Constructing a Dataset for Hallucination Detection in Japanese Summarization with Fine-grained Faithfulness Labels

Hikari Tanaka and Atsushi Keyaki and Mamoru Komachi
Hitotsubashi University, Japan
dm240015@g.hit-u.ac.jp, {a.keyaki,mamoru.komachi}@r.hit-u.ac.jp

Abstract

Large language models (LLMs) can generate fluent text, but the quality of generated content crucially depends on its consistency with the given input. This aspect is commonly referred to as faithfulness, which concerns whether the output is properly grounded in the input context. A major challenge related to faithfulness is that generated content may include information not supported by the input or may contradict it. This phenomenon is often referred to as hallucination, and increasing attention has been paid to automatic hallucination detection, which determines whether an LLM's output is hallucinated. To evaluate the performance of hallucination detection systems, researchers use evaluation datasets with labels indicating the presence or absence of hallucinations. While such datasets have been developed for English and Chinese, Japanese evaluation resources for hallucination detection remain limited.

Therefore, we constructed a Japanese evaluation dataset for hallucination detection in summarization by manually annotating sentence-level faithfulness labels in LLM-generated summaries of Japanese documents. We annotate 390 summaries (1,938 sentences) generated by three LLMs with sentence-level multi-label annotations for faithfulness with respect to the input document.

Beyond binary labels, our dataset includes fine-grained hallucination and faithfulness error types. The taxonomy extends a prior classification scheme and captures distinct patterns of model errors, enabling both binary hallucination detection and fine-grained error-type analysis of Japanese LLM summarization.

1 Introduction

In recent years, large language models (LLMs) have been applied to a wide range of natural language processing tasks such as question answering, document summarization, and machine translation. Meanwhile, hallucination, in which the generated text is not supported by or contradicts the given input (i.e., the provided context), has become a major challenge (Huang et al., 2025; Ji et al., 2023a).

To address this issue, we focus on hallucinations in non-open-ended generation tasks such as open-book QA (where the model is required to answer questions based explicitly on a given reference document), document summarization, and machine translation, where consistency with the given input (e.g., the provided context, document, or source text) is essential. In this work, we evaluate generation quality from the perspective of *faithfulness* (Li et al., 2022), which concerns consistency between the output and the given input. Within this framework, we focus on hallucinations, defined as cases where the output introduces content that is not supported by the input or contradicts it. Table 1 provides concrete examples of such hallucinations in document summarization. We explicitly distinguish this setting from factuality errors with respect to external world knowledge. Hereafter, we use the term *hallucination* to refer to context inconsistency.

Automatic hallucination detection aims to determine whether LLM outputs contain such hallucinations, using the generated text and/or model-derived signals. This line of work has been actively explored, and various approaches have been proposed (Es et al., 2024; Manakul et al., 2023; Sun et al., 2025).

To evaluate hallucination detection systems, researchers apply detectors to labeled texts and measure agreement with the labels; such collections are released as benchmark datasets for hallucination evaluation (Zhang et al., 2023; Lattimer et al., 2023). For English, datasets have been developed that cover not only open-book QA but also tasks such as document summarization and data-to-text generation (Niu et al., 2024). For Chinese, a dataset has been constructed in which hallucinations in LLM outputs for open-book QA are manually annotated, including both their presence and types (Ji

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 4: Student Research Workshop, pages 207–218

et al., 2024). However, for Japanese, hallucination evaluation datasets remain insufficient, making it difficult to propose and evaluate hallucination detection methods for Japanese or to assess cross-lingual transfer of existing approaches.

In addition, some Japanese resources are constructed using automatically generated hallucination examples (Iwamoto and Shimada, 2024). While useful for large-scale construction, such datasets may not fully reflect the characteristics of hallucinations produced by actual LLMs. This limitation motivates the need for manually annotated datasets based on real LLM outputs.

Building on this background, this study constructs a dataset for evaluating faithfulness in Japanese document summarization, enabling hallucination detection and fine-grained error analysis. We annotated the outputs of three LLMs with sentence-level faithfulness labels, including hallucination categories and paraphrase-related errors. The hallucination taxonomy used in this study is an extension of the categorization proposed by Maynez et al. (2020), which classifies types of hallucinations based on how models make errors in summarization tasks.

During preliminary investigations, we found recurring faithfulness issues that did not fit into the original intrinsic/extrinsic hallucination taxonomy for summarization proposed by Maynez et al.. We hypothesize that these issues become more visible with modern LLMs and longer, more abstractive summaries, where paraphrasing is more prevalent. Accordingly, we extend the taxonomy by adding a new category for paraphrase-related faithfulness errors. We then formally define the annotation task and evaluation settings.

Task and evaluation settings. Given a source document D and an LLM-generated summary S, we split S into a sequence of sentences $\{s_1, \ldots, s_n\}$ and annotate each sentence s_i with faithfulness labels drawn from our label taxonomy, based on its consistency with D. Our dataset supports (i) binary hallucination detection, (ii) binary faithfulness detection (faithful vs. any non-faithful label), and (iii) fine-grained sentence-level error analysis (see Section 3 for the label definitions).

The contributions of this study are as follows:

1. We extend the intrinsic/extrinsic hallucination taxonomy for summarization by introducing Paraphrase Error, a non-hallucination faithfulness label, which frequently appears in modern LLM outputs.

2. We construct a dataset with sentence-level, fine-grained faithfulness labels for evaluating hallucination detection methods targeting Japanese document summarization.

3. We analyze hallucination patterns across multiple LLMs on the same inputs, showing that hallucination occurrence is largely independent across models and highlighting strong model-specific effects.

2 Related Work

2.1 Hallucination in LLMs

The term *hallucination* originates in psychology, referring to perceiving something that does not exist. In LLM research, it describes outputs containing false or unfounded information. However, the definition and taxonomy of hallucination vary across studies, and the criteria for what constitutes hallucination vary depending on the target task (Zhang et al., 2023; Xiao and Wang, 2021; Ji et al., 2023b).

In this study, we adopt the taxonomy proposed by Huang et al. (2025). According to their framework, phenomena referred to as hallucination in LLM outputs can be divided into *factuality hallucination* and *faithfulness hallucination*.

Factuality hallucination refers to cases in which an LLM generates content that does not align with real-world facts. Examples include outputting factually incorrect information or fabricating entities or events that do not exist in reality. Accordingly, whether an output is a factuality hallucination is determined by comparing it against external real-world facts.

Faithfulness hallucination refers to cases in which an LLM generates content that deviates from the input text, violates the given instructions, or contains logical inconsistencies within the generated text. Among these, the phenomenon in which the model generates information that diverges from the input it is expected to be grounded in is categorized in Huang et al. (2025) as **context inconsistency**, a subcategory of faithfulness hallucination.

This study focuses on context inconsistency within the broader category of faithfulness hallucination. Accordingly, the dataset constructed in this work is annotated solely from the perspective of faithfulness hallucination, and does not include annotations concerning factuality hallucination. For clarity, throughout Section 3 and beyond, the term

Source Document	Generated Summary	Explanation
ルイスさんが断ると、女性は 3 人をルイスさん宅まで追跡し、警察に通報した。... ルイスさんは警察に職務質問された。(When Lewis refused, the woman followed them home and called the police. Lewis was then questioned by the police.)	ルイスさんが子供とサンドイッチ店を出たところ、女性に職務質問され、警察に通報された。(When Lewis left a sandwich shop with the children, he was questioned by a woman and reported to the police.)	*Intrinsic hallucination.* The person who conducted the questioning was not the woman; the factual relations in the source document are altered.
クラブで働く警官と撃ち合いになった後、人質をとり立てこもったが、午前 5 時ごろ、突入した警官 11 人と銃撃戦になり死亡したという。(After exchanging gunfire with an officer working at the club, the suspect took hostages and barricaded himself inside. Around 5 a.m., an armed confrontation with 11 police officers occurred, resulting in his death.)	ナイトクラブで乱射事件が発生し、50 人以上が死亡、53 人が負傷した。(A mass shooting occurred at a nightclub, and over 50 people were killed and 53 injured.)	*Extrinsic hallucination.* The summary introduces information that is not present in the source.
パキスタンが態度を変えなければ、米国から受けてきた優遇措置を失う可能性があると示唆した。(It was suggested that Pakistan may lose the preferential treatment it has received from the United States if it does not change its stance.)	優遇措置を取り消す可能性があると警告した。(It was warned that the preferential treatment might be revoked.)	*Paraphrase error.* The meaning shifts due to paraphrasing.

Table 1: Sentence-level examples of faithfulness labels used in this study. Each generated summary sentence is annotated independently with respect to the source document. Note that Paraphrase Error indicates a faithfulness issue that does not constitute hallucination.

hallucination refers specifically to **context inconsistency**.

2.2 Hallucination Detection Datasets

Several existing datasets have been proposed for evaluating hallucination or context inconsistency detection. These datasets are typically constructed through manual annotation, which has been shown to be inherently challenging for faithfulness assessment due to its subjective and fine-grained nature (Durmus et al., 2020). This section reviews representative datasets in English and Japanese, and clarifies how they differ from the dataset constructed in this work. Unlike many prior datasets that assume a single label per sentence or span, our annotations allow multiple faithfulness issues to be assigned to a single sentence.

RAGTruth (Niu et al., 2024) is an English dataset covering open-book QA, document summarization, and data-to-text generation, in which hallucination spans are manually annotated with fine-grained labels capturing contradictory and unsupported content. While RAGTruth provides detailed span-level annotations, its label design focuses on error severity and type, which differs from the perspective adopted in our study.

ANAH (Ji et al., 2024) is a dataset for the Generative Question Answering (GQA) task in English and Chinese, annotated at the sentence level. It employs a semi-automatic annotation pipeline in which GPT-4 assigns initial labels that are subsequently reviewed by human annotators, and additionally provides reference fragments and suggested corrections for each annotation.

For Japanese summarization, Iwamoto and Shimada (2024) construct a dataset for factual inconsistency detection by automatically generating inconsistent summaries using methods such as FactCC (Kryscinski et al., 2020) and SumFC (Zhang et al., 2021). The resulting dataset mainly consists of synthetic inconsistencies and is designed for training detection models. In contrast, our study builds a manually annotated dataset of LLM-generated summaries produced in a standard summarization setting, which is well suited for evaluating hallucination detection methods that leverage information available during the generation process, including internal model states (Chen et al., 2024; Ren et al., 2023).

In addition to automatically constructed datasets, JHARS (Kamei et al., 2025) is a manually annotated Japanese dataset for generative question answering. Each sentence is labeled by multiple annotators as having no hallucination, intrinsic hallucination, or extrinsic hallucination, provided that sufficient agreement is achieved. While JHARS targets GQA and contains a limited number of hallucination instances, our dataset focuses on document summarization and includes a larger set of hallucination examples, enabling a more detailed analysis of hallucination phenomena in summarization.

3 Dataset Construction

In this study, we adopt a framework for analyzing faithfulness errors in summaries, focusing on hallucinations and classifying them based on how the generation model makes errors. Specifically, we build upon the taxonomy proposed by Maynez et al. (2020) for evaluating hallucinations in abstractive summarization (Intrinsic / Extrinsic hallucination), and extend it by introducing an additional category, *Paraphrase Error*. Together with *Faithful*, which represents sentences without issues, we refer to these four labels collectively as *faithfulness labels*. Although Paraphrase Error is not a hallucination, it represents sentences that are problematic from the perspective of faithfulness, and such errors frequently appear in LLM-generated summaries. A notable feature of the taxonomy by Maynez et al. (2020) is that hallucination is not simply treated as "output inconsistent with the input," but rather categorized based on *how* the model makes errors.

3.1 Faithfulness Labels

In this study, we categorize each sentence in the output into one of the following four labels based on faithfulness. Each sentence may receive multiple labels when multiple faithfulness issues co-occur. Table 1 provides concrete examples corresponding to each label.

Intrinsic hallucination refers to errors in which the output is constructed using expressions that appear in the input, but the relationships among those expressions are incorrectly described, resulting in semantic inconsistency with the input. Such errors are often observed in the form of incorrect relationships or reversed temporal order. In the example in Table 1, the target sentence states that "a woman conducted a police questioning." Although the source document contains the expressions "woman" and "police questioning," the source text describes that the *police* conducted the questioning. Thus, intrinsic hallucination involves the use of expressions that appear in the source document, but with incorrect interpretation or description of their roles or relationships.

Extrinsic hallucination refers to errors in which the model inserts information that does not exist in the input. Such errors arise when the model generates content that cannot be inferred from the input, based on knowledge or patterns learned during training. In the example in Table 1, the target sentence states that "more than 50 people were killed and 53 were injured," but no such information, or anything that could imply it, is present in the source document.

Paraphrase Error refers to inappropriate paraphrasing that does not constitute hallucination but is problematic from the perspective of faithfulness. Our preliminary analysis revealed that LLMs often paraphrase words or phrases in ways that subtly alter the meaning. Such semantic shifts through paraphrasing are explicitly mentioned in Maynez et al. (2020) as not being hallucinations, since they neither contradict the input directly nor introduce new unsupported information. However, because paraphrasing can result in a discrepancy between the meaning understood from the output and the meaning obtainable from the input, these cases are considered unfaithful outputs. Therefore, we treat Paraphrase Error as an independent faithfulness label, distinct from hallucination.

Faithful denotes outputs that do not contain any of the aforementioned errors and are consistent with the input from the perspective of faithfulness. This label is also assigned to sentences that are unrelated to the summary content itself, such as generic statements like "Here is the summary," which sometimes appear in LLM outputs. [1]

Throughout this paper, Paraphrase Error is treated as a faithfulness issue but not as hallucination. Accordingly, depending on the evaluation setting, Paraphrase Error can be excluded from hallucination detection or included when broader faithfulness issues are of interest.

3.2 Dataset for Document Summarization

As the data source, we used the Japanese portion of XL-Sum,[2] a multilingual summarization dataset constructed from BBC News[3] articles.

In the standard construction of XL-Sum, the first paragraph of each news article is treated as the target summary, and the remaining paragraphs are treated as the source document to be summarized. This structure sometimes results in unnatural source documents or cases in which the target summary contains information not present in the source, which can artificially increase apparent hallucination rates when faithfulness is assessed against the source. To

[1] Only two sentences in the entire dataset consist of generic, content-independent statements (e.g., "Here is the summary"), and one sentence contains a degenerate repetition (an unintentionally repeated phrase).

[2] https://github.com/csebuetnlp/xl-sum

[3] https://www.bbc.com/japanese

mitigate this artifact, we reconstructed the source document by concatenating the headline and the full article body.[4]

In the Japanese subset used in this study, the source documents and target summaries contain approximately 1,700 and 150 characters, respectively.

3.3 Generation of Data for Annotation

3.3.1 Summary Generation

In this study, summaries were generated using the following three LLMs. The first is GPT-4o (gpt-4o-2024-11-20[5]), a black-box model provided by OpenAI and accessible via API. The second is Swallow (Llama-3.1-Swallow-8B-Instruct-v0.2[6] [7]), a white-box model obtained by continued pretraining of Llama on Japanese data. The third is LLM-jp (llm-jp-3-13b-instruct[8]), a white-box model pretrained primarily on Japanese, English, and source code.

All models were given the same source document and prompt, and generation was performed using greedy decoding. This choice is motivated by downstream hallucination detection tasks that require a reproducible generation process. Accordingly, we adopt greedy decoding, which yields deterministic generation behavior.

This procedure yielded three target summaries for each source document. The exact prompts used for generation are provided in Appendix A.

3.3.2 Filtering the Generated Responses

Previous work reports that only a small fraction of summaries produced by current LLMs in English contain hallucinations (Vectara, 2024). A preliminary investigation under our Japanese summarization setting similarly showed that hallucinations occur only in a small portion of generated summaries. To ensure that the dataset contains a sufficient number of hallucination cases, we used GPT-4o[5] to identify and extract only those target summaries that were likely to contain hallucinations and subjected them to annotation. The prompt used for this extraction is listed in Appendix A.

Specifically, for each source document, if at least one of the three generated summaries was judged to contain an error from the perspective of faithfulness, all three summaries were included as annotation targets. To avoid bias toward any particular model, we performed random sampling to balance the number of erroneous outputs contributed by each model.

Through this procedure, we collected a total of 390 summary outputs from the three models for 130 source documents. When split into sentences, this resulted in 1,938 sentences subject to annotation.

It should be noted that this filtering step is introduced solely to construct a benchmark suitable for evaluating hallucination detection systems, rather than to estimate the natural frequency of hallucinations in LLM-generated summaries. The filtering is recall-oriented, and all final labels are determined by human annotators based on the source documents.

3.3.3 Annotation Procedure

We built an annotation system using doccano[9], as shown in Appendix Figure 4. Annotations were conducted through the pipeline illustrated in Figure 1, following the steps below:

1. First, annotators read the news article corresponding to each source document on a website using a browser.

2. Next, annotators read the three summaries generated by the LLMs as displayed in doccano. The order of summaries is randomly shuffled in doccano so that annotators cannot identify which system produced which summary.

3. Annotators then compare the source content with each generated summary and identify descriptions that are inappropriate from the perspective of faithfulness. For each such problematic segment, they assign a "Reason" label (an auxiliary annotation label) to mark the textual span that supports their judgment. Note that at this stage, annotators only enumerate all inappropriate descriptions; they do *not* assign the final faithfulness labels.

4. Finally, for each sentence containing an inappropriate description, annotators determine which of the faithfulness labels defined in

[4] As a result, our use of XL-Sum differs from the standard practice of treating it as a train–test dataset. Additionally, the target documents originating from XL-Sum are not used as references in this study.

[5] https://platform.openai.com/docs/models/gpt-4o

[6] https://huggingface.co/tokyotech-llm/Llama-3.1-Swallow-8B-Instruct-v0.2

[7] At the time of our experiments, larger LLaMA 3.1 Swallow models were evaluated on Japanese language benchmarks and reported competitive results(Swallow LLM Project, 2024). The 8B model represents a smaller and more accessible variant within the same model family.

[8] https://huggingface.co/llm-jp/llm-jp-3-13b-instruct

[9] https://github.com/doccano/doccano

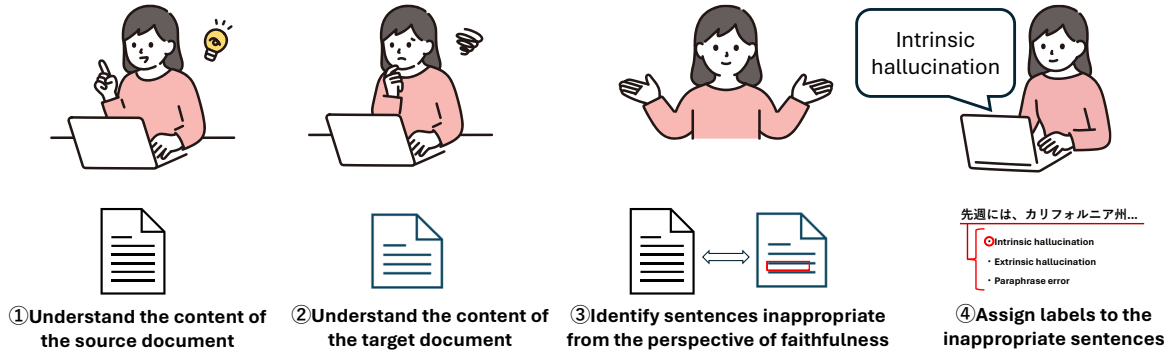

Figure 1: Annotation pipeline for sentence-level faithfulness assessment. Annotators assign one or more faithfulness labels to each sentence in a multi-label manner (with auxiliary "Reason" spans to mark supporting evidence).

this study best represents the type of error. Sentences judged not to contain inappropriate descriptions are assigned the Faithful label. Since some sentences contain multiple types of errors, annotation is conducted in a multi-label format.

Of the 390 generated summaries, 270 were annotated under a setting that included a second-pass review (**verification**), while 120 were annotated independently by each annotator without review (**non-verification**) to assess the baseline level of inter-annotator agreement.

3.3.4 Verification of Annotations

In the verification setting, annotators first conducted independent annotations. Then, for each target sentence, they were shown the distribution of labels assigned by all annotators, allowing them to reflect on differences between their own decisions and those of others. This step was introduced because preliminary experiments revealed ambiguous cases in distinguishing hallucination from Paraphrase Error. We expected that providing annotators the opportunity to align their interpretations would help improve label consistency.

During the review stage, annotators were explicitly told that they did not need to adjust their labels merely to reach full agreement. This ensures that sentences with diverging labels can be interpreted as cases in which the categorization is inherently ambiguous or cases where the labels defined in this study may not fully capture the nature of the error.

3.3.5 Annotation Results

Annotations were performed by six native Japanese-speaking university and graduate students. Since annotators can assign one or more of the four faithfulness labels to each sentence, the annotation results can be aggregated into counts such as: *Faithful: 4 annotators, Intrinsic hallucination: 2 annotators* where counts are computed independently for each label because annotators may assign multiple labels to a sentence. For evaluation convenience, we additionally derive a single sentence-level label by aggregating the multi-annotator, multi-label annotations, using the following rules:[10]

- If at least two annotators assigned a non-faithful label (i.e., a label other than Faithful), and a majority among them selected the same label, that label is adopted.
- If only one annotator assigned a non-faithful label, the sentence is marked as *Unresolved* and no label is assigned.
- If multiple annotators assigned non-faithful labels but no label achieved a majority, the sentence is marked as *Unresolved* and no label is assigned.
- If all annotators judged the sentence to contain no error, the sentence is assigned the *Faithful* label.

The threshold of two annotators was chosen in consideration of the difficulty of faithfulness judgments, which tend to be prone to oversight. Applying these rules resulted in 1,750 out of 1,938 sentences receiving a label.

Among the 188 sentences without a label, 149 cases involved only a single annotator identifying

[10]The publicly released dataset includes all annotations from the six annotators.

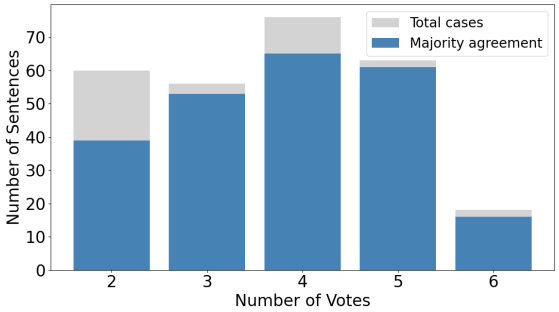

Figure 2: Distribution of sentences by the number of annotators who assigned at least one non-faithful label to a sentence.

	Intrinsic	Extrinsic	Paraphrase	Faithful
GPT-4o	43	21	30	483
Swallow	68	17	11	490
LLM-jp	40	25	3	539

Table 2: Distribution of faithfulness labels at the sentence-level. Counts are label occurrences (a sentence may contribute to multiple labels).

an error, and 39 cases involved disagreement among annotators such that no majority label emerged. Figure 2 presents a histogram showing the distribution of the number of annotators who assigned a non-faithful label, as well as the number of sentences that remained unlabeled due to disagreement.

4 Dataset Analysis

This section analyzes the dataset at both the sentence and summary levels, characterizing faithfulness issues and hallucination phenomena captured by our labels.

Key takeaway. Across different models, hallucination occurrence for the same input document exhibits near-zero mutual information, indicating little shared tendency to hallucinate on specific inputs. This suggests that hallucination generation may be influenced more by model-specific characteristics than by input difficulty, highlighting the importance of evaluating hallucination detection methods across diverse models.

4.1 Sentence-Level Faithfulness Labels

Table 2 shows the sentence-level distribution of faithfulness labels for each model. Since our annotations allow multiple faithfulness labels to be assigned to a single sentence, we report label frequencies by counting each annotated label independently when aggregating the statistics.

	Hallucinated	Paraphrased	Faithful
GPT-4o	56	18	43
Swallow	56	3	71
LLM-jp	45	1	84

Table 3: Distribution of faithfulness labels at the summary-level. (Each output is categorized as Hallucinated, Paraphrased, or Faithful based on sentence-level labels.)

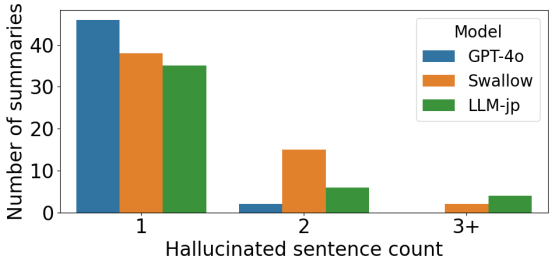

Figure 3: Distribution of the number of hallucinated sentences (Intrinsic or Extrinsic) in each generated summary output labeled as Hallucinated (i.e., containing at least one Intrinsic or Extrinsic sentence).

From Table 2, we observe that Swallow most frequently produces hallucinations (Intrinsic + Extrinsic), while GPT-4o and LLM-jp exhibit similar tendencies in terms of hallucination generation. We also find that the proportion of Paraphrase Error is higher in GPT-4o than in the other models.

A manual inspection of generated summaries suggests that GPT-4o produces a larger amount of paraphrasing than Swallow and LLM-jp. Since extensive paraphrasing typically reduces bigram overlap with the source document, we employ ROUGE-2 Recall to quantify the degree of content reuse. As shown by the ROUGE-2 Recall scores (0.611 for GPT-4o vs. 0.871/0.843 for Swallow and LLM-jp, respectively)[11], GPT-4o reuses fewer bigrams from the source document. This quantitative result is consistent with the higher rate of Paraphrase Error observed for GPT-4o.

4.2 Summary-Level Faithfulness Labels

Table 3 shows the distribution of faithfulness labels at the summary (output) level for each model. Based on the sentence-level labels, we classify each

[11] ROUGE-2 Recall is computed after morphological analysis using MeCab (version 0.996) with the IPA dictionary (IPADIC, version 102, UTF-8).

Model pair	Mutual information
GPT-4o × Swallow	0.002
Swallow × LLM-jp	0.015
GPT-4o × LLM-jp	0.006

Table 4: Mutual information between model pairs regarding whether their summaries are labeled Hallucinated (i.e., containing at least one Intrinsic or Extrinsic sentence) for the same input document.

output as *Hallucinated* (contains any Intrinsic/Extrinsic), *Paraphrased* (contains only Paraphrase Error among non-faithful sentences), or *Faithful* (all sentences are Faithful).

Figure 3 shows the distribution of hallucinated sentence counts within outputs labeled as Hallucinated. Most such outputs contain only a single hallucinated sentence, although in some cases multiple hallucinated sentences appear within the same output. Previous work (Varshney et al., 2023) reports chain-like hallucination phenomena, in which hallucinations, once initiated, trigger subsequent hallucinations. In our dataset, the number of hallucinated sentences per summary varies across models, and in particular, the proportion of summaries containing two or more hallucinated sentences is lower for GPT-4o than for the other two models. This observation suggests that the occurrence of chain-like hallucinations may be associated with model capability.

Table 4 reports the mutual information between pairs of models with respect to whether their outputs for the same input document were labeled Hallucinated. A mutual information of zero indicates statistical independence between the two outputs. Across all three model pairs, the mutual information values were close to zero, suggesting that hallucination occurrence is independent across models. This result implies that hallucination generation may be influenced more by model-specific characteristics than by input difficulty factors such as topic or complexity.

4.3 Inter-Annotator Agreement

Table 5 reports Fleiss' kappa coefficients for inter-annotator agreement in our study under both the verification and non-verification settings. We report two types of agreement scores: *faithfulness decision agreement*, which reflects binary classification of whether a sentence is faithful or not, and *label-type agreement*, which reflects agreement on the four-way classification of the specific label assigned to each sentence.

	Verified	Non-Verified
Faithfulness decision	0.51	0.39
Label-type	0.45	0.33

Table 5: Inter-annotator agreement measured by Fleiss' kappa. Faithfulness decision agreement is based on a binary classification (Faithful vs. any non-Faithful). Agreement is computed over six annotators.

Label pair / Triple	**Count**
Intrinsic–Extrinsic	5
Extrinsic–Paraphrase	6
Intrinsic–Paraphrase	14
Tie across three labels	14

Table 6: Distribution of annotation disagreement cases, showing label count patterns assigned by six annotators for sentences that remained Unresolved.

Faithfulness decision agreement. While verified agreement is moderate (0.51), Fleiss' kappa in the non-verified setting falls below 0.4. These scores are lower than those reported in prior work (Pagnoni et al., 2021) (Fleiss' kappa = 0.58), which annotates summary outputs from the perspective of faithfulness. This may reflect differences in annotation setup (e.g., the absence of annotator screening), as well as differences in the properties of the annotation targets, such as the nature of the source documents and the characteristics of summaries generated by decoder-only LLMs.

Label-type agreement. Fleiss' kappa values are even lower, indicating that maintaining consistent judgments becomes increasingly difficult as the annotation task requires finer-grained label distinctions. This highlights the challenge of fine-grained faithfulness labeling; we discuss representative unresolved cases in Limitations.

4.4 Qualitative Analysis of Annotation Disagreements

Table 6 presents the label count patterns assigned by the six annotators for cases in which no final label could be determined due to disagreement. Below, we provide a qualitative analysis of cases where multiple annotators assigned at least one label other than Faithful.

Source Document	Generated Summary	Labels
アメリカの宇宙開発企業スペース X（エックス）は... 米航空宇宙局（NASA）の宇宙飛行士 2 人を乗せた宇宙船「クルードラゴン」の打ち上げに成功した。アメリカ国内からアメリカ人が宇宙に飛び立つのは、スペースシャトルの有人飛行計画を 9 年前に中止して以来のこととなる。... (U.S. space development company SpaceX successfully launched the Crew Dragon spacecraft, carrying two astronauts from the National Aeronautics and Space Administration (NASA). It marks the first time that Americans have traveled into space from within the United States since the termination of the Space Shuttle program nine years ago.)	... スペース X の有人宇宙船が打ち上げられたのは初めてで、9 年ぶりにアメリカ人が宇宙に飛び立つことになります。(...It is the first crewed spacecraft launch by SpaceX, and for the first time in nine years, Americans will go into space.)	(a) *Non-verified* Intrinsic: 2 Extrinsic: 2 Paraphrase: 1 Faithful: 1
中国東部の青島市で配達係として働いている 7 歳の少年の話がソーシャルメディアで話題になり... 中国版ツイッター「微博（ウェイボー）」では、少年の話がトレンド入りして話題になっている。... (The story of a seven-year-old boy working as a delivery worker in Qingdao, a city in eastern China, has gone viral on social media. ... On Weibo, the Chinese version of Twitter, the story has become a trending topic.)	... 中国版ツイッター「微信」では、少年の話がトレンド入りして話題になっている。... (...On WeChat, the Chinese version of Twitter,the story has become a trending topic.)	(b) *Verified* Intrinsic: 0 Extrinsic: 2 Paraphrase: 2 Faithful: 2

Table 7: Examples of cases with annotator disagreement.

Cases that do not fit any of the defined labels. In example (a) of Table 7, the source document states that "it will be the first time in nine years that an American launches into space *from within the United States*," whereas the target sentence states, "it will be the first time in nine years that an American launches into space," without the restriction "from within the United States." As a result, the meaning conveyed by the source and target differs, making the output problematic from the perspective of faithfulness.

Errors of this kind, where a restrictive expression in the source is omitted in the target, resulting in a shift in meaning, do not fall under any of the faithfulness labels defined in this study. Because annotators attempted to map such cases into one of the provided categories, disagreement naturally arose.

Cases that are difficult to determine as Extrinsic or Paraphrase. In example (b) of Table 7, the social media platform in question is referred to as "WeChat (微信)" in the target sentence, while the source document mentions "Weibo (微博)." In reality, "Weibo" is often described as a Twitter-like microblogging platform in China, whereas "WeChat" is a distinct communication platform resembling LINE; thus, the two refer to different services.

If an annotator is unfamiliar with "Weibo," they may incorrectly interpret "WeChat" as a paraphrase and classify the case as a Paraphrase Error. However, annotators aware of the difference between the two platforms may judge that the model introduced information not present in the source and classify it as Extrinsic. These cases illustrate how the distinction between Paraphrase Error and Extrinsic hallucination can depend on an annotator's prior knowledge.

5 Conclusion

In this study, we constructed a benchmark dataset for hallucination detection in Japanese summarization, providing sentence-level annotations of multiple LLM outputs with four faithfulness labels: Intrinsic, Extrinsic, Paraphrase Error, or Faithful.

Our analysis at the output level revealed that GPT-4o exhibits a lower hallucination rate compared to the other models. At the same time, it tends to produce more paraphrased, i.e., more abstractive, summaries. Cross-model comparisons showed that hallucinations were largely independent across models, suggesting that factors other than input difficulty (e.g., model-specific characteristics) may contribute to hallucination generation.

For future work, we plan to refine the annotation framework based on the insights obtained in this study and benchmark existing and newly proposed hallucination detection methods on the dataset.

Limitations

Our study has several limitations. First, the analysis is limited to three LLMs, and the observed patterns may not generalize to all models. Second, the finding that hallucination occurrence appears largely independent across models should be interpreted in light of the input setting used in this study. Our dataset is based on news articles, which

are relatively well-structured and factually consistent; different tendencies may emerge for noisier inputs, other domains, or languages. Finally, evaluations of hallucination detection methods that rely on internal model states are currently limited to open-source models, as API-based models such as GPT-4o do not provide access to internal generation information.

Acknowledgments

This work was supported by the National Institute of Information and Communications Technology (NICT) under the "Research and Development of externally controllable modeling of multimodal information to enhance the accuracy of automatic translation."

References

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. 2024. INSIDE: LLMs' internal states retain the power of hallucination detection. In *The Twelfth International Conference on Learning Representations*.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, St. Julians, Malta. Association for Computational Linguistics.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2).

Keisuke Iwamoto and Kazutaka Shimada. 2024. Dataset construction and verification for detecting factual inconsistency in Japanese summarization. In *2024 16th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, pages 243–248.

Ziwei Ji, Yuzhe Gu, Wenwei Zhang, Chengqi Lyu, Dahua Lin, and Kai Chen. 2024. ANAH: Analytical annotation of hallucinations in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8135–8158, Bangkok, Thailand. Association for Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023a. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023b. Towards mitigating LLM hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843, Singapore. Association for Computational Linguistics.

Ryohei Kamei, Masaki Sakata, Asahi Hentona, Kentaro Kurihara, and Kentaro Inui. 2025. JHARS: Construction and analysis of a Japanese hallucination evaluation benchmark in rag settings [in Japanese]. In *Proceedings of the Thirty-first Annual Meeting of the Association for Natural Language Processing*, pages 833–838, Nagasaki, Japan. The Association for Natural Language Processing.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Barrett Lattimer, Patrick Chen, Xinyuan Zhang, and Yi Yang. 2023. Fast and accurate factual inconsistency detection over long documents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, page 1691 – 1703. Association for Computational Linguistics.

Wei Li, Wenhao Wu, Moye Chen, Jiachen Liu, Xinyan Xiao, and Hua Wu. 2022. Faithfulness in natural language generation: A systematic survey of analysis, evaluation and optimization methods. arXiv preprint arXiv:2203.05227 [cs.CL]. *Preprint*, arXiv:2203.05227.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting*

of the Association for Computational Linguistics (Volume 1: Long Papers), pages 10862–10878, Bangkok, Thailand. Association for Computational Linguistics.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. Association for Computational Linguistics.

Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J Liu. 2023. Out-of-distribution detection and selective generation for conditional language models. In *The Eleventh International Conference on Learning Representations*.

Zhongxiang Sun, Xiaoxue Zang, Kai Zheng, Yang Song, Jun Xu, Xiao Zhang, Weijie Yu, Yang Song, and Han Li. 2025. ReDeEP: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. In *The Thirteenth International Conference on Learning Representations*.

Swallow LLM Project. 2024. Swallow LLM evaluation: Japanese LLM benchmark. Online evaluation page visualizing Japanese LLM benchmark results across multiple tasks using scatter plots.

Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. 2023. A stitch in time saves nine: Detecting and mitigating hallucinations of LLMs by validating low-confidence generation. arXiv preprint arXiv:2307.03987 [cs.CL]. *Preprint*, arXiv:2307.03987.

Vectara. 2024. Hallucination leaderboard.

Yijun Xiao and William Yang Wang. 2021. On hallucination and predictive uncertainty in conditional language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2734–2744, Online. Association for Computational Linguistics.

Jiaxin Zhang, Zhuohang Li, Kamalika Das, Bradley Malin, and Sricharan Kumar. 2023. SAC^3: Reliable hallucination detection in black-box language models via semantic-aware cross-check consistency. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15445–15458, Singapore. Association for Computational Linguistics.

Sen Zhang, Jianwei Niu, and Chuyuan Wei. 2021. Fine-grained factual consistency assessment for abstractive summarization models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 107–116, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Prompts used for summary generation and data extraction

Tables 8 and 9 show the prompts used for summary generation and data extraction. Each table displays the actual Japanese prompts used and their English translations.

B Annotation UI

Figure 4 shows the annotation workflow. The source document is displayed and referenced in a browser, while the target document is viewed in doccano for annotation. If the source document contains photographs, the annotator can view them.

Although the actual source document input to the model is a plain text string, the web articles contain structured layout and images that help annotators better understand the content. We therefore expected that reading the original news page would lead to improved content comprehension.

次の文書を要約してください。 `{article}`
`Summarize the following document.` `{article}`

Table 8: Prompt used for summary generation.

以下に与える要約が元の文書に忠実であるかどうかを判断してください。要約が忠実であるとは、要約の内容が元の文書によって裏付けられていることを意味します。元の文書に反する内容や、元の文書で述べられていない内容などが含まれていないということです。 元の文書：`{input_text}` 要約：`{generated_summary}` 忠実であるかどうかの理由を段階的に説明し、最後に忠実である場合は「Yes」を、そうでない場合は「No」を出力してください。
Determine whether the summary provided below is faithful to the original document. A faithful summary means its content is supported by the original document. It does not contain content that contradicts the original document or content not stated in the original document. Original document:`{input_text}` Summary:`{generated_summary}` Explain step-by-step why it is faithful or not. Finally, output “Yes” if faithful, or “No” if not.

Table 9: Prompt used to extract summaries potentially containing hallucinations.

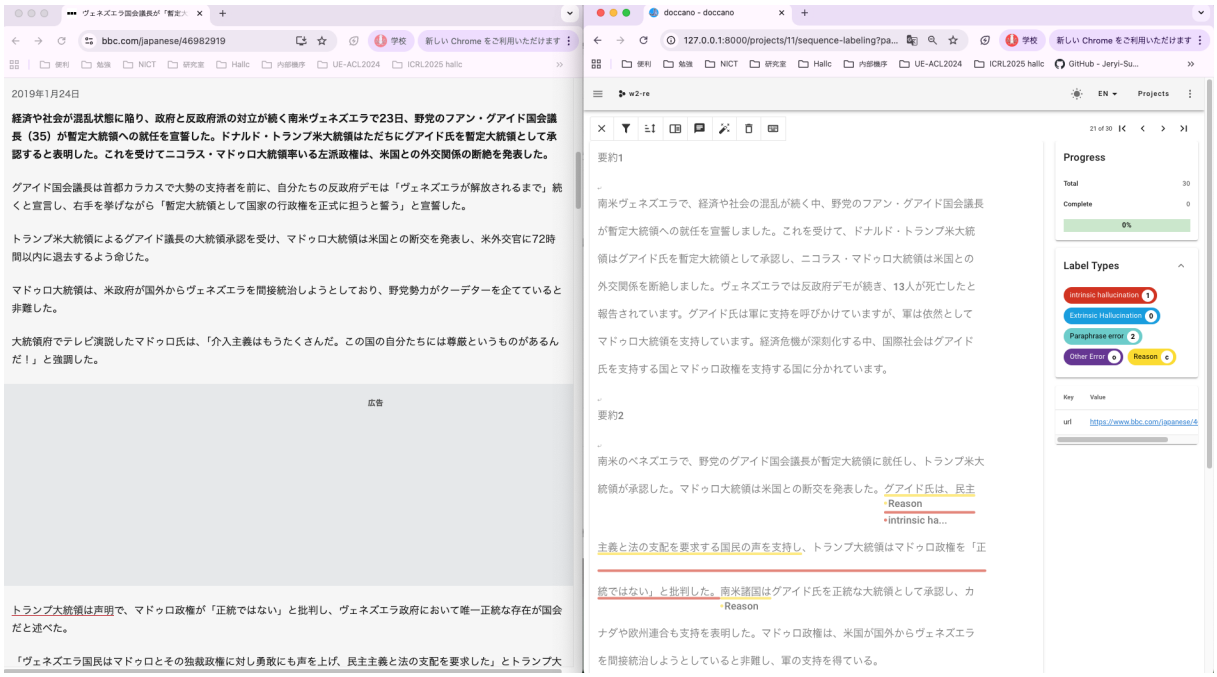

Figure 4: Annotation UI. The source news article is displayed in a web browser (left), while LLM-generated summaries are shown in doccano (right) for sentence-level faithfulness annotation, with auxiliary “Reason” spans used to mark supporting evidence for annotators’ judgments.