである。ハイダイナミックレンジは様々な信号、例えば、ビデオ、もしくは画像、(...)

応答文: ハイダイナミックレンジ (HDR) は、通常よりも高いダイナミックレンジを持つ技術や手法を指します。これは主に (...)

このように、HDR は信号の特性を向上させ、より広範な明暗の情報を捉えることができる機能です。

応答モデル名: Llama-3.1-Swallow-8B-Instruct-v0.1

質問文: タヌキとアライグマを見分ける方法はありますか？

参考文献: 灰褐色の体毛をもち、眼のまわりから頬にかけて黒い斑紋が (...)

応答文: タヌキとアライグマを見分ける方法はいくつかあります。(...)

一方で、ニホンアナグマはタヌキよりも小さく、体の毛並みが粗く、尾の毛が短いという特徴があります。

## C 実験結果の詳細

GPT-4o を用いた Extrinsic hallucination 検出の適合率と再現率を表 7 に示した。Intrinsic hallucination は生成モデルが GPT-4o-mini の場合に 1 件のみ、かつ検出できていなかったため省略している。

表 7 Extrinsic hallucination の検出の適合率と再現率 (%).

Detector	Generator: GPT-4o		Generator: GPT-4o-mini		Generator: Llama-3.1-Swallow-8B-Instruct-v0.1	
	適合率	再現率	適合率	再現率	適合率	再現率
GPT-4o (zero shot)	5.26	57.14	<b>11.01</b>	75.00	<b>6.25</b>	<b>100</b>
GPT-4o w/ 5-shot	<b>9.33</b>	<b>100</b>	9.70	81.25	5.41	<b>100</b>
GPT-4o w/ 10-shot	7.00	<b>100</b>	9.40	<b>87.50</b>	5.88	<b>100</b>
GPT-4o w/ 30-shot	5.83	<b>100</b>	8.86	<b>87.50</b>	<b>6.25</b>	<b>100</b>