Constructing a Japanese Verdict Prediction Dataset for Fact-Checking of LLM-Generated Texts

Miwa Masano[1, 2*], Hirokazu Kiyomaru[2], Atsushi Keyaki[1*], Kaito Horio[3], Rei Minamoto[3, 2], Ribeka Keyaki[4], Kouta Nakayama[2], Hideyuki Tachibana[2], Daisuke Kawahara[3, 2]

[1]Hitotsubashi University, [2]National Institute of Informatics
Research and Development Center for Large Language Models,
[3]Waseda University, [4]Tokyo University of Technology

∗**Corresponding authors:** 5123053k@hit-u.ac.jp, a.keyaki@r.hit-u.ac.jp

Abstract

The development of fact-checking systems for verifying the factuality of text generated by large language models (LLMs) has been advancing. In the verdict prediction step of such systems, the system determines whether claims in the generated text are supported by retrieved evidence, formulated as a natural language inference (NLI) task. This study extends the label set for verdict prediction to capture claim-evidence relationships that humans would commonly interpret as supported or refuted, even in the absence of strict logical entailment or contradiction. It also constructs a Japanese dataset comprising 28,147 instances from two sources based on this extended label set. We analyze the causes of annotation disagreement and find that ambiguity in the boundary of acceptable inference, interpretive characteristics of negative cases, and incomplete information in the evidence affect annotation variability. Using this dataset, we evaluate the performance of prompt-based verdict prediction methods and show that prompts that explicitly elicit chain-of-thought reasoning improve F1 by 4 percentage points compared to baseline.

1 Introduction

Text generated by Large Language Models (LLMs) may contain plausible but incorrect information known as hallucinations (Huang et al., 2025). Against this background, research and development have been advancing to apply fact-checking systems (Guo et al., 2022; Zeng et al., 2021; Kotonya and Toni, 2020; Kamoi et al., 2023) to text generated by LLMs (Wang et al., 2024). A fact-checking system consists of three steps: claim decomposition, evidence retrieval, and verdict prediction. In claim decomposition, the input text is decomposed into claims, which are atomic units of information that express properties or relations about a single entity or event. Next, in evidence retrieval, the system uses each decomposed claim as a query to search an evidence database that stores supporting information and retrieves evidence for each claim. Finally, in the verdict prediction step shown in Figure 1, the system performs Natural Language Inference (NLI) on each pair of a claim and its retrieved evidence and determines whether the claim is supported (entailed) by the evidence.

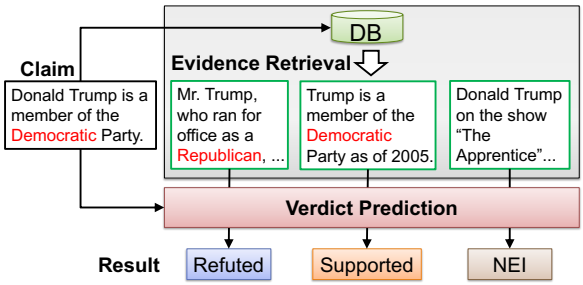

Figure 1: Overview of verdict prediction

This paper focuses on verdict prediction. Verdict prediction directly affects the final output of a fact-checking system and largely determines its overall performance. To develop a high-performance verdict prediction method in a fact-checking system, an evaluation dataset that includes claims, evidence, and labels is essential.

In this study, we construct a Japanese evaluation dataset for verdict prediction on claims obtained from generated text in a previously developed claim decomposition dataset (Masano et al., 2026). To enable flexible evaluation of the relationship between claims and evidence, we introduce new labels, inferentially-supported and inferentially-refuted. We add these to the labels used in prior work (Wang et al., 2024), namely supported, partially-supported, refuted, and not enough information (NEI), where NEI is assigned to cases that do not fall under any of the other labels. We develop annotation guidelines for verdict prediction and construct the dataset by annotating claim-evidence pairs accordingly.

We analyze the causes of disagreement in anno-

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 1139–1151

tation in the constructed dataset. The results show that ambiguity in the boundary of acceptable inference, interpretive characteristics of negative cases, and incomplete information in the evidence can lead to disagreement.

We construct a prompt-based verdict prediction method using GPT-4o and evaluate its performance based on the constructed dataset. The results show that using a Chain-of-Thought prompt that explicitly describes the procedure for verdict prediction improves performance.

2 Related Work

2.1 Natural Language Inference (NLI)

Natural Language Inference (NLI) is a task that analyzes the relationship between a given premise and hypothesis and determines whether the hypothesis can be logically inferred from the premise. Representative datasets for NLI include SNLI (Bowman et al., 2015), MultiNLI (Williams et al., 2018), and XNLI (Conneau et al., 2018). These datasets have been widely used as benchmarks for evaluating reasoning ability even after the emergence of LLMs. In these datasets, the relationship between a premise and a hypothesis is represented by three labels: entailment, when the hypothesis can be logically inferred from the premise; contradiction, when the hypothesis conflicts with the premise; and neutral, when neither condition holds. In addition to the above labels, some studies introduce partial entailment, which indicates that the hypothesis supports part of the premise (Nielsen et al., 2008; Dzikovska et al., 2013).

These NLI datasets are constructed by first providing a premise and then generating a hypothesis based on a predefined label. In other words, the premise and hypothesis are intentionally created within the same context, and assumptions such as topic, entity, and time are treated as aligned even when they are not explicitly stated. As a result, the relationship between the two is relatively clear, and ambiguity is less likely to arise.

In contrast, in verdict prediction in fact-checking, a claim, which corresponds to the hypothesis, is given first, and evidence, which corresponds to the premise, is subsequently retrieved through evidence retrieval, after which a label is assigned based on the relationship between them. Since evidence is obtained from external sources such as web content, the claims and the evidence originate form independently collected and different sources. Therefore, they are not intentionally created to refer to the same topic, event, time, or location.

As a result, there are cases where differences in expression or lack of information make it difficult to determine whether a claim is fully supported or refuted, even when it appears plausible. However, in constructing a fact-checking system, it is not desirable from the perspective of system usefulness to label all such non-strict cases as not enough information. To address this issue, we introduce inferentially-supported and inferentially-refuted for cases where support or refutation can be determined flexibly by supplementing missing information through inference.

2.2 Fact-checking System

Research on fact-checking has traditionally focused on domains such as fake news (Guo et al., 2022; Zeng et al., 2021; Kotonya and Toni, 2020). Subsequently, FEVER (Thorne et al., 2018), a representative dataset for fact-checking, and its successor datasets (Thorne et al., 2019; Aly et al., 2021; Schlichtkrull et al., 2023) have been introduced and have driven research on fact-checking systems. In these datasets, the labels for verdict prediction are supported, refuted, and not enough information.

WiCE (Kamoi et al., 2023) assigns verdict prediction labels at the token level in addition to claim-level labels, which enables fact-checking at a finer granularity. In WiCE, claims are extracted from Wikipedia articles, and the web pages cited in those sentences are used as evidence. The labels for verdict prediction are supported, partially-supported, and not-supported.

Factcheck-Bench (Wang et al., 2024), a framework for fact-checking targeting text generated by LLMs, proposes a fact-checking framework that consists of more fine-grained steps than the aforementioned fact-checking system and constructs a benchmark. Factcheck-Bench uses four labels for verdict prediction: support, partially support, refute, and irrelevant. The label irrelevant indicates that the evidence contains no information related to the claim.

3 Construction of the Verdict Prediction Dataset

We construct a dataset for the quantitative evaluation and analysis of verdict prediction. Each instance in the dataset consists of a triplet of a claim, evidence, and a verdict prediction label.

3.1 Verdict Prediction Labels

We use the following label set:

- **Fully-supported**: The evidence contains explicitly stated information that supports the entire claim.
- **Inferentially-supported**: The evidence contains information that supports the entire claim, but only through implicit inference or background knowledge.
- **Partially-supported**: The evidence contains explicitly stated information that supports the main part of the claim, but not all of its details.
- **Fully-refuted**: The evidence contains explicitly stated information that contradicts the claim.
- **Inferentially-refuted**: The evidence contains information that contradicts the claim, but only through implicit inference or background knowledge.
- **Not enough information (NEI)**: The evidence provides neither explicitly stated nor inferable information sufficient to assign any support or refutation label.

Fully supported and fully refuted correspond to entailment and contradiction in NLI, respectively. Inferentially-supported, partially-supported, inferentially-refuted, and NEI are all assigned the neutral label in NLI, but this dataset treats them as distinct labels for more flexible evaluation. Example (1) shows an instance of inferentially-supported.

(1) **Claim**: 天平文化は聖武天皇の時代に最盛期を迎えました。

'Tenpyo culture reached its peak during the reign of Emperor Shomu.'
Evidence: 天平文化は聖武天皇のころの文化です。

'Tenpyo culture is the culture of the period of Emperor Shomu.'
Label: inferentially-supported

The evidence in Example (1) does not explicitly support the claim. However, from a pragmatic perspective, the expression "culture of the period of " in the evidence typically refers to the central period of that culture and is unlikely to refer to its formative or declining stages. Therefore, it is reasonable as a natural human interpretation to infer that the reign of Emperor Shomu coincided with the peak of Tenpyo culture. By distinguishing such cases from NEI, the dataset enables evaluation of both logically strict verdict prediction and verdict prediction that incorporates the kinds of inferences humans naturally make.

Example (2) shows an instance of partially-supported.

(2) **Claim**: ボブサップはアメリカ合衆国コロラド州出身です。

'Bob Sapp is from Colorado, United States.'
Evidence: ボブサップはアメリカ合衆国出身です。

'Bob Sapp is from the United States.'
Label: partially-supported

Partially-supported is used for cases where the entire claim cannot be fully supported, but there is no evidence that explicitly contradicts the additional details included in the claim. By distinguishing such cases, the dataset enables evaluation of verdict prediction that accounts for differences in the granularity of information. For refutation, if any part of a claim contradicts the evidence, the entire claim is considered contradictory. Therefore, we do not introduce a label for partial refutation.

3.2 Annotation Procedure

The verdict prediction dataset includes claims and evidence. The claims are taken from a previously constructed Japanese claim decomposition dataset (Masano et al., 2026). The Japanese claim decomposition dataset is created by annotating claim decomposition on text generated by LLM-jp-3 13B Instruct[1]. The model takes as input questions from the Japanese QA dataset "AI Official Dataset Version 2.0[2]" (AIO) and user utterances from the Japanese dialogue dataset "LLM-jp Chatbot Arena Conversations[3]" (CBA).[4]

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

[2] https://sites.google.com/view/project-aio/dataset

[3] https://huggingface.co/datasets/llm-jp/llm-jp-chatbot-arena-conversations

[4] Because the input is in Japanese, we selected a model specialized for Japanese. We also considered that the model is open and allows its training corpus to be used as a source of evidence in verification. Furthermore, since the goal of this study is to detect hallucinations, we employed a model that is prone to generating hallucinations to some extent.

To assign verdict prediction labels, we first retrieve evidence for each claim. We use the LLM-jp Corpus v3[5] as the source of evidence. To obtain diverse documents while maintaining relevance to each claim, we rerank the results of full-text search based on BM25 (Robertson et al., 1995) using Maximal Marginal Relevance (Carbonell and Goldstein, 1998) and select the top five documents.[6] In addition, the input text to the LLM may contain information that serves as evidence for the generated text, such as source text for summarization. Taking this into account, we add the input text corresponding to each claim to the five retrieved documents and use the resulting set of six documents as the evidence set for the claim.

We manually annotate verdict prediction labels for each claim-evidence pair. First, two annotators independently assign verdict prediction labels to each pair as primary annotators. When their judgments do not agree, a third annotator reviews their annotations and determines the final label.[7] When annotators are unable to understand the target data, for example when specialized knowledge such as chemistry is required, they assign the label not understandable. Instances assigned this label are excluded from the evaluation dataset. In addition, to prevent annotation errors due to inattention, annotators also mark the supporting spans in the evidence when assigning supported or refuted labels. This allows the dataset to be used for evaluation of fact-checking systems that present not only verdict prediction but also the supporting evidence.

In total, we obtain verdict prediction annotations for 13,642 claim-evidence pairs in AIO and 14,505 pairs in CBA, excluding instances labeled as not understandable. We split the constructed dataset into development and test sets at a 1:1 ratio while maintaining similar label distributions. As a result, the AIO dataset contains 6,818 instances in the development set and 6,824 in the test set, and the CBA dataset contains 7,256 instances in the development set and 7,249 in the test set.

[5] https://gitlab.llm-jp.nii.ac.jp/datasets/llm-jp-corpus-v3

[6] While we observed multiple cases in which the evidence discussed the same topic as the claim but did not contain the main subject term, evidence completely unrelated to the claim was rarely selected.

[7] The third annotator has comparable expertise to the primary annotators; however, because they can refer to both annotators' judgments, they are expected to make the final decision from a more comprehensive perspective.

	AIO		CBA	
Label	Count	Ratio	Count	Ratio
Fully-supported	3,983	0.292	1,798	0.124
Inferentially-supported	1,198	0.088	1,033	0.071
Partially-supported	555	0.041	436	0.030
Fully-refuted	173	0.013	62	0.004
Inferentially-refuted	70	0.005	64	0.004
NEI	7,663	0.562	11,112	0.766

Table 1: Label distribution of the dataset

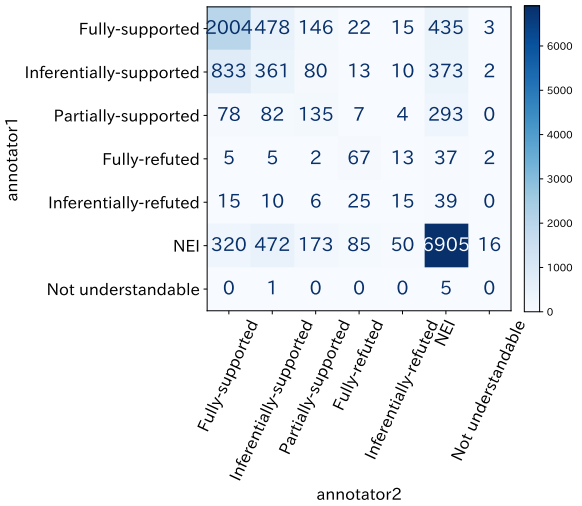

Figure 2: Confusion matrix of the primary annotators' annotations for AIO

3.3 Dataset Statistics

This section reports the label distribution and inter-annotator agreement in order to examine the characteristics of the constructed verdict prediction dataset.

Table 1 shows the distribution of verdict prediction labels in the dataset, excluding instances labeled as not understandable. Overall, NEI is the most frequent label, followed by fully-supported.

Next, to evaluate annotation reliability, we analyze the agreement between the two primary annotators. Cohen's κ coefficient for their annotations is 0.48 for AIO and 0.34 for CBA. These results reflect the inherent difficulty of the task. In verdict prediction, some cases require inference to interpret the relationship between a claim and evidence. In such cases, the correct judgment is not uniquely determined. This also suggests room for improvement in the annotation guidelines, particularly in the definitions of labels and the decision criteria.

The confusion matrices of the primary annotators' annotations are shown in Figure 2 (AIO) and Figure 3 (CBA). Figure 2 shows that, in AIO, the diagonal elements are not the largest in either the

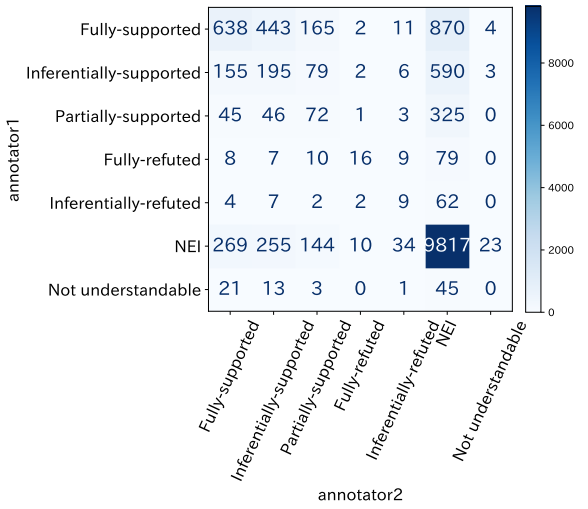

Figure 3: Confusion matrix of the primary annotators' annotations for CBA

row or column direction for inferentially-supported, partially-supported, and inferentially-refuted. In other words, for these labels, even when one annotator assigns a given label, the other annotator often assigns a different label.

For example, in the case of inferentially-supported, many instances that Annotator 1 labels as inferentially-supported are labeled as fully-supported by Annotator 2, followed by cases labeled as NEI. A similar tendency is observed when examining instances that Annotator 2 labels as inferentially-supported.

Meanwhile, Figure 3 shows that, in CBA, disagreements with NEI are observed for many labels. When using Annotator 1's judgments as the reference, the label most frequently assigned by Annotator 2 is NEI in all cases. A similar tendency is observed when using Annotator 2's judgments as the reference, where a large proportion of instances are labeled as NEI.

In the next section, we analyze the factors that lead to annotation disagreement by examining instances where the annotators' labels do not match.

4 Analysis of Annotation Disagreement

We analyze the factors that lead to annotation variability by examining cases where the two primary annotators do not agree. The results suggest that annotation disagreement mainly arises from the following three factors: (1) ambiguity in the acceptable range of inference, (2) characteristics of refutation labels, and (3) incompleteness of information in the evidence. For readability, only English translations of the examples are provided below; the original Japanese examples are given in the Appendix D.

4.1 Ambiguity in the Boundary of Acceptable Inference

The newly introduced inferential labels, inferentially-supported and inferentially-refuted, require judgment based on inference. However, the criteria for what level of inference is acceptable are ambiguous. While shallow inference is permitted, inference involving substantial leaps may not be allowed. This ambiguity in the boundary of acceptable inference leads to label disagreement.

4.1.1 Boundary between Fully-Supported and Inferentially-Supported

At the boundary between fully-supported and inferentially-supported, we observe variability in judgments regarding the acceptable range of inference. In this study, the criterion for fully-supported is whether the evidence contains a statement that expresses content equivalent to the claim, allowing for differences in word order and synonymous expressions. However, we do not provide a clear definition of the scope of synonymous expressions. As a result, annotators differ in their judgments on whether a given expression should be regarded as substantially equivalent to the claim.

(3) **Claim**：The Great Fire of Meireki burned down Edo.
Evidence: This fire, known as the Great Fire of Meireki, is famous as the largest fire in the history of Edo, burning many samurai residences and townhouses, including the main enclosure of Edo Castle.

In Example (3), the point at issue is whether the claim "burned down Edo" is sufficiently supported by the evidence stating that it "burned many samurai residences and townhouses, including the main enclosure of Edo Castle." The expression "burned down" literally implies that the entire target was completely destroyed by fire. In contrast, the evidence only states that many samurai residences and townhouses were burned and does not explicitly state that all of Edo was completely destroyed. Therefore, if "burned down" is interpreted literally, the evidence cannot be said to express exactly the same content.

However, the expression "burned down" in the claim can also be interpreted as a figurative ex-

pression that emphasizes the scale of the damage caused by the fire. Based on this interpretation, the label fully-supported may have been assigned. On the other hand, if "burned down" and "burned many" are not regarded as synonymous, the label inferentially-supported is likely to have been assigned.

A similar type of disagreement due to the acceptable range of inference is also observed at the boundary between fully-refuted and inferentially-refuted. However, the number of such cases is limited, and for refutation labels, issues related to determining whether the claim and evidence are mutually exclusive, as discussed later, are more prominent.

4.1.2 Boundary between Inferential Labels and NEI

The acceptable range of inference also causes disagreement at the boundary between inferential labels and NEI. In cases where the evidence is not entirely unrelated to the claim but requires substantial inference to determine support or refutation, annotators differ in their judgments.

(4) **Claim**：The Rust Belt developed through automobile manufacturing in the past.
Evidence: It is that he won almost all states in the industrial region known as the "Rust Belt"(note). Note: At the time of writing, the final results for Michigan have not been announced, but based on the vote count, Trump's victory in the state is almost certain. And this is precisely the reason for Trump's victory.
The main factor behind Trump's victory
States such as Pennsylvania, Ohio, Michigan, and Indiana, with a concentration of factories in the automobile and steel industries

In Example (4), the first part of the evidence mentions an industrial region called the "Rust Belt," while the latter part states that several states once had a concentration of factories in the automobile and steel industries. However, it is not explicitly stated that these states are part of the Rust Belt. Therefore, to regard the claim as supported, the following inference is required.

1. Interpreting Pennsylvania, Ohio, Michigan, and Indiana as states included in the Rust Belt
2. Interpreting "concentration" at the end of the evidence as indicating "had a concentration"

Annotators who regard these inferences as valid assign inferentially-supported, while those who do not assign NEI. This example is characterized by the fact that the evidence is not unrelated to the claim, yet multiple steps of inference are required to establish a support relation. In situations where there is relevance but insufficient explicitness, differences in the acceptable level of inference lead to disagreement between inferentially-supported and NEI. Similar cases are also observed between inferentially-refuted and NEI.

4.2 Interpretive Characteristics of Refutation Labels

For refutation labels (fully-refuted and inferentially-refuted), we observe a different type of disagreement from that in support labels (fully-supported, inferentially-supported, and partially-supported).

In the case of support labels, it is sufficient to check whether the evidence contains statements that could support the claim. In contrast, for refutation labels, when the evidence contains statements that describe facts different from the claim, it is necessary to determine whether those statements are logically compatible with the claim, that is, whether they are mutually exclusive. Therefore, the judgment process is more complex than for support labels, as it requires a two-step decision.

When logical incompatibility is obvious, annotators tend to agree and assign fully-refuted. For example, this occurs when the name of the youngest person in a group or the specific date of an event differs between the claim and the evidence. In these cases, where uniquely determined facts clearly differ, mutual exclusivity is clear, leading to consistent judgments among annotators.

On the other hand, when it is not clear whether the statements are logically compatible, label disagreement occurs.

(5) **Claim**：The milt of Alaska pollock is processed into mentaiko [seasoned pollock roe].
Evidence: The salted tarako [pollock roe] is marinated in a seasoning made with red chili peppers is called "karashi mentaiko" [spicy seasoned pollock roe].
The eggs of Alaska pollock have fine grains and a moist, sticky texture, and are very

tasty whether boiled or grilled.

The milt of Alaska pollock is also called suketachi. It is more affordable than cod milt and, although it does not match the quality of cod milt, it is still quite tasty.

In Example (5), the claim states that "the milt of Alaska pollock is processed into mentaiko," while the evidence explains that "karashi mentaiko is made by processing tarako." That is, the evidence indicates that mentaiko is produced by processing tarako, but it does not mention the relationship between milt and mentaiko.

If one assumes that only processed tarako is called mentaiko and that processed milt is not referred to as mentaiko, the claim may be considered incorrect. Based on this interpretation, inferentially-refuted may have been assigned. On the other hand, the evidence does not explicitly state that milt is not processed into mentaiko. Therefore, if it is judged that the claim and the evidence are not necessarily logically incompatible, the label NEI is likely to have been assigned.

4.3 Incomplete Information in the Evidence

Even when the information contained in the evidence is fragmentary, fully-supported or fully-refuted may be assigned if the content of the claim can be clearly supported or refuted by interpreting multiple statements in relation to each other. However, due to limitations in the retrieval process, evidence may contain missing fragments, lack contextual information, or appear in an inappropriate format. As a result, annotators may differ in their judgments depending on how they interpret information that is not explicitly stated in the evidence. This type of error is not limited to specific labels and is observed across all labels.

First, some expressions in the evidence are partially missing. In some cases, named entities or expressions corresponding to keywords in the claim appear at the beginning or end of the evidence span and are partially missing. For example, a fragment such as "-uglena" may appear instead of the full noun "Euglena" mentioned in the claim. In such cases, it is unclear whether the expression refers to the same entity as in the claim, and differences in interpretation lead to annotation disagreement.

Second, we observe cases in which the evidence lacks sufficient contextual information. Fragmented paragraphs or the absence of explicit relationships between elements can make it difficult to determine relationships between statements, as the contextual information needed to connect sentences or words is insufficient. For example, when the content of a claim is described across multiple paragraphs in the evidence, the relationships between those paragraphs may not be clearly indicated. In such cases, judgments vary depending on whether connections between sentences are considered part of natural reading or require inference.

Third, we observe cases where inappropriate evidence is retrieved. Some evidence does not preserve the original document structure and includes mixed elements such as headings and fragmented phrases, failing to meet the requirements for reliable evidence. This is likely caused by the loss of structural information and the inclusion of fragments such as tags during HTML-to-plain-text conversion. Such text may contain words related to the claim, but the intent of the information is unclear and it cannot be regarded as reliable evidence.

The three types of cases described above are all errors caused by formal deficiencies in the evidence. In these cases, annotators differ in their judgments depending on how much they compensate for incomplete information in the evidence, which leads to annotation disagreement.

5 Prompt-Based Verdict Prediction

In this section, we evaluate prompt-based methods using the constructed verdict prediction dataset.

5.1 Experimental Setup

We design four types of prompts and conduct experiments to assign six labels: fully-supported, partially-supported, inferentially-supported, fully-refuted, inferentially-refuted, and NEI.

First, we define a **base** prompt that includes only the task instruction and the definitions of claims and labels. In addition, we evaluate few-shot prompting and Chain-of-Thought (CoT) prompting, which are representative prompting methods. In few-shot prompting, examples consisting of a claim, evidence, and the correct label are added to the base prompt. In this study, we use examples extracted from the development set and prepare two settings, **6-shot** and **12-shot**, to analyze the relationship between the number of examples and verdict prediction performance. In **CoT**, the prompt includes the content of the base prompt as well as instructions for the reasoning procedure and examples of step-by-step reasoning that follow

Prompt	Acc.	Prec.	Rec.	F1
base	0.669	0.408	0.475	0.418
+ 6-shot	0.650	0.387	0.482	0.402
+ 12-shot	0.653	0.388	0.482	0.401
+ CoT 6-shot	**0.704**	**0.440**	**0.507**	**0.459**

Table 2: Prompt-based verdict prediction

label	base	6-shot	12-shot	CoT
Fully-supported	0.754	0.760	0.762	**0.773**
Inferentially-supported	0.157	0.231	0.223	**0.278**
Partially-supported	0.202	0.191	0.201	**0.299**
Fully-refuted	0.435	0.391	0.386	**0.481**
Inferentially-refuted	**0.129**	0.068	0.071	0.077
NEI	0.817	0.771	0.773	**0.818**

Table 3: F1 score for each label for each prompt

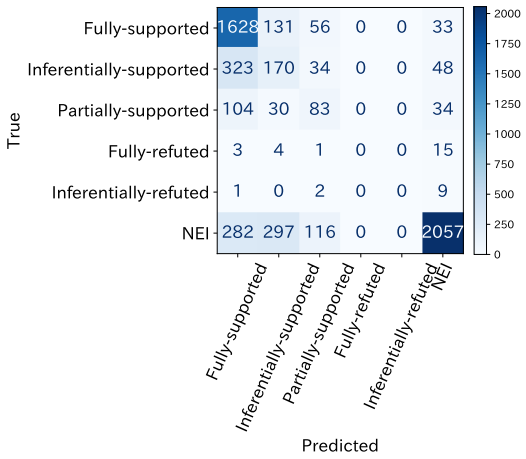

Figure 4: Confusion matrix for the CoT prompt

those instructions. In this study, the authors create reasoning processes for the examples used in the 6-shot setting. The prompt that specifies the reasoning process is shown in Appendix A.

For evaluating verdict prediction performance, we compare the labels assigned by the model with the gold labels in the dataset and report accuracy, precision, recall, and F1 as macro averages. The LLM used in the experiments is `gpt-4o`[8] [9]. Each method is run three times on the AIO test set, and the average results are reported[10].

5.2 Experimental Results

As shown in Table 2, CoT achieves the best performance among the four prompts. Table 3 shows the F1 score for each label for all methods. Compared with 6-shot, CoT shows an improvement trend for all labels, with particularly large improvements in partially-supported and fully-refuted. Figure 4 presents a confusion matrix constructed from the outputs of CoT for the first run. Based on a comparison between Figure 4 and the confusion matrix for 6-shot, the tendency to incorrectly predict inferentially-supported and fully-supported as partially-supported is reduced, and the number of cases where NEI is incorrectly predicted as fully-refuted decreases. As a result, the performance for these two labels improves.

Few-shot prompting with added examples to the base prompt (6-shot and 12-shot) results in decreased performance in all cases. A detailed analysis is provided in Appendix B.1.

Increasing examples from 6-shot to 12-shot does not improve performance. Analyzing the factors limiting performance gains with few-shot prompting remains future work.

According to Table 3, even with CoT, which achieves the best overall performance, inferentially-refuted shows the lowest prediction performance. Figure 4 shows that errors for the inferentially-refuted label often involve predicting it as fully-refuted or predicting NEI as inferentially-refuted.

In cases where inferentially-refuted is incorrectly predicted as fully-refuted, we observe instances where the model assumes that the texts refer to the same entity and makes a judgment, even though the relevant conditions are not explicitly stated. We also observe cases where the model makes a judgment without considering ambiguity in sentence interpretation. In cases where NEI is incorrectly predicted as inferentially-refuted, the model sometimes draws a premature conclusion by assuming that entities not mentioned in the evidence do not exist. Specific examples of these cases are provided in Appendix B.2.

In addition, analysis of the recall for refutation labels and the precision for support labels in Appendix C suggests that CoT is an effective method for the purpose of hallucination detection in fact-checking systems.

6 Discussion

In the dataset analysis described in Section 3.3, a high degree of annotation disagreement is observed

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

[9] As this model is known to exhibit generally high performance, it was adopted as a baseline for comparison in this study.

[10] The designed prompts did not function well for the input text to the LLM, so they are excluded from subsequent experiments.

for inferentially-supported, partially-supported, and inferentially-refuted. In addition, in the experiments on verdict prediction using LLMs described in Section 5, inferentially-refuted shows consistently low performance across all methods. As shown in Table 3, inferentially-supported and partially-supported also exhibit relatively low performance in addition to inferentially-refuted.

The similarity between human annotations and LLM predictions suggests that the difficulty of inference-based cases in this dataset may also affect the experimental results. Furthermore, as discussed in Section 4, lack of contextual information in the evidence and ambiguity in the acceptable range of inference contribute to annotation disagreement. These factors may also affect LLM predictions. Therefore, in constructing a verdict prediction dataset, it is necessary to consider improvements in evidence retrieval methods and to develop annotation guidelines that clearly define the acceptable range of inference.

7 Conclusion

We construct a Japanese verdict prediction dataset for fact-checking of LLM-generated text. We analyze annotation disagreement and clarify the difficulty of verdict prediction. We also confirm that prompts that include reasoning process improve verdict prediction performance. However, errors remain frequent in cases that involve ambiguity. Future work will aim to improve performance by refining annotation guidelines, expanding the dataset, and improving prompt design.

Limitations

This study has several limitations. First, the dataset construction is limited to AIO and CBA. It remains unclear whether the verdict prediction labels and annotation criteria proposed in this study are effective for generated text and claims derived from other datasets. In addition, the scope of this study is limited to Japanese, and its applicability to other languages has not been examined.

Second, this study uses LLM-jp-3 13B Instruct for response generation and GPT-4o for verdict prediction experiments, with both dataset construction and performance evaluation are conducted using a single model in each case. Therefore, differences in performance due to model selection and robustness across different models have not been thoroughly examined.

Third, this study does not conduct a quantitative evaluation of the retrieval results used to obtain evidence, and the impact of retrieval performance on verdict prediction has not been sufficiently analyzed.

Finally, as shown in Section 4, annotation variability exists among annotators in this dataset, and this disagreement may affect not only the quality of the dataset but also the evaluation of model performance.

Acknowledgements

We thank the anonymous reviewers for their valuable and constructive feedback. The work was partially supported by the JSPS the Grant-in-Aid for Scientific Research (B) (#23K28375, #25K03178) and Scientific Research (C) (#24K15066).

References

Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. The Fact Extraction and VERification Over Unstructured and Structured information (FEVEROUS) Shared Task. In *Proc. of the Fourth Workshop on Fact Extraction and VERification (FEVER)*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Myroslava Dzikovska, Rodney Nielsen, Chris Brew, Claudia Leacock, Danilo Giampiccolo, Luisa Bentivogli, Peter Clark, Ido Dagan, and Hoa Trang Dang. 2013. SemEval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 263–274, Atlanta, Georgia, USA. Association for Computational Linguistics.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A Survey on Automated Fact-Checking. *Transactions of the Association for Computational Linguistics (TACL)*, 10:178–206.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. 2023. WiCE: Real-World Entailment for Claims in Wikipedia. In *Proc. of the EMNLP 2023*.

Neema Kotonya and Francesca Toni. 2020. Explainable Automated Fact-Checking: A Survey. In *Proc. of the COLING 2020*.

Miwa Masano, Ribeka Keyaki, Atsushi Keyaki, Rei Minamoto, Kaito Horio, Hirokazu Kiyomaru, Kouta Nakayama, Hideyuki Tachibana, and Daisuke Kawahara. 2026. Constructing a Japanese Claim Decomposition Dataset for Fact-Checking of LLM-Generated Texts. In *International Conference on Language Resources and Evaluation (LREC2026)*.

Rodney D. Nielsen, Wayne Ward, James Martin, and Martha Palmer. 2008. Annotating students' understanding of science concepts. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, and 1 others. 1995. *Okapi at TREC-3*. British Library Research and Development Department.

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. The Fact Extraction and VERification Over Unstructured and Structured information (FEVEROUS) Shared Task. In *Proc. of the NeurIPS 2023*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *Proc. of the NAACL 2018*.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2019. The FEVER2.0 Shared Task. In *Proc. of the Second Workshop on Fact Extraction and VERification (FEVER)*.

Yuxia Wang, Revanth Gangi Reddy, Zain Muhammad Mujahid, Arnav Arora, Aleksandr Rubashevskii, Jiahui Geng, Osama Mohammed Afzal, Liangming Pan, Nadav Borenstein, Aditya Pillai, Isabelle Augenstein, Iryna Gurevych, and Preslav Nakov. 2024. Factcheck-Bench: Fine-Grained Evaluation Benchmark for Automatic Fact-checkers. In *Proc. of the Findings of the EMNLP 2024*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana. Association for Computational Linguistics.

Xia Zeng, Amani S. Abumansour, and Arkaitz Zubiaga. 2021. Automated fact-checking: A survey. *Language and Linguistics Compass*, 15.

A Instructions for the Reasoning Procedure in the CoT Prompt

Please perform reasoning according to the following procedure and output the reasoning process and the answer.

1. Analyze the meaning of the claim and identify what is being asserted.
2. Determine whether the evidence contains content relevant to the claim. If it does, extract that content. If it does not, stop the reasoning, assign "NEI," and proceed to the answer.
3. Determine what stance the content extracted from the evidence takes toward the claim: "fully-refuted," "inferentially-refuted," "fully-supported," "inferentially-supported," "partially-supported," or "NEI."
4. Determine which of the following labels applies: "fully-refuted," "inferentially-refuted," "fully-supported," "inferentially-supported," "partially-supported," or "NEI."

Note that the prompts used in the experiments are in Japanese.

B Analysis of Results

B.1 Label-wise Analysis of base and few-shot

As shown in Table 2, few-shot does not improve performance compared with the base prompt. A label-wise analysis shows that, the error of predicting inferentially-supported as partially-supported, which is prominent in the base prompt, is reduced in few-shot, leading to improved performance for inferentially-supported. On the other hand, for inferentially-refuted, fully-refuted, and NEI, both base and few-shot show tendencies to predict inferentially-refuted as fully-refuted and to predict NEI as inferentially-refuted or fully-refuted, and these errors increase in few-shot.

B.2 Error Analysis of Inferentially-Refuted

Analysis of cases where inferentially-refuted is predicted as fully-refuted reveals the following two types of cases.

The first type involves cases where the model makes a judgment by assuming that the claim and the evidence refer to the same entity, without considering ambiguity in their presuppositions or referents. For example, when a claim refers to the "FIFA World Cup" and the evidence uses only the term "World Cup" to present contradictory information, annotators assign inferentially-refuted by considering the possibility that "World Cup" refers to a different competition. In contrast, the LLM, guided by the prompt, does not consider this ambiguity in the subject and interprets both as referring to the same entity, resulting in a fully-refuted prediction.

The second type involves cases where the model makes a judgment without considering ambiguity in sentence interpretation. For example, consider a claim stating that "four moons of Mars have been confirmed," along with an evidence stating that "Phobos and Deimos are also the names of two moons of Mars." In the evidence, it is ambiguous whether Mars has exactly two moons or at least two moons. However, the LLM interprets this as Mars having two moons and predicts fully-refuted. For this instance, the two primary annotators also differed in their judgments between fully-refuted and inferentially-refuted.

We also observe cases where NEI is incorrectly predicted as inferentially-refuted, in which the model assumes that entities not mentioned in the evidence do not exist and draws a premature conclusion. For example, given a claim that "Dostoevsky wrote a work titled *Fathers and Sons*" and an evidence that lists several of Dostoevsky' s works, there are cases where the model predicts inferentially-refuted solely because the title *Fathers and Sons* does not appear in the evidence.

C Evaluation with Hallucination Detection in Mind

In this paper, we have mainly discussed performance differences between methods based on F1. However, given that the purpose of the fact-checking system constructed in this study is hallucination detection, recall for refutation labels (fully-refuted and inferentially-refuted) is an important evaluation metric. In addition, incorrectly predicting support for a claim that should be judged as refuted means failing to detect a hallucination, which is a serious problem. For this reason, precision is important for support labels (fully-supported, inferentially-supported, and partially-supported).

Based on this perspective, for CoT, which achieves the best performance in terms of F1, we calculate recall for refutation labels and precision for support labels, and evaluate their arithmetic mean. We conduct the same evaluation for the base prompt and compare the results. As shown in Table 4, CoT outperforms the base prompt on both metrics, which suggests that it is an effective method for the purpose of hallucination detection in this system. However, the margin of improvement is limited, and further improvement in verdict prediction methods remains necessary.

D Original Japanese Examples

(3) **Claim**：「明暦の大火」は江戸を焼き尽くしました。
Evidence: この火事は「明暦の大火」と呼ばれ、江戸城本丸をはじめ多くの武家屋敷・町屋を焼いた江戸史上最大の大火として有名です。

(4) **Claim**：「Rust Belt(ラストベルト)」はかつては自動車製造で発展しました。
Evidence: ベルト (Rust Belt)」と呼ばれる工業地帯のほぼすべての州を制したことだ（注）。注：本稿執筆時点でミシガン州の最終結果は公表されていないが、集計経過を見ると、同州でもトランプ氏の勝利は確実であろう。そして、これこそがトランプ氏勝利の理由である。
トランプ氏が勝利した最大の要因
ペンシルベニア、オハイオ、ミシガン、インディアナの諸州は、かつて、自動車産業や鉄鋼産業の工場が集積

(5) **Claim**：スケトウダラの白子は明太子として加工されます。
Evidence: そして塩漬けしたタラコを赤唐辛子を使った調味料に漬けこんだものが、「からし明太子」と呼ばれるものになります。スケトウダラの卵は、卵の粒が細かくしっとりねっとりしており、煮ても焼いても大変美味しいです。スケトウダラの白子は助タチとも呼ばれます。マダラの白子よりも庶民的な値段で買うことができ、マダラの白子にはかないませんがこちらも十分美味しいです。

	Recall (refutation labels)	Precision (support labels)
base	0.3898	0.3657
+ CoT 6-shot	0.4075	0.4162

Table 4: Recall for refutation labels and precision for support labels