Got It! Prompting Readability Using ChatGPT to Enhance Academic Texts for Diverse Learning Needs

Elias Hedlin
KTH Royal Institute of Technology
Stockholm, Sweden
ehedlin@kth.se

Ludwig Estling
KTH Royal Institute of Technology
Stockholm, Sweden
lestling@kth.se

Jacqueline Wong
Utrecht University
Utrecht, Holland
l.y.j.wong@uu.nl

Carrie Demmans Epp*
EdTekLA
University of Alberta (Canada),
Digital Futures (Stockholm)
Edmonton, Canada
demmanse@ualberta.ca

Olga Viberg*
Human-Centered Design
KTH Royal Institute of Technology,
Digital Futures
Stockholm, Sweden
oviberg@kth.se

Abstract

Reading skills are crucial for students' success in education and beyond. However, reading proficiency among K-12 students has been declining globally, including in Sweden, leaving many underprepared for post-secondary education. Additionally, an increasing number of students have reading disorders, such as dyslexia, which require support. Generative artificial intelligence (genAI) technologies, like ChatGPT, may offer new opportunities to improve reading practices by enhancing the readability of educational texts. This study investigates whether ChatGPT-4 can simplify academic texts and which prompting strategies are most effective. We tasked ChatGPT to re-write 136 academic texts using four prompting approaches: Standard, Meta, Roleplay, and Chain-of-Thought. All four approaches improved text readability, with Meta performing the best overall and the Standard prompt sometimes creating texts that were less readable than the original. This study found variability in the simplified texts, suggesting that different strategies should be used based on the specific needs of individual learners. Overall, the findings highlight the potential of genAI tools, like ChatGPT, to improve the accessibility of academic texts, offering valuable support for students with reading difficulties and promoting more equitable learning opportunities.

CCS Concepts

• **Applied computing**; • **Applied computing Education**; • **Applied computing Document management and text processing**; • **Computing methodologies Artificial intelligence**;

Keywords

Large language models, Prompt engineering, Literacy, Readability, Analytics, Equity

*Olga Viberg and Carrie Demmans Epp have contributed equally to this work.

LAK 2025, March 03–07, 2025, Dublin, Ireland

ACM ISBN 979-8-4007-0701-8/25/03
https://doi.org/10.1145/3706468.3706483

ACM Reference Format:
Elias Hedlin, Ludwig Estling, Jacqueline Wong, Carrie Demmans Epp, and Olga Viberg. 2025. Got It! Prompting Readability Using ChatGPT to Enhance Academic Texts for Diverse Learning Needs. In *LAK25: The 15th International Learning Analytics and Knowledge Conference (LAK 2025), March 03–07, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3706468.3706483

1 Introduction

Reading is critical to students' success, given the high demand for text processing and meaning construction across academic disciplines [1]. The importance of reading competence has been recognised in K-12 education worldwide, including the Swedish educational system "through the demands for higher standard in reading in all subjects stated in the national curriculum" [2, p.145]. However, an increasing number of students have been diagnosed with various types of learning disabilities [3], including dyslexia, a prevailing reading disorder [4], and thus require continual reading support to succeed in their studies. The provision of such educational support is the responsibility of various stakeholders, including the municipality in the Swedish context. However, how and when such support should be provided, as well as for how long, are rarely specified [2]. Moreover, instructors today are expected to provide feedback to a larger number of students [5], which can be a challenging task to achieve at an individual level due to time constraints and other teacher tasks. Consequently, students may need to take more responsibility to support themselves in their reading practices.

Students who experience reading difficulties have two avenues they can take: developing their underlying reading skills or using existing support tools and services [6]. Those who gain admission to post-secondary programs have typically already developed their underlying cognitive or reading skills through their formal education or programs like RAVE-O [7], Wilson reading system [8], and PHAST reading program (also called EMPOWER) [9]. In Sweden, the government's "guarantee" (i.e., a program for early support measures to address the needs of individual students who are at risk of not meeting the assessment criteria in Swedish, Swedish as a second language, and mathematics [10]) has not been realized as shown by the results of the PISA 2022 assessment of students'

reading skills in schools [11]. The drop in reading skills among students in Swedish schools was found to be close to the gains observed in 2012, i.e., the lowest ever identified [11]. These results led to a new approved program (i.e., "guarantee") on the part of the Swedish Ministry of Education in 2024 [10]. The aim is to ensure that support measures are implemented early and tailored to each student's needs. The key target groups to be supported are teachers in pre-school and primary education, special education teachers (those who support students with special needs), and school principals [10]. However, how such support should be provided at the individual level, and what exact tools and services should be used are not specified, leaving teachers, students, and schools in a situation similar to 2019. With the increasing accessibility of advanced tools (e.g., ChatGPT) powered by large language models, students can assume responsibility for their own reading practices supported by such tools.

Students who have weak reading skills can benefit from support tools that allow them to access the domain content in a text. Such support tools are currently necessary because the expository texts that are common in the sciences and other domains require higher levels of reading skills than those generally required [12]. Moreover, post-secondary students are expected to read large quantities of these more complex texts, which is difficult if their reading abilities are limited [13]. Simple tools like text-to-speech can support students with dyslexia or those who otherwise struggle to map sounds to characters in a text (e.g., [14]). However, this does not address the needs of students who struggle with reading complex texts because of the advanced or uncommon vocabulary in these texts or the complex grammatical structures that are more common in advanced expository texts. This is where *text simplification* plays an important role in enabling struggling readers to learn and later demonstrate their subject-area knowledge.

Text simplification research has predominantly focused on a narrow subset of languages, primarily European ones such as English, French, German, Portuguese, Spanish, and Italian. This focus is largely due to the availability of corpora in these languages that support the development and evaluation of simplification techniques [15;16], with support for additional languages being encouraged through the creation of new corpora [17]. Despite the considerable body of work conducted on these alphabetic languages, the potential for knowledge transfer to less-studied languages remains underexplored. Given this backdrop, Swedish, a language sharing several linguistic characteristics with these commonly studied languages, presents an interesting case for examining the applicability of existing methods and models, especially in the context of the large language models (LLMs) that are currently accessible to students and teachers.

LLMs, like the one used in ChatGPT, have been trained on data encompassing a diverse range of languages and text genres; they offer new opportunities for advancing text simplification across languages, including Swedish. As a chatbot driven by one such LLM, ChatGPT can generate text based on prompts, potentially offering enhanced readability when guided by specific prompting approaches. However, the effectiveness of these prompting approaches in improving the readability of educational texts has not been rigorously examined, especially in a Swedish context. This gap is particularly relevant for the field of learning analytics given the increasing use of digital tools for creating and modifying educational content, where readability plays a crucial role in learners' comprehension, engagement, and performance across subjects.

To address this gap, this study explores whether and how GPT-4 can be leveraged to improve the readability of Swedish academic texts to support students' reading practices. Readability, is assessed using the *LIX* readability index (abbreviated from the Swedish term 'läsbarhetsindex') [18], a widely used metric that considers sentence, text, and word length as proxies for syntactic and lexical complexity [19]. The study also examines the relative effectiveness of selected prompting approaches–that have been shown to play an important role in the outputs generated by genAI-powered chatbots [20] –to improve text readability using ChatGPT (for more, see Background). The current study is guided by two key research questions:

RQ1: Which of the selected prompting approaches (i.e., meta, roleplay, chain-of-thought, and standard prompting) are the most effective in improving academic texts' readability?

RQ2: To what extent do the selected prompting approaches enhance text readability?

2 Background

2.1 Text readability

Reading is a complex, "long-term developmental process" [21] that relies on a variety of capacities, skills, and knowledge [22]. The simple view of reading (SVR; [23;24]) is the most widely used theory detailing reading processes because it provides a base explanation of the relationships between learner skills, knowledge, and capacities. SVR focuses on the role of word reading and a combination of elements that are referred to as oral language skills, with some of these impacted by the reader's working memory capacity [25]. Word reading is essentially a person's ability to decode or sound words out; learners who cannot decode a text will not be able to identify the words being used and, thus, will not be able to access information from that text. Oral language skills consist of knowledge of word meanings and one's ability to manipulate or reflect on spoken sentences (syntactic awareness) [25]. A learner with less syntactic awareness will struggle to work with the information in a complex sentence [26]. Similarly, a single unknown word can prevent a reader from understanding a sentence [27]. Relationships between these theoretical constructs and children's reading comprehension have been empirically established for alphabetic languages [28;29], such as Swedish.

As suggested by SVR, a text will be more readable if the vocabulary is known to the reader and the sentence structures are simpler, with many features of a text predictive of reading comprehension [30]. As shown by earlier research, students learn more effectively from reading materials that correspond to their reading level, balancing improvement and cognitive load (e.g., [31;32]). Many readability formulae have been developed independently but still embed some approximation of these principles. Most formulae incorporate a representation of sentence length or the number of words per sentence as a proxy for syntactic complexity. This can be seen in the Gunning-Fog Index (GFI; [33]), the Flesch family of indices [34], Dale-Chall [35], the simple measure of gobbledygook (SMOG; [36]), and the LIX readability formula [18].

Most formulae also incorporate a measure to account for the vocabulary requirements of the text. Some, such as Dale-Chall, rely on word lists that represent how common or complex a word is expected to be. Others, like SMOG, Flesh-Kincaid Grade Level, GFI, and LIX, use word length in characters or syllables to represent the complexity of a vocabulary item and through that the likelihood a reader will know the word. Alternative tools, such as Coh-Metrix [27;37] and AMOC [38], have more explicitly integrated underlying theories of reading or writing and tried to extract appropriate proxies for theoretically relevant constructs [39]. However, these more-theoretically grounded tools do not provide a simple measure that is easy for teachers and students to understand, nor do they directly support the automated simplification of text.

Like the above tools and formulae, language models (such as those used in ChatGPT) have been used to predict text difficulty [42] and human reading performance [43]. These predictions depend on the values of an analytic (i.e., surprisal) that can be extracted from language models. These surprisal values represent the likelihood of the next word in a sequence, which is a proxy for the familiarity of a word, providing a measure of whether the sentence structure is consistent with those a reader should expect based on their exposure to the language. The relationship between word predictability and reading performance or reading behavior has been established for a variety of languages that include English [44; 45], German [46], Arabic [47], Chinese [48], Spanish [49], and Korean [50]. Given the ability to predict reading performance using language model features and their ability to summarize or simplify text [51], they should be able to appropriately simplify texts to better support struggling readers. Considering a large decrease in the reading ability among school students in Sweden [11], fostering their agency by empowering them with available LLM-powered tools (e.g., ChatGPT or Google Gemini) that they could use to support their reading practices becomes critical.

2.2 Text simplification

Ideally, text simplification rewrites the provided text so that it has simpler grammatical structures and uses words that are simpler or more familiar to the reader [53;54;55] while maintaining the original meaning of the text [15]. It can also aim to reduce conceptual complexity and increase concreteness [15]. For highly nuanced texts, it may not be possible to maintain the exact meaning of the original text [57] because the substitution of near-synonyms changes the meaning of the text in subtle ways. However, this substitution is expected to be beneficial because it can make the text accessible much in the same way as the vocabulary glossary [57;58] has been shown to facilitate learner comprehension of expository texts [59;60]. In contrast to lexical simplification, structural simplification can maintain the meaning of the text at the risk of increasing its length.

As is typical in natural language processing and text-simplification work, much of the work has been evaluated using readability measures, expert assessment of produced texts, intrinsic measures of language model performance (e.g., perplexity), or model performance metrics on existing corpora (e.g., F1) [15]. Argumentation is then used to claim potential benefits to underserved populations such as those with dyslexia [61], learners with autism [62], and second-language learners [53]. In some cases, studies involving members of the targeted population have been conducted to develop appropriate datasets for evaluating these methods [62]. In rare cases, lexical simplification has been tested for its effect on reader errors, with fewer errors seen when people with dyslexia were reading the manually simplified text. From an ethical perspective, a reliance on intrinsic evaluation [63] makes sense as it enables a preliminary evaluation of new computational methods before subjecting learners to output that may not be helpful to them [64].

Recently, the ability of LLMs to simplify different kinds of texts has been explored in varied contexts and with mixed results. For example, a study examined the application of ChatGPT for the simplification of Dutch government letters, aiming to enhance their comprehensibility without comprising legal accuracy [65]. The results reveal that ChatGPT significantly improves the readability of the government letters with a 20% increase in comprehensibility scores. Scholars have also examined the use of ChatGPT and Google Bard to improve the readability of written patient information and found that ChatGPT could improve the readability of the texts but could not achieve the recommended 6th-grade reading level [66]. Bard reached that level but oversimplified texts by omitting up to 83% of the content. The examination of genAI-powered chatbots to simplify different kinds of texts has been largely conducted in the context of medical education (e.g., [67; 68]) as compared to other educational settings, for which there are few examples despite a considerable amount of text simplification work being motivated from the perspective of supporting developing readers. An exception is a recent study [69], in which GPT-3.5, LLaMA-2 70B, and Mixtral 8x7B were tested for their ability to generate educational materials in English (N = 100 texts) at various readability levels through zero-shot and few-shot prompting. The findings from this study revealed that few-shot prompting significantly improves performance in readability manipulation and information preservation; LLaMa-2 70B was found to perform better in achieving the desired difficulty range; and GPT 3.5 maintained original meaning. Other examples include a rule-based approach to simplifying Brazilian Portuguese [70] and the use of reinforcement learning to manipulate the grade level of English texts [71].

2.3 Prompt engineering

LLMs are trained by processing large amounts of data that contribute to the LLM's representation of information [72]. To optimize an LLM, there are several options. Two are directly related to model development and the third relates to how the model is used. When developing models, you can either train a model that has more parameters [72] which should support better performance assuming the model does not overfit to the data or provides a better representation of the target knowledge processing behavior, which is not always guaranteed [73]. The second option involves fine-tuning an existing model to solve a specific problem [74]. Both options require considerable computational resources [75]. The alternative to these resource-intensive options, prompt engineering, instead optimizes how the model is interacted with. The advantage of prompt engineering is that it does not require any additional resources since the performance of the existing model can be improved by how it

is prompted [75]. It has also been shown that prompt engineering with a general model, such as GPT-4, can perform better than a model trained for a specific purpose [76;77].

Since not all prompts result in the desired LLM behaviors, it is important to design a prompting strategy that aligns with the task. While prompts typically need to be tailored to a specific task, they can often be generalized within the context of a prompt framework. An example of the types of guidelines that exist in these frameworks is to ask the LLM to demonstrate its reasoning process or to break the problem into steps. Research has shown that LLMs like GPT-3.5 and Bard can improve the readability of patient information and radiology reports with simple prompts like "make this text more readable" [78]. Prompting can also be combined with other approaches. For example, studies (e.g., [79]) have simplified and clarified math tasks using few-shot learning (providing the LLM with several examples of correct outcomes) in combination with prompt strategies like Chain-of-Thought [80], where the LLM is guided on how to reason toward a solution. Additionally, newer prompt strategies like Meta-prompting and Roleplay-prompting have been used to create educational AI agents to support specific tasks [81]. However, these strategies have not yet been explored in the context of readability, which is of interest given LLMs' ease of use and effectiveness in other contexts.

We explore the use of simple prompting approaches since their simplicity is likely to enable their later independent use by students. Standard prompt is one such approach where the LLM is prompted with a concrete task without further inputs. We also considered the three prompting strategies that are presented below.

Chain-of-Thought (CoT) is one of the earliest well-researched prompt strategies [80]. The concept behind CoT is based on what is known as few-shot learning [72], providing the LLM with examples of correct answers for questions similar to the one it is then asked to solve. CoT develops this concept and provides suggestions on how to reason to solve the problem. In the absence of CoT prompting, the LLM often provides output that does not match what the user expects, and the user has no guidance on how to get the LLM to fix the problem. By showing the reasoning steps, CoT prompting allows users to see how the LLM has solved a task [80], which can allow users to intervene when the LLM has made an error. CoT prompting has also been seen to increase output accuracy [e.g., 80].

Roleplay prompting is a strategy in which the LLM is asked to roleplay as a character [19]. A study that has further developed the strategy to include two prompts has shown that this strategy improves the LLM's reasoning [95].

Meta-prompting is a strategy where the user gets the LLM to generate its own prompts for a specific task [82]. There are several types of meta-prompt. A serializing meta-prompt asks the LLM to solve a problem by breaking it down into steps [83]. Another way is called a "fill-in-the-blank" meta-prompt where the user leaves gaps in their prompt for the LLM to fill in. An example of such a prompt is: "To make this text easier to read, we analyze the text and search for "–", which makes this text difficult to read." Both are examples of letting the model 'think' (i.e., make statistical inferences) to get better output. In this study, only the serializing meta-prompt is used and referred to as Meta.

3 Method

The current study explores the effect of the prompt approach on the readability of text that was simplified using GPT-4, which has been shown to outperform other models in several studies (e.g., [84;85]). We followed the guidelines of the Swedish Ethical Review Authority and GDPR (https://gdpr-info.eu/). All materials, scripts, and data can be accessed via OSF[1].

3.1 Data

Reading comprehension texts were sourced from previous university entrance exams (SWESAT), which were publicly accessible as of March 2024 in the University and Swedish Council for Higher Education archive (https://www.studera.nu/hogskoleprov/forbered/tidigare-hogskoleprov/). The difficulty level of these texts aligns with the reading and comprehension skills expected of university students. A total of 136 texts were available at the time of the study, and all were included in the analysis. These texts varied in length, vocabulary, and readability (see Table 1).

3.2 Refining prompting approaches

ChatGPT-4 was used to refine the four prompting approaches (i.e., Standard, CoT, Meta, and Roleplay). For each approach, the question and task were included in the same prompt, with a description of the LIX and a request for the information from the original text to be preserved. Taking Roleplay-prompting as an example, we start by defining ChatGPT's role and then assign the task: "You are now a Swedish teacher who is extremely good at rewriting texts to make them more readable. You have now been given the task of reworking a text and lowering its LIX as much as you can. Here is the text".

Each prompting approach was refined to make sure the produced text was of reasonable quality and that the output was formatted in a way that supported analysis. The refinement process involved trial and error where the goal was to write a prompt that could make ChatGPT achieve a sufficient readability level (LIX < 40). The template for these prompts can be seen in Table 2. For all approaches, except the CoT strategy, prompt refinement was a straightforward process where we focused on writing a prompt that produced reasonable text. The Standard, Meta, and Roleplay prompts succeeded in achieving the required LIX-score almost immediately and no further refinement was needed. Additional refinement was needed for CoT because there was no clear way to reason how a text should be processed to lower the LIX. An example of the required adjustment was included in the CoT prompt. A subset of the chat logs are available via OSF1.

3.3 Study procedure: Obtaining the simplified texts & calculating readability

The GPT-4 API was used to create the simplified versions of each input text. Specifically, the GPT-4 Turbo model, gpt-4-0125-preview, was used because it was the latest and least expensive version of

[1]https://osf.io/n4u9k/?view_only$=$b280a45fcb604ebcbef3f14df7c34687 https://osf.io/trfu3/?view_only$=$dae38204389b4f009429610219087a09

Table 1: Characteristics of the original text

	Min	Mdn	Max	Mean	SD
No. Sentences	14	42.5	88	40.9	18.10
No. Words	199	813.0	1193	718.4	309.98
No. Long Words	31	249.0	394	226.3	109.58
Readability (LIX Score)	24.05	50.03	65.29	48.8	7.57

Table 2: Prompt design for each prompting strategy

Prompt Approach	Prompt Design
Standard	You are now going to help me rewrite a text that is difficult to read into a text that is easier to read. [This is defined according to the readability index, LIX. LIX = NUMBER OF WORDS/NUMBER OF SENTENCES + NUMBER OF WORDS LONGER THAN 6 LETTERS * 100/NUMBER OF WORDS. I want to get a text that has a maximum LIX value of 40][b] The information from the original text should be preserved as far as possible in the rewritten, easy-to-read text. You will be provided with a text to rewrite. You do not have to answer anything, just rewrite the text. [TEXT]
Roleplay	You are a very good writer. You work at the Swedish Agency for Accessible Media (MTM). You are extremely good at rewriting texts that are difficult to read into texts that are easier to read, this is your specialty! You always manage to rewrite texts with high LIX values into texts with LIX values below 40. [Explanation of LIX You have been given the task by the Swedish Council for Higher Education (UHR) to rewrite texts from the university entrance examination to make them easier to read. The purpose of your task is to provide those with reading difficulties with easier texts that they can read. The information from the original text should be preserved, to the extent it is possible, in the rewritten, easier-to-read text. Your goal is to reduce the LIX value of the text to below 40. [TEXT]
Meta	[Standard prompt.] The problem with this text is that it has too high a LIX: [TEXT] Rework the text so that it has a lower LIX. Solve the problem by breaking it down into steps.
Chain-of-Thought (CoT)	[Standard prompt]. To do this, we replace long words, words over 6 letters long with shorter synonyms and possibly replace a long word with several short ones. We also try to break up long sentences into several short ones. Example: "It has parallel elements of progressive opportunity and reactionary trap." We see that the sentence contains many words that have more than 6 letters in them. If we paraphrase the sentence and replace these long words with shorter synonyms, this will lower the LIX. We can write "It has both elements of being progressive and of not being progressive." So this is the right way to do it. [TEXT]

[a] The original prompts were crafted in Swedish. We have translated them into English. [b] Bracketed content is the base of the standard prompts that is used in all other strategies.

GPT at the time. For each input text, four simplified texts were created, each using a different prompting strategy: Standard prompting, CoT prompting, Meta-prompting, and Roleplay prompting. This process resulted in 544 simplified texts.

The script that connects to the API was designed to generate output that mimics ChatGPT and calculates the LIX readability score [19] for each of the produced texts and each of the input texts. We used the standard LIX score [19] because it is an established measure for assessing the level of readability of a text and has been extensively trialed in different contexts [19]. This includes recent studies that assessed the readability of texts generated by genAI-powered chatbots (e.g., [40;41]).

LIX was developed in Sweden, and it accounts for lexical and syntactic factors [19], as noted in the Background section. The lowest possible LIX score is 1 and it occurs when you have a sentence containing a single word that is less than 6 letters long.

3.4 Analysis

Since language data, especially word counts [86], have a long-tailed distribution, we report minima (Min), maxima (Max), median (Mdn), mean (M), and standard deviation (SD). These descriptive statistics are used to detail the readability of the original and GPT-generated texts. These statistics are also used to describe the change in readability from the original to the generated texts.

Table 3: LIX readability scores for all texts

Text	LIX score				
	Min	***Mdn***	***Max***	***M***	***SD***
Original	24.1	50.0	65.3	48.8	7.57
Standard	17.5	36.0	67.6	36.0	7.37
Roleplay	17.0	36.1	53.5	36.4	7.20
Meta	16.8	32.1	45.9	32.3	5.55
CoT	14.2	33.0	58.3	33.5	6.89

This change in readability is represented through an adaptation of the normalized gain score [94]. This adaptation accommodates the maximum observable LIX score, based on our data. LIX scores are governed by the length of the longest sentence, making them theoretically infinite but bounded in practice. We draw inspiration from the common practice of using min-max scaling and instead use 68 (the ceiling of the maximum observed LIX score from our data). This is equivalent to the use of the common ceiling for test scores (100) that is employed in the original formula.

$$normalized\ gain\ =\ (Post\ -\ Pre)\,/(68\ -\ Pre)$$

Statistical testing was used to determi if there were differences in text readability or characteristics across sources. When the normality assumption was met, we used a one-way repeated measures ANOVA with post-hoc paired t-tests. When using an ANOVA, we also checked the sphericity assumption. When sphericity was violated, we reported the Greenhouse-Geisser corrected values. When the normality assumption was not met, we used Friedman and Wilcoxon signed-rank tests. We applied Bonferroni correction to all pairwise comparisons, and the Kolmogorov-Smirnov test was used to determine whether data met the normality assumption because we had more than 50 texts.

Our analyses go beyond simple comparisons of the prompting strategies using the readability score to investigations of differences in the characteristics of the generated texts and how the characteristics of the input text relate to those of the generated text. This includes the use of Pearson's correlation to characterize how the input text might constrain the characteristics (i.e., number of words, number of long words, and number of sentences) of the generated text for each approach.

4 Results

4.1 What differences exist in the readability of the texts following simplification?

In most cases, the LIX scores of the simplified texts (Table 3) had a narrower range than that of the original texts. The one exception to this was the texts produced using the Standard prompting strategy. The measures of central tendency all suggest that the produced texts had greater readability (i.e., lower LIX scores) than the original texts on average, and Meta was the only prompting strategy where all texts were simplified to the point where the LIX score was below the median LIX score from the original texts.

Since the data met the normality assumption, we conducted a one-way repeated measures ANOVA. This test revealed a large difference in LIX scores across prompting conditions: $F(4, 540) = 280.367$, $p < .001$, $\eta^2 = .675$. Post-hoc paired t-tests showed that all prompting strategies improved the readability of the text ($p < .001$). Differences were not found between the LIX scores achieved through Standard-prompting and those obtained using the Roleplay strategy ($p = 1.0$). Also, no difference in LIX score was found between the Meta and CoT strategies ($p = .245$). Both Standard and Roleplay were outperformed by Meta and CoT ($p < .001$ in all cases).

This difference in performance can also be seen through the number of texts that fall within different readability categories (Table 4). Meta-prompting simplified most texts to an easy reading level, which is defined as being easier to read than most newspapers. Meta simplified 8.82% more of the texts to this level than the next closest prompting strategy (CoT). The results for improving readability to the level typical in newspapers was similar (LIX < 50). This time the difference between Meta-prompting and the next best performing strategy (CoT) was smaller (5.14%).

4.2 How much did the readability of texts change following simplification?

A Friedman test found differences in how much the LIX score changed following text simplification: $X^2(3) = 57.874$, $p < .001$, $W = .142$. Pairwise comparisons of the observed gain (Table 5) between prompting approaches found no differences between the those that had the highest average gains ($P = .280$, i.e., Standard and Roleplay) and those that had lower average gains ($p = .107$, i.e., Meta and CoT). They also identified that Roleplay and Standard had slightly larger average gains than Meta and CoT ($P < .001$ in all cases). It is worth noting that Standard prompting was the only approach where the LIX scores were seen to increase. This is also the approach where the most variability was seen in the gain scores.

4.3 How do the generated text characteristics differ across prompting approaches?

Since LIX scores are determined using specific characteristics of a text and much of the work on text simplification has aimed to perform one of lexical or syntactic simplification, we also looked for changes in which of the text features differed across prompting approaches. The descriptive statistics for the ChatGPT-generated text features can be seen in Table 6.

As expected, we found large differences in the number of sentences: $F(1.534, 207.036) = 204.820$, $p < .001$, $\eta^2 = .603$. Comparing the original text length to that of the simplified text shows a reduction of around 50% across all prompting approaches ($p < .001$). The

Table 4: The number (%) of texts with improved readability for each prompting strategy

	Standard	Roleplay	Meta	CoT
Easy (LIX < 40)	101 (74.26%)	100 (73.53%)	125 (91.91%)	113 (83.09%)
Medium (LIX < 50)	113 (83.09%)	114 (83.82%)	131 (96.32%)	124 (91.18%)

Table 5: The amount of improvement observed by prompting approach

Approach	Normalized LIX Gain				
	Min	*Mdn*	*Max*	*M*	*SD*
Standard	-70.2	34.4	49.0	34.3	10.98
Roleplay	16.5	34.6	49.7	34.8	6.62
Meta	16.2	30.5	43.7	30.9	5.20
CoT	13.6	31.5	52.1	32.0	6.33

Table 6: Characteristics of the generated text by prompting approach

Approach		Statistics				
		Min	*Mdn*	*Max*	*M*	*SD*
Sentence (No.)						
	Standard	7	15.0	37	15.7	4.83
	Roleplay	8	16.0	35	16.5	4.73
	Meta	6	17.0	43	17.7	5.52
	CoT	9	21.0	47	21.7	6.92
Words (No.)						
	Standard	112	237.0	472	237.1	59.34
	Roleplay	88	265.0	462	264.6	64.86
	Meta	78	225.5	508	227.4	67.38
	CoT	123	273.5	534	278.5	68.14
Long Words (No.)						
	Standard	12	46.5	94	47.6	16.18
	Roleplay	8	51.0	114	52.4	18.72
	Meta	14	40.0	112	43.0	16.08
	CoT	14	52.0	114	55.4	19.74

Standard prompting approach produced the shortest texts, outperforming Roleplay (p = .007), Meta (p < .001), and CoT (p < .001). The Meta and Roleplay strategies did not produce texts with a measurably different length (p = .248), and CoT produced the longest of the simplified texts (p < .001).

Consistent with the sentence length analysis, the Friedman test found differences in the number of words across texts, $X^2(4)$ = 313.598, p < .001, W = .576. Once again, all prompting approaches produced texts with fewer words than the original text (p < .001). These texts were approximately 25% of the length (in number of words) of the original text; this is around a 75% reduction. The post-hoc comparisons did not identify differences in the number of words produced between Meta and Standard (p= .510) or CoT and Roleplay (p= .370), with Meta and Standard producing texts that were around 5 words shorter on average than those produced by the CoT and Roleplay strategies (p < .001).

The last variable used is the number of long words, which differed across texts, $X^2(4)$= 322.166, p< .001, W = .592. Once again, all prompting strategies produced texts with fewer long words than the original text (p < .001). The ratio of long words eliminated was consistent with the number of words by which the text was reduced (~80-75%). The Meta strategy produced texts with the fewest long words (p < .001). No differences were found between the CoT and Roleplay strategies (p = 1.0), both of which included more long words than the Standard and Meta strategies (p < .001).

4.4 What is the role of the original text's features in the readability of the produced texts?

Not surprisingly, the LIX readability of the original text seems to limit the LIX score of the simplified text, as indicated by the moderate to strong relationship between the original texts' LIX score and that for the texts produced by each of the Standard (r = .48, p < .001), Roleplay (r = .61, p < .001), Meta (r = .61, P < .001), and CoT (r = .64, p < .001) approaches. The moderately strong

Table 7: Correlations between original LIX, input variables and LIX of the GPT-generated text (* $p < .001$)

	Standard LIX	Roleplay LIX	Meta LIX	CoT LIX	LIX
Number of words	.23*	.28*	.12	.30*	.28*
Long words (> 6 characters)	.35*	.43*	.30*	.47*	.54*
Sentences	.11	.13	−.02	.14	-.01

relationship between the number of long words in the original text and the LIX readability score of the GPT-simplified text (Table 7) suggests that the presence of more long words is driving this relationship given that the number of words has a weak relationship with the readability of the simplified text in most cases and the number of sentences do not have a relationship with the LIX score. Together, the number of words and sentences are used as the proxy for syntactic complexity within LIX. From this analysis, it would appear this proxy is driven by the text length in number of words.

5 Discussion

The current study aimed to improve our understanding of whether we can use ChatGPT to increase the readability of educational texts, and which prompting approach/es perform better than others on this task. Improving text readability is important since students' reading skills have been gradually declining across countries [11]. This is problematic because students who struggle to read can be at a disadvantage in achieving learning goals, given how text complexity increases as students' progress through the education system [6]. Different support techniques have been examined to assist students in addressing this challenge (e.g., [6-8]), with varying results. Even when these techniques are successful, gaining access to them can be a challenge [87]. GenAI tools trained on LLMs such as ChatGPT, could help to create a more even playing field by simplifying complex texts (i.e., text simplification) for students with reading challenges. Such simplification could support their reading practices in schools and beyond. However, little is known about how well ChatGPT can improve the readability of texts, much less for lower-resource languages, like Swedish [96]. Furthermore, the type of prompting approaches used to request the simplified text could have an impact on the output of ChatGPT [20]. So, we examined which of the selected prompting strategies (i.e., Meta, Roleplay, and CoT) along with a standard prompting approach is the most effective for improving academic texts' readability (RQ1) and the extent to which the selected prompting strategies enhanced text readability (RQ2).

5.1 Meta- and CoT prompting improved text readability

The results show that while all three prompt strategies and the Standard prompting approach improved the readability of texts as measured by LIX, the Meta and CoT prompting strategies outperform Standard and Roleplay. Both Meta and CoT improved text readability to the same level as that of newspapers for over 90% of the texts (LIX < 40), and using the Meta prompting strategy made more than 90% of the texts readable at a middle-school level (LIX < 50). This marked improvement is likely why the texts produced through Meta-prompting had the weakest relationship with the characteristics of the original text. That is, Meta-prompting simplified more of the texts for the targeted measure and was less constrained by the linguistic characteristics of the original text.

When comparing the characteristics of simplified texts across the prompt strategies, results show that Meta-prompting produced texts with the fewest long words but did not reduce the text length substantially more than the other strategies. Therefore, Meta-prompting appears to improve readability by replacing long words and could be best suited to helping readers who struggle with advanced vocabulary. This characteristic of the output could be a result of the information contained in the prompt. The LLM was given the LIX formula and instructed to solve the problem of texts with high LIX scores by breaking this analytic down into steps targeting different aspects of the text.

This finding is particularly relevant for educators seeking to support students with varying levels of reading proficiency, as it underscores the need for tailored strategies that can effectively bridge the gaps among complex academic language, processing speed, and student comprehension [88]. To empower educators with this knowledge and relevant prompting skills, there is a need to provide professional-development programs at a higher level (e.g., the state or the municipality level) or at the school level, where such training could be initiated by the school principal or the group responsible for assisting students with differentiated support needs. Furthermore, this result could inform the practices of students, who could use this prompting strategy to take responsibility for improving their own reading practices. That is, with the information about how to effectively prompt ChatGPT or similar LLM-based tools, they should be able to adapt educational texts to their needs and abilities. In the setting of higher education, student support services groups could provide prompting templates and training to students so that they can learn this skill and obtain support when they need it both within their programs and afterwards.

5.2 Risks and benefits of Standard Prompting

Instead of reducing the LIX, Standard prompting resulted in increased scores for some texts, highlighting that caution should be taken when using standard prompting to improve the general readability of texts. Results from the comparison of text characteristics show that the standard prompt produced the shortest texts, with texts created through Meta-prompting being only slightly longer. If the primary concern is processing or reading time, the findings suggest that the Standard approach to prompting could be an effective strategy for shortening texts, thus saving the student time while giving them access to the main information. In contrast, if those needing support are concerned with both the amount of text

and the complexity of the vocabulary, Meta-prompting is likely a better choice because it balances these two key text characteristics.

5.3 Implications for analytics & LLM use

As the use of analytics and AI-based tools become more prevalent in education, it is essential to address the ethical implications of their use. Ensuring that all students have access and that these tools are used responsibly is vital to preventing widening educational disparities. Moreover, questioning the underlying analytics and algorithms employed by these tools is becoming increasingly important as their inclusion in large-scale technologies, like ChatGPT, has the potential to cause considerable harm and fossilize practices that are not grounded in current knowledge of how people learn.

To this end, we consider both the readability analytic (LIX), and how it may have been used to simplify the text. Our analyses indicate potential limitations of the LIX score as a measure of readability. The lack of relationship between the number of sentences in the original text and its LIX score, which should be correlated by definition, suggests that this aspect of the measure provides little information. This pragmatic but limited representation of syntactic complexity is common across many readability measures and may have been a reasonable choice when these measures were first defined. However, we now have better proxies for syntactic complexity (e.g., parse tree depth [30]) that can be easily obtained using simple techniques from natural language processing (e.g., grammars and parsing) that are embedded within ChatGPT or available through common libraries. Given the availability of such tools and the fact that readability measures do not capture other important aspects of the text (e.g., decodability; [88]), we must revisit the formulation of these analytics to ensure their appropriateness for assessing who may or may not be able to access information from a text.

While these reformulations may be supported by LLMs, there are other aspects of their use that may pose problems. As widely acknowledged [89], LLMs do not excel at mathematics nor are they meant to. The two worst-performing prompting approaches (Standard and Roleplay) gave the LLM information about the formula to be used for improving readability but no additional information about how to change the linguistic characteristics of the text to meet the objective of reducing the value of that formula. In contrast, the two best-performing prompting approaches (Meta and Chain-of-Thought) provided additional information about how to approach reducing the LIX score. For Meta, this was guidance on addressing one aspect of the formula at a time. For CoT, this was explicit guidance on reducing words that exceeded a character count. Such additional guidance seems to have helped the LLM improve text readability, suggesting that providing procedural or complementary information supports LLMs in optimizing text based on a particular metric. Considering how rapidly AI and LLMs are evolving, GPT and other models should be continually tested and compared going forward to determine generalizability and adapt practices as LLM capabilities change.

6 Conclusion

This study underscores the potential of genAI tools, like ChatGPT, to enhance the readability of academic texts and address a critical need in supporting students with reading difficulties. Given the increasing demands for higher reading standards in education and the challenges faced by students with learning disabilities, effective text simplification becomes essential. We have tested four different prompting approaches, which all fall into the category of single prompt strategies, on a large academic dataset. In the future, other types of prompt-engineering techniques should be explored. They include multiple prompt techniques, such as Prompt Chaining, Least-to-Most, Self-Consistency, and Tree of Thoughts, and using LLMs with external tools such as RAG, ReAct, and Reflexion. More important is the need to explore how these approaches could be integrated into existing pedagogical frameworks, and how teachers, students, and parents can be trained and supported to implement them effectively.

Our findings suggest that specific prompting strategies, particularly Meta-prompting, significantly improve text readability, enabling better access to complex subject matter. However, there remains a need for further research to explore the effectiveness of these tools across different languages and educational contexts. Additionally, while the potential of large language models is promising, it is crucial to support educators in employing these tools effectively and help identify specific methodologies that can be adopted to support diverse learners [90]. In sum, by leveraging advanced AI technologies, there is a potential to create more inclusive learning environments that could empower teachers in designing learning materials that address the individual needs of students and students in taking charge of their reading practices, ultimately leading to improved educational outcomes. Future studies should consider the long-term effects of using such tools on students' reading skills, including their comprehension abilities, ultimately providing a coherent understanding of their impact in educational settings. This is important for the provision of an inclusive approach to education, which "acknowledges that all children can learn, and every child has unique characteristics, interests, abilities and learning needs" [91]. As recently stressed by scholars, educational institutions across the world aim for a quality education and lifelong learning opportunities that are inclusive and equitable for all learners [92].

This study contributes both to the growing body of literature on text simplification and interest of the learning analytics community in understanding and enabling the practical application of LLMs in improving technology-driven instructional support. The findings offer insight into best practices for using AI-powered tools in educational contexts, particularly for teachers and learners who work in under-resourced languages.

Acknowledgments

We would like to acknowledge Digital Futures (Stockholm, Sweden) for funding a research stay of one of the study's authors under the "Scholar-in-Residence" program during Fall 2024.

References

[1] Delgadova, E. (2015). Reading literacy as one of the most significant academic competencies for the university students. *Procedia-Social and Behavioral Sciences, 178*, 48-53.
[2] Knospe, Y., Sturk, E., & Gheitasi, P. (2021). Additional support for pupils with reading difficulties – a case study. Education Inquiry, 14(1), 145–161.
[3] National Center for Educational Statistics (2024). Students with disabilities. Retrieved August 22, 2024 from https://nces.ed.gov/programs/coe/pdf/2024/CGG_

508c.pdf
[4] Taneja-Johansson, S. (2021). Facilitators and barriers along pathways to higher education in Sweden: a disability lens. *International Journal of Inclusive Education*, *28*(3), 311–325.
[5] Carless, D., & Boud, D. (2018). The development of student feedback literacy: enabling uptake of feedback. *Assessment & Evaluation in Higher Education*, *43*(8), 1315-1325.
[6] Pirttimaa, R., Takala, M., & Ladonlahti, T. (2015). Students in higher education with reading and writing difficulties. *Education Inquiry*, *6*(1), 24277.
[7] Wolf, M. (2010). The RAVE-O reading intervention program.
[8] Wilson, B. (1988). Wilson Reading System. Millbury, MA: Wilson Language Training.
[9] Lovett, M. W., Lacerenza, L., & Borden, S. L. (2000). Putting struggling readers on the PHAST track: A program to integrate phonological and strategy-based remedial reading instruction and maximize outcomes. *Journal of learning disabilities*, *33*(5), 458-476.
[10] Skolverket (2024). Garantin för tidiga stödinsatser i anpassade grundskolan. Retrieved August 22, 2024 from https://www.skolverket.se/skolutveckling/leda-och-organisera-skolan/ge-extra-stod-till-elever/garantin-for-tidiga-stodinsatser-i-anpassade-grundskolan
[11] OECD (2023). PISA 2022 Results (Volume I and II) - Country Notes: Sweden. Retrieved August 22, 2024 from https://www.oecd.org/en/publications/pisa-2022-results-volume-i-and-ii-country-notes_ed6fbcc5-en/sweden_de351d24-en.html
[12] Bernholt, S., Härtig, H., & Retelsdorf, J. (2023). Reproduction rather than comprehension? Analysis of gains in students' science text comprehension. *Research in Science Education*, *53*(3), 493-506.
[13] Savolainen, H., Ahonen, T., Aro, M., Tolvanen, A., & Holopainen, L. (2008). Reading comprehension, word reading and spelling as predictors of school achievement and choice of secondary education. *Learning and Instruction*, *18*(2), 201-210. https://doi.org/10.1016/j.learninstruc.2007.09.017
[14] Bonifacci, P., Colombini, E., Marzocchi, M., Tobia, V., & Desideri, L. (2022). Text-to-speech applications to reduce mind wandering in students with dyslexia. *Journal of Computer Assisted Learning*, *38*(2), 440-454.
[15] Al-Thanyyan, S. S., & Azmi, A. M. (2022). Automated Text Simplification: A Survey. ACM Computing Surveys, 54(2), 1–36.
[16] Anschütz, M., Oehms, J., Wimmer, T., Jezierski, B., & Groh, G. (2023). Language models for German text simplification: Overcoming parallel data scarcity through style-specific pre-training. *arXiv preprint arXiv:2305.12908*.
[17] Alhafni, B., Hazim, R., Liberato, J. P., Khalil, M. A., & Habash, N. (2024). The SAMER Arabic Text Simplification Corpus. *preprint arXiv:2404.18615*.
[18] Björnsson, C. H. (1968). Läsbarhet, Stockholm: Liber.
[19] Anderson, J. (1983). LIX and RIX: Variations on a little-known readability index. *Journal of Reading*, *26*(6), 490-496.
[20] Knoth, N., Tolzin, A., Janson, A., & Leimeister, J. M. (2024). AI literacy and its implications for prompt engineering strategies. *Computers and Education: Artificial Intelligence*, 6, 100225.
[21] Snow, C. (2002). Reading for understanding: Toward an R&D program in reading comprehension. Rand Corporation.
[22] Woolley, G., & Woolley, G. (2011). Reading comprehension (pp. 15-34). Springer Netherlands.
[23] Gough, P.B. and Tunmer, W.E. 1986. Decoding, Reading, and Reading Disability. *Remedial and Special Education*. 7, 1 (Jan. 1986), 6–10.
[24] Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. Reading and Writing: An Interdisciplinary Journal, 2(2), 127–160.
[25] Cain, K. (2007). Syntactic awareness and reading ability: Is there any evidence for a special relationship? *Applied Psycholinguistics*, 28(4), 679–694.
[26] Graesser, A. C. (2006). Question Understanding Aid (QUAID): A Web Facility that Tests Question Comprehensibility. *Public Opinion Quarterly*, 70(1), 3–22.
[27] Graesser, A. C., McNamara, D. S., & Louwerse, M. M. (2003). What do readers need to learn in order to process coherence relations in narrative and expository text. In *Rethinking reading comprehension* (pp. 82-98). Guilford Press.
[28] Kirby, J. R., & Savage, R. S. (2008). Can the simple view deal with the complexities of reading? *Literacy*, *42*(2), 75-82.
[29] Verhoeven, L., & Van Leeuwe, J. (2008). Prediction of the development of reading comprehension: A longitudinal study. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *22*(3), 407-423.
[30] Nie, Y., Deacon, H., Fyshe, A., & Epp, C. D. (2022). Predicting Reading Comprehension Scores of Elementary School Students. In *Proceedings of the 15th International Conference on Educational Data Mining* (p. 158).
[31] Namaziandost, E., Esfahani, F. R., & Ahmadi, S. (2019). Varying levels of difficulty in L2 reading materials in the EFL classroom: Impact on comprehension and motivation. *Cogent Education*, *6*(1), 1615740.
[32] Brüggemann, T., Ludewig, U., Lorenz, R., & McElvany, N. (2023). Effects of mode and medium in reading comprehension tests on cognitive load. *Computers & Education*, *192*, 104649.
[33] Gunning, R. (1952). The technique of clear writing.
[34] Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., & Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and Flesch reading ease formula) for navy enlisted personnel.
[35] Dale, D. (1948). The Dale-Chall formula for predicting readability. *Educational Research Bulletin*, *27*, 11-20.
[36] McLaughlin, G. H. (1969). SMOG grading - a new readability formula. Journal of Reading, 12(8):639–646
[37] Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher*, *40*(5), 223-234.
[38] Corlatescu, D. G., Dascalu, M., & McNamara, D. S. (2021). Automated model of comprehension V2. 0. In *International Conference on Artificial Intelligence in Education* (pp. 119-123). Cham: Springer International Publishing.
[39] Crossley, S. A., Skalicky, S., & Dascalu, M. (2019). Moving beyond classic readability formulas: New methods and new models. *Journal of Research in Reading*, *42*(3-4), 541-561.
[40] Deveci, C. D., Baker, J. J., Sikander, B., & Rosenberg, J. (2023). A comparison of cover letters written by ChatGPT-4 or humans. *Danish Medical Journal*, 70(12), A06230412.
[41] Skrzypczak, T., Skrzypczak, A. & Szepietowski, J.C. Readability of Patient Electronic Materials for Atopic Dermatitis in 23 Languages: Analysis and Implications for Dermatologists. Dermatol Ther (Heidelb) 14, 671–684 (2024). https://doi.org/10.1007/s13555-024-01115-1
[42] Olney, A. M. (2022, July). Assessing Readability by Filling Cloze Items with Transformers. In *International Conference on Artificial Intelligence in Education* (pp. 307-318). Cham: Springer International Publishing.
[43] Hofmann, M. J., Remus, S., Biemann, C., Radach, R., & Kuchinke, L. (2022). Language models explain word reading times better than empirical predictability. *Frontiers in Artificial Intelligence*, *4*, 730570.
[44] Goodkind, A., & Bicknell, K. (2018, January). Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th workshop on cognitive modeling and computational linguistics (CMCL 2018)*, 10-18.
[45] Hale, J., Dyer, C., Kuncoro, A., & Brennan, J. R. (2018). Finding syntax in human encephalography with beam search. arXiv preprint arXiv:1806.04127.
[46] Kliegl, R., Grabner, E., Rolfs, M., & Engbert, R. (2004). Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology*, *16*(1-2), 262-284.
[47] AlJassmi, M. A., Warrington, K. L., McGowan, V. A., White, S. J., & Paterson, K. B. (2022). Effects of word predictability on eye movements during Arabic reading. *Attention, Perception, & Psychophysics*, 1-15.
[48] Rayner, K., Li, X., Juhasz, B. J., & Yan, G. (2005). The effect of word predictability on the eye movements of Chinese readers. *Psychonomic Bulletin & Review*, *12*, 1089-1093.
[49] Fernández, G., Shalom, D. E., Kliegl, R., & Sigman, M. (2014). Eye movements during reading proverbs and regular sentences: The incoming word predictability effect. *Language, Cognition and Neuroscience*, *29*(3), 260-273.
[50] Yun, H., Lee, D., Hong, U., & Nam, Y. (2017). The predictability effect on eye movement in reading Korean dative sentences. *Language and Information*, 21(1), 73-99.
[51] Kew, T., Chi, A., Vásquez-Rodríguez, L., Agrawal, S., Aumiller, D., Alva-Manchego, F., & Shardlow, M. (2023). BLESS: Benchmarking large language models on sentence simplification. *arXiv preprint arXiv:2310.15773*.
[52] OECD. (2022). PISA 2022 Results (Volume I and II) - Country Notes: Sweden. Retrieved from https://www.oecd.org/en/publications/pisa-2022-results-volume-i-and-ii-country-notes_ed6fbcc5-en/sweden_de351d24-en.html
[53] Paetzold, G. H., & Specia, L. (2017). A survey on lexical simplification. *Journal of Artificial Intelligence Research*, *60*, 549-593.
[54] Qiang, J., Li, Y., Zhu, Y., Yuan, Y., & Wu, X. (2020). Lexical Simplification with Pretrained Encoders. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05), 8649–8656.
[55] Zhou, W., Ge, T., Xu, K., Wei, F., & Zhou, M. (2019). BERT-based Lexical Substitution. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3368–3373.
[56] Agrawal, S., & Carpuat, M. (2024). Do Text Simplification Systems Preserve Meaning? A Human Evaluation via Reading Comprehension. *Transactions of the Association for Computational Linguistics*, *12*, 432-448.
[57] Ko, M. H. (2012). Glossing and Second Language Vocabulary Learning. TESOL Quarterly, 46(1), 56–79.
[58] Wang, S., & Lee, C. I. (2021). Multimedia Gloss Presentation: Learners' Preference and the Effects on EFL Vocabulary Learning and Reading Comprehension. *Frontiers in Psychology*, 11, 602520.
[59] Lofgren, A. (2022). Unraveling Contradictions: Which Glosses Facilitate Reading Comprehension Among ELLs, and Why? *Journal of Language Teaching and Research*, 13(1), 1–11.
[60] Varol, B., & Erçetin, G. (2021). Effects of gloss type, gloss position, and working memory capacity on second language comprehension in electronic reading. *Computer Assisted Language Learning*, 34(7), 820–844.
[61] Rello, L., Saggion, H., Baeza-Yates, R., & Graells, E. (2012, June). Graphical schemes may improve readability but not understandability for people with dyslexia.

Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations. 25-32.

[62] Evans, R., Orasan, C., & Dornescu, I. (2014). An evaluation of syntactic simplification rules for people with autism. Association for Computational Linguistics.

[63] Gala, N., Tack, A., Javourey-Drevet, L., François, T., & Ziegler, J. C. (2020, May). Alector: A parallel corpus of simplified French texts with alignments of misreadings by poor and dyslexic readers. In *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 1353-1361).

[64] Demmans Epp, C., & Makos, A. (2013). Using simulated learners and simulated learning environments within a special education context. *Workshop on Simulated Learners at Artificial Intelligence in Education (AIED), 4*, 1–10.

[65] van Raaij, N. B., Kolkman, D., & Podoynitsyna, K. (2024, May). Clearer Governmental Communication: Text Simplification with ChatGPT Evaluated by Quantitative and Qualitative Research. In *Proc. of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context@ LREC-COLING 2024* (pp. 152-178).

[66] Moons, P., & Van Bulck, L. (2024). Using ChatGPT and Google Bard to improve the readability of written patient information: a proof of concept. *European Journal of Cardiovascular Nursing*, *23*(2), 122-126.

[67] Ayre, J., Mac, O., McCaffery, K. *et al.* New Frontiers in Health Literacy: Using ChatGPT to Simplify Health Information for People in the Community. *J GEN INTERN MED* 39, 573–577 (2024).

[68] Gill, B., Bonamer, J., Kuechly, H., Gupta, R., Emmert, S., Kurkowski, S., ... & Grawe, B. (2024). ChatGPT is a promising tool to increase readability of orthopedic research consents. *Journal of Orthopaedics, Trauma and Rehabilitation*, 22104917231208212.

[69] Huang, C. Y., Wei, J., & Huang, T. H. K. (2024). Generating Educational Materials with Different Levels of Readability using LLMs. *arXiv preprint arXiv:2406.12787*.

[70] Candido Jr, A., Maziero, E. G., Specia, L., Gasperin, C., Pardo, T., & Aluisio, S. (2009, June). Supporting the adaptation of texts for poor literacy readers: a text simplification editor for Brazilian Portuguese. In *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 34-42).

[71] Yanamoto, D., Ikawa, T., Kajiwara, T., Ninomiya, T., Uchida, S., & Arase, Y. (2022, November). Controllable text simplification with deep reinforcement learning. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)* (pp. 398-404).

[72] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P.,... & Amodei, D. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

[73] Oh, B. D., & Schuler, W. (2023). Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times?. *Transactions of the Association for Computational Linguistics*, *11*, 336-350.

[74] Jamet, H., Shrestha, Y. R., & Vlachos, M. (2024, June). Difficulty Estimation and Simplification of French Text Using LLMs. In *International Conference on Intelligent Tutoring Systems* (pp. 395-404). Cham: Springer Nature Switzerland.

[75] Liu, J., Shen, D., Zhang, Y., Dolan, B., Carin, L., & Chen, W. (2021). What Makes Good In-Context Examples for GPT-3?. *arXiv preprint arXiv:2101.06804*.

[76] Nori, H., Lee, Y. T., Zhang, S., Carignan, D., Edgar, R., Fusi, N., ... & Horvitz, E. (2023). Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*.

[77] Tao, Y., Viberg, O., Baker, R. S., & Kizilcec, R. F. (2024). Cultural bias and cultural alignment of large language models. *PNAS Nexus*, *3*(9), 346.

[78] Butler, J. J., Harrington, M. C., Tong, Y., Rosenbaum, A. J., Samsonov, A. P., Walls, R. J., & Kennedy, J. G. (2024). From jargon to clarity: Improving the readability of foot and ankle radiology reports with an artificial intelligence large language model. *Foot and Ankle Surgery*, *30*(4), 331-337.

[79] Patel, N., Nagpal, P., Shah, T., Sharma, A., Malvi, S., & Lomas, D. (2023). Improving mathematics assessment readability: Do large language models help?. *Journal of Computer Assisted Learning*, *39*(3), 804-822.

[80] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, *35*, 24824-24837.

[81] Lan, Y. J., & Chen, N. S. (2024). Teachers' agency in the era of LLM and generative AI. *Educational Technology & Society*, *27*(1), 1-18.

[82] Korzynski, P., Mazurek, G., Krzypkowska, P., & Kurasinski, A. (2023). Artificial intelligence prompt engineering as a new digital competence: Analysis of generative AI technologies such as ChatGPT. *Entrepreneurial Business and Economics Review*, *11*(3), 25-37.

[83] Reynolds, L., & McDonell, K. (2021, May). Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended abstracts of the 2021 CHI conference on Human Factors in Computing Systems* (pp. 1-7).

[84] Dempere, J., Modugu, K., Hesham, A., & Ramasamy, L. K. (2023). The impact of ChatGPT on higher education. *Frontiers in Education* (Vol. 8, p. 1206936). Frontiers Media SA.

[85] Herbold, S., Hautli-Janisz, A., Heuer, U. *et al.* A large-scale comparison of human-written versus ChatGPT-generated essays. *Sci Rep* **13**, 18617 (2023).

[86] Wiegand, M., Nadarajah, S., & Si, Y. (2018). Word frequencies: A comparison of Pareto type distributions. *Physics Letters A*, *382*(9), 621-632.

[87] McNicholl, A., Casey, H., Desmond, D., & Gallagher, P. (2021). The impact of assistive technology use for students with disabilities in higher education: A systematic review. *Disability and Rehabilitation: Assistive Technology*, 16(2), 130-143.

[88] Elleman, A. M., & Oslund, E. L. (2019). Reading comprehension research: Implications for practice and policy. *Policy Insights from the Behavioral and Brain Sciences*, *6*(1), 3-11.

[89] Saha, N., & Cutting, L. (2019). Exploring the use of network meta-analysis in education: Examining the correlation between ORF and text complexity measures. *Annals of Dyslexia,* 69(3), 335–354.

[90] Frieder, S., Pinchetti, L., Griffiths, R. R., Salvatori, T., Lukasiewicz, T., Petersen, P., & Berner, J. (2024). Mathematical capabilities of ChatGPT. *Advances in Neural Information Processing Systems*, *36*.

[91] Olakanmi, O. A., Akcayir, G., Ishola, O. M., & Demmans Epp, C. (2020). Using technology in special education: Current practices and trends. *Educational Technology Research and Development*, *68*(4), 1711-1738.

[92] UNESCO. (2023). What you need to know about inclusion in education. Retrieved from https://www.unesco.org/en/inclusion-education/need-know?hub$=$70285

[93] Viberg, O., Kizilcec, R. F., Wise, A. F., Jivet, I., & Nixon, N. (2024). Advancing equity and inclusion in educational practices with AI-powered educational decision support systems (AI-EDSS). *British Journal of Educational Technology*, *55*(5), 1974-1981.

[94] Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American journal of Physics*, *66*(1), 64-74.

[95] Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, and Xin Zhou. 2023. Better Zero-Shot Reasoning with Role-Play Prompting. Retrieved March 14, 2024 from http://arxiv.org/abs/2308.07702

[96] Felix Morger. 2023. Are There Any Limits to English-Swedish Language Transfer? A Fine-grained Analysis Using Natural Language Inference. In *Proceedings of the Second Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2023)*, pages 30–41, Tórshavn, the Faroe Islands. Association for Computational Linguistics.