Adapting Sentence-Level Automatic Metrics for Document-Level Simplification Evaluation

Mounica Maddela*
Bloomberg
mmaddela3@bloomberg.net

Fernando Alva-Manchego
School of Computer Science and Informatics
Cardiff University, UK
alvamanchegof@cardiff.ac.uk

Abstract

Text simplification aims to enhance the clarity and comprehensibility of a complex text while preserving its original meaning. Previous research on the automatic evaluation of text simplification has primarily focused on sentence simplification, with commonly used metrics such as SARI and advanced metrics such as LENS being trained and evaluated at the sentence level. However, these metrics often underperform on longer texts. In our study, we propose a novel approach to adapt sentence-level metrics for paragraph- or document-level simplification. We benchmark our approach against a wide variety of existing reference-based and reference-less metrics across multiple domains. Empirical results demonstrate that our approach outperforms traditional sentence-level metrics in terms of correlation with human judgment. Furthermore, we evaluate the sensitivity and robustness of various metrics to different types of errors produced by existing text simplification systems.

1 Introduction

Text simplification involves rewriting a text to improve its ease of understanding, while maintaining the original meaning (Saggion, 2017). This refinement greatly improves the readability of documents, making them more accessible to diverse audiences, including children (Kajiwara et al., 2013), non-native speakers (Petersen and Ostendorf, 2007; Pellow and Eskenazi, 2014), and individuals with learning disabilities (Rello et al., 2013). Text simplification also makes specialized documents, such as medical articles (Elhadad and Sutaria, 2007; Devaraj et al., 2021) and legal texts (Garimella et al., 2022), easier to understand for non-expert readers.

One major obstacle for text simplification is reliable automatic evaluation of simplified texts.

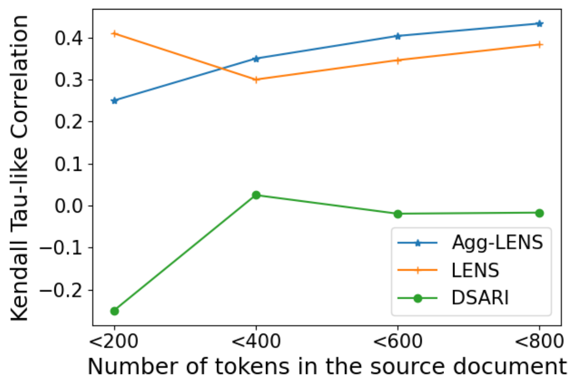

Figure 1: Kendall Tau-like correlation of the simplification metrics DSARI and LENS, along with our proposed Agg-LENS, as a function of the number of tokens in the source text. The source texts and references are from the *Cochrane* dataset (Devaraj et al., 2021), with human judgments for various simplifications collected by Flores et al. (2023). Agg-LENS outperforms both LENS and DSARI on texts longer than 200 tokens.

While the simplification of long texts, such as documents and paragraphs, holds practical utility, existing research has primarily focused on the automatic evaluation of sentence simplification (Xu et al., 2016; Alva-Manchego et al., 2021; Cripwell et al., 2023; Heineman et al., 2023). Commonly used metrics for text simplification, such as SARI (Xu et al., 2016), which measures n-gram overlap between the simplified text and human references, and BERTScore (Zhang et al., 2020), which assesses semantic similarity using BERT embeddings (Devlin et al., 2019), are primarily designed for sentence-level evaluation. Although Sun et al. (2021) propose a variant of SARI for longer texts called DSARI, lexical overlap metrics like SARI and DSARI struggle to effectively capture paraphrases (Alva-Manchego et al., 2021). In contrast, semantic similarity metrics such as BERTScore focus on meaning preservation, often lacking correlation with simplicity (Maddela et al., 2023).

*Work done outside Bloomberg.

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 6444–6459
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

On the other hand, the rise of pretrained language models has led to the development of supervised metrics such as LENS (Maddela et al., 2023) and REFEREE (Huang and Kochmar, 2024), which are fine-tuned on human judgments and effectively capture diverse styles of simplification. However, these metrics are primarily trained and evaluated for sentence simplification, resulting in suboptimal performance on longer texts. For instance, Figure 1 demonstrates that LENS, the state-of-the-art metric for sentence simplification, exhibits the highest correlation with human judgments on shorter texts (under 200 tokens) but shows diminished performance on longer texts. Here, source texts and references are from the Cochrane dataset (Devaraj et al., 2021), and human judgments for the simplified texts are collected by Flores et al. (2023). Section 3 and Table 1 provide further details and results related to the experiment. Additionally, these metrics are constrained by the length limitations of the underlying pretrained language models.

To address these limitations, we propose a simple yet effective method to adapt sentence-level metrics for paragraph- or document-level simplification. Our approach first decomposes long texts into shorter segments using a specialized semantic similarity model and a graph-based alignment strategy. It then employs a sentence-level metric to compute evaluation scores for these shorter texts and aggregates the results. This method can be applied to any reference-based or reference-less metric, with or without the source document. We compare our proposed approach to a wide variety of existing simplification metrics in terms of correlation with human judgments, robustness, and sensitivity to minor errors. Empirical results demonstrate that our approach enhances the correlation of sentence-level metrics across three domains: Wikipedia, news, and medical texts. Our approach also boosts the robustness of existing metrics on longer texts as illustrated in Fig 1 (Agg-LENS vs LENS).

Our main contributions include: (a) a novel approach for adapting sentence-level metrics to long text simplification; (b) benchmarking this approach alongside a comprehensive set of reference-based and reference-less simplification metrics across multiple domains; and (c) evaluating the sensitivity and robustness of automatic metrics to various types of errors produced by existing text simplification systems.

2 Aggregating Sentence-level Metrics for Long Texts

In this section, we present a novel method for adjusting sentence-level metrics to evaluate paragraph- or document-level simplification. We employ a specialized sentence alignment model (Jiang et al., 2020) alongside a graph alignment strategy to identify smaller units of related text across the input, simplified, and reference texts. These related texts can encompass multiple sentences, allowing our graph alignment strategy to effectively handle multi-sentence simplification edits, such as sentence reordering, fusing sentences, and selecting relevant content across sentences (Laban et al., 2023). We then calculate metric values for these smaller units and average them to derive the final metric value.[1]

Step 1: Construct similarity matrices. Given a complex text $C = (c_1, \ldots, c_i, \ldots, c_m)$, its simplification $S = (s_1, \ldots, s_j, \ldots, s_n)$, a reference $R = (r_1, \ldots, r_k, \ldots, r_p)$, and a sentence-level metric $M(.)$, the goal is to compute a score z that captures the overall quality of S. Here, c_i, s_j, and r_k correspond to sentences in C, S, and R respectively. First, we compute two sentence similarity matrices: $A_{cs} \in R^{m \times n}$ with sentence pairs (c_i, s_j) and $A_{cr} \in R^{m \times p}$ with sentence pairs (c_i, r_k). We utilize Jiang et al. (2020)'s sentence pair similarity model to construct A_{cs} and A_{cr}. This model was trained to measure similarity of sentences in parallel complex articles and their simplified versions.[2] However, our approach is agnostic to the type of sentence aligner and can be replaced with any aligner that better suits the target dataset or task, offering flexibility for different contexts.

Step 2: Extract smaller units of related text. We use the similarity matrices in a graph-based alignment approach to construct smaller segments of related texts across C, S, and R. We construct an undirected graph G with the sentences as vertices and sentence pairs with similarity > 0.5 as edges:

$$\begin{aligned} V =& C \cup S \cup R \\ E =& \{(c_i, s_j) \mid A_{cs}(i,j) > 0.5\} \\ & \cup \{(c_i, r_k) \mid A_{cr}(i,k) > 0.5\} \end{aligned}$$

[1] Our code for the approach and experiments is available at https://github.com/cardiffnlp/document-simplification

[2] We use the BERT model trained to align sentences between English Wikipedia and Simple Wikipedia articles.

We extract connected components $cc(G) = \{g_1, \ldots, g_l, \ldots, g_o\}$ from G using breadth-first search. Note that each component g_l contains a subset of sentences in C, S, and R. We partition g_l into three sets $({g_l}^c, {g_l}^s, {g_l}^r)$, each containing sentences from C, S, and R respectively, followed by concatenation within each set. In section 5, we delve deeper into the impact of different similarity threshold choices and alignment strategies.

Step 3: Compute and aggregate metric scores. Finally, we compute the metric value $M({g_l}^c, {g_l}^s, {g_l}^r)$ for each component g_l and average them across all the components. In scenarios involving multiple references, we choose the reference with maximum z. For reference-less metrics, we omit R in the approach while keeping the rest of the steps the same. We provide further implementation details in Appendix C.

3 Evaluating Correlation with Human Judgements

In this section, we benchmark existing simplification metrics and their aggregated versions constructed using our proposed approach.

3.1 Datasets

We evaluate different metrics on the following three publicly available human ratings datasets:

COCHRANE-HUMAN (Flores et al., 2023) includes 120 binary comparisons of simplified English texts for overall readability by three human judges.[3] Given an original document and its two corresponding simplified versions, generated by two different systems, the dataset contains human ratings indicating which system is better at simplifying the original text. We use the majority rating from these judges as the final score for each comparison. We use the majority rating from these judges as the final score for each comparison. The original English texts belong to the Cochrane simplification dataset (Devaraj et al., 2021) that consists of abstracts from the Cochrane Database of Systematic Reviews and their corresponding plain language versions written by domain experts, following Cochrane's PLS standards. This human ratings dataset contains outputs from GPT4 (OpenAI et al., 2024) and four BART-based systems namely vanilla BART (Lewis et al., 2020), BART trained using unlikelihood loss (Li et al., 2020), BART trained to simplify using a two step summarize-then-simplify strategy (Lu et al., 2023), and BART with a readability enhanced decoding approach (Flores et al., 2023).

D-WIKIPEDIA (Sun et al., 2021) consists of 5-point Likert scale ratings on fluency, meaning preservation, and overall simplicity for 500 simplifications across five systems[4] including fine-tuned BART, a BERT-based extractive summarization system (Liu and Lapata, 2019), a human-written simplification, and a vanilla Transformer model and its variant that enhances contextual information. Three human judges rate each simplification and we take the average as the final rating. The original texts are derived from the D-Wikipedia test set (Sun et al., 2021), which consists of paragraphs from Wikipedia articles and their corresponding aligned paragraphs from Simple Wikipedia.

ONESTOPQA (Agrawal and Carpuat, 2024) evaluates the meaning preservation ability of 9 simplification systems using a reading comprehension task.[5] Given a simplified text from a news article and three questions that are answerable by the original text, the dataset contains human annotations capturing if the questions can be answered by the simplified version. This study calculates two scores for each simplified text: accuracy, the percentage of correctly answered questions, and answerability, the percentage of questions deemed unanswerable. The complex texts and their human references for computing the metrics are extracted from the OneStopEnglish dataset (Vajjala and Lučić, 2018).[6] The simplification systems in ONESTOPQA include ChatGPT, an unsupervised system trained using reinforcement learning (Laban et al., 2021), a fine-tuned BART model with control tokens to adapt to different readability levels (Martin et al., 2022), a fine-tuned T5 model with similar control tokens (Sheang and Saggion, 2021), and two supervised edit-based non-autoregressive models (Agrawal and Carpuat, 2022).

We provide statistics for each dataset in Appendix A.

[3]https://github.com/ljyflores/simplification-project

[4]D-WIKIPEDIA contains human ratings along four dimensions: fluency, meaning preservation, overall simplicity, and word-level simplicity. We report the results on the first three.

[5]https://github.com/sweta20/ATS-EVAL

[6]OneStopEnglish contains documents written at three readability levels: advanced, intermediate, and elementary. We use the advanced version as the complex text and the elementary version as the human reference

3.2 Automatic Evaluation Metrics

We benchmark the following metrics:

BLEU (Papineni et al., 2002) is a precision-based metric calculating n-gram overlap between a candidate and its reference along with a brevity penalty.

SARI (Xu et al., 2016), the widely utilized metric for text simplification, calculates F1/precision scores for the n-grams added, removed, and retained in comparison to human references.

BERTScore (Zhang et al., 2020) is a semantic similarity metric that measures word-level similarity used BERT (Devlin et al., 2019) embeddings.

LENS (Maddela et al., 2023) is a learned simplification metric based on RoBERTa (Liu et al., 2019) that computes a quality score given a complex sentence, its simplified version, and a set of references.

LENS-SALSA (Heineman et al., 2023) is a learned reference-less metric that trains the LENS metric on phrase-level simplification edits.

REFEREE (Huang and Kochmar, 2024) is another supervised reference-less metric that measures the overall quality of the simplified sentence. It first pretrains a DeBERTa (He et al., 2021) with existing metrics and finetunes it on human ratings.

SLE (Cripwell et al., 2023) is a learned reference-less metric that focuses on measuring the raw simplicity of the simplified sentence, or the relative simplicity gain when compared to the input complex sentence. It is based on a finetuned RoBERTa.

DSARI (Sun et al., 2021) is a variant of SARI that computes the same F1/precision scores as SARI but also includes length penalties.

QuestEval (Rebuffel et al., 2021) measures the meaning preservation of a simplification by comparing answers to a list of questions on the simplification and its corresponding source document.

Llama3-based metric We use Llama3-8B-Instruct (Dubey et al., 2024) to evaluate the generated simplified text along three dimensions: meaning preservation, fluency, and simplicity. Following Liu et al. (2023), we first provide the task description to the model and generate intermediate rating instructions. We then augment the original instructions with these intermediates, along with the source and simplified text, and ask the model to predict a score between 1 and 5 for the specified dimension. The final score is derived from a probability-weighted summation of the output scores. We provide more details prompts in Appendix B.

Note that DSARI, QuestEval, and Llama3-based metrics are not sentence-level metrics. Therefore, we skip the application of our aggregation strategies on these three metrics.

3.3 Evaluation Setup

Following previous work in machine translation (Bojar et al., 2017; Ma et al., 2018) and evaluation of sentence-level simplification metrics (Maddela et al., 2023; Huang and Kochmar, 2024), we report **Kendall Tau-like correlation** on the COCHRANE-HUMAN dataset to capture the relative ranking of two systems. Given an input c and its simplifications from 2 systems s_1 and s_2, we calculate Kendall Tau-like coefficient τ as:

$$\tau = \frac{|Concordant| - |Discordant|}{|Concordant| + |Discordant|} \quad (1)$$

where *Concordant* is the set of pairs where the metric ranked (s_1, s_2) in the same order as humans and *Discordant* is the set of the pairs where the order is different. For DWIKI and ONESTOPQA, we report **Pearson correlation** (ρ) between the metric scores and the human ratings.

3.4 Results

Table 1 shows the correlation with human ratings on COCHRANE-HUMAN, D-WIKIPEDIA, and ONESTOPQA datasets. We summarize the trends below:

Aggregation improves the performance of reference-based metrics across multiple dimensions. *Agg-LENS* and *Agg-SARI* outperform their corresponding non-aggregated versions on the readability dimension of COCHRANE-HUMAN, simplicity dimension of D-WIKIPEDIA, and meaning-based dimensions of ONESTOPQA and D-WIKIPEDIA. For *BERTScore*, the aggregated version (*Agg-BERTScore*) shows an improvement on D-WIKIPEDIA and ONESTOPQA.

Aggregation helps only with the meaning preservation dimension for reference-less metrics. *Agg-REFEREE*, *Agg-LENS-SALSA*, and *Agg-SLE* outperform their non-aggregated counterparts in

		COCHRANE	D-WIKIPEDIA			ONESTOPQA	
		Readability	Fluency	Meaning	Simplicity	Accuracy	Answerability
Sentence-level, Reference-based	BLEU	-0.183	0.251	0.119	0.432	0.261	0.240
	SARI	-0.083	0.257	-0.023	0.386	0.150	0.136
	BERTScore	0.183	0.017	0.037	0.012	<u>0.365</u>	0.344
	LENS	<u>0.383</u>	**0.623**	-0.114	0.461	0.196	0.179
Sentence-level, Reference-less	SLE	0.150	-0.067	0.259	-0.073	-0.119	-0.111
	LENS-SALSA	0.200	0.611	0.041	0.489	0.123	0.118
	REFEREE	0.316	0.417	0.200	0.440	0.083	0.046
Document-level	DSARI	-0.017	0.331	-0.138	0.414	0.068	0.041
	QuestEval	0.016	0.227	**0.557**	0.389	0.300	0.317
	LLAMA3	0.117	<u>0.619</u>	0.416	0.452	0.309	0.280
Metrics aggregated using our approach							
Document-level, Reference-based	Agg-SARI	0.0 ↑	0.338 ↑	0.067 ↑	<u>0.498</u> ↑	0.172 ↑	0.149 ↑
	Agg-BERTScore	0.167 ↓	0.235 ↑	0.418 ↑	<u>0.498</u> ↑	**0.366** ↑	**0.375** ↑
	Agg-LENS	**0.433** ↑	0.573 ↓	-0.003 ↑	**0.506** ↑	0.360 ↑	<u>0.353</u> ↑
Document-level, Reference-less	Agg-SLE	-0.05 ↓	-0.142 ↓	<u>0.498</u> ↑	-0.102 ↓	0.041 ↑	0.078 ↑
	Agg-LENS-SALSA	0.217 ↑	0.520 ↓	0.142 ↑	0.370 ↓	0.183 ↑	0.176 ↑
	Agg-REFEREE	0.017 ↓	0.258 ↓	0.455 ↑	0.329 ↓	0.326 ↑	0.328 ↑

Table 1: Correlation results of automatic metrics on three human ratings datasets: COCHRANE-HUMAN, D-WIKIPEDIA, and ONESTOPQA. We report the Kendall Tau-like correlation on COCHRANE-HUMAN and Pearson correlation for the rest. The best values are marked in **bold** and the second best values are <u>underlined</u>. ↑ and ↓ represent if the aggregated version of the metric improves or degrades the performance when compared to the original sentence-level version.

meaning preservation on D-WIKIPEDIA and in accuracy and answerability on ONESTOPQA. However, these variants exhibit a decrease in correlation regarding the simplicity dimension on both COCHRANE-HUMAN and D-WIKIPEDIA datasets.

Reference-based aggregated metrics outperform document-level metrics. *Agg-LENS* and *Agg-BERTScore* outperform document-level metrics such as *DSARI*, *QuestEval*, and *Llama3* on COCHRANE-HUMAN, ONESTOPQA, and simplicity dimension of D-WIKIPEDIA. *Agg-SARI* outperforms *DSARI* on all the three datasets.

Learned metrics perform reasonably well even without any aggregation. Although trained at a sentence-level, *LENS*, *LENS-SALSA*, and *REFEREE* outperform DSARI. *REFEREE*, a reference-less metric, shows correlation results close to *LENS* with respect to readability on COCHRANE-HUMAN and simplicity on D-WIKIPEDIA.

Learned metrics perform better than lexical and semantic metrics on challenging domains. COCHRANE-HUMAN is a challenging dataset for metrics as it contains medical abstracts with extremely lengthy sentences and complex terminology. Supervised metrics based on ROBERTA namely *Agg-LENS* and *LENS* show the best and the second best correlations on COCHRANE-HUMAN.

Metrics need to be used with caution to evaluate deletion-based simplifications. We observe conflicting results between the meaning preservation dimensions of D-WIKIPEDIA and ONESTOPQA. In the former dataset, reference-less metrics like *Agg-REFEREE* and *Agg-SLE* outperform reference-based metrics such as *Agg-LENS* and *Agg-BERTScore*. However, this trend reverses on ONESTOPQA. The discrepancy arises because deletion is heavily penalized in D-WIKIPEDIA, where reference-less metrics, which focus solely on differences with respect to the complex text, are better suited to capture missing information. In contrast, ONESTOPQA incorporates factuality and imposes less stringent penalties for deletion, allowing reference-based metrics like *Agg-BERTScore* and *Agg-LENS* to perform better.

Recommendations. Based on the results, we recommend *Agg-LENS* to evaluate readability and fluency and *Agg-BERTScore* to evaluate meaning preservation for long text simplification. For reference-less cases, *Agg-REFEREE* is suitable to evaluate meaning preservation and *REFEREE* for the other dimensions.

		Deletion	In-Document Hallucination	Out-Document Hallucination	Grammar	Coherence	Copy
Sentence-level, Reference-based	BLEU	73.3	80.0	83.3	96.7	91.0	11.6
	SARI	**88.3**	45.0	63.3	81.7	81.7	43.3
	BERTScore	83.3	96.7	**100.0**	**100.0**	90.0	15.6
	LENS	60.0	53.3	88.3	**100.0**	73.3	83.3
Sentence-level, Reference-less	SLE	43.3	86.7	91.7	50.0	45.0	26.7
	LENS-SALSA	65.0	51.7	60.0	**100.0**	66.3	90.0
	REFEREE	10.7	**98.1**	98.6	**100.0**	83.7	**91.7**
Document-level	DSARI	43.3	88.3	88.3	83.3	81.7	60.0
	QuestEval	63.3	20.0	86.7	65.0	80.0	21.6
	LLAMA3	83.3	96.7	98.3	**100.0**	90.0	80.0
Metrics aggregated using our approach							
Document-level, Reference-based	Agg-SARI	80.0 ↓	88.3 ↑	**100.0** ↑	78.3 ↓	91.7 ↑	63.0 ↑
	Agg-BERTScore	81.1 ↑	95.0 ↓	96.7 ↓	90.0 ↓	**98.3** ↑	8.6 ↑
	Agg-LENS	81.7 ↑	68.3 ↑	80.0 ↓	**100.0**	81.6 ↑	46.7 ↓
Document-level, Reference-less	Agg-SLE	85.0 ↑	60.3 ↓	60.0 ↓	55.0 ↑	76.7 ↑	31.7 ↑
	Agg-LENS-SALSA	71.6 ↑	46.7 ↓	43.3 ↓	**100.0**	86.7 ↑	68.3 ↓
	Agg-REFEREE	76.3 ↑	88.3 ↓	85.0 ↓	**100.0**	90.0 ↑	70.3 ↓

Table 2: Consistency (%) of automatic simplification metrics on ONESTOPQA. The best values are marked in **bold** and the second best values are underlined. ↑ and ↓ represent if the aggregated version of the metric improves or degrades the performance when compared to the original sentence-level version.

4 Evaluating Robustness using Adversarial Attacks

While correlation with human ratings gauges the capability of automatic metrics to evaluate the overall quality of simplification, it may not effectively capture how sensitive metrics are to minor errors in simplification system outputs. Considering that state-of-the-art generation models such as GPT4 (OpenAI et al., 2024) are capable of producing high-quality text, it is crucial for evaluation metrics to detect even minor errors. In this section, we outline various types of errors typically generated by paragraph-level simplification systems and suggest methods for perturbing a well-written simplification to introduce each error with minimal alterations. Subsequently, we present results obtained by different automatic metrics in detecting these minor errors.

4.1 Common Simplification System Errors

We highlight the common errors made by long text simplification systems and the techniques we employed to generate each error type.

Deletion of Salient Information. At times, simplification systems fail to retain crucial information from the input text. To replicate this error, we omit the longest 20% of sentences from the text, as they are likely to contain vital information.

Hallucinations. Generating text that deviates from the intended context is a prevalent error generated by state-of-the-art generation systems (Huang et al., 2023). Following Guo et al. (2024) that introduces hallucinations into well-composed summaries for evaluation of metrics, we propose modifications to simplified texts to induce two types of hallucinations: (a) **In-Document hallucinations**, which relate to the input text's topic, and (b) **Out-Document hallucinations**, which introduce unrelated information. To create in-document hallucinations, we add two random sentences from the same article as the input paragraph. For out-document hallucinations, we append two sentences from a randomly selected article in the dataset.

Grammatical Errors. These errors include mistakes in the use of grammar that disrupt the flow of a sentence. We swap the order of 4-5 words in 20% of the sentences in the simplified paragraph to simulate grammatical errors.

Coherence. While a simplified text may exhibit fluency, it can still pose reading challenges due to poor logical arrangement of sentences in the text, a characteristic known as textual coherence. We generate incoherent texts by swapping the order of 20% of the sentences in the coherent simplifications.

Copying with Minimal Paraphrasing. Occasionally, simplification systems make minor changes that do not affect text complexity. To simulate this, we select the complex text as its own simplification and paraphrase 10% of its sentences using a T5-based model (Raheja et al., 2023) that preserves the original complexity.

4.2 Evaluation Setup

We apply our perturbations to 60 simplifications generated by ChatGPT[7] in the ONESTOPQA dataset. We selected ChatGPT due to its high-quality simplifications. For each error type, we create a modified erroneous version based on the original ChatGPT simplification. We then calculate the consistency of each metric across all simplifications. **Consistency** of a metric refers to the percentage of simplifications in which the perturbed version is ranked lower than the generated simplification by the metric. This measure has been used previously to assess the robustness of factuality metrics (Ma et al., 2023; Gabriel et al., 2021).

4.3 Results

Table 2 shows the sensitivity of metrics to different types of errors. We summarize the trends below:

SARI is the most sensitive towards deletion. SARI heavily penalizes deletion by computing deleted n-grams relative to the input. Additionally, aggregated versions of all metrics, with the exception of SARI, demonstrate superior performance compared to their sentence-level counterparts in capturing deletion.

In-Document hallucinations are more challenging than Out-Document hallucinations. All metrics show lower consistency scores while identifying in-document hallucinations that include new information from the same topic than out-document hallucinations that incorporate information from irrelevant topics.

Aggregated metrics underperform their non-aggregated counterparts on hallucinations. Aggregated versions of all metrics, with the exception of SARI, show a drop in consistency when compared to their original versions. This is because aggregated versions treat good sentence-level simplifications and hallucinations equally, whereas original metrics penalize hallucinations more.

Aggregated metrics outperform their non-aggregated versions on incoherent text. *Agg-SARI* and *Agg-BERTScore* are the best at capturing coherence errors. Aggregation improves the consistency scores for all the metrics.

Most metrics effectively penalize grammatical errors. We observe 100% consistency scores for *BERTScore*, *LENS*, *LENS-SALSA*, *Agg-LENS*, and *REFEREE* in identifying fluency errors. However, *SLE* and *SARI* display lower scores compared to the others due to their emphasis on simplicity.

5 Ablation Analysis

We analyze the design decisions that are essential for the effective performance of our approach: (a) a threshold of 0.5 for sentence pair similarity, (b) graph-based alignment of sentences, and (c) sentence pair similarity model trained on parallel simplification corpora.

Sentence Similarity Threshold. Figure 2 illustrates the correlation with human ratings for Agg-LENS, Agg-BERTScore, and Agg-SARI across different thresholds for sentence pair similarity. In our graph-based alignment approach, sentence pair similarity is represented by the edges between sentence nodes. We add edges between sentences only if their similarity value exceeds the threshold. A higher threshold results in fewer edges for alignment and consequently more sub-units of text. The results indicate minimal variance with respect to the threshold, with a threshold of 0.5 serving as a reasonable choice. This is primarily because the similarity values generated by the sentence pair similarity model are mostly clustered near the extreme ends of the scale (close to 0 or 1),

Graph-based Alignment of Sentences. Table 3 compares our alignment approach, which supports many-to-many alignments between sentences in complex and simplified texts, with more restricted variants that match each simplified sentence with the most similar complex sentence, or vice versa. The results show that our graph-alignment approach outperforms the one-to-many and many-to-one alignment methods. This is because the latter methods fail to account for multi-sentence simplification operations such as sentence reordering, content selection and fusing sentences, which frequently occur in long text simplification.

Sentence Pair Similarity Model. Table 4 demonstrates that our method, which employs the sentence pair similarity model from Jiang et al.

[7] https://openai.com/index/chatgpt/

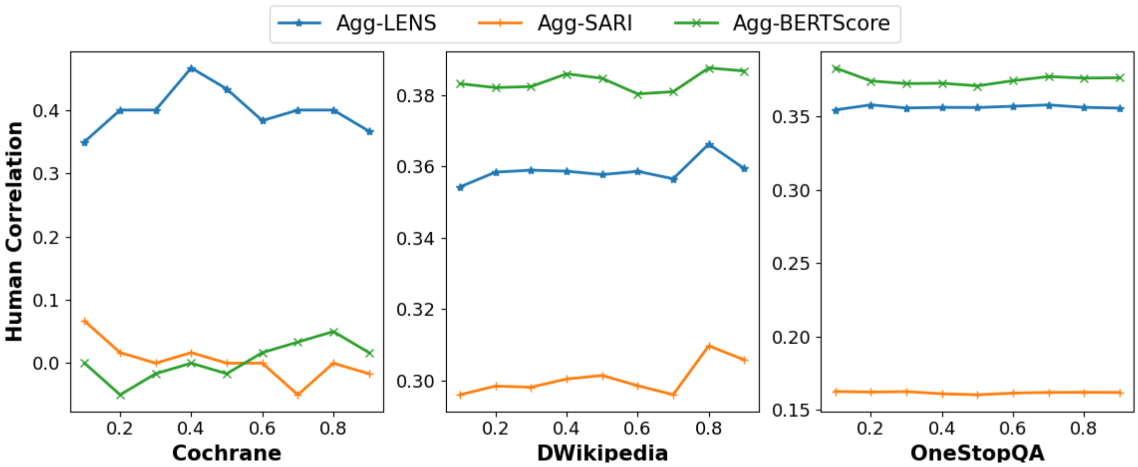

Figure 2: Ablation results of Agg-LENS, Agg-BERTScore, and Agg-SARI with different thresholds for sentence pair similarity. We plot the Kendall correlation value for COCHRANE-HUMAN and average Pearson correlation value across all the dimensions for D-WIKIPEDIA and ONESTOPQA respectively.

	COCHRANE	DWIKI	ONESTOPQA
Proposed approach with many-to-many alignments			
Agg-SARI	0.0	0.301	0.160
Agg-BScore	0.167	0.384	0.371
Agg-LENS	0.433	0.358	0.356
Simplified sentence aligned to the best complex sentence			
Agg-SARI	0.033	0.181	0.116
Agg-BScore	0.1	0.301	0.337
Agg-LENS	0.367	0.267	0.252
Complex sentence aligned to the best simplified sentence			
Agg-SARI	-0.017	0.192	0.118
Agg-BScore	0.0	0.303	0.341
Agg-LENS	0.383	0.252	0.244

Table 3: Ablation results comparing our graph-based alignment approach with their variants allowing only one-to-many and many-to-one alignments. We report the Kendall correlation value for COCHRANE-HUMAN and average Pearson correlation value across all the dimensions for D-WIKIPEDIA and ONESTOPQA respectively.

(2020), outperforms both BERTScore[8] and SentenceBERT[9] (Reimers and Gurevych, 2019). Jiang et al. (2020) fine-tuned BERT on manually aligned complex-to-simple article pairs from the Wikipedia corpus (Xu et al., 2015). This indicates that an in-domain sentence similarity model fine-tuned on simplification corpora surpasses generalized similarity approaches.

6 Related Work

Automatic Evaluation of Text Simplification. Existing automatic metrics for simplification are primarily designed for sentence simplification and can be broadly divided into three types: (a) lexical similarity metrics (Xu et al., 2016; Papineni et al., 2002); (b) semantic similarity metrics (Zhang et al., 2020; David et al., 2023); and (c) learned metrics (Maddela et al., 2023; Huang and Kochmar, 2024; Heineman et al., 2023; Cripwell et al., 2023), which fine-tune pretrained language models on human judgments. There has been limited exploration of evaluation metrics for documents. Readability metrics, such as Flesch-Kincaid Grade Level (Kincaid, 1975), have also been used to assess the simplicity dimension of simplified texts. However, studies have shown that these metrics do not correlate well with the overall quality of the generated simplification (Maddela et al., 2023; Alva-Manchego et al., 2021; Devaraj et al., 2021). Sun et al. (2021) introduced a new metric for document-level simplification that incorporates length penalties into SARI. Rebuffel et al. (2021) proposed QuestEval, a paragraph-level evaluation metric that generates questions based on simplifications and measures the similarity of their answers to the source text. However, this metric focuses solely on meaning preservation. Conversely, our approach adapts sentence-level metrics for long text simplification.

Evaluation of Long-Text Generation. Document-level automatic metrics for common text generation tasks such as machine translation (Papineni et al., 2002; Sellam et al., 2020; Agarwal and Lavie, 2008) and summarization (Lin, 2004; Vasilyev et al., 2020) focus on meaning preservation. Splitting long texts into shorter chunks has been explored for summariza-

[8]We used "roberta-large" model for BERTScore.
[9]We used "all-mpnet-base-v2" for SentenceBERT.

	COCHRANE	DWIKI	ONESTOPQA
Our work with similarity model from Jiang et al. (2020)			
Agg-SARI	0.033	0.301	0.160
Agg-BScore	0.167	0.384	0.371
Agg-LENS	0.433	0.358	0.356
BERTScore as similarity model			
Agg-SARI	0.033	0.198	0.111
Agg-BScore	0.17	0.295	0.321
Agg-LENS	0.267	0.225	0.202
SentenceBERT as similarity model			
Agg-SARI	-0.05	0.259	0.124
Agg-BScore	-0.033	0.319	0.226
Agg-LENS	0.233	0.321	0.290

Table 4: Ablation results of Agg-SARI, Agg-BERTScore, and Agg-LENS with different sentence pair similarity models. We report the Kendall correlation value for COCHRANE-HUMAN and average Pearson correlation value across all the dimensions for D-WIKIPEDIA and ONESTOPQA respectively.

tion, focusing on two directions: (1) decomposing a summary into smaller facts (min; Nawrath et al., 2024) and (2) breaking a summary into sentences and aligning them with the sentences in source text (Amplayo et al., 2022). Our approach aligns more closely with the second category. However, these metrics are not suitable for simplification, as they are specifically designed for summarization and prioritize factuality. Additionally, they typically allow only one-to-one mappings between the generated text and the source, which do not capture multi-sentence simplification operations such sentence splitting, sentence fusion, and sentence reordering. In contrast, our approach enables many-to-many alignments among the source, simplified, and reference texts, effectively capturing such operations.

7 Conclusion

In this work, we propose a novel approach for adapting sentence-level automatic metrics for long text simplification. Results show that our approach enhances the correlation with human judgments of sentence-level metrics across multiple domains. We also conduct the first systematic study of automatic evaluation metrics for document-level simplification by benchmarking a comprehensive range of metrics, spanning traditional lexical and semantic measures to recent learned approaches. Finally, we evaluate the robustness of simplification metrics using adversarial attacks that simulate different errors made by long text simplification systems.

8 Limitations

Limited to English Language. Our work evaluates simplification metrics exclusively for the English language, as all selected human rating datasets are available only in English. This limitation restricts the generalizability of our findings to other languages, where linguistic structures and simplification challenges may differ significantly. Further research is essential to investigate the application of automatic simplification metrics for non-English languages.

Subjectivity of Human Ratings in the Datasets. Human judgments in the selected datasets come from annotators with diverse backgrounds. While COCHRANE-HUMAN and D-WIKIPEDIA include annotations from non-native speakers, ONESTOPQA features annotations from native speakers. This variation may introduce biases, and there is also a degree of inter-annotator disagreement in the ratings. Therefore, the findings of this paper should be interpreted with this subjectivity in mind, as it may influence the overall assessment of simplification metrics and their applicability across different contexts and populations. Further research could benefit from addressing these subjective elements to enhance the reliability of the evaluations.

Lack of Elaborative Simplification Evaluation. Elaboration is a simplification operation that aims to enhance clarity and readability by adding context, explanations, or definitions to complex texts (Srikanth and Li, 2021). However, our study focuses on simplification operations that transform text without adding new content, such as paraphrasing, deleting irrelevant information, fusing sentences, sentence reordering, and sentence splitting. Consequently, the findings of this paper are limited to these operations. Further research is needed to explore the applicability of the proposed approach and existing metrics for elaborative simplification.

9 Ethics Statement

We use publicly available datasets and will make our code available upon publication. As mentioned in the limitations around non-English languages and possible biases from human annotators, further work is needed to apply the proposed approach to specific target audiences.

References

Abhaya Agarwal and Alon Lavie. 2008. METEOR, M-BLEU, and M-TER: Evaluation Metrics for High-Correlation with Human Rankings of Machine Translation Output. In *Proceedings of the Third Workshop on Statistical Machine Translation*. Association for Computational Linguistics.

Sweta Agrawal and Marine Carpuat. 2022. An Imitation Learning Curriculum for Text Editing with Non-Autoregressive Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Sweta Agrawal and Marine Carpuat. 2024. Do Text Simplification Systems Preserve Meaning? A Human Evaluation via Reading Comprehension.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2021. The (Un)Suitability of Automatic Evaluation Metrics for Text Simplification. *Computational Linguistics*, 47(4).

Reinald Kim Amplayo, Peter J. Liu, Yao Zhao, and Shashi Narayan. 2022. SMART: Sentences as Basic Units for Text Evaluation.

Ondřej Bojar, Yvette Graham, and Amir Kamran. 2017. Results of the WMT17 Metrics Shared Task. In *Proceedings of the Second Conference on Machine Translation*. Association for Computational Linguistics.

Liam Cripwell, Joël Legrand, and Claire Gardent. 2023. Simplicity Level Estimate (SLE): A Learned Reference-Less Metric for Sentence Simplification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Beauchemin David, Saggion Horacio, and Khoury R. 2023. MeaningBERT: Assessing Meaning Preservation between Sentences. In *Frontiers of Artificial Intelligence.*, Online. PubMed Central.

Ashwin Devaraj, Iain Marshall, Byron Wallace, and Junyi Jessy Li. 2021. Paragraph-level Simplification of Medical Texts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, GabrielleMe Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei

Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The Llama 3 Herd of Models.

Noemie Elhadad and Komal Sutaria. 2007. Mining a Lexicon of Technical Terms and Lay Equivalents. In *Biological, translational, and clinical language processing*.

Lorenzo Jaime Flores, Heyuan Huang, Kejian Shi, Sophie Chheang, and Arman Cohan. 2023. Medical Text Simplification: Optimizing for Readability with Unlikelihood Training and Reranked Beam Search Decoding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics.

Saadia Gabriel, Asli Celikyilmaz, Rahul Jha, Yejin Choi, and Jianfeng Gao. 2021. GO FIGURE: A Meta Evaluation of Factuality in Summarization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Online. Association for Computational Linguistics.

Aparna Garimella, Abhilasha Sancheti, Vinay Aggarwal, Ananya Ganesh, Niyati Chhaya, and Nandakishore Kambhatla. 2022. Text Simplification for Legal Domain: Insights and Challenges. In *Proceedings of the Natural Legal Language Processing Workshop 2022*.

Yue Guo, Tal August, Gondy Leroy, Trevor Cohen, and Lucy Lu Wang. 2024. APPLS: Evaluating Evaluation Metrics for Plain Language Summarization.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with Disentangled Attention.

David Heineman, Yao Dou, Mounica Maddela, and Wei Xu. 2023. Dancing Between Success and Failure: Edit-level Simplification Evaluation using SALSA. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 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. 2023. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions.

Yichen Huang and Ekaterina Kochmar. 2024. REFeREE: A REference-FREE Model-Based Metric for Text Simplification. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13740–13753, Torino, Italia. ELRA and ICCL.

Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu. 2020. Neural CRF Model for Sentence Alignment in Text Simplification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

Tomoyuki Kajiwara, Hiroshi Matsumoto, and Kazuhide Yamamoto. 2013. Selecting Proper Lexical Paraphrase for Children. In *Proceedings of the 25th Conference on Computational Linguistics and Speech Processing (ROCLING 2013)*.

Kincaid. 1975. Derivation Of New Readability Formulas (Automated Readability Index, Fog Count And Flesch Reading Ease Formula) For Navy Enlisted Personnel. *research branch report*.

Philippe Laban, Tobias Schnabel, Paul Bennett, and Marti A. Hearst. 2021. Keep It Simple: Unsupervised Simplification of Multi-Paragraph Text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online. Association for Computational Linguistics.

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)*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pretraining for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Margaret Li, Stephen Roller, Ilia Kulikov, Sean Welleck, Y-Lan Boureau, Kyunghyun Cho, and Jason Weston. 2020. Don't Say That! Making Inconsistent Dialogue Unlikely with Unlikelihood Training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A Package for AutomaticEevaluation of Summaries. In *Text summarization branches out*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Yang Liu and Mirella Lapata. 2019. Text Summarization with Pretrained Encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv*, abs/1907.11692.

Junru Lu, Jiazheng Li, Byron Wallace, Yulan He, and Gabriele Pergola. 2023. NapSS: Paragraph-level Medical Text Simplification via Narrative Prompting and Sentence-matching Summarization. In *Findings of the Association for Computational Linguistics: EACL 2023*. Association for Computational Linguistics.

Liang Ma, Shuyang Cao, Robert L Logan IV, Di Lu, Shihao Ran, Ke Zhang, Joel Tetreault, and Alejandro Jaimes. 2023. BUMP: A Benchmark of Unfaithful Minimal Pairs for Meta-Evaluation of Faithfulness Metrics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Qingsong Ma, Ondřej Bojar, and Yvette Graham. 2018. Results of the WMT18 Metrics Shared Task: Both characters and embeddings achieve good performance. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*. Association for Computational Linguistics.

Mounica Maddela, Yao Dou, David Heineman, and Wei Xu. 2023. LENS: A Learnable Evaluation Metric for Text Simplification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Louis Martin, Angela Fan, Éric de la Clergerie, Antoine Bordes, and Benoît Sagot. 2022. MUSS: Multilingual Unsupervised Sentence Simplification by Mining Paraphrases. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. European Language Resources Association.

Marcel Nawrath, Agnieszka Nowak, Tristan Ratz, Danilo Walenta, Juri Opitz, Leonardo Ribeiro, João Sedoc, Daniel Deutsch, Simon Mille, Yixin Liu, Sebastian Gehrmann, Lining Zhang, Saad Mahamood, Miruna Clinciu, Khyathi Chandu, and Yufang Hou. 2024. On the Role of Summary Content Units in Text Summarization Evaluation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. GPT-4 Technical Report.

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 on Association for Computational Linguistics*.

David Pellow and Maxine Eskenazi. 2014. An Open Corpus of Everyday Documents for Simplification Tasks. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*.

Sarah E. Petersen and Mari Ostendorf. 2007. Text Simplification for Language Learners: A Corpus Analysis. In *Slate*.

Vipul Raheja, Dhruv Kumar, Ryan Koo, and Dongyeop Kang. 2023. CoEdIT: Text Editing by Task-Specific Instruction Tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics.

Clément Rebuffel, Thomas Scialom, Laure Soulier, Benjamin Piwowarski, Sylvain Lamprier, Jacopo Staiano, Geoffrey Scoutheeten, and Patrick Gallinari. 2021. Data-QuestEval: A Referenceless Metric for Data to Text Semantic Evaluation. *arXiv preprint arXiv:2104.07555*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics.

Luz Rello, Ricardo Baeza-Yates, and Horacio Saggion. 2013. The Impact of Lexical Simplification by Verbal Paraphrases for People with and without Dyslexia. In *Computational Linguistics and Intelligent Text Processing*.

Horacio Saggion. 2017. Automatic Text Simplification. *Synthesis Lectures on Human Language Technologies*.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning Robust Metrics for Text Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

Kim Cheng Sheang and Horacio Saggion. 2021. Controllable Sentence Simplification with a Unified Text-to-Text Transfer Transformer. In *Proceedings of the 14th International Conference on Natural Language Generation*. Association for Computational Linguistics.

Neha Srikanth and Junyi Jessy Li. 2021. Elaborative Simplification: Content Addition and Explanation Generation in Text Simplification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Online.

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*.

Sowmya Vajjala and Ivana Lučić. 2018. OneStopEnglish corpus: A New Corpus for Automatic Readability Assessment and Text Simplification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics.

Oleg Vasilyev, Vedant Dharnidharka, and John Bohannon. 2020. Fill in the BLANC: Human-Free Quality Estimation of Document Summaries. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, Online. Association for Computational Linguistics.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics (TACL)*.

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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

A Dataset Statistics

	Avg. word length	
	Complex	Simple
COCHRANE-HUMAN	396.7	213.4
ONESTOPQA	172.3	92.4
D-WIKIPEDIA	137.1	123.9

Table 5: Dataset Statistics

B Llama3 Prompts and Evaluation Details

We attempted five evaluations and averaged the results. We used the default temperature of Llama3 (0.6). We evaluated Llama3 in a zero-shot setting without a reference and a one-shot setting with a human reference. We reported results on the zero-shot setting as it performed the best.

We use the following prompts for the *Llama3-8B-Instruct*[10] model under a zero-shot setting.

B.1 Prompt for fluency:

You will be given a text and its simplified version written by an AI system. Your task is to rate the simplified version in terms of fluency on a scale of 1-5. Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria for Fluency:
Coherence: How well does the simplified version

[10] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

flow logically and smoothly, with each sentence building on the previous one to convey a clear and cohesive message?
Grammar and Syntax: Are the sentences in the simplified version grammatically correct, with proper sentence structure, verb tense consistency, and subject-verb agreement?
Vocabulary: Are the words and phrases used in the simplified version appropriate, accurate, and concise, without unnecessary complexity or ambiguity?
Readability: How easy is the simplified version to read and understand, with a natural flow and rhythm that facilitates comprehension?
Naturalness: How well does the simplified version sound like natural language, with a tone and style that is engaging and clear?

Rating Scale:
1: Very Poor (simplified version is difficult to follow, with significant grammatical errors and awkward phrasing).
2: Poor (simplified version is clumsy, with noticeable errors in grammar, syntax, or vocabulary).
3: Fair (simplified version is understandable, but with some awkward phrasing, minor errors, or slightly unnatural language).
4: Good (simplified version is clear and coherent, with good grammar, syntax, and vocabulary, and a natural flow).
5: Excellent (simplified version is highly fluent, with a smooth and natural flow, accurate vocabulary, and no noticeable errors).

Now, rate the simplification:
Source Text: ||complex||
Simplified Text: ||simplification||
Please write only the numeric rating in the next line:

B.2 Prompt for meaning preservation:

You will be given a text and its simplified version written by an AI system. Your task is to rate the simplified version in terms of meaning preservation on a scale of 1-5. Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria for Meaning Preservation:
Accuracy: How well does the simplified version maintain the original meaning and content of the paragraph?
Completeness: Does the simplified version cover all the main points and essential information from the original paragraph?
Fidelity: How faithful is the simplified version to the tone, style, and intent of the original paragraph?
Clarity: Is the simplified version clear and easy to understand, without introducing ambiguity or confusion?
Omissions: Are any important details or context omitted from the simplified version that alter its meaning or impact?

Rating Scale:
1: Very Poor (significant meaning lost or distorted).
2: Poor (some meaning preserved, but with notable omissions or distortions).
3: Fair (most meaning preserved, with minor omissions or distortions).
4: Good (meaning well-preserved, with high fidelity and clarity).
5: Excellent (meaning perfectly preserved, with no omissions or distortions).

Now, rate the simplification:
Source Text: ||complex||
Simplified Text: ||simplification||
Please write only the numeric rating in the next line:

B.3 Prompt for simplicity:

You will be given a paragraph and its simplified version written by an AI system. Your task is to rate the simplified version in terms of simplicity or readability on a scale of 1-5. In other words, the text needs to be easier to understand. Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria for Simplicity/Readability:
Clarity: How easy is the simplified version to understand, with a clear and concise message?
Vocabulary: Are the words used in the simplified version simple, common, and easy to understand?
Sentence structure: Are the sentences in the simplified version short, straightforward, and easy to follow?
Complexity reduction: Has the simplified version successfully reduced the complexity of the original paragraph, making it easier to comprehend?
Overall readability: How easy is the simplified

version to read and understand, with a natural flow and rhythm?

Rating Scale:
1: Very Poor (simplified version is still difficult to understand, with complex language and structures).
2: Poor (simplified version is somewhat easier to understand, but still uses some complex vocabulary or sentence structures).
3: Fair (simplified version is easier to understand, but may still have some clarity issues or slightly complex language).
4: Good (simplified version is clear and easy to understand, with simple vocabulary and straightforward sentence structures).
5: Excellent (simplified version is very easy to understand, with a natural flow and rhythm, and no complexity or clarity issues).

Now, rate the simplification:
Source Text: ||complex||
Simplified Text: ||simplification||
Please write only the numeric rating in the next line:

C Implementation Details

We implement our approach using PyTorch. We utilize the publicly available code released by the authors to execute each metric within our framework. We ran our experiments on one NVIDIA A10 GPU.