Is it Possible to Modify Text to a Target Readability Level? An Initial Investigation Using Zero-Shot Large Language Models

Asma Farajidizaji, Vatsal Raina, Mark Gales
ALTA Institute, University of Cambridge, UK
farajiasma@gmail.com, vr311@cam.ac.uk, mjfg@cam.ac.uk

Abstract

Text simplification is a common task where the text is adapted to make it easier to understand. Similarly, text elaboration can make a passage more sophisticated, offering a method to control the complexity of reading comprehension tests. However, text simplification and elaboration tasks are limited to only relatively alter the readability of texts. It is useful to directly modify the readability of any text to an absolute target readability level to cater to a diverse audience. Ideally, the readability of readability-controlled generated text should be independent of the source text. Therefore, we propose a novel readability-controlled text modification task. The task requires the generation of 8 versions at various target readability levels for each input text. We introduce novel readability-controlled text modification metrics. The baselines for this task use ChatGPT and Llama-2, with an extension approach introducing a two-step process (generating paraphrases by passing through the language model twice). The zero-shot approaches are able to push the readability of the paraphrases in the desired direction but the final readability remains correlated with the original text's readability. We also find greater drops in semantic and lexical similarity between the source and target texts with greater shifts in the readability.

Keywords: Language control, Readability, Text modification

1. Introduction

Natural language consists of information that is conveyed for a targeted audience. In order to make the text appropriate for a diverse set of readers, the source text needs to be modified accordingly. Automatic text simplification is a popular natural language processing (NLP) task where the source text is adapted to make the content easier to understand by reducing its linguistic complexity (Siddharthan, 2014; Sikka and Mago, 2020). Typically such simplification solutions are valuable for various audiences including younger readers (De Belder and Moens, 2010), foreign language speakers (Bingel et al., 2018), dyslexics (Rello et al., 2013), sufferers of autism (Evans et al., 2014) and aphasics (Carroll et al., 1998). Similarly, text elaboration offers methods to make content more challenging for reading comprehension tasks and hence cater to higher level students (Ross et al., 1991).

However, both text simplification and elaboration are able to only relatively control the readability of the text. This means that the generated text is simplified/elaborated relative to the original text document but it does not guarantee the text itself is at an appropriate readability level for the target audience. In an ideal setting, it should be possible to modify a text document to a precise and absolute readability level. Pertinently, the readability of the modified text should be *independent* and *uncorrelated* with the source text's readability. Hence, regardless of the nature of the source text, it can be modified to any other readability level.

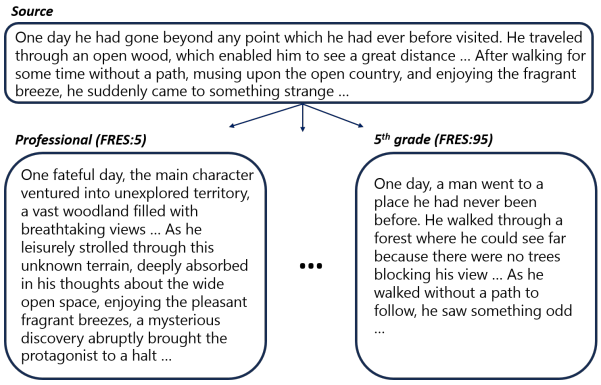


Figure 1: Example for the readability-controlled text modification task. The source text from CLEAR (Crossley et al., 2023) is paraphrased at various target readability levels according to the Flesch reading ease score (FRES) (Flesch, 1948).

To address the relative nature of current text modification approaches, we propose a novel text modification task to control text readability (Harris and Hodges, 1995). Given a set of text documents across the whole spectrum of readability levels, generate 8 versions for each document corresponding to different target readability levels. Precisely, the target readability scores are ranging from being readable for a 5^{th} grade student to understandable for university graduates (Flesch, 1948).

Paraphrasing is a common NLP task where a source text is modified to convey the same mean-

Asma Farajidizaji did the research in collaboration with the ALTA Institute

LREC-COLING 2024, pages 9325–9339

ing but using different words or sentence structures (Zhou and Bhat, 2021). Hence, automated paraphrasing solutions offer an opportunity to modify text to various target readability levels. However, standard solutions do not attempt to control the readability of the generated paraphrase and usually aim to maintain consistency with the source text (Kumar et al., 2020; Chen et al., 2020). Despite the lack of flexibility of these paraphrasing models, the remarkable growth of large-scale autoregressive foundation models (Zhou et al., 2023) have demonstrated capabilities across a broad range of NLP tasks with simple prompting (Sanh et al., 2021). Thus, the baseline solutions for the novel task in our work use zero-shot (Brown et al., 2020) prompting of such models as their backbones for readability-controlled paraphrasing.

The text modification approaches that generate the eight adaptations of each source text document are assessed for their ability to control the readability. The proposed metrics assess both the readability control at an individual example level and the population level. At the individual scale, we assess, with various metrics, whether the readability of a text approaches the target value. At the population scale, we explore the extent to which the measured readability of a generated text document is conditional on the source text document's readability. Additionally, we explore the behaviour of the modified texts for each target readability level according to standard paraphrasing metrics. A good paraphrase can expect to be lexically divergent but semantically similar to the source.

Our contributions can be summarized as follows:

- Introduction of a novel task for readability-controlled text modification.
- Definition of appropriate evaluation metrics for controlling readability.
- In-depth analysis of zero-shot large language model solutions for controlling text readability with paraphrasing.

2. Related Work

In this work we focus on controllability in text modification for readability. Previous works have explored similar approaches and tasks to control various attributes across a diverse range of natural language tasks. Here, we discuss the control of attributes in machine translation (Logeswaran et al., 2018), automatic summarization and text generation (Zhang et al., 2022).

Machine translation is a natural language generation task that translates a source text into a different language. Kikuchi et al. (2016) investigates the ability to control the length of generated sentences such that translations can vary from brief summaries to longer texts. Beyond structural control, Yamagishi et al. (2016) controls the voice of the translation while Sennrich et al. (2016) controls the honorifics and politeness as the selected attributes.

Summarization is a standard natural language generation task where a source text must be condensed whilst maintaining the core elements of the original passage. Automatic summarization has observed the control of various attributes including length, entity-centric, source-specific and linked to a particular portion of the source document (Fan et al., 2018).

In text generation, Zhang et al. (2022) states that the attributes to control are grouped into 3 distinct categories: semantic, structural and lexical. Semantic control involves the control of emotion (Chen et al., 2019; Dathathri et al., 2019) such as sentiment of the generated text as well the choice of the topic (Khalifa et al., 2020) being discussed and the degree of toxicity (Krause et al., 2021; Liu et al., 2021) in the text. Structural control typically looks at defining the syntax in the generated text and the occurrence of graphs and tables (Puduppully et al., 2019; Ribeiro et al., 2021). Finally, lexical control in text generation focuses on attributes such as the inclusion of keywords or phrases (Carlsson et al., 2022; He, 2021).

Besides controlling of attributes, there have been attempts to control text simplification to specific target readability levels. For example, Alkaldi and Inkpen (2023) makes use of the Newsela dataset (Xu et al., 2015) to simplify challenging news articles to four different readability levels. In contrast, the work in this paper emphasises the need to be able to take text of any source readability to any target readability. Hence, text elaboration is important alongside text simplification.

3. Text Readability

Text readability assesses how easy a piece of text is to read. Several standard measures exist for measuring the readability of text including the Flesch-Kincaid Grade Level (Kincaid et al., 1975), Dale Chall Readability (Dale and Chall, 1949), Automated Readability Index (ARI) (Senter and Smith, 1967), Coleman Liau Index (Coleman and Liau, 1975), Gunning Fog (Gunning et al., 1952), Spache (Spache, 1953) and Linsear Write (Klare, 1974).

In this work, the Flesch reading-ease (Flesch, 1948) score (FRES) is used where higher scores indicate material that is easier to read while lower scores are reflective of more challenging passages. The score accounts for the ratio of the number of words to the number of sentences and the ratio of the number of syllables to the number of words to determine the overall readability as indicated in

Equation 1 [1].

$$\text{FRES} = 206.835 - 1.015\left(\frac{n_w}{n_{se}}\right) - 84.6\left(\frac{n_{sy}}{n_w}\right) \quad (1)$$

where n_w denotes the total number of words, n_{se} denotes the total number of sentences and n_{sy} denotes the total number of syllables.

Range	Level (US)	Description
0-10	Professional	Extremely difficult to read. Best understood by university graduates.
10-30	College graduate	Very difficult to read. Best understood by university graduates.
30-50	College	Difficult to read.
50-60	10-12th grade	Fairly difficult to read.
60-70	8-9th grade	Plain English. Easily understood by 13- to 15-year-old students.
70-80	7th grade	Fairly easy to read.
80-90	6th grade	Easy to read. Conversational English for consumers.
90-100	5th grade	Very easy to read. Easily understood by an average 11-year-old student.

Table 1: Interpretable meaning of FRES (Flesch, 1948).

FRES is selected as a simple measure for readability because it has highly interpretable ranges for the score as well as a high correlation with human comprehension as measured by reading tests (DuBay, 2007). For example, Table 1 shows that a FRES score below 10 indicates the text is readable by university graduates, FRES in the fifties is targeted for $10-12^{\text{th}}$ grade while FRES above 90 is readable for 5^{th} grade students. Such well defined ranges allows an exploration of the ability for controlling the readability of text. Note, FRES is not strictly constrained to be in the range of 0 to 100.

[1] Implementation available at: https://pypi.org/project/py-readability-metrics/

4. Readability-Controlled Text Modification

4.1. Task definition

The readability-controlled text modification task is defined as follows:

"Given a text paragraph x, a function $\mathcal{F}$ for calculating readability scores, and K pre-defined readability scores r_1, r_2, ..., r_K, generate K versions of x (y_1, y_2, ..., y_K), such that $\mathcal{F}(y_1) = r_1$, $\mathcal{F}(y_2) = r_2$, ..., $\mathcal{F}(y_K) = r_K$."

In this work, $K = 8$ with $r_1 = 5$, $r_2 = 20$, $r_3 = 40$, $r_4 = 55$, $r_5 = 65$, $r_6 = 75$, $r_7 = 85$, $r_8 = 95$ and FRES (see Equation 1) is selected as the readability function, $\mathcal{F}$. This task is applied for every text in a dataset of text paragraphs. The target readability scores are selected as the halfway values for each range of FRES from Table 1.

4.2. Evaluation

The quality of the readability-controlled text modifications generated are assessed according to individual and population scale control in readability as well as additional analysis with standard paraphrasing metrics.

Individual-scale readability control: For each example in a test set, 8 paraphrases are generated. The individual-scale readability control metrics assess the ability to appropriately control the readability of these paraphrases for each individual example. Broadly, the ranking, regression and classification abilities of a readability-controlled paraphrase generator are assessed.

Let x denote the original text sequence, $y_{(r)}$ denote the generated paraphrase with target readability score of $r \in \mathcal{R} = \{5, 20, 40, 55, 65, 75, 85, 95\}$. Let $\mathcal{F}(\cdot)$ represent the function for calculating FRES from Equation 1.

The ranking ability is assessed by calculating the Spearman's rank correlation coefficient, ρ, between the 8 values of $\mathcal{F}\left(y_{(r\in\mathcal{R})}\right)$ and $\mathcal{R}$. Hence, here we only assess whether the order of the generated paraphrases aligns with their target readabilities.

Given the target readability scores, the regression ability of the model is assessed by calculating the root mean square error (rmse) between the actual and target readability scores of the paraphrases.

$$\text{rmse} = \left[\frac{1}{8}\sum_{r\in\mathcal{R}}\left(\mathcal{F}(y_{(r)}) - r\right)^2\right]^{1/2} \quad (2)$$

Finally, the classification ability checks the ability

of the paraphrase generator to control the readability of the generated text into the target range as defined in Table 1. For example, a paraphrase with a target readability of 65 is deemed correct if the measured generated text readability is in the range of 60-70 and incorrect otherwise. Therefore, the classification accuracy can be calculated according to Equation 3.

$$\text{accuracy} = \frac{1}{8} \sum_{r \in \mathcal{R}} \mathbf{1}_{\mathcal{A}_{(r)}}(\mathcal{F}(y_{(r)})) \quad (3)$$

where $\mathcal{A}_{(5)} \in [0, 10]$, $\mathcal{A}_{(20)} \in [10, 30]$, $\mathcal{A}_{(40)} \in [30, 50]$, $\mathcal{A}_{(55)} \in [50, 60]$, $\mathcal{A}_{(65)} \in [60, 70]$, $\mathcal{A}_{(75)} \in [70, 80]$, $\mathcal{A}_{(85)} \in [80, 90]$, $\mathcal{A}_{(95)} \in [90, 100]$.

For the ranking, regression and classification metrics, the mean is reported across the test set of examples.

Population-scale readability control These metrics assess the actual readability of each target readability across a whole population (test set) rather than considering each example individually. In particular, an important aspect of readability control requires the controlled readability of the generated text to be decorrelated and independent with the source passage readability. In principle, the original text should not have any influence on the readability of the generated text if the control of the paraphrase generator is ideal.

First, we report the Pearson's correlation coefficient (pcc) between the source readability and the calculated generated text readability separately for each target readability class. Ideally, a decorrelated score should expect pcc=0.

Additionally, a linear regression line of the form of $y = ax + b$ is calculated for each target readability class between the source and generated text readability scores. In an ideal setting, the regression line should approach a gradient $a = 0$.

Standard paraphrasing A good paraphrase should be lexically divergent but semantically similar to the original text (Gleitman and Gleitman, 1970; Chen and Dolan, 2011; Bhagat and Hovy, 2013). In line with Lin et al. (2021), we assess lexical divergence using self-WER (Och, 2003) [2]. Semantic similarity is assessed using BERTScore (Zhang et al., 2019) [3].

Self-WER calculates the word error rate (WER) (inspired from automatic speech recognition (Malik et al., 2021) and machine translation (Lee et al., 2023)) between the source and generated text respectively. A lexically divergent paraphrase can expect to have a high self-WER. BERTScore compares the semantic similarity of the source and paraphrase by calculating the pairwise cosine similarities between pre-computed BERT (Kenton and Toutanova, 2019) token embeddings of each of the texts. Hence, the F1 metric is reported as the harmonic mean of precision and recall.

[2]Implementation available at: https://github.com/belambert/asr-evaluation

[3]Implementation available at: https://github.com/Tiiiger/bert_score

5. Experiments

5.1. Data

CLEAR (Crossley et al., 2023, 2021) is a large-scaled corpus for assessing text readability. Here, it is used as a test set of input passages on which readability-controlled text modification is performed. Table 2 outlines the main statistics. There are roughly 5000 different texts with a mean of 10 sentences, allowing the text modification task to be performed at the passage-level rather than at the sentence-level.

# examples	# words	# sentences	# paragraphs
4,724	$179_{\pm 18}$	$9.6_{\pm 4.6}$	$2.5_{\pm 1.9}$

Table 2: CLEAR dataset statistics.

Other standard datasets exist for text simplification but these are generally at the sentence-level (Sun et al., 2021) while we focus on longer texts. Alternatively, various popular passage-level datasets exist in reading comprehension and paraphrasing literature. Figure 2 compares the distribution of the FRES for the passages within CLEAR, the SQuAD (Rajpurkar et al., 2016) development set and News-Commentary (Lin et al., 2021) test set. Due to the presence of texts across the whole spectrum of FRES scores, CLEAR is an attractive choice for investigating readability-controlled text modification. Hence, the experiments here are conducted on the CLEAR dataset only.

5.2. Zero-shot

Large-scale generative foundation models (Brown et al., 2020; Chowdhery et al., 2022; Scao et al., 2022), including the popularized ChatGPT, have demonstrated state-of-the-art performance across a large range of natural language tasks in zero-shot and few-shot settings. Despite not having been specifically trained on certain tasks, these models are capable of successfully performing novel tasks with natural language prompting. Therefore, our baseline solutions for readability-controlled text modification involve zero-shot solutions using ChatGPT and Llama-2 (Touvron et al., 2023). Specif-

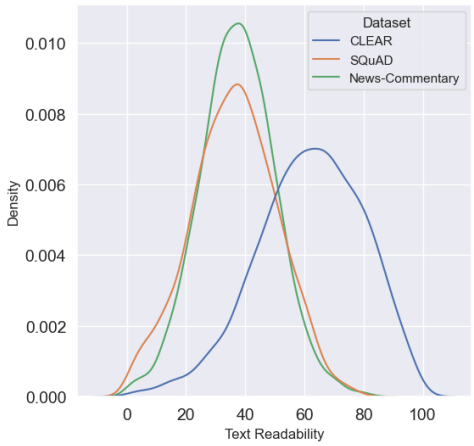


Figure 2: Distribution of text readability scores.

Target	Prompt
5	Paraphrase this document for a professional. It should be extremely difficult to read and best understood by university graduates.
20	Paraphrase this document for college graduate level (US). It should be very difficult to read and best understood by university graduates.
40	Paraphrase this document for college level (US). It should be difficult to read.
55	Paraphrase this document for 10th-12th grade school level (US). It should be fairly difficult to read.
65	Paraphrase this document for 8th/9th grade school level (US). It should be plain English and easily understood by 13- to 15-year-old students.
75	Paraphrase this document for 7th grade school level (US). It should be fairly easy to read.
85	Paraphrase this document for 6th grade school level (US). It should be easy to read and conversational English for consumers.
95	Paraphrase this document for 5th grade school level (US). It should be very easy to read and easily understood by an average 11-year old student.

Table 3: Model prompts for each target readability level.

ically, we use `gpt-3.5-turbo` [4] and `Llama-2-7b-chat-hf` [5] respectively.

Note, Llama-2 model weights are open-sourced, allowing future solutions to further finetune the zero-shot solution specifically for readability controlled text modification. Inference with ChatGPT only requires API requests while for Llama-2, generating 8 paraphrases per example passage takes approximately 45 seconds on an Nvidia A100 GPU. All experiments conducted in this work are based on publicly accessible datasets and models for reproducibility [6].

Vanilla The zero-shot solutions using ChatGPT and Llama-2 require natural language prompts to control the generated paraphrases. As the models do not have an inherent understanding of the FRES, explicit prompts are required to control the readability appropriately. Table 3 summarizes the prompts corresponding to each target readability level as defined in Section 4.1. The prompts are selected in relation to the descriptions in Table 1.

It is observed that often the outputs from the Llama-2 zero-shot solution for certain input passages are random incoherent string of tokens. Therefore, a simple garbage detector checks whether the generated paraphrase is coherent English and in the situation garbage is detected, the corresponding paraphrase is replaced by the original text [7].

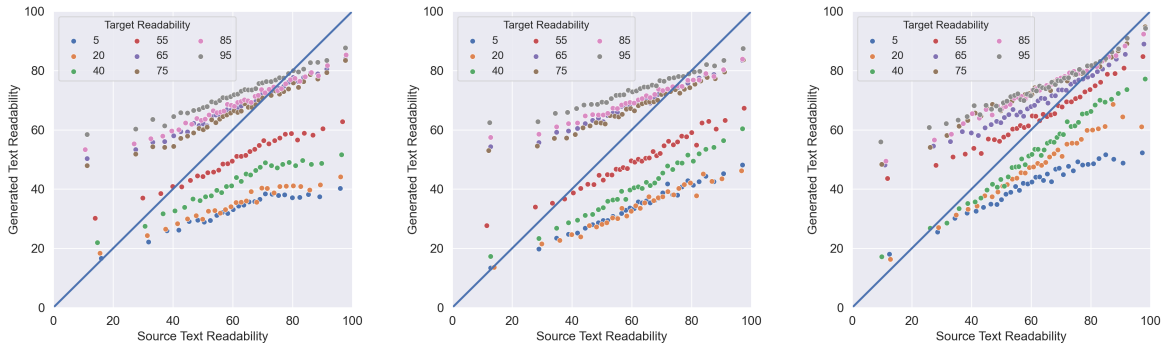
Two-step Unlike many other natural language generation tasks (such as question generation (Lu and Lu, 2021), summarization (Widyassari et al., 2022) and question-answering (Baradaran et al., 2022)), the nature of the output matches the input for text modification. Therefore, paraphrasing based zero-shot approaches to control readability using large language models can sequentially be applied multiple times on a source text. Here, the two-step process is as follows: 1. the selected large language model is prompted to generate a paraphrase at the target readability level as according to Table 3 with the source text at the input; 2. the model is then again prompted (with the identical prompt) to generate a new text but instead with the output from the previous step at the input. The intuition for this approach is motivated by the concept that it is possible to shift closer to a target readability if the source readability is closer to the target value. Here, we explore the two-step process for ChatGPT as it's

[4] API access through https://platform.openai.com/docs/models/gpt-3-5

[5] Available at: https://huggingface.co/meta-llama/Llama-2-7b-chat-hf

[6] Experiments available at: https://github.com/asma-faraji/text-readability-control

[7] This has only been observed to occur in under 1% of generated paraphrases.



(a) ChatGPT 1-step (b) ChatGPT 2-step (c) Llama-2

Figure 3: Generated text readability against source text readability as a binned scatterplot.

Approach	ρ (↑)	rmse (↓)	accuracy (↑)
Copy	0.0	35.4	12.5
ChatGPT 1-step	$\mathbf{87.5}_{\pm 9.2}$	$19.4_{\pm 4.9}$	$23.1_{\pm 12.9}$
ChatGPT 2-step	$86.0_{\pm 9.0}$	$\mathbf{19.2}_{\pm 4.5}$	$\mathbf{24.2}_{\pm 13.2}$
Llama-2	$73.3_{\pm 24.1}$	$23.6_{\pm 8.0}$	$20.6_{\pm 12.6}$

Table 4: Individual-scale readability control with Spearman's rank correlation coefficient (as %), ρ, regression ability with rmse and classification accuracy. The mean across all the examples is reported for each performance metric as well as one standard deviation.

the higher performing model (see Table 4).

6. Results and Discussion

As described in Section 5.2, several baseline solutions are considered for the readability controlled to target values. Table 4 presents the performance of these solutions for the individual-scale metrics averaged across all examples in the CLEAR test set (see Section 4.2). The *copy* system represents the setup where the source text is simply copied for each of target readability levels 5, 20, 40, 55, 65, 75, 85 and 95. Hence, the *copy* system offers a lowerbound on performance according to each of the metrics.

According to the Spearman's rank correlation coefficient, all ChatGPT and Llama-2 implementations are effective at relatively controlling the readability of the text documents with ChatGPT 1-step attaining the highest correlation of 87.5% while Llama-2 lags behind by about 15%. In contrast, the models struggle to directly map the readability of texts to absolute target readability levels with rmse values spanning typically two readability ranges (see Table 1) and the classification accuracies below 25%. Given all approaches are zero-shot implementations where the language models do not have an exact understanding of FRES, it is understandable the models are able to achieve a sensible readability ranking for the 8 generated texts but are incapable of matching the exact target readability values. Additionally, it can be noted that the 2-step process on ChatGPT observes incremental improvements (roughly 0.2 rmse and 1.0 classification accuracy) in achieving the absolute target readability values compared to the 1-step process. This perhaps is because there are two attempts to push the model towards the desired numeric readability score.

Figure 3 presents the relationship between the source text readability and the generated text measured readability for each of the target readability classes 5 to 95. The relationship is plotted as a binned scatterplot where the average measured readability is plotted for each bin on the source text readabilities. For all 3 models, it is observed that the readabilities of each class are generally in a sensible order where the measured readabilities for the target class 5 run along the bottom and the scores for the target class 95 act as the highest curve. Llama-2 also appears to be better than ChatGPT at disentangling the 5 and 20 classes but struggles at the higher classes. Additionally, the 2-step process is able at the lower target readabilities to push down to lower measured readabilities compared to the 1-step process.

However, it is also apparent from Figure 3 that the measured readability of the generated texts, albeit correctly ordered, is highly correlated with the source text readability. Table 5 further quantifies the ability of the models to decorrelate the measured readability of the generated text with the

Target	ChatGPT 1-step				ChatGPT 2-step				Llama-2			
	pcc (↓)	a (↓)	b	r^2	pcc (↓)	a (↓)	b	r^2	pcc (↓)	a (↓)	b	r^2
Source	100	1	0	1	100	1	0	1	100	1	0	1
5	36.9	0.29	14.8	0.14	52.3	0.41	8.7	0.27	47.7	0.41	16.4	0.23
20	39.9	0.32	15.3	0.16	50.8	0.40	8.8	0.26	51.1	0.61	9.9	0.26
40	46.0	0.38	18.2	0.21	58.6	0.52	8.2	0.34	70.0	0.73	6.9	0.49
55	56.0	0.41	24.7	0.31	62.2	0.47	20.4	0.39	65.1	0.51	33.0	0.42
65	69.7	0.42	41.7	0.49	65.8	0.38	45.9	0.43	67.2	0.48	32.9	0.45
75	68.8	0.44	39.4	0.47	65.0	0.38	44.2	0.42	62.6	0.47	44.3	0.39
85	67.8	0.39	44.9	0.46	63.4	0.33	49.6	0.40	63.5	0.47	44.4	0.40
95	66.3	0.36	50.3	0.44	61.3	0.31	54.8	0.38	61.3	0.42	48.1	0.38

Table 5: Population-scale readability control with Pearson's correlation coefficient, pcc, and linear regression, $y = ax + b$, between the source and generated text readability with $0 < r^2 < 1$ denoting the quality of the fit of the regression line with 0 as worst fit and 1 as best fit.

source text readability according to the population-scale metrics (see Section 4.2). An ideal system can expect to have a Pearson's correlation coefficient of 0 and a regression line of best fit with gradient 0 and y-intercept corresponding to the absolute target readability level. It is observed that the relationship for all target readability classes remains highly dependent with the source text readability with ChatGPT 1-step achieving best results for lower target readability classes and better results for ChatGPT 2-step and Llama-2 for higher readability classes. It is also seen that the lower target readability classes are closer to the ideal performance for all models compared to the higher target readability classes.

For further analysis, we look at the behaviour of the generated texts at various target readability metrics according to lexical divergence and semantic similarity from paraphrasing literature (see Section 4.2). We present the analysis here specifically for the best performing model overall: ChatGPT 2-step.

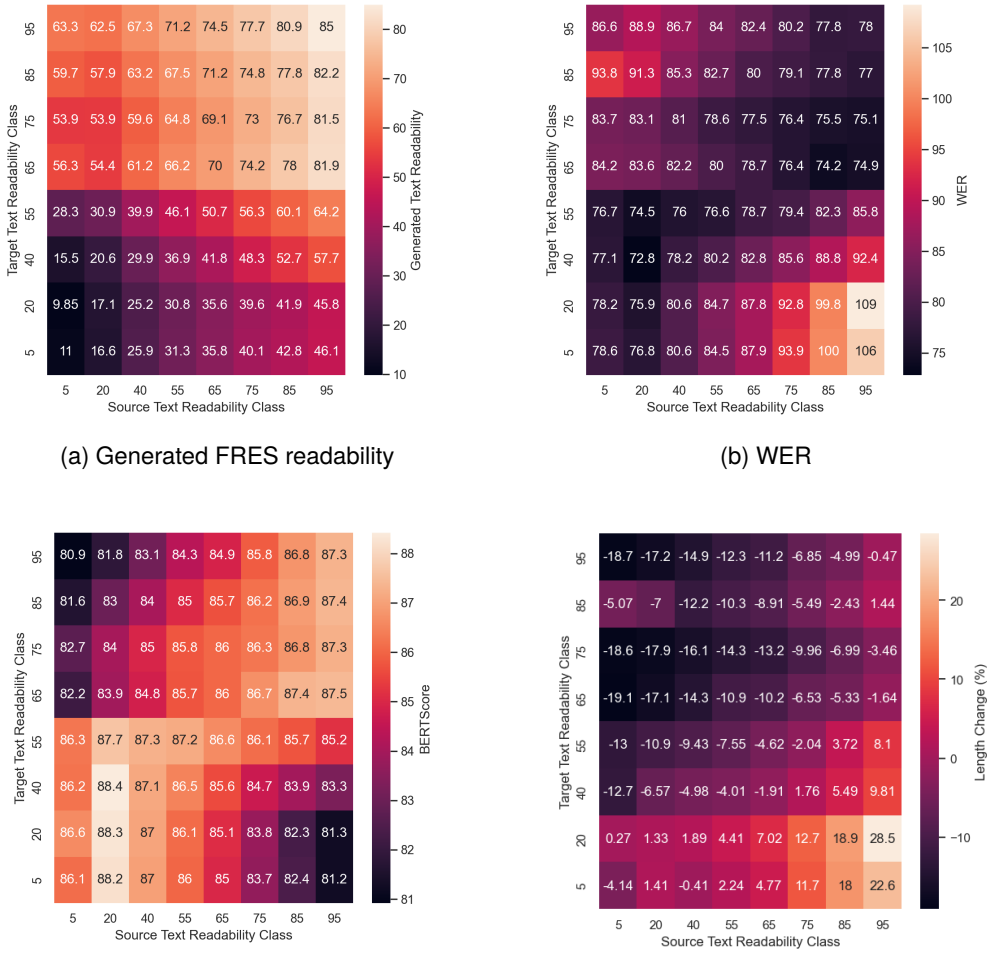
Figures 4a, 4b and 4c display how the generated text readability, WER (measure of lexical divergence) and BERTScore F1 (measure of semantic similarity) of the generated texts respectively vary with shifts between the target and source readability score classes. In order to plot the heatmap, each source text has its readability classed into one of the 8 readability ranges defined by Table 1. Hence the source and target text readability classes fall into one of following classes: $\{5, 20, 40, 55, 65, 75, 85, 95\}$. The heatmap depicts the mean of the selected variable (the variable for the heatmap to plot includes either generated text readability, WER or BERTScore) for each pairing of going from the source readability class (several source texts fall into each class) and the corresponding target readability class.

First, we see Figure 4a reinforces the observations from Figure 3 as the lightest colours are observed in the top right while the darkest in the bottom left. This means that the highest generated text readability scores are observed for when the target readability class is high but also when the source text readability is high.

From Figure 4b it is noticeable that along the leading diagonal we have the darkest shades and lighter shades on the peripheries of the off-diagonal. Conversely, from Figure 4c we see the lighter shades in the leading diagonal. Hence, keeping a matched readability between the source and the target leads to lower lexical divergence and higher semantic similarity. It can further be noted that there is an asymmetry for the WER. For the WER, we see that changing from a very high source text readability to a very low target readability has a greater WER compared to a modification from a low source text readability to a high text readability. This suggests that it is more challenging to maintain the same lexical language for text elaboration compared to text simplification.

We can conclude from the variations in WER and BERTScore that greater the change between the source and target texts, the lower their semantic similarity and greater their lexical divergence.

Additionally, Figure 4d presents the percentage change in the length of the paraphrase relative to the original text for each binned pair of source and target readability levels. The resulting heatmap is in agreement with the observations from Figure 4b where the least absolute change in length is on the leading diagonal while greater variations in lengths are observed when the paraphrasing model is requested to make a greater change in the readability. Moreover, Figure 4d presents a further asymmetry in the sign of the length change. It is observed that length changes are positive in the bottom right corner while the length change is negative in the top left. Hence, decreasing the readability of the text leads to an increase in the overall text length while requesting an increase in the readability causes a reduction in the text length.



(a) Generated FRES readability

(b) WER

(c) BERTScore F1

(d) Length change (words)

Figure 4: Heatmaps of averaged select variables for each pair of source and target text readability classes. Each cell value is the mean of the specified variable for all texts that have a certain source readability and are modified to a certain target readability.

7. Ablations

The core results focus on zero-shot approaches. In this section, an ablation study investigates the potential of finetuning a system for readability controlled text modification. Note, finetuning a system on the task enables absolute understanding of the readability measure, which is not inherently possible with zero-shot approaches.

Here, we look at finetuning Llama-2. Hence, we partition the CLEAR dataset into train and test splits with 2,834 and 1,890 examples respectively. For each text passage in the train split, we generate 8 paraphrases at the set target readability levels described in the core results with ChatGPT. This allows us to form a labeled training dataset of input-output pairs where we have annotations for the source readability for the source passage and the measured readability (not the target readability) for each generated passage. Then Llama-2 is finetuned according to the following prompt: `{source_text} Paraphrase the following document changing the readability from {source_score} to {target_score}.`

We use parameter efficient finetuning (due to compute limits) with quantized low rank adapters (QLoRA) (Dettmers et al., 2023). Final hyperparamters include: learning rate of 1e-3, batch size of 4, lora rank of 8, lora α of 32 and dropout 0.1. Training takes 2 hours for 3 epochs on NVIDIA A100.

The finetuned system achieves a Spearman's rank correlation of 61.4 (scaled by 100), rmse of 29.4 and accuracy of 16.3%. Despite the Llama-2 finetuned system being the worst performing, it has the most exciting results because it is the only model that actually understands the quantitative meaning of the readability score. Hence, it offers opportunity for improvement while the zero-shot

models can only be prompted qualitatively. The quantitative model also offers opportunity to generalize to finetune to control other metrics which don't necessarily have equivalent qualitative prompts like FRES.

8. Conclusions

This work introduces the readability-controlled text modification task. Our task challenges controllable language models to generate eight versions of a text, each targeted for specific readability levels, in a manner independent from the source text readability. Novel metrics, inspired by paraphrasing, assess the quality of readability-controlled text modification. Zero-shot adaptations for ChatGPT and Llama-2 show potential in steering readability but retain some correlation with the source text readability. A two-step process of generating paraphrases sequentially offers modest gains over one-step approaches. Notably, more significant shifts in readability lead to reduced semantic and lexical similarity between source and target texts, highlighting the challenge of balancing readability control and content preservation.

9. Ethics Statement

There are no ethical concerns with this work.

10. Limitations

The main insights drawn from this work for controllable text modification are based upon a single dataset, CLEAR; some of the observations may not generalize to datasets in other domains. Additionally, the current work uses FRES as the readability function. There are alternative options too.

11. Acknowledgements

This research is funded by the EPSRC (The Engineering and Physical Sciences Research Council) Doctoral Training Partnership (DTP) PhD studentship and supported by Cambridge Assessment, University of Cambridge and ALTA.

12. Bibliographical References

Wejdan Alkaldi and Diana Inkpen. 2023. Text simplification to specific readability levels. *Mathematics*, 11(9):2063.

Razieh Baradaran, Razieh Ghiasi, and Hossein Amirkhani. 2022. A survey on machine reading comprehension systems. *Natural Language Engineering*, 28(6):683–732.

Rahul Bhagat and Eduard Hovy. 2013. What is a paraphrase? *Computational Linguistics*, 39(3):463–472.

Joachim Bingel, Gustavo Paetzold, and Anders Søgaard. 2018. Lexi: A tool for adaptive, personalized text simplification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 245–258.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Fredrik Carlsson, Joey Öhman, Fangyu Liu, Severine Verlinden, Joakim Nivre, and Magnus Sahlgren. 2022. Fine-grained controllable text generation using non-residual prompting. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6837–6857.

John Carroll, Guido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait. 1998. Practical simplification of english newspaper text to assist aphasic readers. In *Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, pages 7–10. Association for the Advancement of Artificial Intelligence.

David Chen and William B Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200.

Huimin Chen, Xiaoyuan Yi, Maosong Sun, Wenhao Li, Cheng Yang, and Zhipeng Guo. 2019. Sentiment-controllable chinese poetry generation. In *IJCAI*, pages 4925–4931.

Wenqing Chen, Jidong Tian, Liqiang Xiao, Hao He, and Yaohui Jin. 2020. A semantically consistent and syntactically variational encoder-decoder framework for paraphrase generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1186–1198.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.

Scott Crossley, Aron Heintz, Joon Suh Choi, Jordan Batchelor, Mehrnoush Karimi, and Agnes Malatinszky. 2023. A large-scaled corpus for assessing text readability. *Behavior Research Methods*, 55(2):491–507.

Scott A Crossley, Aron Heintz, Joon Choi, Jordan Batchelor, Mehrnoush Karimi, and Agnes Malatinszky. 2021. The commonlit ease of readability (clear) corpus. In *EDM*.

Edgar Dale and Jeanne S Chall. 1949. The concept of readability. *Elementary English*, 26(1):19–26.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2019. Plug and play language models: A simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164*.

Jan De Belder and Marie-Francine Moens. 2010. Text simplification for children. In *Prroceedings of the SIGIR workshop on accessible search systems*, pages 19–26. ACM; New York.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

William H DuBay. 2007. *Smart Language: Readers, Readability, and the Grading of Text.* ERIC.

Richard Evans, Constantin Orasan, and Iustin Dornescu. 2014. An evaluation of syntactic simplification rules for people with autism. Association for Computational Linguistics.

Angela Fan, David Grangier, and Michael Auli. 2018. Controllable abstractive summarization. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 45–54.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Lila R Gleitman and Henry Gleitman. 1970. Phrase and paraphrase: Some innovative uses of language.

Robert Gunning et al. 1952. Technique of clear writing.

Theodore L Harris and Richard E Hodges. 1995. *The literacy dictionary: The vocabulary of reading and writing.* ERIC.

Xingwei He. 2021. Parallel refinements for lexically constrained text generation with bart. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8653–8666.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Muhammad Khalifa, Hady Elsahar, and Marc Dymetman. 2020. A distributional approach to controlled text generation. *arXiv preprint arXiv:2012.11635*.

Yuta Kikuchi, Graham Neubig, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. 2016. Controlling output length in neural encoder-decoders. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1328–1338.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

George R Klare. 1974. Assessing readability. *Reading research quarterly*, pages 62–102.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2021. Gedi: Generative discriminator guided sequence generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4929–4952.

Ashutosh Kumar, Kabir Ahuja, Raghuram Vadapalli, and Partha Talukdar. 2020. Syntax-guided controlled generation of paraphrases. *Transactions of the Association for Computational Linguistics*, 8:330–345.

Seungjun Lee, Jungseob Lee, Hyeonseok Moon, Chanjun Park, Jaehyung Seo, Sugyeong Eo, Seonmin Koo, and Heuiseok Lim. 2023. A survey on evaluation metrics for machine translation. *Mathematics*, 11(4):1006.

Zhe Lin, Yitao Cai, and Xiaojun Wan. 2021. Towards document-level paraphrase generation with sentence rewriting and reordering. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1033–1044.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A Smith, and Yejin Choi. 2021. Dexperts: Decoding-time controlled text generation with experts and anti-experts. 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)*, pages 6691–6706.

Lajanugen Logeswaran, Honglak Lee, and Samy Bengio. 2018. Content preserving text generation with attribute controls. *Advances in Neural Information Processing Systems*, 31.

Chao-Yi Lu and Sin-En Lu. 2021. A survey of approaches to automatic question generation: from 2019 to early 2021. In *Proceedings of the 33rd Conference on Computational Linguistics and Speech Processing (ROCLING 2021)*, pages 151–162.

Mishaim Malik, Muhammad Kamran Malik, Khawar Mehmood, and Imran Makhdoom. 2021. Automatic speech recognition: a survey. *Multimedia Tools and Applications*, 80:9411–9457.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st annual meeting of the Association for Computational Linguistics*, pages 160–167.

Ratish Puduppully, Li Dong, and Mirella Lapata. 2019. Data-to-text generation with content selection and planning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6908–6915.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Luz Rello, Ricardo Baeza-Yates, Laura Dempere-Marco, and Horacio Saggion. 2013. Frequent words improve readability and short words improve understandability for people with dyslexia. In *Human-Computer Interaction–INTERACT 2013: 14th IFIP TC 13 International Conference, Cape Town, South Africa, September 2-6, 2013, Proceedings, Part IV 14*, pages 203–219. Springer.

Leonardo FR Ribeiro, Martin Schmitt, Hinrich Schütze, and Iryna Gurevych. 2021. Investigating pretrained language models for graph-to-text generation. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 211–227.

Steven Ross, Michael H Long, and Yasukata Yano. 1991. Simplification or elaboration? the effects of two types of text modifications on foreign language reading comprehension.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40.

RJ Senter and Edgar A Smith. 1967. Automated readability index. Technical report, Cincinnati Univ OH.

Advaith Siddharthan. 2014. A survey of research on text simplification. *ITL-International Journal of Applied Linguistics*, 165(2):259–298.

Punardeep Sikka and Vijay Mago. 2020. A survey on text simplification. *arXiv preprint arXiv:2008.08612*.

George Spache. 1953. A new readability formula for primary-grade reading materials. *The Elementary School Journal*, 53(7):410–413.

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.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Adhika Pramita Widyassari, Supriadi Rustad, Guruh Fajar Shidik, Edi Noersasongko, Abdul Syukur, Affandy Affandy, et al. 2022. Review of automatic text summarization techniques & methods. *Journal of King Saud University-Computer and Information Sciences*, 34(4):1029–1046.

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*, 3:283–297.

Hayahide Yamagishi, Shin Kanouchi, Takayuki Sato, and Mamoru Komachi. 2016. Controlling the voice of a sentence in japanese-to-english neural machine translation. In *Proceedings of the 3rd Workshop on Asian Translation (WAT2016)*, pages 203–210.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2022. A survey of controllable text generation using transformer-based pre-trained language models. *ACM Computing Surveys*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, Lifang He, et al. 2023. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. *arXiv preprint arXiv:2302.09419*.

Jianing Zhou and Suma Bhat. 2021. Paraphrase generation: A survey of the state of the art. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 5075–5086.

A. Examples

The below example uses ChatGPT 2-step to generate the paraphrases at each of the target readability scores.

Source

When the young people returned to the ballroom, it presented a decidedly changed appearance. Instead of an interior scene, it was a winter landscape. The floor was covered with snow-white canvas, not laid on smoothly, but rumpled over bumps and hillocks, like a real snow field. The numerous palms and evergreens that had decorated the room, were powdered with flour and strewn with tufts of cotton, like snow. Also diamond dust had been lightly sprinkled on them, and glittering crystal icicles hung from the branches. At each end of the room, on the wall, hung a beautiful bear-skin rug. These rugs were for prizes, one for the girls and one for the boys. And this was the game. The girls were gathered at one end of the room and the boys at the other, and one end was called the North Pole, and the other the South Pole. Each player was given a small flag which they were to plant on reaching the Pole. This would have been an easy matter, but each traveller was obliged to wear snowshoes.

Readability: 74.5

Target 5

Upon the return of the youthful attendees to the grand hall, they noticed a conspicuous transformation had occurred. The traditional interior arrangement had been replaced with a wintry scene. The floor was covered with a pristine canvas resembling a snow-covered terrain, complete with irregular elevations and ripples. The palm trees and evergreen decorations in the room had also been transformed to resemble a snowy landscape, with fine grains of flour and fluffy clusters of cotton creating a snow-like effect. Additional decorations, such as diamond dust sprinkled on the flora, added a glittering icicle-like appearance. On opposite ends of the room, there were ethereal and beautifully displayed bear-skin rugs. These rugs were designated as coveted prizes for both the fair maidens and gallant gentlemen, creating a captivating rivalry. The game involved the young maidens gathering at one end of the room, while the young men gathered at the other. Referred to as the North Pole and South Pole, each participant was given a small flag as an accessory upon their arrival. Despite the seemingly simple task, achieving success in this quest proved to be a challenging endeavor. The participants had to navigate the terrain while wearing snowshoes, adding to the difficulty but also serving as a helpful aid. In summary, upon returning to the grand hall, the attendees were surprised to find a winter wonderland. The floor resembled a snowy landscape, and the decorations had been transformed to create a snowy effect. The bear-skin rugs at either end of the room served as prizes, and the participants had to navigate the terrain while wearing snowshoes.

Readability: 41.9
WER: 127.6%
BERTScore F1: 84.8%
Length change: 47.5%

Target 20

Upon the return of the youthful attendees to the ballroom, a noticeable transformation had occurred. The environment had completely changed, resembling a wintry landscape. The dancing surface was covered in pristine, rumpled snow, imitating the uneven terrain of real snowfields. The palm trees and evergreen plants that had decorated the room were coated with flour and cotton, resembling snow. Diamond dust was delicately sprinkled to give an iridescent effect, while crystal icicles hung from the tree branches, adding to the overall charm. On opposite walls of the room, there were two beautiful rugs made from the highly valued skins of bears. These rugs were meant as desirable prizes, with one given to the female participants and the other to the male contenders. This set the stage for the game, which engaged the participants in an exciting pursuit. The young ladies gathered at one end of the room, while the young gentlemen assembled at the other end, representing the North Pole and the South Pole, respectively. Each player was given a small flag, marking their goal of planting it at their designated pole. However, this apparently simple task presented a significant challenge. Each adventurer had to wear snowshoes on their feet, adding a layer of complexity and skill to the endeavor.

Readability: 47.7
WER: 96.0%
BERTScore F1: 87.1%
Length change: 17.3%

Target 40

Upon the return of the young people to the ballroom, they were met with a notable change. The previous interior had been transformed into a wintry landscape. The floor was covered in an uneven layer of white canvas, resembling real snow with its bumps and hills. The once-adorned palms and evergreens had been altered to look like snow-covered trees, with the addition of flour and cotton. These decorations were lightly sprinkled with diamond dust, and crystal icicles hung from the branches. Magnificent bearskin rugs adorned the walls at each end of the room. These rugs were prizes to be awarded to one girl and one boy. The game involved the girls gathering at the North Pole and the boys assembling at the South Pole. Each participant was given a small flag to plant when they reached their destination. However, this challenge wasn't as simple as it appeared, as every traveler had to wear snowshoes.

Readability: 62.1
WER: 77.0%
BERTScore F1: 88.1%
Length change: -14.0%

Target 65

When the young people returned to the ballroom, it looked completely different. Instead of being indoors, it appeared to be a snowy winter scene. The floor was covered in bumpy and uneven white material, like a real snow field. The trees and plants in the room were made to look snowy with flour and bits of cotton, and had shiny ice hanging from them. There were nice bear-skin rugs on each wall of the room. These rugs were going to be given as prizes, one for the girls and one for the boys. Here's how the game worked: the girls stood on one side of the room and the boys stood on the other side. One side was called the North Pole and the other side was called the South Pole. Each person received a small flag that they had to stick into the floor when they reached their side. But here's the twist - everyone had to wear snowshoes!

Readability: 81.6
WER: 64.9%
BERTScore F1: 88.7%
Length change: -10.6%

Target 55

When the young people returned to the large room for the ball, there was a noticeable change. Instead of being inside, the room now looked like a wintry landscape. The floor was covered with a fabric that looked like snow, but it was not perfectly flat. Instead, it was crumpled and uneven, resembling a real snowfield. The various palm and evergreen trees that had originally decorated the room were coated with flour and had pieces of cotton attached to them to make them look like they were covered in snow. Also, a light sprinkle of diamond dust was put on, and shiny crystal icicles hung from the branches. On the walls at each end of the room, there were two magnificent and beautiful rugs made from the skin of bears. These rugs were meant to be given as prizes - one for the girls and another for the boys. And this is how the game worked: the girls were positioned at one end of the room, while the boys were gathered at the opposite end. These ends were called the North Pole and the South Pole, respectively. Every player was given a small flag that they had to put in the ground when they reached their assigned end. However, there was a challenge involved - each participant had to wear special shoes that make it easier to walk on snow.

Readability: 69.6
WER: 86.8%
BERTScore F1: 88.8%
Length change: 28.5%

Target 75

When the young people returned to the ballroom, it looked completely different. Instead of being in a regular room, it looked like a scene from winter. The floor was covered in white cloth that resembled snow. It wasn't flat, but had bumps and hills like a real snow field. The plants and trees that had been used for decoration were covered in flour and bits of cotton to make them look like they were covered in snow. There was also some shiny material that looked like ice sprinkled on them, and there were shiny ice formations hanging from the branches. On each side of the room, there were beautiful bear-skin rugs hanging on the walls. These rugs were going to be given as prizes, one for the girls and one for the boys. Now, let me explain the game they were going to play. The girls were all standing on one side of the room, and the boys on the other. One side was called the North Pole, and the other was called the South Pole. Each player was given a small flag that they had to stick in the ground when they reached the Pole. But here's the catch - everyone had to wear snowshoes, which made it a bit more challenging.

Readability: 80.9
WER: 75.3%
BERTScore F1: 89.8%
Length change: 19.0%

Target 85

When the young people returned to the ballroom, it looked very different. Instead of being inside, it seemed like a snowy winter scene. The floor was covered with bumpy white cloth that looked like snow, and the trees and plants were decorated with flour, bits of cotton, and shiny pieces that resembled ice. On the walls, there were two beautiful rugs made from bear skins. These rugs would be awarded as prizes, with one for the girls and one for the boys. Here's how the game worked: the girls were at one end of the room, called the North Pole, while the boys were at the other end, called the South Pole. Each player had to wear snowshoes and put a little flag in the ground when they reached their end of the room.

Readability: 76.9
WER: 74.1%
BERTScore F1: 87.4%
Length change: -25.1%

Target 95

When the young people returned to the ballroom, it looked completely different. Instead of being inside, it looked like a snowy winter scene. The floor was covered in white fabric that looked like real snow. The trees and plants were covered in flour and bits of cotton to make them look snowy. They even had shiny ice hanging from the branches. On each wall, there was a beautiful bear-skin rug that would be given as a prize to the boys and girls. The game was for the girls to gather on one side and the boys on the other side. One side was called the North Pole and the other was the South Pole. Each person had to wear snowshoes and carry a small flag. The goal was to reach their assigned pole and plant their flag. This would have been easy, but the snowshoes made it a bit tricky.

Readability: 82.5
WER: 75.3%
BERTScore F1: 87.3%
Length change: -16.2%