Constructing a Japanese Claim Decomposition Dataset for Fact-Checking of LLM-Generated Texts

Miwa Masano[1,2*], **Ribeka Keyaki**[2], **Atsushi Keyaki**[1*]
Rei Minamoto[3,4], **Kaito Horio**[3], **Hirokazu Kiyomaru**[4]
Kouta Nakayama[4], **Hideyuki Tachibana**[4], **Daisuke Kawahara**[3,4]
[1]Hitotsubashi University
2-1 Naka, Kunitachi, Tokyo, Japan
*Corresponding authors: 5123053k@g.hit-u.ac.jp, a.keyaki@r.hit-u.ac.jp
[2]National Institute of Informatics
1-1-1 Hitotsubashi, Chiyoda-ku, Tokyo, Japan
[3]Tokyo University of Technology
1404-1 Katakuramachi, Hachioji City, Tokyo, Japan
[4]Waseda University
Ookubo, Shinjuku-ku, Tokyo, Japan

Abstract

Since texts generated by large language models (LLMs) may contain hallucinations (misinformation), developing fact-checking systems capable of assessing their veracity has become increasingly important. One of the mainstream approaches to fact-checking is the claim-based one, which first decomposes a generated text into *claims*, i.e., independent and atomic units of information. Each claim is then used as a query to retrieve supporting evidence, and a verdict is predicted for each claim–evidence pair. Conducting fact-checking at the claim level enhances the explainability of verification results. However, achieving highly accurate verification requires that the text be decomposed into claims at an appropriate level of granularity. To address this, we constructed a dataset for Japanese claim decomposition. As part of this dataset construction, we design detailed guidelines for claim decomposition, ensuring that the extracted claims are in a form useful for fact-checking and that the decomposition rules mitigate annotator variability. Quantitative evaluation confirmed that the constructed dataset is of high quality. Additionally, experiments on prompt-based claim decomposition using the constructed dataset demonstrated that adding high-quality few-shot examples and guidelines to prompts improved performance.

Keywords: fact-checking, claim decomposition, guideline, LLM

1. Introduction

While the societal adoption of large language models (LLMs) continues to expand, their susceptibility to generating seemingly plausible misinformation known as *hallucinations* remains a major concern (Huang et al., 2025).

Preventing the occurrence of hallucination is extremely challenging; for example, even if the training data consist solely of factual statements, hallucinations may still emerge (Kalavasis et al., 2025; Kalai et al., 2025).

Against this background, recent research has focused on developing fact-checking systems that verify the factual consistency of LLM-generated text, with the aim of improving their reliability and facilitating broader adoption (Kotonya and Toni, 2020; Zeng et al., 2021; Guo et al., 2022; Kamoi et al., 2023; Wang et al., 2024).

Current mainstream fact-checking systems typically consist of the following three steps: (1) claim decomposition, (2) evidence retrieval, and (3) verdict prediction. An overview of this process is illustrated in Figure 1. In this framework, the LLM-generated input text is first decomposed into claims, which are independent and atomic units of information representing a single property or relation.[1] Subsequently, each claim serves as a query to extract relevant evidence passages from a database, followed by predicting a verdict on each claim–evidence pair using natural language inference (NLI) to assess whether the evidence supports (i.e., entails) the claim's content.

Conducting claim-level fact-checking enhances result explainability. For example, when the sentence "Donald Trump is the Democratic President of the United States." is judged to be false as a whole, it remains unclear whether the error lies in the party affiliation or the position title. In contrast, verifying at the claim level enables finer-grained detection of hallucinations. However, if the claim decomposition is inaccurate, the resulting information units may not correspond appropriately to the actual facts, making it difficult to produce reliable verdict predictions. In fact, prior research reported that the accuracy of claim decomposition critically affects fact-checking performance (Wanner et al., 2024). Thus, quality of claim decomposition is cru-

[1]Atomic claims are also referred to as "sub-claims."

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 2362–2375
11–16 May 2026.

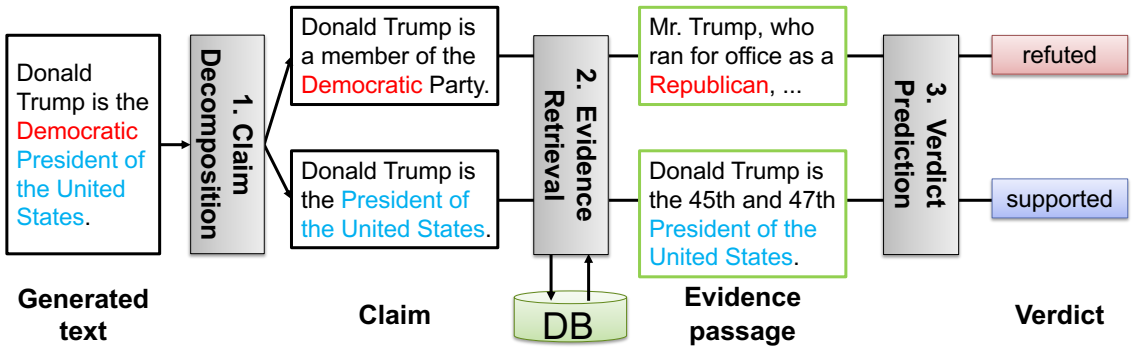

Figure 1: Overview of a fact-checking system.

cial for accurate fact-checking systems.

We aim to improve the quality of claim decomposition in Japanese fact-checking systems by constructing a Japanese claim decomposition dataset, consisting of LLM-generated texts and their corresponding claims. Since English is a high-resource language where LLMs generally achieve their best performance, constructing a dataset in Japanese, a middle-resource language, enables the evaluation of LLMs' cross-lingual robustness by examining the performance gap between English and Japanese datasets. This dataset enables (i) quantitative evaluation of claim decomposition methods, (ii) analysis of inherent challenges, and (iii) subsequent methodological improvements based on these results.

This paper presents the design of claim decomposition guidelines developed as part of constructing a claim decomposition dataset. The primary objective of the guidelines is to ensure that decomposed claims are useful for fact-checking. Specifically, we established rules to enhance performance in evidence retrieval and reduce ambiguity in verdict prediction. We also introduced rules to mitigate variability in decomposed results. Since the guidelines are designed to be highly generalizable, they enable non-experts to construct datasets across a wide range of languages.

Quantitative evaluation confirmed that the constructed dataset is of high quality. Additionally, experiments on prompt-based claim decomposition using the constructed dataset demonstrated that adding high-quality examples and guidelines to prompts improved performance[2].

[2]We will make our datasets and guidelines publicly available upon acceptance.

2. Related Work

2.1. Fact-checking System

Research on claim-based fact-checking has traditionally focused on detecting misinformation such as fake news (Kotonya and Toni, 2020; Zeng et al., 2021; Guo et al., 2022). Subsequently, the release of benchmark datasets such as FEVER (Thorne et al., 2018) and its successors (Thorne et al., 2019; Aly et al., 2021; Schlichtkrull et al., 2023) has driven progress in the development of fact-checking systems. In these datasets, each claim–evidence pair is annotated with one of the following labels: *supported*, *refuted*, and *not enough information*.

WiCE (Kamoi et al., 2023) enables fact-checking at a finer level of granularity by assigning not only claim-level verdict labels but also token-level labels. The claim decomposition process is automated using GPT-3. In WiCE, the verdict prediction labels consist of *supported*, *partially supported*, and *not supported*.

On the other hand, recent advances in LLMs have raised concerns about hallucinations, which are instances of misinformation generated by LLMs. Therefore, research on fact-checking systems aimed at detecting hallucinations has also been actively conducted. Factcheck-Bench (Wang et al., 2024), a fact-checking framework for LLM-generated text, comprises finer-grained steps than the aforementioned systems and provides a benchmark for evaluating the fact-checking of LLM outputs. Moreover, leveraging the verification results, the framework regenerates responses to improve the generated text.

None of the above datasets that involve claim decomposition performed manual decomposition directly on the texts; instead, they annotated claims that had been decomposed by an LLM. Moreover, the annotation trial in Factcheck-Bench found that it is hard to reach a high agreement between annotators for claim decomposition. In contrast,

this study develops guidelines for constructing a dataset that mitigates annotator variability while ensuring that each claim satisfies the properties defined in Section 2.2.

2.2. Claim Decomposition

Wanner et al. (2024) experimentally demonstrated that the quality of claim decomposition has a substantial impact on the performance of factuality-related evaluation tasks conducted at the claim level, such as factual consistency evaluation of generated texts (Min et al., 2023), natural language inference (Chen et al., 2023), and fact-checking (Kamoi et al., 2023; Wang et al., 2024). This is because claim decomposition determines both the number of decomposed claims and the information contained within each claim, meaning that errors in decomposition directly affect the accuracy of factuality assessments.

It has been argued that claim decomposition involves several important perspectives. For example, Wanner et al. (2024) identified three essential aspects for achieving high-quality claim decomposition:

1. **Coverage**: how much of the information in the text is present in the claims.
2. **Coherence**: whether the information in the claims accurately reflects what is stated in the text.
3. **Atomicity**: how separated the information in each claim is.

In FactLens (Mitra et al., 2025), six evaluation criteria are proposed for assessing the quality of decomposed claims:

1. **Atomicity**: each sub-claim should refer to a single factual unit within the original claim.
2. **Sufficiency**: each claim needs to be independently verifiable.
3. **Fabrication**: the decomposition process must not introduce additional information or attempt to correct factual errors.
4. **Coverage**: the list of claims must cover all factual assertions in the original text, leaving no claims missing.
5. **Redundancy**: measuring whether the claims, as a whole, contain redundant facts.
6. **Readability**: assessing how readable the claims are to the end-user.

Among these, the primary evaluation criteria are (1)–(4). While (5) and (6) do not directly affect verifiability or accuracy, they are considered desirable complementary qualities that contribute to the overall usability of the decomposed claims.

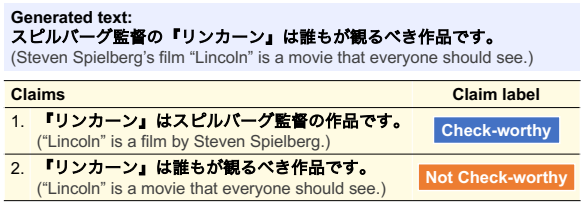

Figure 2: Components of the Dataset

3. Dataset Overview

The claim decomposition dataset constructed in this study comprises three components, as illustrated in Figure 2: (i) text generated by an LLM, (ii) claims extracted from the LLM-generated text through claim decomposition, and (iii) claim check-worthiness labels assigned to each claim.[3]

The construction process involves two tasks: claim decomposition and claim check-worthiness labeling. Both tasks were manually performed by humans to extract claims useful for subsequent stages of evidence retrieval and verdict prediction.

Claim Decomposition. Building on prior research, we define a claim as the finest independent unit of information, each of which expresses a property of, or a relationship concerning, a single entity or event. During the claim decomposition step, the LLM-generated text is segmented into individual claims at the finest possible granularity, while preserving the original meaning entailed by the source text.

Check-worthiness Labeling. Not all claims obtained through the claim decomposition process are equally suitable for fact-checking. For instance, statements expressing personal opinions are difficult to verify, and interrogative sentences lack a well-defined truth value. After decomposition, to ensure that fact-checking is conducted only on claims that can be meaningfully evaluated for veracity, each claim is classified into one of the following two categories (Hassan et al., 2017).

- Check-worthy: Claims experessing objective facts or evaluations, suitable for fact-checking.
- Not check-worthy: Claims expressing subjective opinions or personal experiences, not suitable for fact-checking.

Note that this paper focuses on claim decomposition and therefore does not provide detailed guidelines for claim check-worthiness labeling.

[3]The actual dataset consists entirely of Japanese texts and claims. In the following discussion, however, only English translations are presented for clarity when language-specific syntactic features of Japanese are not the focus. The translations are carefully designed to preserve structural correspondence with the original Japanese sentences to the extent possible.

4. Claim Decomposition Guidelines

Since our claim decomposition process is performed by human workers, guidelines are essential to ensure consistent and high-quality outcomes. To this end, we developed guidelines, which are described in this section. These guidelines are based on three guiding principles.

Principle 1: Each claim should be interpretable on its own, without any context dependency.

Principle 2: The original information should be recoverable from the decomposed claims.

Principle 3: The expressions used in each claim should, as much as possible, remain consistent with those in the source text.

These principles ensure that each claim represents independent and atomic information and that the resulting dataset supports accurate evidence retrieval and verdict prediction.

In what follows, we present the main rules of the guidelines.

4.1. Granularity of Decomposition

Claim decomposition aims to construct claims as independent and atomic units of information (*Sufficiency* and *Atomicity*; § 2.2) by dividing the generated text into the finest possible granularity.[4] Accordingly, sentences, clauses, and parallel structures[5] are all subject to decomposition. Conversely, claim decomposition must remain within the scope of the original text to ensure semantic consistency (*Coherence* and *Fabrication*). Even when a claim can be decomposed structurally, it may not be decomposable semantically. The following rules were designed to ensure that the original information is recoverable from the decomposed claims (*Coverage*).

4.1.1. Parallel Structures

Although parallel structures are generally subject to decomposition, some cases are better left semantically undecomposed. The following example illustrates a parallel structure of four work titles:

(1) Text: *The Rhinegold*, *The Valkyrie*, *Siegfried*, and *Twilight of the Gods* are each independent works, yet together they **form a single grand narrative**.

When creating a claim of the boldfaced portion, decomposing the underlined subject expression would alter the original meaning. This happens because each title, if treated as a separate subject, would fail to convey the collective meaning that the four works together form one grand narrative. This is an instance of a linguistic phenomenon known as *collective reading* (see, e.g., Link 1983).

Therefore, the following claim is created so as to retain the parallel structure.

(2) 1. *The Rhinegold*, *The Valkyrie*, *Siegfried*, and *Twilight of the Gods* form a single grand narrative.

4.1.2. Scope-Limiting Expressions

As a general rule, modifiers and the modified elements should be separated during claim decomposition. However, separating scope-limiting expressions such as those indicating constraints, conditions, or limitations may alter the meaning of a claim or introduce ambiguity.

For example, for the following text (3), the claim decomposition (4) is inappropriate.

(3) Text: In Japan's television industry, "prime time" refers to the time period from 7:00 p.m. to 11:00 p.m., commonly known as golden time.

(4) Incorrect:
1. "Prime time" is commonly known as golden time.
2. "Prime time" refers to the period from 7:00 p.m. to 11:00 p.m.
3. "Prime time" is a term used in Japan's television indsutry.

In (4), the scope-limiting expression "in Japan's television industry" appears in claim (4-3). However, as a result of removing the scope-limiting expression, claims (4-1) and (4-2) have become ambiguous. Since "prime time" may refer to different time slots across countries, claim (4-2) is not necessarily true even if the original text (3) is true, and the entailment between the generated text and the claim no longer holds. Accordingly, the correct decomposition result is as follows.

(5) Correct:
1. In Japan's television industry, "prime time" is commonly known as golden time.
2. In Japan's television industry, "prime time" refers to the period from 7:00 p.m. to 11:00 p.m.

[4] It has been reported that fine-grained decomposition can accumulate noise and degrade the aggregated results (Hu et al., 2025). However, we perform fine-grained claim decomposition because presenting results at the claim level allows users to select the information themselves.

[5] We use *parallel structure* as a cover term for constructions whose elements are syntactically or semantically coordinate. In our annotation scheme, this includes both overt coordination and non-overt predicate linkage, such as Japanese *-te* constructions, which we treat as parallel based on prior analyses of their semantic coordination (Yuasa and Sadock, 2002; Haspelmath, 2004).

4.2. Expression of Claims

In order to further promote unique decompositions from ambiguous expressions e.g., pronoun, we introduce the concept of "focus term." To motivate this, let us look at the boldfaced portion of (6).

(6) Text: In Japan's television industry, "prime time" refers to the time period from 7:00 p.m. to 11:00 p.m., commonly known as golden time. **This time slot has a wide range of viewer age groups**, and therefore tends to achieve high audience ratings.

Here, several candidate noun phrases could replace the context-dependent expression *this time slot*: *prime time*, *golden time*, and *the time period from 7:00 p.m. to 11:00 p.m.* However, the appropriate choice cannot be determined solely from the sentence structure of (6), which may lead to variability in decomposed results.

The concept of "focus term" could mitigate such variability. A focus term refers to a topical expression expected to be highly discriminative in evidence retrieval. In the previous example, the term "prime time" can be regarded as a focus term.

When decomposing a sentence or a part of it into claims, the focus term should first be identified, and the decomposition should be carried out around it. In the resulting claims, the focus term should generally serve as the subject, although naturalness in Japanese should take precedence. The following examples show the claims created from the text (6), decomposed around the focus term "prime time."

(7)
1. In Japan's television industry, "prime time" is commonly known as golden time.
2. In Japan's television industry, "prime time" refers to the period from 7:00 p.m. to 11:00 p.m.
3. In Japan's television industry, "prime time" has a wide range of viewer age groups.
4. In Japan's television industry, "prime time" tends to achieve high audience ratings.

By centering claim decomposition on a focus term, the claim candidates can be narrowed, thereby mitigating annotator variability. Furthermore, among potential claim candidates, it becomes possible to select those expected to have higher discriminative power for retrieval, that is, claims that are more likely to be useful for subsequent evidence retrieval.

4.3. Information to Be Omitted

Principle 2 (§ 4) states that "the original information should be recoverable from the decomposed claims." However, certain types of information, such as those listed below, are omitted during claim decomposition. Instead, priority is given to achieving finer granularity.

4.3.1. Discourse Relations

Following the principle of constructing atomic claims, sentences containing two or more clauses are divided at clause boundaries. Any conjunctions that appear at those boundaries are removed in the process. As a result, conjunctions functioning as discourse markers (e.g., *but*, *because*) are also removed, which may lead to the loss of discourse relations such as contrastive or causal ones. Nevertheless, priority is given to producing more independent and fine-grained claims.

4.3.2. Illustrative Expressions

In general, illustrative expressions such as "for example, A" or "including A" imply the existence of instances other than A. When performing verdict prediction, however, the corresponding evidence sentences may lack such illustrative expressions. Consequently, such evidence does not technically support these illustrative claims, as it does not entail instances beyond A.

Since verdict prediction is conducted independently for each claim–evidence pair, it suffices to fact-check only the portion concerning A in such claims. Therefore, illustrative expressions of this kind are excluded during claim decomposition. Accordingly, the claim extracted from the boldfaced portion of text (8) is (9).

(8) Text: **The term "tender offer," which is conducted, for example, during corporate acquisitions**, is expressed with the three alphabetic letters "TOB."

(9)
1. "Tender offer" is conducted during corporate acquisitions.

4.3.3. Information Related to Enumeration

From an expression such as "The entities that have property A are B and C," we create only two claims: "B has property A" and "C has property A." In contrast, "there are two entities that have property A," is not extracted as a separate claim.

The boldfaced portion of (10), for example, can be interpreted as meaning that there are two terms used to refer to a number composed entirely of the digit 1. In the decomposition shown in (11), that information is represented as the third claim.

(10) Text: **A number in which all digits are 1 is called a "unit digit number" or a "repeating unit number,"** an abbreviation derived from "repeated unit."

(11) Incorrect:
1. A number in which all digits are 1 is called a "unit digit number."

2. A number in which all digits are 1 is called a “repeating unit number.”
3. There are two terms used to refer to a number in which all digits are 1.

Here, (11-3) presents several issues. The text (10) merely implies that there are *at least two* terms, and the corresponding claim (11-3) should be created under that interpretation. Nevertheless, some workers may create a seemingly identical claim while interpreting it as *exactly two*. Preventing such ambiguity would require workers to clearly distinguish between scalar implicature (Grice, 1975) and logical entailment, which is unrealistic in practice. Even resolving this ambiguity would not eliminate the problem: during evidence retrieval or verdict prediction, when only one term is found, it is unclear whether the verdict should be *partially supported* or *not supported*.

Taking these issues into account, we do not create a claim such as (11-3), and regard only (11-1) and (11-2) as the correct decomposed claims.

5. Construction of the Claim Decomposition Dataset

Using the guidelines described in § 4, we constructed a claim decomposition dataset.

5.1. LLM-based Text Generation

We first generated texts using `llm-jp/llm-jp-3-13b-instruct` (LLM-jp, 2024) from the following two datasets.

AIO: AIO Official Dataset Version 2.0 (Suzuki et al.) is a Japanese open-domain question–answering dataset presented in quiz format. Most of the questions are designed to probe factual information.

First, answers to 1,000 questions from the development set of the dataset were generated. A total of 864 questions successfully generated answers. Extremely short answers frequently lacked sufficient context (for example: “It is called e-sports.”); therefore, responses containing fewer than 24 characters were excluded. In addition, in the first version of the dataset, responses exceeding 151 characters were also excluded. As a result, a total of 420 generated answers were collected.

CBA: LLM-jp Chatbot Arena Conversations Dataset (LLM-jp, 2025) is obtained through an online platform designed to fairly compare and evaluate various LLMs. This dataset contains a wide range of user inputs, including those that do not necessarily inquire about factual information, thereby representing a more realistic conversational scenario.

First, responses to 920 inputs from the dataset were generated. A total of 632 inputs successfully generated responses. In CBA, most generated texts were long responses exceeding 151 characters. Therefore, we used a different criterion from AIO and excluded responses longer than 1,024 characters. CBA includes many inputs that fall outside the scope of fact-checking, such as greetings (“Happy New Year!”) or numerical calculations (“If USD/JPY is 123, how much is $100?”). To prioritize the inclusion of check-worthy texts in the dataset, we filtered the data using two LLMs (Qwen3-32B (Alibaba, 2025) and gemma-3-27b-it (Google, 2025)) and retained only texts identified by both LLMs as containing check-worthy information.

As a result, 267 generated texts were identified as check-worthy, from which 100 samples were randomly selected for use.

5.2. Manual Claim Decomposition and Claim Classification Labeling

Claim decomposition and claim classification labeling were conducted by non-expert human workers with prior experience in annotation tasks. The workers were provided with sample materials prepared by the authors and were instructed to carefully review the guidelines before beginning the task. The workflow consisted of two stages: one worker performed claim decomposition and labeling, and a separate checker reviewed the results and discussed any issues or ambiguities. During the process, the authors provided feedback whenever questions emerged and updated the guidelines as necessary. The workers were also asked to include comments or notes for cases in which their decisions had been uncertain.

During the verification process of decomposed results conducted by the authors, it was determined that three generated texts from CBA could not be appropriately decomposed into claims; therefore, these samples were excluded from the dataset. Consequently, the constructed dataset comprises 420 generated texts from AIO and 97 from CBA, each decomposed into claims and assigned check-worthiness labels. The representative statistics of the dataset are presented in Table 1.

We divided the constructed datasets into development and test sets at a 1:1 ratio. Notably, the AIO development set includes the generated texts used for explaining the rules in the guidelines, while the remaining texts were randomly assigned.

5.3. Analysis

We evaluate and analyze the quality of the constructed dataset.

	AIO	CBA
Number of generated texts	420	97
Average number of sentences	2.00	15.32
Average number of claims	5.55	32.11
Standard deviation of claim count	2.89	20.02
Average claim length (words)	15.71	17.42
Average claim length (content words)	6.90	8.23
Ratio of check-worthy claims	0.98	0.87

Table 1: Dataset Statistics

5.3.1. Quantitative Analysis

Inspired by DecompScore (Wanner et al., 2024), a metric for assessing claim decomposition quality, we compute the quality of claim decomposition using the proposed Normalized DecompScore (NDS) defined as follows:

$$NDS = \frac{1}{N}\sum_{t=1}^{N}\frac{C_t^{\text{entailed}}}{C_t} \quad (1)$$

where N denotes the total number of generated texts used for evaluation, C_t denotes the set of claims contained in the generated text t, C_t^{entailed} represents the set of claims in generated text t that are judged as entailed by the generated text. Accordingly, NDS measures the ratio of entailed claims of generated texts, serving as an indicator of the overall quality of claim decomposition. The judgment of whether the generated text entails each claim was conducted using Qwen3-32B (Alibaba, 2025).

AIO achieved an exceptionally high NDS score of 0.980, demonstrating its high quality. CBA also achieved an exceptionally high NDS score of 0.975, demonstrating its high quality.

To examine the reliability of LLM entailment judgments in NDS, we randomly sampled 25 cases labeled TRUE (entailed) and 25 labeled FALSE (not entailed) and compared them with manual judgments.

As a result, all 25 cases labeled TRUE were also judged TRUE by manual evaluation, confirming the high reliability of entailment judgments.

In contrast, among the 25 cases labeled FALSE, 14 (56%) were also judged FALSE in the manual evaluation. These cases mainly resulted from inappropriate decomposition of parallel structures or errors in handling of focus term, resulting in claims not entailed by the generated text. The rules related to these issues should be reconsidered in the annotation guidelines. The remaining 11 cases were judged TRUE in the manual evaluation, indicating errors in the LLM's entailment judgments. The causes of these misclassifications included interpretation errors during translation in the judgment process and cases where the generated text contained contradictory information. A detailed analysis is provided in Appendix A.

These results indicate that some FALSE judgments contain errors. However, since FALSE judgments account for only a small proportion of the entire dataset (1.59%), their impact on NDS is expected to be limited.

5.3.2. Qualitative Analysis

We evaluated the annotators' claim decomposition results using the decomposition results independently created by the authors for 17 generated texts from AIO. The results showed that the decompositions were largely consistent, even for generated texts with complex structures. However, some variations were observed due to differences in the interpretation of parallel structures, the selection of the focus term, and individual background knowledge. The details are provided in Appendix B. Further examination will be conducted regarding the introduction of additional rules expected to mitigate such inconsistencies.

In CBA, during the decontextualization process in claim decomposition, some samples were found where the context could not be reconstructed appropriately without referring to the user's input information.[6] These observations highlight the need to introduce new rules to handle more realistic scenarios and to employ procedures that allow workers to refer to contextual information in the input when necessary during claim decomposition.

6. Prompt-Based Claim Decomposition

In this section, we implement prompting baselines for claim decomposition and evaluate their performance on the constructed dataset. Due to space limitations, we primarily report the experimental results for AIO.

6.1. Claim Decomposition Method

First, a base prompt containing only task instructions and the claim definitions is referred to as `base`. As a proposed method, we construct a prompt that includes the annotation guidelines developed during dataset construction (`guideline`). `guideline` consists of claim decomposition rules and examples of claim decomposition, accompanied by explanatory notes intended to enhance the

[6] For example, when the user input was "<task> Please propose locations to install vending machines that would dramatically increase the sales of *Seventeen Ice*, along with the reasons," some generated texts did not include the term *Seventeen Ice* (a Japanese ice cream brand sold mainly through vending machines).

understanding of the decomposition process. In total, the guideline includes 12 examples of claim decomposition. For the sake of simplifying the explanation of the rules, some examples were created by extracting partial segments from actual generated texts; consequently, the number of decomposed claims in these examples tends to be smaller than that in full generated texts.

We also evaluate the performance of few-shot prompting (`few-shot`). To assess the impact of the number of examples (shots) on performance, the shot counts were set to 2, 4, 6, and 8. The examples are drawn from the development set and were carefully selected to ensure a diverse pattern of decomposition rules. For comparison, we also evaluate `guideline-based shot` in which the 12 claim decomposition examples included in the guidelines are used as few-shot examples.

Additionally, we examined the combinations of `guideline` and `8-shot` (`guideline+8-shot`).

To enable comparison with existing claim decomposition methods, we also evaluated prompts of prior studies, *FactScore* (Min et al., 2023) and *R-ND* (Wanner et al., 2024). These prompts employ an 8-shot setting. The eight examples used in these prompts are identical to those used in our methods.

`gpt-4o` (OpenAI, 2024) was used to perform the decomposition.

6.2. Metric

The evaluation was conducted by comparing the claims decomposed by each prompt (predicted claims) with those contained in the constructed dataset (gold claims). The claim-matching procedure and evaluation metrics followed the proposition-level natural language inference (NLI) framework proposed in (Chen et al., 2023).

Exact match Pairs of predicted and gold claims that are completely identical are regarded as *correct matches*.

Fuzzy match To assess the similarity between the set of predicted claims $\{p_i\}$ and the set of gold claims $\{g_j\}$, we first calculate a similarity matrix of word-level Jaccard similarity $J_{ij} = |p_i \cap g_j| / |p_i \cup g_j|$. The Hungarian algorithm (Kuhn, 1955) is then employed to identify optimal bipartite matches between predicted and gold claims. Pairs of predicted and gold claims whose Jaccard similarity exceeds the threshold θ are considered *correct matches*. We set the Jaccard similarity threshold θ as 0.8 following Chen et al. (2023). Here, we report the version of the fuzzy match calculated for content words, namely nouns, verbs, adverbs, and adjectives.[7]

For each of the matching methods described above, we compute precision, recall, and F1 scores. Precision is defined as the ratio of predicted claims that are judged to be correct matches, while recall is defined as the ratio of gold claims that are judged to be correct matches.

For each method, the reported performance represents the average over five independent runs.

6.3. Experimental Results

As presented in Table 2, `few-shot` generally improved performance as the number of shots increased. `guideline-based shot`, which is based on examples from the guidelines, performs worse than the `few-shot` with fewer but higher-quality examples, further highlighting the importance of selecting high-quality exemplars. This result is consistent with the findings of Wanner et al. (2024), which demonstrated that providing high-quality few-shot examples is crucial for effective claim decomposition. Although `guideline` improved performance, its effect was limited. In contrast, `guideline+8-shot` achieved the highest among all evaluated methods. In addition, `FactScore` and `R-ND` demonstrated performance comparable to that of `8-shot`.

Across all methods, recall was found to be lower than precision. To investigate this, we calculated the decomposition rate, defined as the ratio of the number of predicted claims to the number of gold claims. The results indicated that the overall decomposition rate was approximately 70 to 80%. Although decomposition rate and performance were generally correlated, `guideline+8-shot` outperformed the `6-shot`, which shows the highest decomposition rate. These findings indicate the necessity of achieving a balance between decomposing text into claims at the finest possible granularity and ensuring that each decomposed claim remains semantically appropriate.

For example, in the generated text “This strait connects the Black Sea and the Sea of Marmara,” the worker judged that “the Black Sea and the Sea of Marmara” constituted a non-decomposable parallel structure. Similarly, `guideline+8-shot` correctly avoided decomposing the phrase “the Black Sea and the Sea of Marmara.”

Therefore, since more fine-grained decomposition is still required, improving the prompts to achieve finer decomposition remains an important direction for future work.

[7] The results of fuzzy match based on all words are presented in Appendix C. The analysis revealed many cases in which differences in function words between the claim pairs lowered the matching scores.

Prompt	# of shots	Exact match			Fuzzy match			Decomp. Rate
		Prec.	Rec.	F1	Prec.	Rec.	F1	
`base`	0	0.009	0.007	0.008	0.412	0.358	0.374	0.797
`+2-shot`	2	0.150	0.119	0.130	0.529	0.416	0.457	0.720
`+4-shot`	4	0.197	0.169	0.179	0.560	0.479	0.507	0.796
`+6-shot`	6	0.184	0.163	0.171	0.562	0.500	0.520	**0.834**
`+8-shot`	8	0.202	0.173	0.184	0.561	0.486	0.512	0.804
`+guideline-based shot`	12	0.156	0.132	0.141	0.545	0.442	0.480	0.724
`guideline`	1	0.159	0.129	0.140	0.509	0.399	0.439	0.681
`+8-shot`	8	**0.223**	**0.192**	**0.204**	**0.587**	**0.504**	**0.534**	0.807
`FactScore` (Min et al., 2023)	8	0.193	0.169	0.177	0.540	0.472	0.494	0.806
`R-ND` (Wanner et al., 2024)	8	0.195	0.169	0.179	0.545	0.476	0.499	0.809

Table 2: Evaluation of Prompt-Based Claim Decomposion

Due to space limitations, we omit the detailed experimental results for CBA; however, similar to AIO, the combination of the guideline and few-shot achieved the highest performance. Nevertheless, the overall performance was lower than that of AIO (the F1 score of `guideline+8-shot` of Fuzzy match was 0.431), and increasing the number of shots resulted in only limited performance improvements. These results indicate that CBA is more challenging than AIO.

Thus, the construction of the claim decomposition dataset enables quantitative evaluation of the performance of decomposition methods.

6.4. Analysis of Matching Results

For the analysis of the matching results, the authors manually annotated those obtained from `guideline+8-shot`, which achieved the highest performance. The annotation labels were defined as follows:

Match The matching is correct (53%).
Not match The matching is incorrect (20%).
Less The predicted claim is decomposed more coarsely than the gold claim (11%).
Over The predicted claim is decomposed more finely than the gold claim (3%).
Others Cases involving inconsistent focus term or claims that violates the claim decomposition rules (13%).

The annotation was conducted on 50 generated texts, comprising a total of 295 claims.

The annotation results are shown in Figure 3. The values shown after each label description represent the ratio of that label. The average Jaccard similarity for *Match* labels is close to 1.0, while the average scores for incorrect match types, *Not match*, *Less*, *Over*, and *Others*, are all below 0.8. This suggests that setting the threshold θ to 0.8 appropriately distinguishes correct matches from incorrect ones.

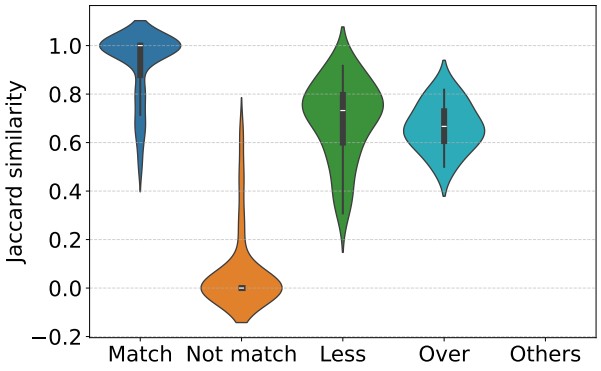

Figure 3: Violin plot of Jaccard similarity distributions.

7. Conclusion

In this study, we developed guidelines for claim decomposition and construct a Japanese claim decomposition dataset with the aim of building Japanese fact-checking systems for LLM-generated texts. These guidelines were designed to ensure that the decomposed claims are useful for downstream processes in fact-checking systems. The quantitative evaluation confirmed that the constructed dataset is of high quality. Furthermore, adding high-quality examples and guidelines to prompts improved the performance of prompt-based claim decomposition.

Ethical Considerations

This study constructs and evaluates datasets based on LLM-generated texts. Because of the inherent nature of AI-generated content and our objective of fact-checking, the data may include factual inaccuracies.

Human tasks such as claim decomposition and check-worthy labeling were outsourced to a reliable vendor that upholds worker protection and ethical standards. Workers provided informed consent, received fair compensation, and no personally identifiable information was collected.

The resulting datasets are intended to support future research on fact-checking and factual consistency in LLM-generated documents. While we expect them to contribute to more transparent evaluation of factual accuracy, users should remain aware of potential residual biases or errors. Any released data will be accompanied by documentation outlining appropriate usage guidelines and precautions to prevent misuse.

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).

Bibliographical 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 Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER), pages 1–13, Dominican Republic. Association for Computational Linguistics.

Sihao Chen, Senaka Buthpitiya, Alex Fabrikant, Dan Roth, and Tal Schuster. 2023. PropSegmEnt: A large-scale corpus for proposition-level segmentation and entailment recognition. In Findings of the Association for Computational Linguistics: ACL 2023, pages 8874–8893, Toronto, Canada. Association for Computational Linguistics.

H. Paul Grice. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, Syntax and Semantics, volume 3, pages 41–58. Academic Press, New York.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. Transactions of the Association for Computational Linguistics, 10:178–206.

Martin Haspelmath. 2004. Coordinating constructions: An overview. In Martin Haspelmath, editor, Coordinating Constructions, volume 58 of Typological Studies in Language, pages 3–39. John Benjamins, Amsterdam.

Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017. Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '17, pages 1803–1812, New York, NY, USA. Association for Computing Machinery.

Qisheng Hu, Quanyu Long, and Wenya Wang. 2025. Decomposition dilemmas: Does claim decomposition boost or burden fact-checking performance? In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 6313–6336, Albuquerque, New Mexico. 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).

Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. 2025. Why language models hallucinate.

Alkis Kalavasis, Anay Mehrotra, and Grigoris Velegkas. 2025. On the limits of language generation: Trade-offs between hallucination and mode-collapse. In Proceedings of the 57th Annual ACM Symposium on Theory of Computing, STOC '25, page 1732â 1743, New York, NY, USA. Association for Computing Machinery.

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. 2023. WiCE: Real-world entailment for claims in Wikipedia. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 7561–7583, Singapore. Association for Computational Linguistics.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking: A survey. In Proceedings of the 28th International Conference on Computational Linguistics, pages 5430–5443, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Harold William Kuhn. 1955. The Hungarian method for the assignment problem. Naval Research Logistics Quarterly, 2(1-2):83–97.

Godehard Link. 1983. The logical analysis of plurals and mass terms: A lattice-theoretical approach. In R. Bäuerle, C. Schwarze, and A. von Stechow, editors, Meaning, Use, and Interpretation of Language, pages 302–323. de Gruyter, Berlin.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 12076–12100, Singapore. Association for Computational Linguistics.

Kushan Mitra, Dan Zhang, Sajjadur Rahman, and Estevam Hruschka. 2025. FactLens: Benchmarking fine-grained fact verification. In Findings of the Association for Computational Linguistics: ACL 2025, pages 18085–18096, Vienna, Austria. Association for Computational Linguistics.

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. AVeriTeC: A dataset for real-world claim verification with evidence from the web. In Advances in Neural Information Processing Systems, volume 36, pages 65128–65167. Curran Associates, Inc.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2019. The FEVER2.0 shared task. In Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER), pages 1–6, Hong Kong, China. Association for Computational Linguistics.

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 Findings of the Association for Computational Linguistics: EMNLP 2024, pages 14199–14230, Miami, Florida, USA. Association for Computational Linguistics.

Miriam Wanner, Seth Ebner, Zhengping Jiang, Mark Dredze, and Benjamin Van Durme. 2024. A closer look at claim decomposition. In Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024), pages 153–175, Mexico City, Mexico. Association for Computational Linguistics.

Etsuyo Yuasa and Jerry M. Sadock. 2002. Pseudo-subordination: A mismatch between syntax and semantics. Journal of Linguistics, 38(1):87–111.

Xia Zeng, Amani S. Abumansour, and Arkaitz Zubiaga. 2021. Automated fact-checking: A survey. Language and Linguistics Compass, 15(10):e12438.

Language Resource References

Alibaba. 2025. Qwen/Qwen3-32B. distributed via Hugging Face Hub. PID https://huggingface.co/Qwen/Qwen3-32B.

Google. 2025. google/gemma-3-27b-it. distributed via Hugging Face Hub. PID https://huggingface.co/google/gemma-3-27b-it.

LLM-jp. 2024. llm-jp/llm-jp-3-13b-instruct. distributed via Hugging Face Hub. PID https://huggingface.co/llm-jp/llm-jp-3-13b-instruct.

LLM-jp. 2025. LLM-jp Chatbot Arena Conversations Dataset. https://huggingface.co/datasets/llm-jp/llm-jp-chatbot-arena-conversations.

OpenAI. 2024. GPT-4o System Card. OpenAI. PID https://cdn.openai.com/gpt-4o-system-card.pdf.

Jun Suzuki and others. AIO Official Dataset Version 2.0. https://sites.google.com/view/project-aio/home.

A. Detailed Analysis of Entailment Judgments

A.1. Cases Where FALSE Judgments Agreed with Manual Evaluation

We analyzed 14 cases that were labeled FALSE by the LLM and also judged FALSE in the manual evaluation. In these cases, the decomposition process incorrectly produced claims that were not entailed by the generated text. We observed two main types of such cases.

A.1.1. Inappropriate Decomposition of Parallel Structures

As described in Section 4.1.1, the annotation guidelines generally require parallel structures to be decomposed, but some exceptional cases were incorrectly decomposed, resulting in FALSE labels.

(12) Text: Eggs Benedict is commonly prepared by placing ham and a poached egg on an English muffin and pouring hollandaise sauce over the top.

(13) 1. Eggs Benedict is commonly prepared by placing ham on an English muffin and pouring hollandaise sauce over the top.
2. Eggs Benedict is commonly prepared by placing a poached egg on an English muffin and pouring hollandaise sauce over the top.

Text (12) describes the preparation process of Eggs Benedict. In the claims in (13), the parallel structure "ham and a poached egg" is decomposed. As a result, (13-1) alone can be interpreted as the restrictive interpretation that Eggs Benedict is commonly prepared with only ham. This interpretation is not entailed by the generated text.

A.1.2. Semantic Changes Accompanying the Subjectification of Focus Terms

We observed cases in which using the focus term as the subject changed the meaning of the original text.

(14) Text: **In fractions,** dividing both the numerator and the denominator by the same number to obtain a simpler fraction, **for example, turning two-fourths into one-half, is referred to as "reduction."**

(15) 1. In fractions, "reduction" refers to the operation of turning two-fourths into one-half.

The generated text (14) explains reduction using a concrete example. The claim (15-1) is obtained by reordering the bolded parts of (14), taking "reduction" as the focus term. However, this transformation produces a sentence that can be interpreted as restricting the meaning of reduction to the specific operation of turning two-fourths into one-half. As a result, it does not accurately preserve the original meaning of the text (14).

A.2. Cases Where the LLM's Entailment Judgments Were Incorrect

Among the 25 cases labeled FALSE, 11 were judged TRUE in the manual evaluation, indicating that the LLM's entailment judgments were incorrect. The main causes are as follows.

A.2.1. Interpretation Errors Caused by the Translation Process

We observed cases where the LLM translated the text into English in the judgment process, which led to an incorrect interpretation that affected the judgment result.

(16) Text: **The protein found in cow's milk and breast milk that is called "milk basic protein" in Japanese** is referred to by the three-letter alphabet name **"β-casein."**

(17) 1. "β-casein" is a protein that is called "milk basic protein" in Japanese.

The claim (17-1) is derived from the bolded parts of text (16). In the LLM's reasoning, "β-casein" was incorrectly interpreted as the Japanese name of the protein. This mistranslation of the modifier relationship led the LLM to judge the claim as FALSE.

A.2.2. Misclassification Caused by Contradictory Information in the Generated Text

We observed cases where the generated text itself contained internally contradictory information. In such cases, a claim may be supported by a specific sentence, but the entailment judgment based on the entire text resulted in a FALSE label.

(18) Text: **The bird** that is sometimes written with the kanji meaning "white-headed bird," and whose name is also used for the mountain path said to have been taken by Minamoto no Yoshitsune to launch a surprise attack in the Battle of Ichinotani, **is "Hiyodorigoe."** Hiyodorigoe is located in Kobe, Hyogo Prefecture. It is said that when Yoshitsune launched a surprise attack on the large Taira force, he rode his horse down this steep cliff. The name of this bird is "hiyodori."

(19) 1. "Hiyodorigoe" is a bird.

The claim (19-1) is derived from the first sentence of text (18), which can be interpreted as stating that "Hiyodorigoe" is a bird. Thus, the claim is supported when this sentence is considered alone. However, the final sentence states that "the name of this bird is 'hiyodori,'" which contradicts the

Exact match			Fuzzy match		
Prec.	Rec.	F1	Prec.	Rec.	F1
0.518	0.524	0.521	0.762	0.767	0.764

Table 3: Evaluation of annotators' claim decomposition using the authors' decomposition as reference

claim because Hiyodorigoe is a mountain path. As a result, the entailment judgment based on the entire text was FALSE.

B. Analysis of Decomposition Results by the Authors and Annotators

We evaluated the annotators' claim decomposition using decomposition independently created by authors for 17 generated texts. We used exact match and fuzzy match as evaluation metrics, which were also used in the experiments in Section 6.

As shown in Table 3, both exact match and fuzzy match indicate high level of agreement. Compared with the claim decomposition results obtained using prompting shown in Table 2, the agreement between the authors and the annotators was markedly higher, particularly in terms of exact match. Inspection of the annotation results revealed that, in many manually decomposed claims, the focus term was placed at the beginning of the claim. This likely increased the number of claims that exactly matched between the authors and the annotators. Since subjects in Japanese are often placed at the beginning of the claim, this high level of agreement may be attributable to the instruction that the focus term should generally serve as the subject.

An example in which the decomposition strategy was correctly applied to a complex generated text is shown below.

(20) Text: In an ice hockey game, each team usually has 16 players. Among them, one is a goalkeeper (goalie), and the remaining 15 play as forwards or defensemen.

(21)
1. In an ice hockey game, each team usually has 16 players.
2. Among the 16 players in ice hockey, one is a goalkeeper (goalie).
3. Among the 16 players in ice hockey, 15 play as forwards or defensemen.

The second sentence of the text (20) contains the expression "among them." The claim (21-2), derived from this sentence, the expression is completed as "among the 16 players in ice hockey." In addition, the claim (21-3), which contains the parallel structure "forwards or defensemen" from (20), does not decompose this parallel structure. Both of these decisions are consistent with the decomposition strategy used by the authors.

In contrast, we also observed cases where the decomposition strategies differd. In the following example, the decomposition granularity differed from the authors' approach.

(22) Text: The peace treaty concluded between the Allied Powers and Japan to end World War II is called the "San Francisco Peace Treaty."

(23)
1. The "San Francisco Peace Treaty" is a treaty to end World War II.
2. The peace treaty concluded between the Allied Powers and Japan is called the "San Francisco Peace Treaty."

In the generated text (22), the part corresponding to the claim (23-2) was decomposed differently by the authors. In the authors' decomposition, the modifier attached to "peace treaty" was separated, resulting in the following two claims:

1. The "San Francisco Peace Treaty" was concluded between the Allied Powers and Japan.
2. The "San Francisco Peace Treaty" is a peace treaty.

The guidelines in this study did not sufficiently define explicit criteria for atomic level of decomposition. As a result, in some cases multiple semantically valid decompositions were possible. Consequently, although the entailment relations themselves were not problematic, discrepancies arose due to differences in decomposition granularity.

C. Evaluation of Fuzzy Match Based on All Words

Section 6.3 reported fuzzy match results using only content words. In this section, we also report the performance when all words are used.

As shown in Table 4, fuzzy match using all words resulted in lower overall performance than using only content words. The analysis revealed that minor differences in function words between the claim pairs affected the matching scores, particularly for short claims.

Prompt	# of shots	Fuzzy match (all words)			Fuzzy match			Decomp. Rate
		Prec.	Rec.	F1	Prec.	Rec.	F1	
`base`	0	0.229	0.193	0.204	0.412	0.358	0.374	0.797
`+2-shot`	2	0.499	0.382	0.425	0.529	0.416	0.457	0.720
`+4-shot`	4	0.546	0.454	0.487	0.560	0.479	0.507	0.796
`+6-shot`	6	0.530	0.459	0.484	0.562	0.500	0.520	**0.834**
`+8-shot`	8	0.539	0.454	0.484	0.561	0.486	0.512	0.804
`+guideline-based shot`	12	0.521	0.411	0.450	0.545	0.442	0.480	0.724
`guideline`	1	0.503	0.384	0.427	0.509	0.399	0.439	0.681
`+8-shot`	8	**0.573**	**0.481**	**0.514**	**0.587**	**0.504**	**0.534**	0.807
`FactScore` (Min et al., 2023)	8	0.525	0.446	0.472	0.540	0.472	0.494	0.806
`R-ND` (Wanner et al., 2024)	8	0.533	0.453	0.480	0.545	0.476	0.499	0.809

Table 4: Performance of fuzzy match using all words