Evaluation of Document-Level Text Simplification in Japanese

Iori Yamashita, Hikari Tanaka, Hajime Kiyama,
Kexin Bian, Zhousi Chen, Mamoru Komachi
Hitotsubashi University, Japan
{iori, hikari, hajime, kexin, zhousi, komachi}@scl.sds.hit-u.ac.jp

Abstract

This study establishes an evaluation framework for document-level text simplification in Japanese by constructing a human-annotated dataset and examining the reliability of LLM-based automatic evaluation. We first developed detailed annotation guidelines covering four criteria—*necessity*, *sufficiency*, *sentence-level simplicity*, and *document-level simplicity*—and collected human ratings for 1,128 source–target document pairs derived from the Wikipedia part of the Japanese simplification corpus JADOS. Using this dataset, we conducted extensive experiments comparing human judgments with evaluations from large language models, including GPT, Claude, and Gemini. The results show that GPT-4o and Gemini 2.5 Pro achieve high agreement with human annotators even in the 0-shot setting, demonstrating their potential as reliable automatic evaluators for Japanese simplification. However, LLMs exhibited a consistent tendency to underestimate document-level simplicity, particularly for kanji-dense texts or texts with relatively long sentences and a small number of sentences. This work provides the first benchmark for evaluating document-level text simplification in Japanese and offers practical evidence that LLM-based evaluation can support scalable assessment for Japanese document-level simplification.

Keywords: Document-level simplification, Evaluation, Japanese, LLM-as-a-judge

1. Introduction

Text simplification aims to rewrite complex texts into forms that are easier to understand, thereby enhancing accessibility for language learners, children, and readers with cognitive or linguistic difficulties (Shardlow, 2014). Most prior work has focused on sentence-level simplification, which rewrites individual sentences to reduce lexical or syntactic complexity. However, real-world reading comprehension occurs at the document level, where readers integrate information across sentences and paragraphs (Alva-Manchego et al., 2019). As a result, simplifying sentences in isolation does not always make the entire document easier to understand. Document-level simplification extends beyond lexical and syntactic changes; it requires higher-level processing such as maintaining discourse coherence, preserving narrative flow, and selecting or omitting information—challenges that are distinct from those at the sentence level (Sun et al., 2021; Vásquez-Rodríguez et al., 2023).

In recent years, research on document-level text simplification has received growing attention. For English, several large-scale corpora have been proposed. The D-Wikipedia dataset (Sun et al., 2021) pairs standard Wikipedia articles with their counterparts in Simple English Wikipedia, enabling large-scale document-level simplification. Similarly, SWiPE (Laban et al., 2023) reconstructs the simplification process from Wikipedia's edit history, focusing not only on lexical and syntactic simplification but also on discourse coherence and information reorganization. These resources have facilitated model training and evaluation that capture global textual coherence, pushing the field beyond sentence-level processing.

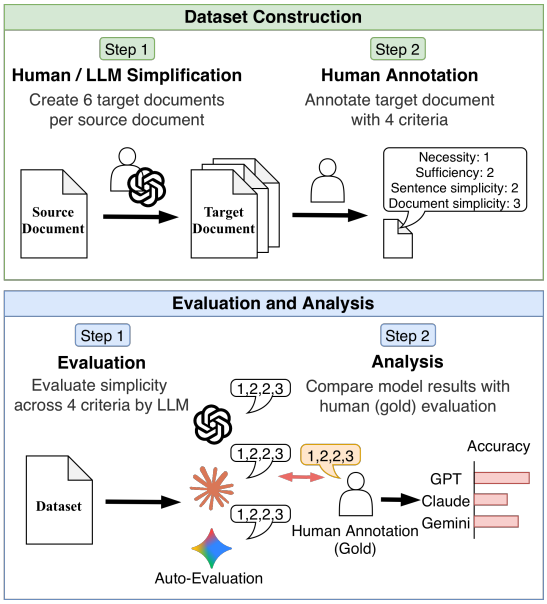

Figure 1: Overview of the study illustrating dataset construction, human evaluation, and LLM-based automatic evaluation for Japanese document-level text simplification.

Document-level simplification research has also expanded to other languages. For instance, the DEplain corpus (Stodden et al., 2023) aligns standard German texts with their "Leichte Sprache" (easy German) counterparts at both sentence and document levels. Such multilingual efforts high-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 1092–1109
11–16 May 2026.

light that simplification must account for language-specific discourse and lexical constraints, positioning document-level simplification as a universal yet language-dependent challenge. However, Japanese research on document-level simplification has been limited due to the lack of corpora that consider discourse-level coherence. This situation changed only recently with the release of the JADOS corpus (Nagai et al., 2024), which provides aligned source and simplified texts for Japanese Wikipedia and news articles.

In contrast, automatic evaluation of document-level text simplification remains largely underexplored. In a pioneering study, Sun et al. (2021) proposed D-SARI, an extension of the sentence-level simplification metric SARI (Xu et al., 2016) to the document level. However, its correlation with human judgments was relatively low (Spearman's $\rho \approx 0.4$), indicating limited reliability for document-level assessment. Furthermore, their experiments focused on Transformer-based systems such as BERT (encoder-based) and BART (encoder–decoder), but did not include human-written simplifications or decoder-only large language models (LLMs), which now play a central role in text generation research.

To address this gap, our study constructs a human-annotated evaluation dataset for Japanese document-level text simplification and investigates the feasibility of automatic evaluation using LLM-as-a-judge (Li et al., 2024; Gu et al., 2025). By comparing human and model-based evaluations, we examine their alignment, divergences, and biases. The main contributions of this work are summarized as follows:

Construction of a human-annotated dataset for Japanese document-level simplification evaluation. We constructed the first human-annotated dataset for evaluating Japanese document-level text simplification. The dataset contains human judgments on outputs from multiple generation models, including LLMs and human-written simplifications. To assess simplification quality, we designed annotation guidelines covering four criteria—necessity, sufficiency, sentence-level simplicity, and document-level simplicity—and conducted annotations by native Japanese speakers. The dataset will be publicly released under the Creative Commons Attribution (CC BY 4.0) license[1].

Comprehensive evaluation combining human and LLM-based judgments. Using the constructed dataset, we conducted a comprehensive analysis comparing human and LLM-based evaluations. GPT-4o and Gemini 2.5 Pro showed strong alignment with human judgments even in the 0-shot setting, demonstrating their reliability as automatic evaluators for document-level simplification. In contrast, Claude 4 Sonnet performed substantially worse on both sentence- and document-level simplicity metrics. We further observed a largely consistent tendency of LLMs to underestimate document-level simplicity, especially in texts with high kanji density or long, complex sentences, suggesting the presence of language-specific evaluation biases.

[1] https://github.com/SDS-NLP/JADOS-eval

2. Related Work

2.1. Document-Level Text Simplification

Traditional research on text simplification has mainly focused on sentence-level transformations, while work addressing document-level simplification remains limited (Shardlow, 2014). Document-level simplification introduces additional challenges such as maintaining inter-paragraph coherence, preserving discourse structure, and deciding which information to retain or omit (Sun et al., 2021; Vásquez-Rodríguez et al., 2023).

In English, large-scale datasets such as SWiPE (Laban et al., 2023) and document-level simplification frameworks such as SIMSUM (Blinova et al., 2023) have advanced this line of research. SWiPE reconstructs simplification operations from Wikipedia edit histories, emphasizing discourse-level reorganization and information flow. SIMSUM integrates summarization and simplification as a two-stage process, modeling simplification as both information compression and linguistic simplification. These studies collectively highlight that document-level simplification requires balancing global coherence with readability, establishing it as an independent subfield within text generation research.

2.2. Japanese Text Simplification

In Japanese, most prior work has focused on sentence-level simplification or expert-edited rewriting. For example, Miyata et al. (2024) released the MATCHA corpus, which contains tourism articles simplified by professional editors and covers a wide range of lexical and syntactic operations. Nagai et al. (2024) introduced JADOS (JApanese DOcument-level text Simplification dataset), the first Japanese document-level simplification corpus aligning Wikipedia and newspaper texts. Similarly, Urakawa et al. (2024) proposed SJNC (Simplified Japanese News Corpus), a news simplification dataset derived from Asahi Shimbun articles with an emphasis on factual faithfulness.

However, these corpora primarily provide reference texts for generation and lack human-labeled evaluation data necessary for systematic assessment of model outputs. Thus, no public resource

has yet been developed for evaluating Japanese document-level simplification quality.

2.3. Automatic Evaluation

Automatic evaluation of text simplification has traditionally relied on metrics such as BLEU (Papineni et al., 2002) and SARI (Xu et al., 2016). However, these metrics focus mainly on surface-level lexical overlap and fail to capture document-level coherence or discourse consistency. For instance, transformations involving paragraph reorganization or selective addition and deletion of information are difficult to assess with BLEU or SARI.

In response, Sun et al. (2021) proposed D-SARI, a document-level extension of SARI that aggregates lexical operations (Add/Delete/Keep) across entire documents. Although D-SARI partially captures global conciseness and information retention, it still cannot evaluate semantic coherence effectively. Nagai et al. (2024) reported that the D-SARI scores for the JADOS dataset show weak correlations for adequacy and simplicity evaluations, suggesting that D-SARI does not fully capture the aspects of meaning preservation and simplicity perceived by human evaluators. They concluded that while D-SARI provides a useful automatic measure for document-level simplification, it should be complemented with human evaluation or alternative metrics to ensure a more comprehensive assessment of simplification quality.

Subsequent research (Cripwell et al., 2024) emphasized the importance of a two-dimensional evaluation perspective—*meaning preservation* and *simplicity*—and incorporated reference-free semantic metrics such as SummaC and QAFactEval for meaning preservation, and $\varepsilon\text{SLE}_{\text{doc}}$ for simplicity. Their findings revealed a trade-off between preserving meaning and increasing simplicity, demonstrating the complexity of document-level simplification evaluation. Building on this perspective, our study adopts four complementary criteria—*necessity*, *sufficiency*, *sentence-level simplicity*, and *document-level simplicity*. While *sufficiency* directly corresponds to meaning preservation by assessing how well the main idea of the source text is retained, *necessity* serves as a content-coverage dimension that ensures essential information is not omitted, which is particularly relevant for encyclopedic texts such as Wikipedia articles. Together, these two criteria provide a fine-grained decomposition of meaning preservation, enabling a more comprehensive evaluation of document-level simplification.

Recently, the LLM-as-a-judge paradigm (Li et al., 2024; Gu et al., 2025) has emerged as a promising alternative. Comparative studies (Bavaresco et al., 2025) show that LLMs can approximate human judgments across diverse NLP tasks. Moreover, reasoning-based techniques such as Chain-of-Thought prompting (Wei et al., 2022) and Self-Consistency decoding (Wang et al., 2023) improve reliability through step-by-step reasoning or output aggregation. However, evaluation consistency varies across prompt languages (Fu and Liu, 2025), with particularly low robustness in low-resource languages such as Japanese.

To our knowledge, no prior study has directly compared human and LLM-based evaluation for Japanese document-level text simplification, which motivates the present work.

3. Dataset Construction

3.1. Dataset Preparation

We constructed a new evaluation dataset for document-level text simplification in Japanese. As source documents, we selected 188 articles from the validation set of the Japanese document-level simplification corpus JADOS (Nagai et al., 2024). The original validation set contains 195 Wikipedia articles categorized as "Good Articles" or "Featured Articles." To ensure compatibility with model context windows, we excluded documents exceeding 2,000 characters. Since JADOS includes manually written simplified versions (around 150 characters for elementary school readers), it provides a reliable foundation for evaluating simplification quality.

For each source document, we collected six types of simplified outputs, including open and commercial LLMs: (1) human-written simplifications from JADOS, (2) a fine-tuned encoder–decoder model (`bart-large-japanese`[2]), (3) an open decoder-only LLM (`Llama-3.1-Swallow-8B-Instruct-v0.2`[3]), (4) another open decoder-only LLM (`gemma-2-9b-it`[4]), (5) a commercial `GPT-4o-2024-11-20` (0-shot), and (6) `GPT-4o-2024-11-20` (1-shot). In total, the dataset comprises $188 \times 6 = 1,128$ source–target document pairs.

Table 1 shows the average target length by model. Human-written JADOS outputs averaged 148 characters, whereas LLM-generated texts tended to be longer, even when instructed to produce approximately 150 characters.

3.2. Annotation Criteria

To comprehensively evaluate the document-level simplification quality of Wikipedia articles, we defined four annotation criteria along two dimensions—*validity* and *simplicity*. Note that "ne-

[2] https://huggingface.co/ku-nlp/bart-large-japanese

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

[4] https://huggingface.co/google/gemma-2-9b-it

Target document	Avg. characters	SD
JADOS-target	147.9	4.0
bart-large-japanese	153.3	21.9
Llama-3.1-Swallow-8B	253.8	121.1
gemma-2-9b-it	240.4	63.7
GPT-4o 0-shot	192.5	33.3
GPT-4o 1-shot	205.6	36.5

Table 1: Average character length and standard deviation (SD) for each model.

cessity" in this context denotes information required to adequately explain the topic within a Wikipedia article, not information that would be necessary in all kinds of documents.

Annotation criteria. Annotators evaluated each simplified document along four criteria, adapted from prior work on text simplification and summarization. Each criterion has a clearly defined scale, summarized as follows:

- **Necessity (0/1):** Whether all essential informational elements of the source document—*Who, What, When, and Where*—are preserved in the simplified version. If an element is absent in the source, its omission in the target is not penalized.
  - **1:** All required elements are included.
  - **0:** One or more essential elements are missing.
- **Sufficiency (1 – 3):** The extent to which the main ideas and overall meaning of the source text are retained in the simplified version.
  - **3:** Fully preserves the main content and intent of the original.
  - **2:** Partially preserves the main ideas; some details are omitted or distorted.
  - **1:** Fails to convey the main content of the original.
- **Sentence Simplicity (1 – 3):** The degree of lexical and syntactic simplicity of individual sentences, relative to a 6th-grade reading level. Evaluated based only on the simplified text (target document), without reference to the source.
  - **3:** Sentences are short, use simple vocabulary, and contain minimal complex syntax.
  - **2:** Mostly simple, but includes some difficult words or compound sentences.
  - **1:** Frequent use of difficult vocabulary or long, complex sentence structures.
- **Document Simplicity (1 – 3):** The overall readability and comprehensibility of the entire document, judged for a typical 6th-grade reader. Evaluated using only the simplified document (target) as reference.
  - **3:** Easily understandable for almost all 6th-grade readers.
  - **2:** Understandable for about half of 6th-grade readers.
  - **1:** Difficult for most 6th-grade readers to understand.

Metric	n/5	Random
Necessity	0.828	0.375
Sufficiency	0.572	0.136
Sentence Simplicity	0.586	0.136
Document Simplicity	0.698	0.136

Table 2: n/5 and random agreement rates (n/5 is the proportion of cases where at least 4 out of 5 annotators assigned the same rating).

3.3. Dataset Statistics

Five native Japanese speakers (university students majoring in interdisciplinary studies) annotated all 1,128 pairs following detailed guidelines.

Inter-annotator agreement was defined as the proportion of documents for which four or more annotators (out of five) assigned the same label. Based on this criterion, agreement indicated high consistency for **Necessity** (0.828) and moderate agreement for the other three criteria—**Sufficiency** (0.572), **Sentence Simplicity** (0.586), and **Document Simplicity** (0.698).

Figure 2 compares annotation distributions across model outputs.[5] Human and `bart-large-japanese` outputs exhibited similar score distributions (within 10%), whereas LLM-based simplifications showed greater variability. Notably, `Llama-3.1-Swallow-8B` and `GPT-4o` achieved higher validity scores, while `Gemma-2-9b-it` and `GPT-4o` tended to receive higher simplicity ratings.

Next, we further examine the reliability of the annotation scheme. Because the distribution of evaluation labels is imbalanced, the rating scale is ordinal, and multiple annotators are involved, conventional agreement measures such as Fleiss'κ do not adequately capture agreement in our setting. Accordingly, we adopt the n/5 agreement rate—defined as the proportion of instances in which at least four out of five annotators assigned the same rating —as our primary agreement metric. This measure is less sensitive to chance agreement and focuses on strong consensus among annotators.

As shown in Table 2, the n/5 agreement rate substantially exceeds the random baseline across all evaluation metrics, confirming that the evaluation criteria are applied consistently and reliably throughout the dataset. These results support the validity of the subsequent comparative analyses

[5]The label distributions for each individual annotator are shown in Figure 6 in the Appendix.

across model outputs.

4. Automatic Evaluation

4.1. Experimental Setup

We evaluated the effectiveness of automatic assessment using both LLM-based and BERT-based models.

LLM-based models. We employed three LLMs: `GPT-4o-2024-08-06`, `Claude-4-Sonnet-20250514-v`, and `Gemini 2.5 Pro`[6]. Inference parameters were unified with `temperature`=0.2 and `max_tokens`=512. These models represent recent multilingual LLMs with high performance on Japanese downstream tasks.

In-context learning. We tested three prompting conditions—*0-shot*, *1-shot*, and *few-shot*—to examine the effect of context examples on evaluation reliability:

- **0-shot:** No example provided.
- **1-shot:** One source–target example shown before evaluation.
- **Few-shot:** One source document together with six simplification examples. [7]

For transparency, detailed prompting examples are provided in Appendix 8. The test documents were excluded from the evaluation set to avoid data leakage. All prompts were written in Japanese, with English prompts additionally used for comparison.

Prompt design. Models were instructed to output four scores—*Necessity*, *Sufficiency*, *Sentence Simplicity*, and *Document Simplicity*—in JSON format according to the human annotation criteria (Section 3). We also tested prompt strategies to improve reliability:

- **Majority voting:** Generate five outputs and adopt the most frequent score (Wang et al., 2023).
- **Reasoning:** Require explicit justifications for each score (Wei et al., 2022).
- **English:** Translate prompts into English to test language effects (Fu and Liu, 2025).

BERT-based models. For comparison, we fine-tuned two Japanese BERT models: `sbintuitions/modernbert-ja-130m`[8] and `cl-tohoku/bert-base-japanese-v3`[9]. Both were trained with 10-fold cross-validation (8:1:1 split) for 20 epochs, batch size 16, and a learning rate of 3×10^{-5}.

Evaluation metric. Accuracy was used as the primary evaluation metric:

$$\text{Accuracy} = \frac{\text{\#correct predictions}}{\text{total evaluations}}$$

because correlation coefficients were less informative given the coarse and imbalanced label distributions.

In addition, we report **Balanced Accuracy** to account for class imbalance. Balanced Accuracy is defined as the average recall across classes:

$$\text{Balanced Accuracy} = \frac{1}{K}\sum_{k=1}^{K} \text{Recall}_k,$$

where K denotes the number of classes and Recall_k is the recall for class k. Unlike standard accuracy, Balanced Accuracy gives equal weight to each class, preventing majority classes from dominating the evaluation.

Automatic metric comparison. In addition to accuracy-based evaluation, we compared two types of automatic metrics: (1) the reference-based D-SARI score, and (2) the criterion-wise auto evaluation, in which GPT-4o (0-shot) directly predicts the four human evaluation scores (*necessity*, *sufficiency*, *sentence simplicity*, and *document simplicity*). For D-SARI, Japanese sentences were analyzed using Sudachi (SplitMode.C, small dictionary) at the morpheme level, and token-level operations (addition, deletion, and keeping) were computed based on 1–4-gram sequences. Multiple reference sentences were concatenated into a single multiset, and, the DEL score was computed using precision only. Length and sentence penalties (LP_1, LP_2, SLP) were applied in the same way as in the original code (Sun et al., 2021). This comparison examines whether LLM-based direct prediction can better capture human-perceived simplicity than traditional lexical-overlap metrics.

4.2. Experimental Results

LLM-as-a-judge evaluation. Table 3 summarizes the results[10]. Overall, GPT-4o and Gemini 2.5 Pro achieved relatively stable performance across

[6] Released on June 17, 2025

[7] Here, the few-shot setting provides one source document together with six annotated simplification examples, rather than multiple independent source-target pairs.

[8] https://huggingface.co/sbintuitions/modernbert-ja-130m

[9] https://huggingface.co/tohoku-nlp/bert-base-japanese-v3

[10] For Gemini 2.5 Pro only, four source documents were excluded from the evaluation owing to "PROHIBITED CONTENT" refusals (1-shot Reasoning × 1, 1-shot English × 1, Few-shot Reasoning × 2).

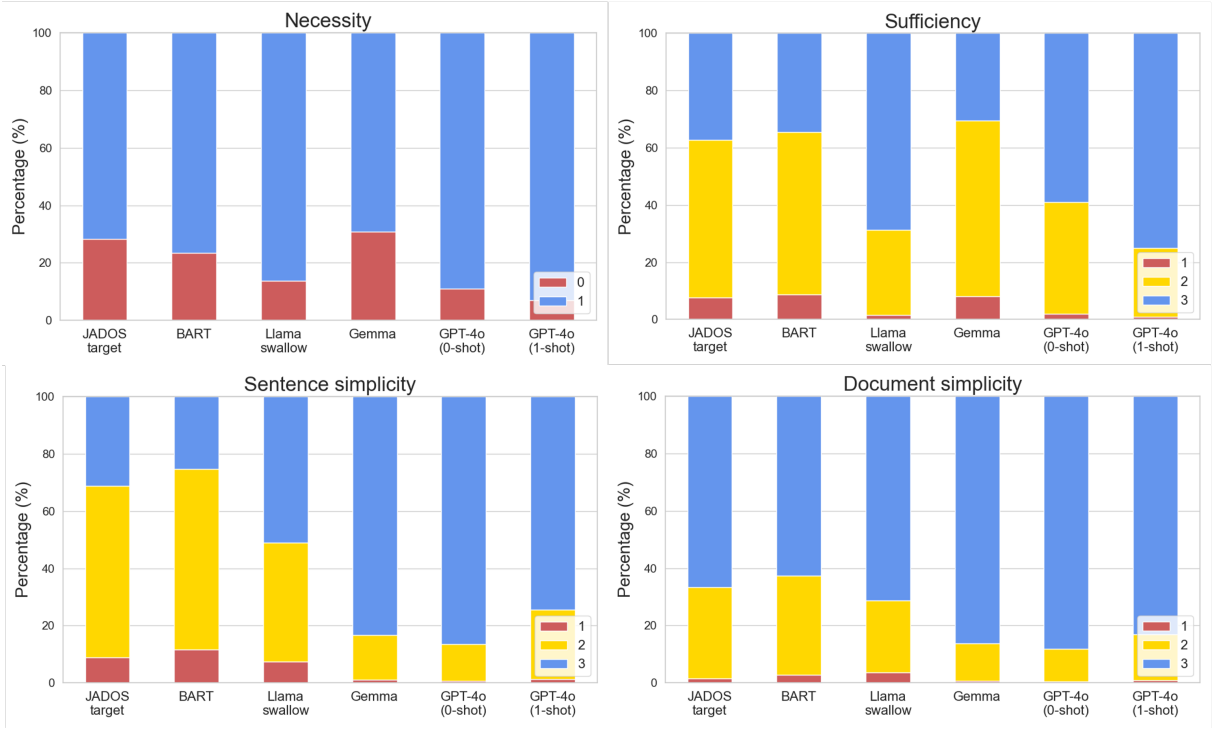

Figure 2: Comparison of label distributions across target document types.

three of the four metrics even in the 0-shot condition, suggesting that modern LLMs have strong potential as automatic evaluators without task-specific training data.

Claude 4 Sonnet showed a different tendency. In particular, its standard accuracy was substantially lower than that of the other LLMs for *Sentence Simplicity* and *Document Simplicity*. However, its balanced accuracy (shown in parentheses) was not as severely degraded. This suggests that Claude predicts minority labels more evenly across classes, while its predictions are less aligned with the majority human labels.

In contrast, the fine-tuned BERT-based models achieved competitive performance. While they were generally slightly worse than the LLM-as-a-judge models on most metrics, they produced comparable results on *Document Simplicity*. Importantly, unlike these supervised models, LLM-as-a-judge does not require task-specific training data, highlighting its advantage in terms of data efficiency.

In the following analyses, we mainly discuss standard accuracy unless otherwise noted, while balanced accuracy is used to confirm that the main trends are not solely driven by class imbalance.

Meta-evaluation. In addition, Table 4 reports the correlations between human judgments and three automatic metrics, D-SARI, BERT-based evaluation and GPT-4o (0-shot). While D-SARI showed weak or sometimes negative correlations with human ratings, GPT-4o exhibited consistently higher

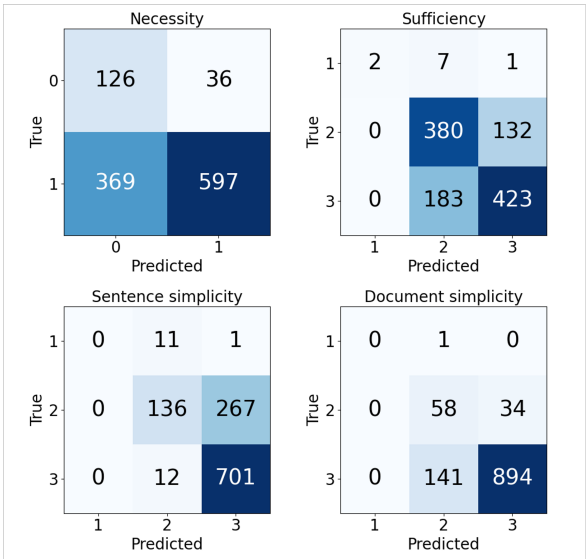

Figure 3: Confusion matrix between human evaluation (gold labels) and GPT-4o (0-shot) .

correlations across most models and criteria. These results suggest that criterion-wise LLM-based automatic evaluation provides a more reliable approximation of human evaluation than reference-based lexical overlap metrics, particularly for document-level simplicity.

4.3. Analysis

To better understand the behavior of LLM-based evaluators, we conducted additional analyses focusing on five aspects: (1) the relationship between automatic metrics and human judgments, (2) the effect of prompting strategies, (3) the influence of

Model	Shot	Setting	Necessity	Sufficiency	Sent. simplicity	Doc. simplicity
GPT-4o	0-shot	Baseline	0.640 (0.698)	0.714 (0.531)	**0.742** (0.435)	0.844 (0.496)
		Majority	0.642 (0.704)	0.723 (0.537)	0.735 (0.430)	0.846 (0.496)
		Reasoning	0.707 (0.705)	0.723 (0.514)	0.730 (0.422)	0.826 (0.564)
		English	**0.781** (0.708)	0.721 (0.537)	0.681 (0.376)	**0.915** (0.475)
	1-shot	Baseline	0.750 (0.700)	0.694 (0.519)	0.750 (0.446)	0.865 (0.565)
		Majority	0.749 (0.695)	0.693 (0.517)	**0.751** (0.446)	0.867 (0.568)
		Reasoning	0.758 (0.672)	0.675 (0.510)	0.709 (0.405)	0.889 (0.502)
		English	**0.780** (0.700)	0.690 (0.522)	0.665 (0.362)	**0.924** (0.402)
	Few-shot	Baseline	**0.856** (0.519)	0.397 (0.441)	**0.771** (0.479)	0.773 (0.563)
		Majority	0.756 (0.670)	**0.725** (0.574)	0.734 (0.435)	**0.879** (0.566)
		Reasoning	0.777 (0.623)	0.711 (0.543)	0.725 (0.468)	0.857 (0.554)
		English	0.802 (0.694)	0.692 (0.558)	0.650 (0.350)	0.922 (0.443)
Claude 4 Sonnet	0-shot	Baseline	0.836 (0.601)	**0.646** (0.603)	0.381 (0.552)	0.174 (0.458)
		Majority	0.840 (0.601)	0.633 (0.615)	0.379 (0.548)	0.176 (0.453)
		Reasoning	0.840 (0.555)	0.627 (0.569)	0.666 (0.612)	0.412 (0.596)
		English	**0.848** (0.687)	0.573 (0.534)	**0.707** (0.691)	**0.481** (0.477)
	1-shot	Baseline	0.811 (0.653)	**0.683** (0.615)	0.730 (0.731)	0.548 (0.537)
		Majority	0.814 (0.660)	0.682 (0.613)	**0.734** (0.734)	0.552 (0.541)
		Reasoning	**0.824** (0.637)	0.663 (0.601)	0.703 (0.709)	0.487 (0.518)
		English	0.725 (0.753)	0.619 (0.544)	0.725 (0.672)	**0.689** (0.563)
	Few-shot	Baseline	0.769 (0.683)	0.607 (0.600)	0.788 (0.729)	0.589 (0.619)
		Majority	0.770 (0.684)	0.610 (0.625)	0.791 (0.731)	0.588 (0.619)
		Reasoning	0.783 (0.623)	0.524 (0.570)	**0.840** (0.632)	0.717 (0.618)
		English	**0.832** (0.723)	**0.635** (0.549)	0.821 (0.679)	**0.751** (0.553)
Gemini 2.5 Pro	0-shot	Baseline	0.697 (0.779)	0.676 (0.508)	**0.822** (0.674)	0.699 (0.530)
		Majority	0.688 (0.782)	0.689 (0.555)	0.809 (0.680)	0.691 (0.551)
		Reasoning	0.757 (0.811)	0.652 (0.537)	0.810 (0.703)	0.711 (0.677)
		English	**0.850** (0.794)	**0.692** (0.531)	0.810 (0.692)	**0.755** (0.622)
	1-shot	Baseline	0.664 (0.783)	0.671 (0.522)	0.814 (0.700)	0.742 (0.603)
		Majority	0.644 (0.772)	0.677 (0.548)	**0.820** (0.698)	0.733 (0.617)
		Reasoning	0.658 (0.769)	**0.682** (0.550)	0.816 (0.700)	0.731 (0.670)
		English	**0.802** (0.811)	0.677 (0.522)	0.812 (0.704)	**0.815** (0.621)
	Few-shot	Baseline	0.747 (0.774)	0.693 (0.513)	0.797 (0.719)	0.708 (0.603)
		Majority	0.743 (0.816)	0.693 (0.545)	**0.801** (0.698)	0.689 (0.566)
		Reasoning	0.751 (0.823)	**0.704** (0.544)	0.794 (0.716)	0.684 (0.552)
		English	**0.824** (0.825)	0.688 (0.564)	0.787 (0.667)	**0.758** (0.584)
ModernBERT-ja			0.690 (0.562)	0.547 (0.371)	0.637 (0.449)	0.835 (0.442)
BERT-Base-ja			0.730 (0.557)	0.614 (0.419)	0.729 (0.493)	0.814 (0.422)

Table 3: Evaluation results for each model under 0-shot, 1-shot, and Few-shot settings. In addition to the baseline configuration, three prompting variants are compared: majority voting over five generations, reasoning-augmented output, and an English-translated prompt. For each metric, standard accuracy is reported, and the corresponding balanced accuracy is provided in parentheses.

human rating variability, (4) systematic directional tendencies in document-level simplicity evaluation, and (5) additional experiment with orthographic features.

Correlation with Automatic Metrics. Table 4 reports the correlations between automatic metrics and human judgments. D-SARI showed near-zero or negative correlations across all evaluation criteria, including a moderate negative correlation for *Sentence Simplicity* ($r = -0.401$). This may occur because D-SARI rewards lexical additions and deletions without considering readability, which can lead to higher scores even when the resulting sentences become more complex.

In contrast, the learned evaluators demonstrated consistently positive correlations with human judgments. The BERT-based evaluator achieved weak to moderate correlations (e.g., $r = 0.445$ for *Sentence Simplicity*), while GPT-4o (0-shot) showed the highest correlations across all criteria (e.g., $r = 0.473$ for *Sentence Simplicity*, $r = 0.463$ for *Sufficiency*, and $r = 0.395$ for *Document Simplicity*).

Target	Metric	D-SARI	BERT	GPT-4o
BART	Necessity	**0.096**	0.087	0.078
	Sufficiency	0.019	0.302	**0.513**
	Sent. simplicity	-0.032	0.054	**0.205**
	Doc. simplicity	0.058	0.212	**0.218**
Llama	Necessity	-0.109	0.077	**0.156**
	Sufficiency	-0.195	0.042	**0.371**
	Sent. simplicity	0.020	0.279	**0.325**
	Doc. simplicity	0.039	**0.336**	0.320
Gemma	Necessity	-0.130	0.112	**0.403**
	Sufficiency	-0.083	0.111	**0.324**
	Sent. simplicity	0.002	-0.049	**0.207**
	Doc. simplicity	0.010	-0.069	**0.511**
GPT-4o (0-shot)	Necessity	-0.067	-0.009	**0.146**
	Sufficiency	**-0.165**	0.117	0.069
	Sent. simplicity	**0.096**	0.038	–
	Doc. simplicity	-0.012	**-0.029**	–
GPT-4o (1-shot)	Necessity	-0.076	0.038	**0.161**
	Sufficiency	**-0.235**	0.114	0.202
	Sent. simplicity	-0.081	**0.094**	–
	Doc. simplicity	–	–	–
Overall	Necessity	-0.048	0.097	**0.280**
	Sufficiency	-0.020	0.256	**0.463**
	Sent. simplicity	-0.401	0.445	**0.473**
	Doc. simplicity	-0.128	0.210	**0.395**

Table 4: Spearman correlations between human ratings and three automatic metrics: reference-based D-SARI, BERT-based evaluation, and GPT-4o (0-shot). Bold indicates the highest absolute correlation value in each row. The **Overall** block reports correlations computed over all target documents combined. Dashes (–) denote undefined correlations caused by zero variance in human ratings.

	Correlation	Logistic Regression	
Metric	Spearman ρ	Coefficient	Std. Error
Necessity	-0.220	-0.53	0.06
Sufficiency	-0.183	-0.39	0.07
Sent. Simplicity	-0.409	-0.81	0.07
Doc. Simplicity	-0.359	-0.64	0.07

Table 5: Relationship between the variance of human evaluation scores and whether the GPT-4o 0-shot (Baseline) judgment agrees with human evaluation (agreement = 1, disagreement = 0). Spearman's rank correlation coefficients and the results of logistic regression are reported.

Notably, for *Sentence Simplicity*, the direction of correlation differed between D-SARI and the learned evaluators: while D-SARI was negatively correlated with human judgments, both BERT and GPT-4o were positively correlated.

These results suggest that reference-based lexical overlap metrics may not reliably reflect human perceptions of simplicity at the document level, whereas model-based evaluators show stronger alignment with human judgments.

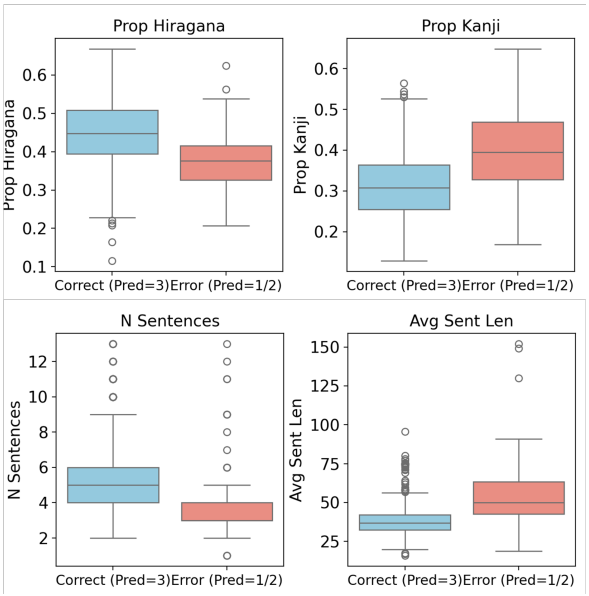

Figure 4: Comparison of text features in correctly and incorrectly evaluated documents labeled as 3 in Document Simplicity by GPT-4o (0-shot) and human annotators.

Prompting strategies. Majority-voting and reasoning strategies did not yield consistent improvements. One possible explanation is that the evaluation task already provides a well-defined scoring rubric, which reduces the benefit of additional reasoning steps or self-consistency sampling. In contrast, accuracy improved notably for *Document Simplicity* when using English prompts (Mondshine et al., 2025; Fu and Liu, 2025). A possible explanation is that English prompts reduce ambiguity in evaluation instructions, while Japanese prompts may allow broader interpretation of simplicity criteria. This finding confirms that prompt language affects evaluation reliability and suggests that translating prompts can mitigate ambiguity in Japanese instructions.

Variability in Human Evaluation. In highly subjective evaluation tasks, it has been reported that LLM classification performance tends to decline as inter-annotator agreement decreases (Lu et al., 2025). To examine whether a similar tendency can be observed in our task, which is also considered highly subjective, we used the automatic evaluation results produced by GPT-4o (Baseline), which achieved the highest 0-shot accuracy for document-level simplicity. For each evaluation criterion, we computed the Spearman rank correlation coefficient between the variance of human evaluation scores for each target document and whether the LLM judgment agreed with the human evaluation

(agreement = 1, disagreement = 0), and conducted logistic regression analysis (Table 5).[11]

The results show that all coefficients were negative and statistically significant ($p < 0.001$) across all four evaluation criteria. These findings indicate that, even in the automatic evaluation of document-level simplicity, LLM evaluation accuracy tends to decrease for documents with greater variability in human judgments.

Underestimation tendency in document-level simplicity. Directional bias was evaluated using the mean difference between predicted and gold scores (prediction minus gold) for each setting. A setting was labeled as showing overestimation or underestimation when the bootstrap confidence interval of the mean difference did not include zero ($\delta = 0$); otherwise, it was labeled as neutral. We additionally report results under a minimal practical difference threshold ($\delta = 0.1$) to assess the practical magnitude of the shift.

Across most evaluation settings, LLMs exhibited an underestimation tendency in document-level simplicity under $\delta = 0$. Although GPT-4o with English prompts showed an overestimation tendency under $\delta = 0$, this shift became neutral under $\delta = 0.1$, suggesting that the magnitude of the deviation is small. Overall, the bias appears primarily directional rather than reflecting large systematic score shifts.

To examine how this tendency manifests in detail, Figure 3 presents the confusion matrix for GPT-4o in the 0-shot (Baseline) setting. The matrix indicates that predictions for higher gold scores are more likely to shift downward than upward, consistent with the observed underestimation tendency.

Errors were more frequent for texts with high kanji density or those consisting of few but lengthy sentences (Figure 4). This pattern suggests that surface-level orthographic density and sentence length may influence document-level judgments.

Other evaluation dimensions—*necessity*, *sufficiency*, and *sentence simplicity*—showed model-dependent directional tendencies, indicating that the underestimation pattern is most pronounced in document-level simplicity.

Effect of Orthographic Features (Kanji Density). To further investigate this possibility, we focused on the documents whose document-level simplicity had been underestimated under the GPT-4o 0-shot (Baseline) setting and conducted a re-evaluation after applying an operation that modified only surface-level simplicity while preserving content-level simplicity.

Specifically, we performed morphological analysis using Sudachi [12], and converted kanji representations of conjunctions, adjectival nouns, adjectives, adverbs, and verbs into their hiragana forms [13]. The modified documents were then re-evaluated under the GPT-4o 0-shot setting.

As a result, underestimation was corrected for 29.4% of the documents. These findings suggest that a characteristic property of Japanese—namely, that kanji representations tend to give a more difficult impression than hiragana or katakana representations—may influence LLM-based evaluation of document-level simplicity.

5. Conclusion

We presented the first evaluation dataset for document-level text simplification in Japanese and conducted a comprehensive comparison between human and LLM-based evaluation. The dataset extends the existing Japanese simplification corpus JADOS with human-annotated evaluation scores covering four criteria: *necessity*, *sufficiency*, *sentence-level simplicity*, and *document-level simplicity*.

Through extensive experiments, we demonstrated that GPT-4o and Gemini 2.5 Pro showed strong alignment with human judgments, whereas Claude 4 Sonnet exhibited lower consistency. Furthermore, a correlation analysis revealed that the reference-based D-SARI metric had almost no correspondence with human ratings, while the label-based auto evaluation showed significantly higher correlations across multiple criteria. This indicates that LLM-based direct label prediction captures human-perceived simplicity more effectively than traditional lexical overlap metrics.

Our findings highlight that large language models can serve as practical and human-aligned automatic evaluators for document-level simplification in Japanese. At the same time, we observed a systematic tendency to underestimate document-level simplicity, particularly for texts with high kanji density or long sentences. These results establish a foundation for scalable and language-aware evaluation frameworks that bridge human and automatic assessment, paving the way toward reliable, model-based evaluation of text simplification.

[11] In the logistic regression analysis, the variance of human evaluation scores was standardized for each evaluation criterion, as the scale of variance differed across criteria.

[12] We used SudachiPy (version 0.6.10) for morphological analysis and SudachiDict-core (version 20250825) as the dictionary. The segmentation mode was set to SplitMode.A.

[13] Concrete examples of the surface-level manipulation are shown in Table 10 in the Appendix.

6. Limitations

Our study has several limitations that should be addressed in future work:

- **Dataset coverage:** The dataset currently includes only 188 documents, primarily sourced from Wikipedia. This limits domain diversity and linguistic variability. Future work should expand coverage to other domains such as news, education, and technical documents.
- **Evaluation dimensions:** The four evaluation metrics—necessity, sufficiency, sentence simplicity, and document simplicity—capture core aspects of simplification. Although document simplicity partially reflects higher-level characteristics, these metrics do not fully address dimensions such as discourse coherence, pragmatic adequacy, or cognitive load. Developing multi-dimensional evaluation frameworks will improve comprehensiveness.
- **Annotator demographics:** Annotations were performed by university students rather than the intended target readers. Inter-annotator agreement was moderate, indicating potential subjectivity. Future work should include feedback from younger or general readers to enhance validity.
- **LLM output stability:** Some LLMs occasionally failed to produce scores due to safety filters or token-length limits, which restricted complete evaluation in certain settings.

Acknowledgments

This work was supported by a commissioned research project from the National Institute of Information and Communications Technology (NICT), titled "Research and Development of Externally Controllable Modeling of Multimodal Information for Improving Machine Translation Accuracy."

Bibliographical References

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2019. Cross-Sentence Transformations in Text Simplification. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 181–184, Florence, Italy. Association for Computational Linguistics.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2025. LLMs instead of Human Judges? a Large Scale Empirical Study across 20 NLP Evaluation Tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–255, Vienna, Austria. Association for Computational Linguistics.

Sofia Blinova, Xinyu Zhou, Martin Jaggi, Carsten Eickhoff, and Seyed Ali Bahrainian. 2023. SIMSUM: Document-level Text Simplification via Simultaneous Summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9927–9944, Toronto, Canada. Association for Computational Linguistics.

Liam Cripwell, Joël Legrand, and Claire Gardent. 2024. Evaluating Document Simplification: On the Importance of Separately Assessing simplicity and meaning preservation. In *Proceedings of the 3rd Workshop on Tools and Resources for People with REAding DIfficulties (READI) @ LREC-COLING 2024*, pages 1–14, Torino, Italia. ELRA and ICCL.

Xiyan Fu and Wei Liu. 2025. How Reliable is Multilingual LLM-as-a-Judge? ArXiv preprint arXiv:2505.12201.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A Survey on LLM-as-a-Judge. ArXiv preprint arXiv:2411.15594.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. LLMs-as-Judges: A Comprehensive Survey on LLM-based Evaluation Methods. ArXiv preprint arXiv:2412.05579.

Junyu Lu, Kai Ma, Kaichun Wang, Kelaiti Xiao, Roy Ka-Wei Lee, Bo Xu, Liang Yang, and Hongfei Lin. 2025. Is LLM an Overconfident Judge? Unveiling the Capabilities of LLMs in Detecting Offensive Language with Annotation Disagreement. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5609–5626, Vienna, Austria. Association for Computational Linguistics.

Itai Mondshine, Tzuf Paz-Argaman, and Reut Tsarfaty. 2025. Beyond English: The Impact of Prompt Translation Strategies across Languages and Tasks in Multilingual LLMs. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1331–1354, Albuquerque, New Mexico. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages

311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Matthew Shardlow. 2014. A Survey of Automated Text Simplification. *International Journal of Advanced Computer Science and Applications (IJACSA) , Special Issue on Natural Language Processing 2014*, 4(1).

Laura Vásquez-Rodríguez, Matthew Shardlow, Piotr Przybyła, and Sophia Ananiadou. 2023. Document-level Text Simplification with Coherence Evaluation. In *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability*, pages 85–101, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *The Eleventh International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing Statistical Machine Translation for Text Simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

7. Language Resource References

Philippe Laban, Jesse Vig, Wojciech Kryscinski, Shafiq Joty, Caiming Xiong, and Chien-Sheng Wu. 2023. SWiPE: A Dataset for Document-level Simplification of Wikipedia Pages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10674–10695, Toronto, Canada. Association for Computational Linguistics.

Rina Miyata, Hinata Koretaka, Hiroki Yamauchi, Daiki Yanamoto, Tomoyuki Kajiwara, Takashi Ninomiya, and Yasuhiro Nishiwaki. 2024. MATCHA: Parallel Corpus for Japanese Text Simplification Based on Professionally Simplified Articles. *Journal of Natural Language Processing*, 31(2):590–609.

Yoshinari Nagai, Teruaki Oka, and Mamoru Komachi. 2024. A Document-Level Text Simplification Dataset for Japanese. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 459–476, Torino, Italia. ELRA and ICCL.

Regina Stodden, Omar Momen, and Laura Kallmeyer. 2023. DEplain: A German Parallel Corpus with Intralingual Translations into Plain Language for Sentence and Document Simplification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 16441 – 16463. Association for Computational Linguistics.

Renliang Sun, Hanqi Jin, and Xiaojun Wan. 2021. Document-Level Text Simplification: Dataset, Criteria and Baseline. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7997–8013, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Toru Urakawa, Yuya Taguchi, Takuro Niitsuma, and Hideaki Tamori. 2024. A Japanese News Simplification Corpus with Faithfulness. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 659–665, Torino, Italia. ELRA and ICCL.

Appendix

A. Simplification Prompt

次の記事を小学生が理解しやすい記事に変換してください。150 字程度の短い記事になるように要約し、難しい表現は簡単な表現に言い換えたり補足の説明をしたりしてください。 `{article}`
English translation (for reference) Please rewrite the following article so that elementary school students can easily understand it. Summarize it into a short article of about 150 Japanese characters, and replace difficult expressions with simpler ones or add supplementary explanations when necessary. `{article}`

Table 6: simplification prompt

B. Distribution of Evaluation Scores

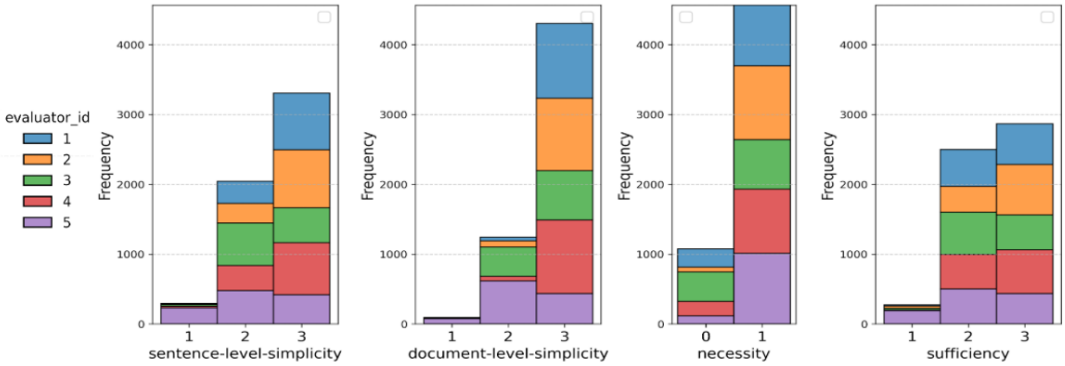

Figure 5: Distribution of scores for each evaluation metric across the full dataset

C. Annotation UI

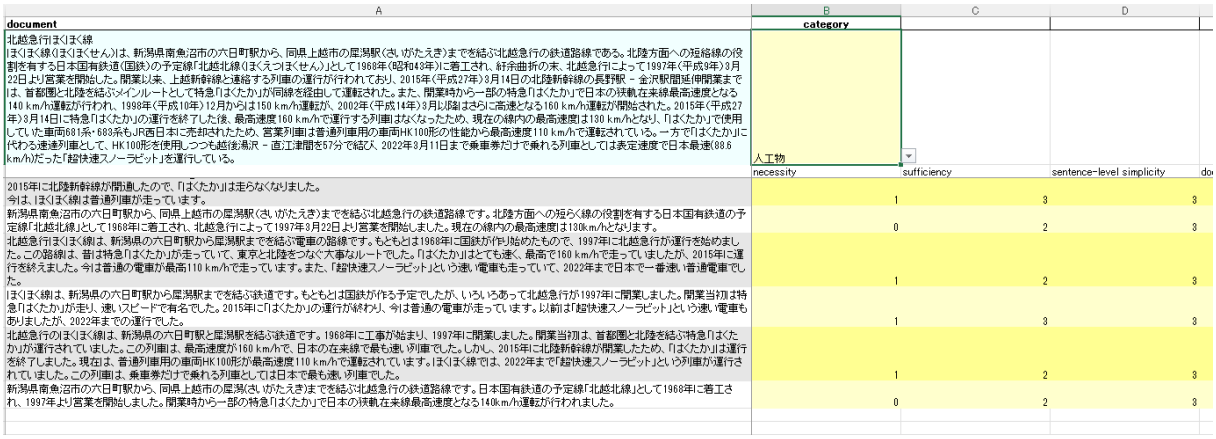

Figure 6: Annotation UI

D. Evaluation Prompt

あなたは日本語で平易化の評価を行うアシスタントです。
元の文:
{original}
簡易化後の文:
{simplified}
以下の項目を評価し，strict な JSON 形式で出力してください。
sentence_simplicity と document_simplicity の評価は簡易化後の文のみを参照して評価してください。
”necessity”: 0 または 1
(文書の説明に必要な要素がすべて含まれるかどうか
「When」「Where」「Who」「What」
ソースに要素が含まれない場合はターゲットに含まれなくてもよい
{1：全て含む, 0：いずれかが欠けている}),
”sufficiency”: 1〜3 の整数
(文書の主旨がターゲットに含まれているかどうか
{3：含まれている, 2：一部含まれている, 1：含まれていない}),
”sentence_simplicity”: 1〜3 の整数
(小学 6 年生基準で，難解な語彙や複雑な文の構造がどれほど含まれるか
{3：ほぼない, 2：一部含まれている, 1：多い}),
”document_simplicity”: 1〜3 の整数
(小学 6 年生基準で，文書全体を通して読んだ時に難解に感じるかどうか
{3：殆どの小学生が理解できる, 2：半数程度の小学生が理解できる, 1：殆どの小学生は理解できない})

English translation
Original text:
`{original}`

Simplified text:
`{simplified}`

Please evaluate the following aspects and output them in strict JSON format:
For sentence_simplicity and document_simplicity, evaluate based only on the simplified text.

”necessity”: 0 or 1 (Whether all necessary elements of the document explanation are included: ”When”, ”Where”, ”Who”, ”What”. If the source does not contain an element, it is acceptable for the target to omit it. {1: All included, 0: Some missing}),

”sufficiency”: An integer from 1 to 3 (Whether the main idea of the document is included in the target. {3: Fully included, 2: Partially included, 1: Not included}),

”sentence_simplicity”: An integer from 1 to 3 (Based on a 6th-grade reading level, how much complex vocabulary or sentence structure is present. {3: Almost none, 2: Some, 1: A lot}),

”document_simplicity”: An integer from 1 to 3 (Based on a 6th-grade reading level, how difficult the document is to understand as a whole. {3: Most 6th graders can understand, 2: About half can understand, 1: Most cannot understand}).

Table 7: evaluation prompt

E. Examples Used in 1-shot and Few-shot Prompting

Table 8: Few-shot examples used in the evaluation prompt. For 1-shot prompting, the original document and Example 1 were provided. For few-shot prompting, the original document and all six examples (Example 1–6) were provided. The same examples were consistently supplied across all evaluated models. The examples were selected randomly, and correspond to original_id=174 in the dataset. English translations are literal translations for reference.

Source (Original Document)

『バッハの旋律を夜に聴いたせいです。』「『バッハの旋律を夜に聴いたせいです。』」（バッハのせんりつをよるにきいたせいです、“Es ist weil ich die Musik von Bach Nachts höre”）は、日本のバンド、サカナクションによる楽曲。バンドのフロントマン山口一郎によって制作されたこの楽曲は、音楽性としてはダンス・ミュージックやオルタナティブ・ロックの要素を持つエレクトロ・ダンスソングであり、4 つ打ちをベースとしたビートで構成されている。楽曲は、ファースト・コーラスはダンス・サウンドで展開されるものの、セカンド・コーラスではバンド・サウンドへと変化する。更にセカンド・コーラス後のブレイクでは、楽曲タイトルにもなっているバッハの音楽がフィーチャーされている。この楽曲は 2011 年 6 月、同名のシングルとしてリリースされた。シングルのアートワークはデザイン集団 Hatos が手がけており、デザインはドイツで行われた。ジャケット写真は、山口の自宅での一場面を写した写真と楽曲タイトルのドイツ語訳を合わせたものとなっている。この楽曲は音楽評論家から賛否両論の評価を受けた。批評家の中にはサウンド、リズム、メロディといった様々な要素をシンプルに凝縮してまとめた点を評価したものもいたが、一方ではこうした作り方によって楽曲の情感や爆発力が不足することになったと指摘する批評家もいた。楽曲は日本のフィジカル、ダウンロードチャート双方にチャート・インした。また、RIAJ 有料音楽配信チャートでは最高位 30 位を記録した。付随する楽曲のミュージック・ビデオでは田中裕介が監督を務め、「ねじれていく世界」が描かれたビデオはエンターテインメント性と芸術性が追求された。ビデオは薄暗い洋室の中での山口本人と 4 体の山口の分身「山口人形」によるダンスや、女優の麻生久美子演じる女性とのラブシーンが繰り広げられる内容となっている。サカナクションはこの楽曲を 2011 年のコンサート・ツアー『SAKANAQUARIUM 2011 “ZEPP ALIVE”』で初披露しており、更に楽曲は『ミュージックステーション』や『ROCK IN JAPAN FESTIVAL』、『SWEET LOVE SHOWER』でも演奏されている。

English translation (literal):

“Bach no Senritsu wo Yoru ni Kiita Sei desu.” “ ‘Bach no Senritsu wo Yoru ni Kiita Sei desu’ ” (Bach no senritsu wo yoru ni kiita sei desu, “Es ist weil ich die Musik von Bach Nachts höre”) is a song by the Japanese band Sakanaction. This song, produced by the band’ s frontman Ichiro Yamaguchi, is an electro-dance song that incorporates elements of dance music and alternative rock, and is composed around a four-on-the-floor beat. The song develops with a dance sound in the first chorus, but shifts to a band sound in the second chorus. Furthermore, in the break after the second chorus, Bach’ s music, which also appears in the song title, is featured. The song was released in June 2011 as a single of the same name. The artwork for the single was created by the design collective Hatos, and the design was carried out in Germany. The jacket photograph combines a scene taken at Yamaguchi’ s home with the German translation of the song title. The song received mixed reviews from music critics. Some critics praised how various elements such as sound, rhythm, and melody were simply condensed and summarized, while others pointed out that this approach resulted in a lack of emotional depth and explosiveness. The song charted on both Japan’ s physical and download charts. It also reached a peak position of 30 on the RIAJ paid music distribution chart. The accompanying music video was directed by Yusuke Tanaka, and the video depicting a “twisting world” pursued both entertainment value and artistic expression. The video features Yamaguchi himself dancing with four alter-egos called “Yamaguchi dolls” in a dimly lit Western-style room, as well as love scenes with actress Kumiko Aso. Sakanaction first performed this song on their 2011 concert tour “SAKANAQUARIUM 2011 ‘ZEPP ALIVE’ ,” and the song was also performed on “Music Station,” “ROCK IN JAPAN FESTIVAL,” and “SWEET LOVE SHOWER.”

Example 1

「バッハの旋律を夜に聴いたせいです。」は、日本のバンド、サカナクションによる楽曲です。メンバーの山口一郎によって制作されたこの楽曲は、音楽性としてはダンス・ミュージックやオルタナティブ・ロックの要素を持つエレクトロ・ダンスソングで、4 つ打ちをベースとしたビートで構成されています。楽曲は日本のフィジカル、ダウンロードチャート双方にチャート・インしました。

English translation (literal):

“Bach no Senritsu wo Yoru ni Kiita Sei desu.” is a song by the Japanese band Sakanaction. This song, produced by member Ichiro Yamaguchi, is an electro-dance song incorporating elements of dance music and alternative rock, and is composed based on a four-on-the-floor beat. The song charted on both Japan’ s physical and download charts.

Annotation

necessity	sufficiency	sentence_simplicity	document_simplicity
0	2	2	3

Example 2

サカナクションの「バッハの旋律を夜に聴いたせいです」は、夜にバッハの音楽を聴くことで生まれた気持ちを描いた曲です。この曲は、ダンスミュージックとロックの要素がミックスされた、リズムが気持ちいい曲です。サビでは、バッハの音楽が使われていて、不思議な雰囲気になります。音楽評論家からは賛否両論の意見がありましたが、多くの人がこの曲を気に入り、チャートにも入りました。ミュージックビデオでは、不思議な世界観が表現されていて、ダンスやラブシーンなど見どころがたくさんあります。

English translation (literal):
Sakanaction’s “Bach no Senritsu wo Yoru ni Kiita Sei desu” is a song that depicts feelings born from listening to Bach’s music at night. This song mixes elements of dance music and rock and has an enjoyable rhythm. In the chorus, Bach’s music is used, creating a mysterious atmosphere. Music critics expressed mixed opinions, but many people liked the song, and it entered the charts. In the music video, a mysterious worldview is expressed, featuring highlights such as dance and love scenes.

Annotation

necessity	sufficiency	sentence_simplicity	document_simplicity
0	2	3	3

Example 3
サカナクションの「バッハの旋律を夜に聴いたせいです。」は、2011 年に発売された曲です。この曲は、ダンスミュージックとロックの要素を持ち、特に夜に聴くと楽しいと言われています。曲の中では、バッハという昔の音楽家のメロディも使われています。ミュージックビデオでは、山口さんが自分の分身と一緒に踊ったり、女優の麻生久美子さんと共演したりしています。この曲は、いろいろな音楽番組やフェスティバルで演奏され、多くの人に楽しまれました。

English translation (literal):
Sakanaction’s “Bach no Senritsu wo Yoru ni Kiita Sei desu.” is a song released in 2011. This song contains elements of dance music and rock and is said to be especially enjoyable when listened to at night. Within the song, melodies by Bach, an old composer, are also used. In the music video, Mr. Yamaguchi dances together with his alter-egos and appears with actress Kumiko Aso. The song was performed on various music programs and at festivals and was enjoyed by many people.

Annotation

necessity	sufficiency	sentence_simplicity	document_simplicity
1	2	3	3

Example 4
『バッハの旋律を夜に聴いたせいです。』は、日本のバンド、サカナクションの曲です。この曲は、ダンス・ミュージックとロックの要素を持つエレクトロ・ダンスソングで、特にバッハの音楽が使われています。2011 年にシングルとして発売され、音楽評論家からは賛否両論の評価を受けました。ミュージック・ビデオでは、山口一郎さんが自分の分身と踊るシーンや、女優の麻生久美子さんとのシーンが見られます。この曲はコンサートやテレビ番組でも演奏されました。

English translation (literal):
“Bach no Senritsu wo Yoru ni Kiita Sei desu.” is a song by the Japanese band Sakanaction. This song is an electro-dance song that has elements of dance music and rock, and it particularly uses Bach’s music. It was released as a single in 2011 and received mixed reviews from music critics. In the music video, scenes can be seen of Ichiro Yamaguchi dancing with his alter-egos and appearing with actress Kumiko Aso. The song was also performed at concerts and on television programs.

Annotation

necessity	sufficiency	sentence_simplicity	document_simplicity
1	2	3	3

Example 5
サカナクションの曲「バッハの旋律を夜に聴いたせいです」は、バンドのフロントマン山口一郎が作った曲です。ダンスミュージックやロックの要素を含むエレクトロダンスソングで、4 つ打ちのビートで構成されています。最初の部分はダンスサウンドで展開されますが、次の部分ではバンドサウンドに変わります。さらに、次の部分ではバッハの音楽が使われています。この曲は 2011 年 6 月にシングルとしてリリースされました。ジャケットのデザインはドイツで行われ、山口の自宅の写真と曲のタイトルのドイツ語訳が使われています。この曲は音楽評論家から賛否両論の評価を受けました。シンプルにまとめられたサウンド、リズム、メロディーを評価する人もいれば、楽曲の情感や爆発力が不足していると指摘する人もいました。日本のチャートで上位にランクインし、RIAJ 有料音楽配信チャートでは最高位 30 位を記録しました。ミュージックビデオは田中裕介が監督し、エンターテインメント性と芸術性を追求した内容となっています。サカナクションはこの曲をコンサートツアーで初披露し、テレビ番組や音楽フェスティバルでも演奏しました。

English translation (literal):

Sakanaction’ s song “Bach no Senritsu wo Yoru ni Kiita Sei desu” was created by the band’ s frontman Ichiro Yamaguchi. It is an electro-dance song containing elements of dance music and rock and is composed with a four-on-the-floor beat. The first section develops with a dance sound, but in the next section it changes to a band sound. Furthermore, in the following section, Bach’ s music is used. The song was released as a single in June 2011. The jacket design was carried out in Germany and uses a photograph taken at Yamaguchi’ s home together with the German translation of the song title. The song received mixed reviews from music critics. Some people praised the simply condensed sound, rhythm, and melody, while others pointed out that the song lacked emotional depth and explosiveness. It ranked high on Japanese charts and reached a peak position of 30 on the RIAJ paid music distribution chart. The music video was directed by Yusuke Tanaka and pursued both entertainment value and artistic expression. Sakanaction first performed the song on their concert tour and also performed it on television programs and at music festivals.

Annotation

necessity	sufficiency	sentence_simplicity	document_simplicity
1	3	2	3

Example 6

日本のバンド、サカナクションによる楽曲です。ファースト・コーラスはダンス・サウンドで展開されるものの、セカンド・コーラスではバンド・サウンドへと変化し、その後のブレイクではバッハの音楽がフィーチャーされています。2011 年 6 月にリリースされました。音楽評論家から賛否両論の評価を受けました。

English translation (literal):

It is a song by the Japanese band Sakanaction. The first chorus develops with a dance sound, but in the second chorus it changes to a band sound, and in the following break Bach’ s music is featured. It was released in June 2011. It received mixed reviews from music critics.

Annotation

necessity	sufficiency	sentence_simplicity	document_simplicity
1	2	2	3

F. Directional tendency

Model	Shot	Setting	Necessity	Sufficiency	Sent. simplicity	Doc. simplicity
GPT-4o	0-shot	Baseline	**U** (U)	**N** (N)	**O** (O)	**U** (N)
		Majority	**U** (U)	**N** (N)	**O** (O)	**U** (N)
		Reasoning	**U** (U)	**N** (N)	**O** (O)	**U** (N)
		English	**U** (N)	**N** (N)	**O** (O)	**O** (N)
	1-shot	Baseline	**U** (U)	**N** (N)	**O** (O)	**U** (N)
		Majority	**U** (U)	**N** (N)	**O** (O)	**U** (N)
		Reasoning	**U** (N)	**U** (N)	**O** (O)	**U** (N)
		English	**U** (N)	**U** (N)	**O** (O)	**O** (N)
	Few-shot	Baseline	**U** (N)	**O** (N)	**O** (O)	**U** (N)
		Majority	**U** (N)	**O** (N)	**O** (O)	**U** (N)
		Reasoning	**U** (N)	**O** (N)	**O** (O)	**U** (N)
		English	**U** (N)	**U** (N)	**O** (O)	**O** (N)
Claude 4 Sonnet	0-shot	Baseline	**O** (N)	**U** (U)	**U** (U)	**U** (U)
		Majority	**O** (N)	**U** (U)	**U** (U)	**U** (U)
		Reasoning	**O** (N)	**U** (U)	**U** (U)	**U** (U)
		English	**N** (N)	**U** (U)	**U** (U)	**U** (U)
	1-shot	Baseline	**U** (N)	**O** (N)	**U** (U)	**U** (U)
		Majority	**U** (N)	**O** (N)	**U** (U)	**U** (U)
		Reasoning	**U** (N)	**U** (N)	**U** (U)	**U** (U)
		English	**U** (U)	**U** (N)	**U** (U)	**U** (U)
	Few-shot	Baseline	**U** (N)	**U** (U)	**U** (N)	**U** (U)
		Majority	**U** (N)	**U** (U)	**U** (N)	**U** (U)
		Reasoning	**U** (N)	**U** (U)	**O** (N)	**U** (U)
		English	**U** (N)	**U** (N)	**U** (N)	**U** (U)
Gemini 2.5 Pro	0-shot	Baseline	**U** (U)	**O** (O)	**O** (N)	**U** (U)
		Majority	**U** (U)	**N** (N)	**O** (N)	**U** (U)
		Reasoning	**U** (U)	**U** (U)	**O** (N)	**U** (U)
		English	**U** (N)	**O** (N)	**N** (N)	**U** (U)
	1-shot	Baseline	**U** (U)	**O** (N)	**N** (N)	**U** (U)
		Majority	**U** (U)	**O** (N)	**O** (N)	**U** (U)
		Reasoning	**U** (U)	**N** (N)	**N** (N)	**U** (U)
		English	**U** (U)	**N** (N)	**N** (N)	**U** (U)
	Few-shot	Baseline	**U** (U)	**U** (U)	**U** (N)	**U** (U)
		Majority	**U** (U)	**U** (U)	**N** (N)	**U** (U)
		Reasoning	**U** (U)	**U** (N)	**U** (N)	**U** (U)
		English	**U** (U)	**N** (N)	**N** (N)	**U** (U)
ModernBERT-ja			**U** (U)	**O** (N)	**U** (U)	**U** (N)
BERT-Base-ja			**U** (N)	**U** (N)	**O** (N)	**U** (N)

Table 9: Directional bias across all models and prompting configurations. Cell color follows the $\delta = 0$ decision. Darker shades indicate agreement between $\delta = 0$ and $\delta = 0.1$, while lighter shades indicate that the bias weakens under $\delta = 0.1$. An underestimation tendency in document-level simplicity is widely observed across configurations.

G. Examples of the surface-level manipulation

	Document Example
Before Modification	ビザンティン建築とは 330 年から 1453 年までのほぼ 1100 年間にも**及ぶ**時代を**指します**．ローマ帝国では国教となったキリスト教の礼拝空間が形成され，初期キリスト教建築と**呼ばれます**．しかし，イスラム帝国や異民族の侵入による国土の縮小，帝国の政治機構の転換などに**伴って**ビザンチン建築も変容し，**特有**の建築形態を獲得しました． Byzantine architecture refers to a period spanning nearly 1,100 years from 330 to 1453. In the Roman Empire, spaces for Christian worship were established after Christianity became the state religion, known as Early Christian architecture. However, along with territorial shrinkage due to invasions by the Islamic Empire and other groups, and changes in the imperial political system, Byzantine architecture transformed and acquired distinctive architectural forms.
After Modification	ビザンティン建築とは 330 年から 1453 年までのほぼ 1100 年間にも**およぶ**時代を**さします**．ローマ帝国では国教となったキリスト教の礼拝空間が形成され，初期キリスト教建築と**よばれます**．しかし，イスラム帝国や異民族の侵入による国土の縮小，帝国の政治機構の転換などに**ともなって**ビザンチン建築も変容し，**とくゆう**の建築形態を獲得しました． Byzantine architecture refers to a period spanning nearly 1,100 years from 330 to 1453. In the Roman Empire, spaces for Christian worship were established after Christianity became the state religion, known as Early Christian architecture. However, together with territorial shrinkage due to invasions by the Islamic Empire and other groups, and changes in the imperial political system, Byzantine architecture transformed and acquired distinctive architectural forms.

Table 10: Example of data used to examine the effect of the surface-level difficulty of kanji on LLM-based document-level simplicity evaluation.