Dyslexia and AI: Do Language Models Align with Dyslexic Style Guide Criteria?

Eleni Ilkou$^{1[0000-0002-4847-6177]}$, Thomai Alexiou[2], Grigoris Antoniou$^{1,3[0000-0003-3673-6602]}$, and Olga Viberg[4]

[1] L3S Research Center, Leibniz University Hannover, Germany
[2] Aristotle University of Thessaloniki, Greece
[3] Leeds Beckett University, UK
[4] KTH Royal Institute of Technology, Sweden
ilkou@l3s.de

Abstract. Dyslexia presents significant challenges in education for students worldwide. While assistive technologies have been used to enhance readability, no study has systematically evaluated the ability of Language Models (LMs) to generate dyslexia-friendly text aligned with established accessibility guidelines. This proof-of-concept study assesses three state-of-the-art LMs on their ability to identify and apply dyslexia-friendly text criteria. Our findings reveal that their knowledge is limited and poses potential risks. To address this, we introduce *DysText*, a novel metric that quantifies dyslexia-friendly text characteristics based on the British Dyslexia Association's Dyslexia Style Guide. Results indicate that while LMs can enhance the dyslexia-friendliness of texts, their responses should not be blindly trusted, underscoring the need for further verification.

Keywords: Special Education · Text Accessibility · Dyslexic Criteria.

1 Introduction

Around 10% of the general population are people with dyslexia [48], with some suggesting that the percentage could be up to 20% [43], making it one of the predominant causes of learning difficulties. Also, among students with special education needs, about 75% face difficulties in reading [9], which can be benefited from assistive technologies used for dyslexia-friendly texts. Yet, the adoption of natural language processing (NLP) tools and Language Models (LMs) in dyslexia-friendly transformations of educational content is still limited [42].

The integration of artificial intelligence (AI) and LMs into special education settings presents both opportunities and challenges for all learners [6], and particularly those with dyslexia [20]. LMs are used to assist learners with dyslexia by simplifying text [30], and assisting in written communication [23,28]. As there is no metric for evaluating the dyslexia-friendliness of texts, the research relies on manual methods [21,24], readability indexes (ie. LIX) [30], and lexical (ie. BLEU) and semantic (ie. BERTScore) metrics [44]. Although these metrics are effective for their intended tasks, they do not align with the criteria outlined in dyslexia-friendly text guidelines and are limited in capturing only a few of the key characteristics of such texts. Thus, there is a need for an evaluation technique that aligns with dyslexia-friendly criteria for text.

Dyslexia-friendly Criteria There are plenty of criteria introduced for making a presented text friendlier to readers with dyslexia. Most of them focus on the user-interface, such as the "easy-to-read" which is supported by the United Nations [35], and the European standards for making information easy to read and understand [45]. In the web, the W3C has a Web Accessibility Initiative (WAI) [58] includes plenty of accessibility criteria, which fit the learners with dyslexia [14]. However, there is no one guideline with specific mandatory characteristics that serves as the gold-standard for dyslexia-friendly text. To overcome this obstacle, we adopt the publicly and specifically described criteria in the Dyslexia Style Guide by the British Dyslexia Association [10], which offers an overview of specific dyslexia-friendly text criteria, as we report in Table 1.

Objective and Research Questions To the best of our knowledge, the present work is the first to quantify guidelines for dyslexia-friendly text, and investigate LMs' understanding of dyslexia-friendly text criteria and measure their ability to produce dyslexia-friendly text. We develop a proof-of-concept study centered on the Dyslexia Style Guide criteria and the LMs Gemma, Phi4 and GPT4-turbo. We summarize our inquiry into specific research questions (RQs), as following:
RQ1: Are LMs effective in recognizing the Dyslexic Style Guide criteria?
RQ2: How well can LMs generate dyslexia-friendly text according to the Dyslexia Style Guide?
RQ3: Can we improve the LMs' performance on producing dyslexia-friendly text if we provide the Dyslexic Style Guide criteria?

To tackle the RQs, we selected a diverse set of representative LMs, namely Gemma, Phi4 and GPT4-turbo, to evaluate their capacity to generate text that aligns with established dyslexia-friendly criteria. We find that LMs recognize around half of the Dyslexic Style Guide criteria (Table 2 Section 4.1), while their recommendations for additional criteria hinders potential risks for inclusivity (Table 3 Section 4.1). We introduce *DysText* (Section 3.4), a novel metric which quantifies the text-related criteria from the Dyslexic Style Guide. We find that the LMs can produce dyslexia-friendly text with *DysText* total scores of 2.88, 3.24, and 2.22 in Gemma, Phi4 and GPT4-turbo respectively (Table 5 Section 4.2). However, by analyzing the generated texts we conclude that the LMs responses require further quality control. The improvements in *DysText* total scores are significant across all prompts (Figure 1 Section 4.3, and online).

2 Background

Assistive technologies and educational tools are recommended for students with dyslexia [11], as they can support in customizing text presentation [46,1,26,56,61], font design [27], spelling and grammar [37,55], offer game-based learning [34] and text simplification [5,53] to address the learners' needs better [3]. The research to support learners with dyslexia has focused on adaptive learning support technologies, such as audio-books [7], text-to-speech [15,16,47], interface adjustments [3], and spell checkers [12], which enable learners to bypass traditional reading and writing barriers [50].

Table 1: The Dyslexia Style Guide by the British Dyslexia Association reported per computation feature. With "C", we indicate the criteria which can be computed from a JSON file, and with "O" the rest.

	Feature	Explanation
O1	Font style	Use sans serif fonts, such as Arial and Comic Sans, as letters can appear less crowded. Alternatives include Verdana, Tahoma, Century Gothic, Trebuchet, Calibri, Open Sans.
O2	Font size	Font size should be 12-14 point or equivalent (e.g. 1-1.2em / 16-19 px). Some dyslexic readers may request a larger font.
O3	Spacing inter-letters	Larger inter-letter / character spacing (sometimes called tracking) improves readability, ideally around 35% of the average letter width. If letter spacing is excessive it can reduce readability.
O4	Spacing inter-words	Inter-word spacing should be at least 3.5 times the inter-letter spacing
O5	Spacing inter-lines	Some dyslexic people find that larger line spacing improves readability. It should be proportional to inter-word spacing; 1.5 / 150% is preferable.
O6	Headings font size	For headings, use a font size that is at least 20% larger than the normal text. If further emphasis is required, then use bold.
O7	Assistive technology	Use formatting tools for text alignment, justification, indents, lists, line and paragraph spacing to support assistive technology users.
O8	Hyperlinks look	Ensure hyperlinks look different from headings and normal text.
O9	Background colour	Use single colour backgrounds. Avoid background patterns or pictures and distracting surrounds.
O10	Background contrast	Use sufficient contrast levels between background and text. Use sufficient contrast levels between background and text.
O11	No green, red/pink	Avoid green and red/pink, as these colours are difficult for those who have colour vision deficiencies (colour blindness).
O12	Alternative to white background	Consider alternatives to white backgrounds for paper, computer and visual aids such as whiteboards. White can appear too dazzling. Use cream or a soft pastel colour. Some dyslexic people will have their own colour preference.
O13	Print paper quality	When printing, use matt paper rather than gloss. Paper should be thick enough to prevent the other side showing through.
O14	Use images	Use images to support text. Flow charts are ideal for explaining procedures. Pictograms and graphics can help to locate and support information in the text
O15	Instructions	Give instructions clearly.
O16	No jargon	Avoid jargon where possible; always provide the expanded form when first used. Provide a glossary of jargon
C1	Spacing paragraphs	Use white space to remove clutter near text and group related content.
C2	No *Italics*	Avoid Underlining and italics as this can make the text appear to run
C3	No Underline	together and cause crowding.
C4	Use **Bold**	Use bold for emphasis.
C5	No uppercase	Avoid using capital letter and uppercase letters for continuous text. Lower case letters are easier to read.
C6	Left align	Left align text, without justification. This makes it easier to find the start and finish of each line and ensures even spacing between words.
C7	Short sentences	Write short simple sentences: 60 to 70 characters is optimal.
C8	Short paragraphs	Be concise; avoid using long, dense paragraphs.
C9	Section headings	Break up the text with regular section headings in long documents
C10	ToC	and include a table of contents.
C11	Active voice	Use active rather than passive voice.
C12	Simple language	Write in simple clear language using every day words.
C13	Bullets or lists	Consider using bullet points and numbering rather than continuous prose.
C14	No double negatives	Avoid double negatives.
C15	No abbreviations	Avoid abbreviations where possible.
C16	Avoid columns	Avoid multiple columns (as used in newspapers).
C17	Headings spacing	Add extra space around headings and between paragraphs.

A widely adopted approach to improve text for learners with dyslexia is the text simplification, which is used to enhance texts' accessibility and inclusion in educational settings. Schicchi and Taibi [51] introduced an automatic text simplification system using automatic text simplification and automatic text complexity evaluation. They suggest that automatic text simplification systems can be useful when limited textual resources are available for teaching a particular topic, allowing the tool to assist in creating multiple versions of the same text to assist text accessibility. AI-powered tools for content summarization condense lengthy texts into manageable summaries, empowering learners with dyslexia to grasp the main ideas efficiently [40]. Further most studies evaluate the accessibility of text based on readability indexes, such as the LIX readability index, which takes into consideration the number of words, letters and periods present in a text [8].

Madjidi and Crick [38] developed a model that modifies text based on dyslexia-friendly guidelines using transfer learning, training T5 and RoBERTa on crowdsourced data. A follow-up study enhanced this by adding syllable and morphological analysis, helping simplify complex word structures [39]. Testing undergraduates with dyslexia showed faster reading times, though comprehension varied. Meanwhile, Ayang et al. [13] created a cloud-based system that reformats Microsoft Word documents into dyslexia-friendly versions by adjusting fonts, spacing, and colors, allowing users to customize modifications for improved readability and privacy. While useful for general readability, the existing research and tools do not fully capture nor measure the features of dyslexia-friendly texts and do not incorporate specific dyslexia-friendly criteria into their approaches.

3 Methodology

3.1 Data Collection

Currently, there is no gold-standard dataset available for converting texts into both original and dyslexia-friendly formats [61]. To ensure the applicability of most dyslexia-friendly criteria which are not applicable to short texts, ie. "C13: Bullets or lists" from Table 1, we require longer texts, hence we focus on the domain of history. Our focus is on educational text found in learning resources, and not AI-generated text to ensure the pedagogical alignment of our study. Due to licensing restrictions and copyright protections, we aimed for widely adopted textbooks with an open license to ensure broad applicability and maximize the impact of our research. Our first choice is the textbook "History Textbook: West African senior school certificate examination" which is openly available under a CC BY-NC 3.0 license [4]. The book claims that *"is aimed at West African students taking the West African Senior School Certificate Examination"*, which makes it a textbook with high impact and everyday outreach. The second choice of textbook is the "U.S. History" from openstax which is openly available under a CC BY 4.0 license [19]. We exclude texts with small paragraphs, annotations and hyperlinks as they already fulfill some of the criteria. We gather a total of 50 chapters which we manually add into JSON file while maintaining the line breaks characteristics.

3.2 Language Models and Prompts

The selection of LMs is based on their applicability to educational settings, as the smaller models have low requirements for local installation in a personal or school laptop, while GPT is a popular alternative for users around the world. Also, our choice is motivated by variations in size, and architecture, which might impact dyslexia-friendly performance. The hyper-parameters used are temperature=0 for consistent outputs, top_p=0.5 to control the diversity of outputs, repeat_penalty=1.1 and no maximum tokens prefixed to allow full paragraph-length responses without artificial thresholds. In our study, we use:

Gemma Gemma [54], developed by Google, is a 7B parameter language model based on the transformer decoder architecture. It is pre-trained on primarily-English data, and it demonstrate strong performance across academic benchmarks for language understanding.

Phi4 Phi4 [2], developed by Microsoft, is a 14B parameter language model based on a decoder-only transformer architecture. It is pre-trained on organic sources and synthetic data, and achieves state-of-the-art performance comparable to large models such as GPT 4o in STEM.

GPT4-turbo GPT4-turbo [41], developed by OpenAI, has hundreds of billions parameters LM based sparse attention and mixture-of-experts architecture. GPT4-turbo has a context window of 128k tokens allowing for big text inputs.

Prompt Engineering Prompt engineering techniques in LMs are widely studied [49]. As we are limited by successful examples of text in its original form and in dyslexia-friendly format, we perform zero-shot prompting. Following a similar study's results [30], we formulate our prompts to be direct and short. To detect the LMs knowledge about the dyslexia-friendly criteria, we deploy four prompts: 1. "Which are the criteria for making a text dyslexia friendly?", 2. "Which are the principles in Dyslexia Style guide that suggests changes in written material for dyslexic readers?", 3. "What guidelines should be followed to make a text more accessible for people with dyslexia?", and 4. "How can I make a text dyslexia-friendly?". To evaluate the LMs ability to generate dyslexia-friendly text, we deploy two prompts without a guide included: 1. "Make the following text dyslexia-friendly :\n [Chapter]", and 2. "Rewrite the following text according to dyslexic style guide:\n [Chapter]". Additionally we utilize 1. the full Dyslexia Style Guide, and 2. only the JSON text related criteria as [Guide] and formulate the prompts "I will provide the Dyslexic Style Guide and a text, I want you to make the text according to the Dyslexic Style Guide. \n Here is the Dyslexic Style guide: \n [Guide]\n Here is the text: \n [Chapter]".

3.3 Data Analysis

We process the LMs responses to retain only answers related to dyslexia-friendly text, as manual cleaning proved necessary even when LMs were instructed to exclude guidelines or explanations. We report the statistics of the minimum (Min.), median (Mdn.), maximum (Max.), average (Av.) and standard deviation (SD) as the descriptive statistics are commonly used to detail the readability of the original versus the generated text by the LMs [30]. To evaluate significance, we deployed paired statistical tests to compare the generated texts of the LMs against the original dataset. The normality was assessed using the Shapiro-Wilk test, a robust method for testing departures from normality. For comparisons where the differences met the assumption of normality, a two-tailored paired t-test was performed to assess whether the mean differences were statistically significant. In cases where the differences violated the assumption of normality, the non-parametric Wilcoxon signed-rank test was utilized as a distribution-free alternative. Statistical significance was evaluated at the threshold of $p < 0.05$.

3.4 New Metric for Dyslexia-Friendly Text - *DysText*

Our proposed metric, *DysText*, is theoretically grounded on the guidelines set forth by the Dyslexia Style Guide from the British Dyslexia Association. *DysText* is pioneer metric that quantifies the 17 out of the 33 criteria introduced by Dyslexia Style Guide, and is able to measure the accessibility of text based inclusivity guidelines. The metric focuses on features present in JSON formats, as extracted by LMs responses. Building on the criteria outlined in Table 1, we examine the characteristics that can be computed by the LMs JSON output marked in "C". We compute each criterion by recognizing patterns in the JSON format, and utilizing NLP libraries. Our data, further results, and the *DysText* metric are available at: `https://github.com/eilkou/DysText`

Semantic Interpretation of Dyslexic Style Guide Criteria Given that the British Dyslexia Association does not assign a priority to any individual criterion, we treat each criterion as equally significant in contributing to the dyslexia-friendliness of the text. To accurately capture the impact of each criterion on text accessibility, we map the corresponding recommendations that when present improve the accessibility to a scale ranging from 0 to 1, where values from 0 to 1 indicate improvements in dyslexia-friendly text accessibility. Conversely, recommendations to avoid are mapped to values between -1 and 0, reflecting their potential to hinder accessibility. The metric is computed as the sum of individual criterion outputs and has a range of values of [-10,11]. We assume that the criteria of left align (C6) and avoid columns (C16) are covered by default.

4 Results

4.1 LMs Knowledge of Dyslexic Style Guide Criteria

We begin our analysis with prompting LMs to provide the criteria for dyslexia-friendly texts. In Table 2, we report the scores of the LMs in recognizing the dyslexia-friendly criteria, accompanied with the wrongly reported criteria per LM. The total score is calculated based on individual criterion scores, with +1 point awarded for a correct response, +0.5 points for a partially correct response (e.g., when the LM recognizes "Font size" as a relevant criterion but fails to suggest appropriate fonts), 0 points for a missed criterion, and -1 for an incorrect answer (e.g., suggesting the use of green when the criteria advise against it).

The negative score serves as a penalty, as incorrect recommendations can reduce dyslexia-friendliness and negatively impact the overall reading experience.

Table 2: The LMs knowledge scores of the Dyslexic Style Guide criteria. They recognize less than half of the 33 criteria reported in Table 1.

	Total Score		Wrongly reported
	Av.	SD	
Gemma	10.13	0.63	C2: No *Italics*, O11: No green, red/pink
Phi4	13.88	0.63	-
GPT4-turbo	13.00	0.41	-

Phi4 is more informed about the dyslexia-friendly criteria than GPT4-turbo. All LMs perform below average, as they identify less than 14 criteria from the 33 included in the Dyslexia Style Guide. A common theme that erased from the criteria the LMs recommended was the font size and style, the short sentences and simple language. Regarding the font style, the LMs recommended additionally the usage of OpenDyslexic font [27], an open-source font which is designed with distinctive letter forms. Although some studies suggest it can enhance reading accuracy and rate [18], it might not always be helpful to learners [59]. It is worth-mentioning that all the LMs included additional criteria, which are not found in the Dyslexic Style Guide, as we report in Table 3.

Table 3: The additional criteria reported by the Language Models.

Additional Criterion	LM	Risk	Example
Accessibility considerations	Gemma	Low	"Keyboard-accessible navigation with clear links and menus"
Audio support	Gemma, Phi4, GPT4-turbo	Low	"Where possible, offer audio versions of texts to provide alternative access methods."
Avoid effects	Gemma, Phi4	Low	"Avoid flickering or animated content"
Consistent layout	Phi4, GPT4-turbo	Low	"Maintain a consistent layout throughout the document to help readers establish familiarity and reduce confusion."
Feedback & Testing	Gemma, Phi4, GPT4-turbo	Medium	" "Feedback from dyslexic individuals on readability and accessibility"
Encouragement & Praise	Gemma	High	"Use of positive language and encouraging messages"
Interactive elements	Gemma, Phi4	Medium	"Interactive activities and games to promote engagement"
Predictability	Gemma	Medium	"Provide predictable syllable patterns and word structures"
Reader's Perspective	Gemma	High	"Imagine you are reading the text from the perspective of someone with dyslexia."
Summaries	Gemma, GPT4-turbo	Low	"Provide summaries or outlines at the beginning of lengthy documents to give an overview of what follows"

Potential Risks from Additional Recommended Criteria In Table 3, we report a qualitative analysis and classification of the additional criteria provided by LMs. Although interactive elements, are not always practical in static text formats [25], they can enhance comprehension when applied judiciously [52]. Predictability in word structure enhances readability but, if oversimplified, may compromise linguistic depth [17]. Although encouragement is beneficial for all students, praise can have negative effects when not accompanied with detailed feedback [36]. This is particularly critical in students with low self-esteem, which receive more person praise rather than process praise and ends up predisposing them for feeling failure shame [29]. Hence, excessive encouragement and praise, may be patronizing or ineffective, especially if they emphasize personal attributes rather than learning processes. Writing from the perspective of a reader with dyslexia also presents risks, as it introduces labeling concerns and may border on unintentional discrimination [57]. While considering diverse perspectives is valuable, inclusivity should be achieved through broader accessibility rather than narrowly defined reader experiences.

4.2 LMs Generating Dyslexia-friendly Text

In Table 4, we report the statistics of the original texts and in Table 5, we present the same statistics over the LMs' generated texts. All three LMs improved their scores in $DysText$ visual, content and total scores, which demonstrates their ability to improve the text into a more dyslexia-friendly format according to the Dyslexia Style Guide. However, their performance in $DysText$ metric shows that there is plenty of room for improvement as their values reached a maximum of 5 (5.05 in Phi4, 5.31 in Gemma, and 5.53 in GPT4-turbo) over the maximum possible score of 11 points. Gemma produced the shorter generated text compared to Phi4 and GPT4-turbo, which could imply that Gemma missed reporting valuable information.

LIX vs *DysText* Scores A crucial element highlighted by the comparison between readability LIX score, in which the smaller scores indicate easier to read texts, and $DysText$. $DysText$ incorporates the elements of LIX with additional criteria, and both seem to predict the LM with the better dyslexia-friendly generated texts. Gemma scores the lowest in LIX score with 43.81 on average across all generated texts, which indicates that the LMs responses are medium in difficulty to read. At the same time, Gemma gains the biggest average content score and the second biggest average total performance in $DysText$, which indicates bigger compliance to the Dyslexic Style Guide criteria and more dyslexia-friendly texts. However, since $DysText$ encompasses more characteristics than just word and sentence count, it demonstrates that relying solely on the LIX score does not provide insight into the dyslexia-friendly features of a text. This is highlighted by Phi4, which shows only a modest improvement in the LIX score, from an average of 49.68 to 48.79, despite achieving the highest average $DysText$ total score. The $DysText$ visual score is more closely correlated with the number of paragraphs (No. Paragraphs) which includes the sum of headers and bullet points, hence, higher values in "No. Paragraphs" indicate high values in $DysText$ visual score.

Qualitative Result: LMs Responses Require Verification Our qualitative analysis reveals that the LMs responses should not be fully trusted, and a supervised or human-in-the-loop approach is suggested to validate the content of the responses. In minor consideration of the responses, the LMs might require small edits as the generated text misses spaces between words in some cases, ie. the response contained *"butalso"* instead of *"but also"*. These cases were more often found in the end of the response rather at the first sentences. In odd observations, Gemma changed the word "the" to *"da"* in two of the chapters, and

Table 4: Analysis of the dataset with the original texts.

	Chapters				
	Min.	Mdn.	Max.	Av.	SD
No. Sentences	8	21.5	59	24.74	1.74
No. Words	184	490	1468	538.84	39.74
No. Long Words (>6 characters)	50	127	381	145.92	10.94
No. Paragraphs	1	4	17	4.82	0.46
Readability LIX score	37.43	50.61	62.07	49.68	0.81
DysText visual score	0	0	0	0	0
DysText content score	-3.08	-0.43	0.39	-0.62	0.10

Table 5: Analysis of LMs generated dyslexia-friendly text over the dataset. The number of paragraphs (No. Paragraphs) includes the headers and bullet points.

	Gemma					Phi4					GPT4-turbo				
	Min.	Mdn.	Max.	Av.	SD	Min.	Mdn.	Max.	Av.	SD	Min.	Mdn.	Max.	Av.	SD
No. Sentences	4	17	36	17.24	6.15	9	22	40	21.37	6.62	6	15	27	15.18	3.50
No. Words	69	263.5	711	280.46	107.87	140	324.5	704	330.56	103.54	135	330.5	3041	356.11	248.48
No. Long Words	14	71	200	76.73	38.38	41	105.5	248	108.44	38.38	36	117.5	1900	140.11	158.61
No. Paragraphs	1	6	24	7.22	4.87	1	13	45	15.35	9.33	2	10	40	11.80	6.95
LIX score	24.10	44.39	81.98	43.81	10.64	33.49	48.12	67.53	48.79	6.92	35.00	57.20	265.21	60.84	21.65
DysText Visual	0	1	3	1.00	0.88	0	2	2	1.45	0.82	0	2	4	1.21	1.07
DysText Content	-1.60	1.92	3.52	1.88	0.93	-0.39	1.91	3.16	1.80	0.72	-8.26	1.25	3.27	1.02	1.37
DysText Total	-1.60	3.05	5.31	2.88	1.46	-0.39	3.44	5.05	3.24	1.08	-7.26	2.51	5.53	2.22	1.82

produced a few spelling mistakes, such as *"Diplomcy"* instead of *"Diplomacy"*. Also, in the one case GPT4-turbo did not provide a response, and Gemma gave as an answer only recommendations for making the text dyslexia-friendly instead of providing the text; in those cases we re-run the prompt.

Further, potential issues could be raised from specific phrasing. An example is extracted from headings proposed for different chapters by the LMs, the *"Reparation - Easier to Read"* from Gemma, and the *"Making History Easier to Read: Britain's Financial Challenges After War* and *"English Colonization in North America: A Dyslexia-Friendly Overview"* from Phi4. While the latter has a clearly positive tone, and the use of *"Making History Easier to Read"* could be seen as simplifying a complex historical topic, however, it might unintentionally suggest that the audience is incapable of understanding the historical topic unless it is made "easy". Emphasis on sensitivity of language and minor adjustments, such as in Gemma *"[Topic]- Easier to Read"*, can maintain the integrity of the content without risking demeaning the target audience. We highlight that in no prompt the word "easy" was used. Hence, that phrasing can be problematic, interpreted as insensitive and patronizing commentary, and hinder inclusivity.

In addressing more significant concerns, we turn our attention to the qualitative analysis of the generated text responses. Notably, we observe that GPT4-turbo in section "European Trading Communities in West Africa" [4] deviated from its assigned task, and transitioned mid-response into an unprompted discourse on humanity and history. This deviation raises significant concerns regarding the LLM's ability to adhere to task-specific constraints and maintain alignment with instructional prompts, underscoring potential challenges in its reliability for focused educational or research applications.

Another example related to potential risks is raised from the quality of responses. In five cases, the GPT4-turbo responses were initiated with proper context, without raising concerns, however the body context raised concerns. The five cases were split into: one case from the book "U.S. History" the chapter "The Rise of Slavery in the Chesapeake Bay Colonies", and four cases from the book "History Textbook: West African senior school certificate examination" in the sections "European Contact with West Africa" and "Trans-Saharan Trade. Origins, organization and effects in the development of West Africa". After careful consideration of the whole body of responses, we see that the GPT4-turbo repeats nonsense phrases. All these cases had in common the lengthy responses by GPT4-turbo, which could indicate a potential threshold for quality control. However, in a broader perspective, even as isolated events, these incidents highlight the need to examine the generated answers, by focusing not only on technical criteria compliance, but also on the accuracy and quality of the content. It is worth questioning why the misconduct on responses occurred on the topics about colonies and slavery, even if not all chapters on these thematics were affected. This discrepancy may indicate a potential bias or restrictions in topics of responses in GPT4-turbo, warranting further investigation is needed to understand these inconsistencies and their impact on accessibility.

4.3 LMs Improving Dyslexia-friendliness of Texts

All LMs showed statistically significant improvements to the original texts across all prompts. In Figure 1 are the boxplots of the $DysText$ total scores differences between each generated text and its corresponding original. Phi4 improved 100% of the original texts in visual and total $DysText$ score across all prompts. One of the reasons for the 100% improved $DysText$ visual score could be that Phi4, unlike Gemma and GPT-4 Turbo, often provided a glossary and/or a summary.

Prompting the Dyslexic Style Guide criteria which are only applicable to text improves the performance and produces consistent good improvements. This is demonstrated by prompt 4 in Figure 1, which provided the criteria only applicable to text in JSON format as are in the explanations of features "C" in Table 1. Prompt 4 improved the original texts $DysText$ scores in 100% of visual, content and total on Phi4, in 96% of visual and 100% of content and total on Gemma, and in 100% of visual and 98% of content and total on GPT4-turbo. Gemma is reporting its second best improvements in prompt 4, which is probably due to its shorter context length of 8192 tokens, in contrast to 16k of Phi4 and 128k of GPT4-turbo. The context length refers to the maximum number of words and characters the LM can process and retain in a single pass, where exceeding this limit results in the truncation or loss of earlier input, hence, affecting the LM's performance. The second best was the prompt: "Make the following text dyslexia-friendly :\n [Chapter]". The least benefited score was the $DysText$ visual in prompt 2 "Rewrite the following text according to dyslexic style guide:\n [Chapter]". The biggest improvements in scores were noticed in

DysText visual score on GPT4-turbo in prompt 4, with 2.16 points gained on average, in *DysText* content and total scores on Gemma in prompt 1, with 3.21 and 4.75 points gained on average, respectively. This analysis demonstrates the capabilities of small LMs such as Gemma and Phi4, to compete with large LMs.

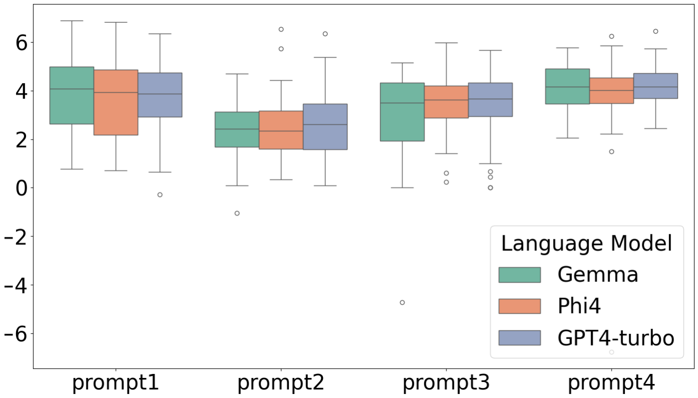

Fig. 1: The boxplots of the improvements for each text over the total DysText scores across the prompts and LMs.

5 Discussion and Conclusion

LMs knowledge of dyslexia-friendly criteria is limited. Our results show that the LMs are not informed about most than half of the Dyslexia Style Guide criteria, as they identify 13 from the 33 criteria. Moreover, even if they are aware of a criterion, ie "O2: Font style", they often recommend additional features for it, which are not included in the guide, not fully supported by research findings.

LMs significantly improve the dyslexia-friendliness of texts. Our study demonstrates that Gemma, Phi4, and GPT-4 Turbo significantly enhance original texts, making them more dyslexia-friendly, as measured by the DysText metric. This metric quantifies adherence to the Dyslexic Style Guide by detecting formatting and linguistic features in JSON-structured outputs. The results indicate that the examined LMs consistently perform better when prompted with text-only criteria alongside the corresponding chapter, suggesting that structured input helps optimize their ability to generate accessible content. This aligns with previous research highlighting the importance of explicit, well-defined constraints in guiding LMs to produce more user-friendly and readable outputs.

LMs responses require verification. However, we emphasize that the findings from the qualitative analysis underscored the need to avoid automatically trusting the LMs' generated text responses. Moreover, the LMs are not a reliable source for informing the users about the dyslexia-friendly recommendations for text, as their recommendations can hinder potential risks in the documentation of reported dyslexia-friendly criteria from Gemma, the headings selection from Phi4, and generated text in historically sensitive topics from GPT4-turbo.

LIX readability score misses dyslexia-friendly features. While Gemma has the lowest LIX score, indicating medium difficulty, it achieves the highest DysText score, showing stronger compliance with the Dyslexic Style Guide. In contrast, Phi4 shows only a slight improvement in LIX despite achieving the best DysText score. This discrepancy highlights a fundamental difference between the two metrics: 1. LIX assesses readability based on sentence length and word complexity, which could be insufficient in capturing accessibility improvements tailored for dyslexic readers. 2. *DysText*, integrates a broader set of dyslexia-friendly criteria, such as the existence of bullet points, offering a more comprehensive evaluation of how well a text aligns dyslexia-friendly features.

Limitations. There are ethical and legal implications when using LMs in education [60], particularly regarding the biases, transparency, fairness, and accountability of the LM-generated content, such as the nonsensical or biased responses generated by GPT4-turbo on texts about charged themes, such as colonization and slavery, raising concerns about their reliability and potential to misinform learners. Our study highlights two key challenges. First, LMs misinterpret or omit critical information, which can compromise dyslexia-friendly content. Second, there are legal concerns about text use without proper processing rights, particularly in relation to licensing agreements and GDPR compliance [22]. The DysText metric is limited to English-language texts, restricting its applicability to other linguistic and cultural contexts and necessitating external validation for context alignment. Also, jargon detection is context- and domain-dependent. For example, such terms as *Senegambia*, require expert validation to ensure accuracy.

Future Work. Since prompting LMs performs best with quality examples, e.g. Chain of Thought reasoning [49], our aim is to develop a benchmark dataset featuring original and dyslexia-friendly texts in collaboration with experts and the DysText metric. This dataset could further support fine-tuning LMs for dyslexia-related learning practices. For multilingual adaptation, NLP libraries for non-English corpora can be utilized alongside cultural and linguistic coverage criteria to adjust language-sensitive features (e.g., C11: Active voice). In addition, DysText could be expanded to incorporate all the criteria of the Dyslexic Style Guide, including the considerations of the graphical user interface [3]. We plan to analyze which criteria LMs handle most effectively and where they struggle in generating dyslexia-friendly texts, forming further improvements in AI-driven accessibility solutions [31,32].

Conclusions. Our study is a proof-of-concept exploring LMs' awareness and application of dyslexia-friendly criteria by LMs, quantifying their ability to generate accessible text. Our findings stress both the potential and the risks of automated dyslexia-friendly text generation using LMs that can be deployed on student or school laptops. These methods could enhance text accessibility, reduce stigma in special education, and address educational inequities, aligning with global efforts towards inclusive and equitable quality education [33].

Acknowledgments This work was partially supported by a 2024 AIED-DEIA Fellowship grant from the IAIED Society (iaied.org). Viberg was partly supported by the Future AI project by KTH.

References

1. A Cambium Learning, K.e.: Kurzweil 3000, https://www.kurzweil3000.com/
2. Abdin, M., Aneja, J., Behl, H., et al.: Phi-4 technical report. arXiv preprint arXiv:2412.08905 (2024)
3. Abdul Aziz, N.I., Husni, H., Hashim, N.L.: Dyslexia-friendly design features for tangible user interfaces: a systematic literature review. The International Journal of Information and Learning Technology **39**(4), 360–372 (2022)
4. Achebe, N., Adu-Gyamfi, S., Alie, J., et al.: History textbook: West African senior school certificate examination (2018)
5. AI, Y.: Dyslexic GPT-Dyslexia-Friendly Reading Tool, empowering readability with ai, https://www.yeschat.ai/gpts-9t55QeOXpjD-Dyslexic-GPT
6. Aliu, T.V.: Artificial intelligence in special education: A literature review. Systemic Analytics **2**(2), 188–199 (2024)
7. Almgren Bäck, G., Lindeblad, E., Elmqvist, C., Svensson, I.: Dyslexic students' experiences in using assistive technology to support written language skills: a five-year follow-up. Disability and Rehabilitation: Assistive Technology **19**(4), 1217–1227 (2024)
8. Anderson, J.: Lix and rix: Variations on a little-known readability index. Journal of Reading **26**(6), 490–496 (1983)
9. Association, I.D.: Frequently Asked Questions how common are language-based learning disabilities? (2025), https://dyslexiaida.org/frequently-asked-questions-2/
10. Association, T.B.D.: Creating a dyslexia friendly workplace, dyslexia friendly style guide, https://www.bdadyslexia.org.uk/advice/employers/creating-a-dyslexia-friendly-workplace/dyslexia-friendly-style-guide
11. Association, T.D.: Assistive technology training (2025), https://www.dyslexia.uk.net/service/assistive-technology-training/
12. el Atawy, S.M., Ahmed, H.M.: Spelling checker for dyslexic second language arab learners. Journal of Theoretical and Applied Information Technology (2021)
13. Ayang, D., John, C., Quadras, J., Nadar, M., Curley, A., Gordon, D., Tierney, B.: Accessibility made easy: The development of a cloud-based service to make documents more dyslexia-friendly. In: EDULEARN24 Proceedings. IATED (2024)
14. Berget, G., Herstad, J., Sandnes, F.E.: Search, read and write: An inquiry into web accessibility for people with dyslexia. In: Universal Design 2016: Learning from the Past, Designing for the Future, pp. 450–460. IOS Press (2016)
15. Bhola, N.: Effect of text-to-speech software on academic achievement of students with dyslexia. Integrated Journal for Research in Arts and Humanities (2022)
16. Bonifacci, P., Colombini, E., Marzocchi, M., Tobia, V., Desideri, L.: Text-to-speech applications to reduce mind wandering in students with dyslexia. Journal of Computer Assisted Learning **38**(2), 440–454 (2022)
17. Borleffs, E., Maassen, B.A., Lyytinen, H., Zwarts, F.: Cracking the code: The impact of orthographic transparency and morphological-syllabic complexity on reading and developmental dyslexia. Frontiers in psychology **9**, 2534 (2019)
18. Broadbent, L.: Comparing the impact of OpenDyslexic and Arial fonts on the reading performance of Key Stage 2 readers with dyslexia. Ph.D. thesis (2023)
19. Corbett, P.S., Janssen, V., Lund, J.M., Pfannestiel, T., Waskiewicz, S., Vickery, P.: US history (2024)
20. Dawson, K., Antonenko, P., Lane, H., Zhu, J.: Assistive technologies to support students with dyslexia. Teaching exceptional children **51**(3), 226–239 (2019)

21. De Marco, V., Sciarrone, F., Temperini, M.: Tutorchat: a chatbot for the support to dyslexic learner's activity through generative ai. In: 2024 IEEE International Conference on Advanced Learning Technologies (ICALT). pp. 155–157. IEEE (2024)
22. Duncan, A., Joyner, D.A.: With or without eu: Navigating gdpr constraints in human subjects research in an education environment. In: Proceedings of the Eighth ACM Conference on Learning@ Scale. pp. 343–346 (2021)
23. D'Urso, S., Sciarrone, F.: Ai4la: An intelligent chatbot for supporting students with dyslexia, based on generative ai. In: International Conference on Intelligent Tutoring Systems. pp. 369–377. Springer (2024)
24. Eroğlu, G., Abou Harb, M.R.: Assessing chatgpt's accuracy in dyslexia inquiry. In: 2024 Medical Technologies Congress (TIPTEKNO). pp. 1–4. IEEE (2024)
25. Fawcett, A., Nicolson, R.: Dyslexia, learning, and the brain. MIT Press (2008)
26. fontconverter4dyslexia@gmail.com: Font converter for dyslexia (2025), https://fontconverterfordyslexia.neocities.org/
27. Gonzalez, A.: OpenDyslexic,a typeface for dyslexia, https://opendyslexic.org/
28. Goodman, S.M., Buehler, E., et al.: Lampost: Ai writing assistance for adults with dyslexia using large language models. Communications of the ACM (2024)
29. Graf-König, N., Puca, R.M.: "wow, you're really smart!"–how children's self-esteem affects teachers' praise. Educational Psychology **44**(6-7), 749–764 (2024)
30. Hedlin, E., Estling, L., Wong, J., Epp, C.D., Viberg, O.: Got it! prompting readability using chatgpt to enhance academic texts for diverse learning needs. In: Proceedings of the 15th Learning Analytics and Knowledge Conference (2025)
31. Ilkou, E., Galletti, M., Dobriy, D., et al.: Edumultikg attains 92% accuracy in k-12 user profiling. In: Proceedings of the ESWC. vol. 2043 (2023)
32. Jaldi, C.D., Ilkou, E., Schroeder, N., Shimizu, C.: Education in the era of neurosymbolic ai. Journal of Web Semantics **85**, 100857 (2025)
33. Johnstone, C.J., Schuelka, M.J., Swadek, G.: Quality education for all? the promises and limitations of the sdg framework for inclusive education and students with disabilities. In: Grading goal four, pp. 96–115. Brill (2020)
34. Learning, N.: Dyslexia quest dyslexia quest, quickly screen and identify those children at risk for dyslexia, https://www.nessy.com/en-gb/product/dyslexia-quest-home
35. of Library Associations, I.F., Institutions: International federation of library association and institutions ifla professional reports: Guidelines for easy-to-read materials (2010), https://www.ifla.org/wp-content/uploads/2019/05/assets/hq/publications/professional-report/120.pdf
36. Lipnevich, A.A., Eßer, F.J., Park, M.J., Winstone, N.: Anchored in praise? potential manifestation of the anchoring bias in feedback reception. Assessment in Education: Principles, Policy & Practice **30**(1), 4–17 (2023)
37. ltd., G.: Ghotit, dyslexia writing & reading assistant, https://www.ghotit.com/
38. Madjidi, E., Crick, C.: Enhancing textual accessibility for readers with dyslexia through transfer learning. In: Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility. pp. 1–5 (2023)
39. MADJIDI, E., CRICK, C.: Towards inclusive reading: A neural text generation framework for dyslexia accessibility (2024)
40. Nandhini, K., Balasundaram, S.: Improving readability of dyslexic learners through document summarization. In: 2011 IEEE International Conference on Technology for Education. pp. 246–249. IEEE (2011)
41. OpenAI: Introducing gpt-4 turbo (2023), https://openai.com/blog/gpt-4-turbo
42. Paudel, S., Acharya, S.: A comprehensive review of assistive technologies for children with dyslexia. arXiv preprint arXiv:2412.13241 (2024)

43. Phillips, B.A.B., Odegard, T.N.: Evaluating the impact of dyslexia laws on the identification of specific learning disability and dyslexia. Annals of Dyslexia **67**, 356–368 (2017)
44. Price, G., Wu, S.: Lost in translation: Benchmarking commercial machine translation models for dyslexic-style text (2024)
45. Programme, L.L.: Information for all european standards for making information easy to read and understand, `https://easy-to-read.inclusion-europe.eu/wp-content/uploads/2014/12/EN_Information_for_all.pdf`
46. Reader, B.: BeeLine Reader, read faster and easier, all day long (2017), `https://www.beelinereader.com/`
47. Rello, L., Baeza-Yates, R., Bott, S., Saggion, H.: Simplify or help? text simplification strategies for people with dyslexia. In: Proceedings of the 10th international cross-disciplinary conference on web accessibility. pp. 1–10 (2013)
48. Roitsch, J., Watson, S.M.: An overview of dyslexia: definition, characteristics, assessment, identification, and intervention. Science Journal of Education **7**(4) (2019)
49. Sahoo, P., Singh, A.K., Saha, S., Jain, V., Mondal, S., Chadha, A.: A systematic survey of prompt engineering in large language models: Techniques and applications. arXiv preprint arXiv:2402.07927 (2024)
50. Schaur, M., Koutny, R.: Dyslexia, reading/writing disorders: Assistive technology and accessibility: Introduction to the special thematic session. In: International Conference on Computers Helping People with Special Needs. pp. 269–274. Springer (2024)
51. Schicchi, D., Taibi, D.: Ai-driven inclusion: Exploring automatic text simplification and complexity evaluation for enhanced educational accessibility. In: International Conference on Higher Education Learning Methodologies and Technologies Online. Springer (2023)
52. Snowling, M.J., Hulme, C.: Annual research review: Reading disorders revisited–the critical importance of oral language. Journal of Child Psychology and Psychiatry **62**(5), 635–653 (2021)
53. Team, D.: Diffit, learning resources for all., `https://web.diffit.me/`
54. Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M.S., Love, J., et al.: Gemma: Open models based on gemini research and technology. arXiv preprint arXiv:2403.08295 (2024)
55. texthelp: Claroread, help neurodiverse students to achieve more with reading and writing, `https://www.texthelp.com/en-gb/solutions/dsa/claroread/`
56. texthelp: Read&Write, help students understand and express themselves, `https://www.texthelp.com/products/read-and-write-education/`
57. Tunmer, W.E., Chapman, J.W.: Does set for variability mediate the influence of vocabulary knowledge on the development of word recognition skills? Scientific Studies of Reading (2012)
58. W3C Web Accessibility Initiative (WAI), E., (EOWG), O.W.G.: W3c web accessibility, introduction to web accessibility, `https://www.w3.org/WAI/fundamentals/accessibility-intro/`
59. Wery, J.J., Diliberto, J.A.: The effect of a specialized dyslexia font, opendyslexic, on reading rate and accuracy. Annals of dyslexia **67**, 114–127 (2017)
60. Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., Gašević, D.: Practical and ethical challenges of large language models in education: A systematic scoping review. British Journal of Educational Technology **55**(1), 90–112 (2024)
61. YAP, J.R., ARUTHANAN, T., CHIN, M.: Artificial intelligence in dyslexia research and education: A scoping review. IEEE Access (2025)