LET AI READ FIRST: ENHANCING READING ABILITIES FOR INDIVIDUALS WITH DYSLEXIA THROUGH ARTIFICIAL INTELLIGENCE

Sihang Zhao
School of Science and Engineering
The Chinese University of Hong Kong, Shenzhen
sihangzhao@link.cuhk.edu.cn

Shoucong Xiong
School of Management
Zhejiang University
carolhsiung@163.com

Bo Pang
Computer Network Information Center
Chinese Academy of Science
bopang@cnic.cn

Xiaoying Tang
School of Science and Engineering
The Chinese University of Hong Kong, Shenzhen
tangxiaoying@cuhk.edu.cn

Pinjia He*
School of Data Science
The Chinese University of Hong Kong, Shenzhen
hepinjia@cuhk.edu.cn

ABSTRACT

Dyslexia, a neurological condition affecting approximately 12% of the global population, poses substantial challenges to reading proficiency and overall quality of life. Existing assistive technologies offer limited utility due to constraints like limited applicability in quiet settings, high costs, and the potential for meaning alteration or non-real-time. To fill this gap, we introduce LARF (Let AI Read First), a novel method leveraging the state-of-the-art artificial intelligence(AI) language model(LM), OpenAI's GPT-4 to annotate text for enhanced readability without altering original content. We evaluate LARF through two distinct experiments, involving 150 participants with dyslexia and 160 typical participants without dyslexia. The results indicate that, compared to the control condition and existing method, LARF significantly improves reading performance and overall reading experience in individuals with dyslexia. We control the severity of dyslexia among participants and observe that LARF significantly enhanced reading performance more pronouncedly for individuals facing greater reading challenges. Our research demonstrates the significant potential of AI and LMs in the accessible design of reading aids for neurodiversity conditions such as dyslexia.

***K*eywords** Human-Computer Interaction Accessible Design Reading Disability Dyslexia GPT-4

1 Introduction

Dyslexia constitutes a category of neurodevelopmental impairments that affect reading abilities, typically manifested as challenges in reading fluency, speed, and comprehension. Such disabilities can significantly impact the lives and learning of affected individuals[1, 2]. Approximately 12% of the global population has dyslexia[3]. Individuals with dyslexia often struggle with word decoding and recognition, which also impairs their comprehension, fluency, and vocabulary. This condition leads to academic difficulties and increases the risk of bullying in children and teenagers[4]. Additionally, adults with dyslexia commonly experience considerable deficits in executive functioning[5]. Despite the critical impact of dyslexia on individuals' lives and learning, current interventions, mainly in the form of accessible designs, are fraught with limitations.

With the rapid development of AI[6], especially in natural language processing(NLP)[7], numerous spelling assistance for dyslexia have demonstrated considerable capabilities[8, 9]. However, assistance for reading abilities tends to focus on only a few areas: converting text to speech[10], videos or games[11, 12], adjusting text font through electronic readers[13, 14] (e.g., character size, colour, spacing between words), and replacing complex words with simpler synonyms[15]. Nevertheless, these efforts often exhibit one or more of the following shortcomings:

1. In scenarios demanding quiet, such as conferences and exams, the use of multimedia-assisted tools presents practical difficulties due to their limited applicability.
2. Reliance on manual annotating. Converting text descriptions into videos or games manually can be both expensive and non-real-time.
3. Simple synonym substitution and rewriting can sometimes alter the original meaning and compromise the aesthetic qualities of the original texts (e.g., emotions, rhymes).

In comparison to the available knowledge on reading difficulties and the demonstrated capabilities of AI models, there are relatively few existing accessible designs addressing these challenges[16]. We have not yet discovered any existing reading assistance tools or research that has utilized or discussed how to integrate state-of-the-art AI techniques to address these issues.

Therefore, to fill these gaps, we propose an AI-based presentation method to assist individuals with dyslexia in reading online texts. We introduce LARF (**L**et **AI** **R**ead **F**irst), a reading assistance software application that uses OpenAI's GPT-4[17] to annotate important information in texts with highlights, bolding, underlining, and other marks. This approach aims to help readers focus more easily on key content, thereby improving their reading performance and experience. Unlike direct AI-generated summaries, LARF's design of annotating the original text preserves the maximum amount of original textual information. In addition, LARF supports both user-customized and default processing modes, allowing it to function as a real-time reading aid as well as an economical tool for pre-processing texts, thereby accommodating a variety of application scenarios.

Our main hypothesis is that LARF can improve the overall reading performance and experience of individuals with dyslexia, including those diagnosed and those self-reporting significant reading difficulties. Consequently, we conducted a randomized controlled trial in a non-clinical setting to evaluate this hypothesis.

We first designed Experiment 1, dividing 150 individuals with dyslexia (N=150) from the UK and USA into three groups. These groups included direct reading of original materials, using a conventional tool (i.e., Bionic Reading), and using LARF-annotated texts. We test the accuracy in recalling and retrieving details, and comprehension level through multiple-choice. The experimental results show that, compared to the traditional method and control groups, participants reading LARF-annotated texts displayed better recall and retrieval performance. Participants are also asked to complete a series of Subjective Evaluations to investigate user experience with LARF or the conventional tool. The results indicate that LARF-annotated texts significantly improved perceived user-friendly, overall satisfaction, perceived helpfulness, future-use and recommendation tendencies. Users also believed this method should be applied as a text expression method for dyslexic populations in more scenarios (e.g., exams, designing accessible websites). In Experiment 2, we further conduct a conceptual replication with general participants from the USA (N = 160). We observe similar but slicer results which indicate that LARF is also broadly applicable to the general population.

As dyslexia is a spectrum disorder, people with dyslexia often experience different subsets of challenges[18], we also construct a series of post hoc evaluations to assess the effects of LARF on different symptoms of reading disabilities.

Furthermore, most existing research on reading assistance design for dyslexia primarily focuses on English and Spanish[12]. However, GPT-4's multilingual capabilities open up possibilities for LARF to serve individuals with reading disabilities in other language environments. So we refine LARF based on feedback from the first two experiments and constructed a follow-up study to discuss the use of LARF in ecologically valid, non-English environments. We provide the software demo of LARF to a population in non-English speaking regions (China) suffering from both ADHD and dyslexia and collect their opinions and hopes for LARF and future reading assistance software.

In summary, this paper has the following contributions:

- Proposing LARF, a novel AI-based presentation method and tool that enhances text readability for dyslexic individuals.
- Empirical evaluation of Reading Assistance: We comprehensive randomized controlled trials to empirically evaluate the effectiveness of LARF. These trials involve both dyslexic and typical readers, providing a broad and diverse assessment base.
- Providing insights for Future Reading Assistance Technologies: This work provides insights for the HCI community and accessibility designers focusing on individuals with dyslexia and other related neurodiversity

populations: future work and research can concentrate on a series of tasks involving the use of AI and language models for text annotation and presentation. A more detailed discussion is available in the Discussion section.

2 Related Work and Background

2.1 Reading Assistance Tools for Dyslexia

In the realm of accessible design interventions to alleviate reading difficulties, myriad solutions have been proposed. A popular trend has been to incorporate text-to-speech conversion[10], enabling individuals with reading difficulties to access written content orally. Parallelly, innovative efforts have been made to employ multimedia elements such as videos and games to facilitate reading comprehension[11]. However, these software solutions often face environmental constraints. In scenarios necessitating quietness, such as conferences or exams, the use of text-to-speech conversion is impractical. Moreover, despite proven effectiveness[19, 20], the high cost of software like Kurzweil3000 limits its widespread adoption [19]. Furthermore, traditional methods of transforming textual information into images, audio, or even games require substantial involvement from experienced annotators, developers, and designers. This significantly escalates costs and eliminates the possibility of real-time use, thus further restricting its application scenarios.

The other trend is using adjustable text presentation, allowing for modifications in character size, colour, and word spacing[13, 14]. Santana et al.[21] created Firefixia, which is a Mozilla Firefox extension that enables dyslexic readers to tailor websites for enhanced readability. Dickinson et al.[22] enlisted 12 dyslexic students to evaluate diverse attributes like colours, sizes, spacings, column widths, and letter highlights in MS Word documents, aiming to enhance subjective readability. Optimal parameters were refined based on feedback from seven dyslexic individuals, resulting in noticeably improved readability. The outcomes were incorporated into the SeeWord tool for MS Word[23]. Text4All[24], an online service for web pages, and the Android IDEAL eBook reader4 for e-books are customization tools informed by prior eye-tracking research on dyslexic individuals[13]. Text4All extends its offerings to include medical language adaptation, terminology annotation, and language analysis. Currently, a popular method called Bionic Reading[25] revises texts so that the most concise parts of words are highlighted. This guides the eye over the text and the brain remembers previously learned words more quickly. While these methods can be applied in a broader range of contexts, they treat all text as a uniform entity, lacking targeted emphasis on key segments such as definitions or summary sentences. This results in substantial room for improvement in enhancing reading performance and experience.

In another approach, complex words have been substituted with simpler synonyms to aid comprehension [15]. However, such an approach can not only fail to guarantee accuracy in the context of substitution (i.e., it may completely distort the original intent of the text), but these alterations might also impact the literary attributes of the text, such as emotional intensity and rhythm.

In juxtaposition with the plethora of knowledge surrounding reading difficulties, the traditional accessible designs addressing these challenges appear scant[16]. Given the swift advancements in artificial intelligence (AI), the incorporation of AI models displaying superior reading comprehension and creativity into the accessible design is an avenue warranting further exploration. As these models become increasingly versatile and powerful, the intersection of AI and accessible design presents a promising opportunity to overcome the limitations of current solutions.

2.2 Language Models

In recent years, the rapid development of artificial intelligence (AI) technology has been evident [6], with the advancement of natural language processing (NLP) tasks, specifically language models (LMs) [26, 27, 28, 29], being particularly prominent.

Figure 1: Language models probabilistically generate text (a completion) based on given text (a prompt).

Despite the intricate structures and profound mathematical foundations underlying language models, they can essentially function as a "simple" black box, as depicted in Figure 1, processing input text to generate appropriate output for the given task[7].

ChatGPT[17] published by OpenAI, is a successor of InstructGPT [30] with fine-tuned via Reinforcement Learning with Human Feedback (RLHF) [31] and has gained phenomenal and huge attention and widespread discussion not only in the NLP community but also around the globe.

In this paper, we use the GPT-4, a similar model to ChatGPT launched by OpenAI, as a tool to generate content that is friendly to individuals with reading challenges. We also undertake a concise introduction about certain fundamental task Question Answering (QA) in NLP and how to deploy our specific objective into it in the subsequent sections.

2.3 Question Answering Tasks

Question answering (QA) serves as a fundamental task in NLP, forming the foundation for numerous practical applications such as web search, chatbots, and personal assistants. The scope of QA is extensive, encompassing a wide range of question types and demanding various skills for answer retrieval. These skills include comprehensive language understanding knowledge integration, and reasoning abilities [32]

In recent years, academia has begun to perceive QA tasks more as a format[33]. There are This implies that we can utilize this QA format to enable LMs to accomplish summarization and key information extraction tasks[34]. In this paper, Our work can be understood as employing this QA format, posing questions to LMs and expecting LMs to return the starting and ending positions of words and sentences within the target texts. The method that lets LMs output our expected text is called "prompting".

2.4 Prompt for Large Language Models

The Prompt is a user input or a set of starting instructions designed to guide the models' response. Large Language Models (LLMs) including GPT-4 variant, use these prompts to generate contextually appropriate text[35]. By processing the prompts, these models infer the desired information and format required in the output. Each input word or phrase in a prompt contributes to adjusting the model's internal state, which directly affects its response. The complexity of prompts varies widely, from single-word prompts to complex sequences, all tailored to the required tasks[28]. A simple example was given by Fig. 2.

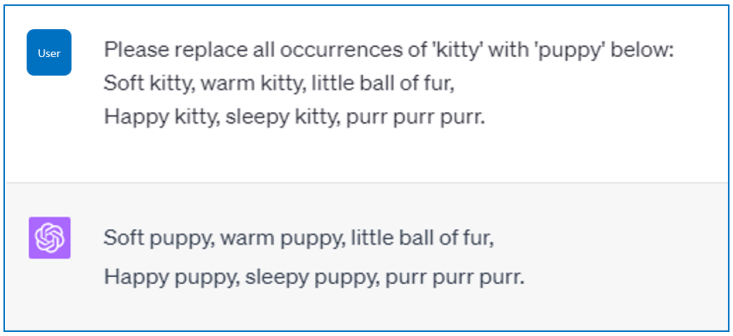


Figure 2: In the above example, the user asks GPT-4 to replace all occurrences of 'kitty' with 'puppy' in the text. The model completes the correct replacements for all the target text without any other unexpected changes. In this example, the prompt was 'Please replace all occurrences of "kitty" with "puppy" below.'

2.5 Hyper Text Markup Language (HTML)

In Hyper Text Markup Language (HTML) [36, 37], various tags can be used to manipulate the display of text, such as "bold," highlighting, italics, changing font colour, and adjusting font size. For instance, the "<b><\b>" tags can render text in bold format as shown in Figure 3. We can change the presentation of the text without modifying the textual content by incorporating the tags and displaying them in HTML.

In summary, this section introduces significant limitations and considerable room for improvement in the context of accessible design for Dyslexia. It outlines the approach of utilizing LLMs through prompting in the form of QA to obtain desired text outputs. In the next section, we will provide a more detailed explanation of the workflow of LARF, which employs the prompt engineering methodology to process the dyslexia-friendly text effectively.

HTML tag	Description	Example
<b>	Makes text bold	**Example**
<i>	Makes text italic	*Example*
<u>	Underlines text	Example
	Makes text strong(usually highlight)	Example

Figure 3: An example of the different HTML tags and their display result in HTML format. It is worth noting that in practical programming environments, the display of HTML tags can be highly diverse, including adjustments to font, text colour, highlighting colours, and so forth. The figure provides examples of only a few common default tags and their display results.

3 Method and Data

3.1 Workflow of LARF

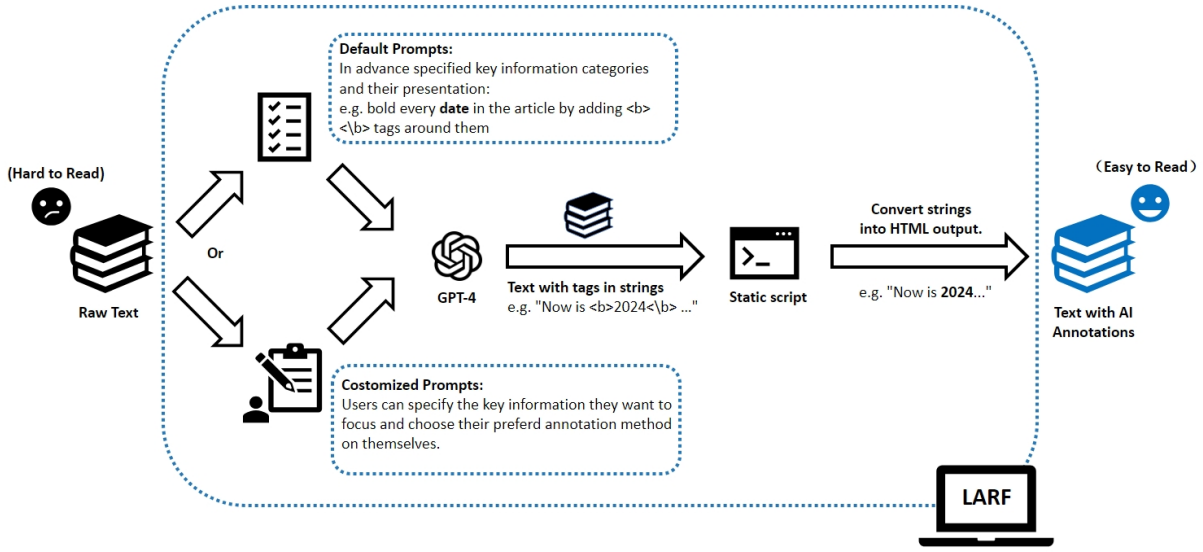


Figure 4: This flowchart explains the main process of the proposed method in this paper. GPT-4 processes the raw input text by combining the default or the customized prompts to generate a paragraph of text with HTML tags. Then A Python script automatically converts this string into an HTML file and outputs it.

The workflow of LARF is illustrated in Fig. 4, wherein the original text is input as a string. Guided by predetermined prompt combinations, GPT-4 processes the original text, incorporating the aforementioned HTML tags. The resulting text, containing HTML tags, is output as a string. Subsequently, a Python script compiles this HTML-tagged string into an HTML file, serving as the final output. Consequently, users receive a presentation in which specific information has undergone modifications such as bolding or highlighting, while the textual content itself remains entirely unaltered.

A simple example in Fig. 5 shows a segment extracted from Wikipedia. We prompt GPT-4 to highlight sentences that serve a summarizing role using <mark><\mark>tags. We also prompt it to bold important names of people and items using <b><\b>tags. After processing the output of GPT-4 with the same display scripts, we obtain the content with annotation.

When dealing with a large volume of text processing in advance, such as writing books or designing extensive web pages, default prompts can be used for batch processing. This significantly reduces the cost of manual annotation. LARF

Raw Text	GPT-4 Marked Text in HTML
BlackPink is a popular South Korean girl group consisting of members Jisoo, Jennie, Rosé, and Lisa. They are known for their energetic performances, diverse music styles, and fashionable image. With hits like "DDU-DU DDU-DU," and "How You Like That," BlackPink has gained global recognition and a strong fan following. They have become brand ambassadors and actively participate in charitable activities. BlackPink's unique charm and international influence have made them one of the most prominent groups in contemporary pop music.	**BlackPink** is a popular **South Korean girl group** consisting of members **Jisoo, Jennie, Rosé, and Lisa.** They are known for their energetic performances, diverse music styles, and fashionable image. With hits like **"DDU-DU DDU-DU"** and **"How You Like That"**, BlackPink has gained global recognition and a strong fan following. They have become brand ambassadors and actively participate in charitable activities. **BlackPink**'s unique charm and international influence have made them one of the most prominent groups in contemporary pop music.

Figure 5: The left-hand side is the raw text and the right-hand side is the annotation given by GPT-4 in HTML format.

can also serve as a browser or e-reader plugin, allowing users to make fine adjustments in different usage scenarios. For example, if users are reading an article about Blackpink, they can adjust the "key information " they want to focus on by selecting or typing "names of songs" and specifying how this information should be annotated. (Fig. 6).

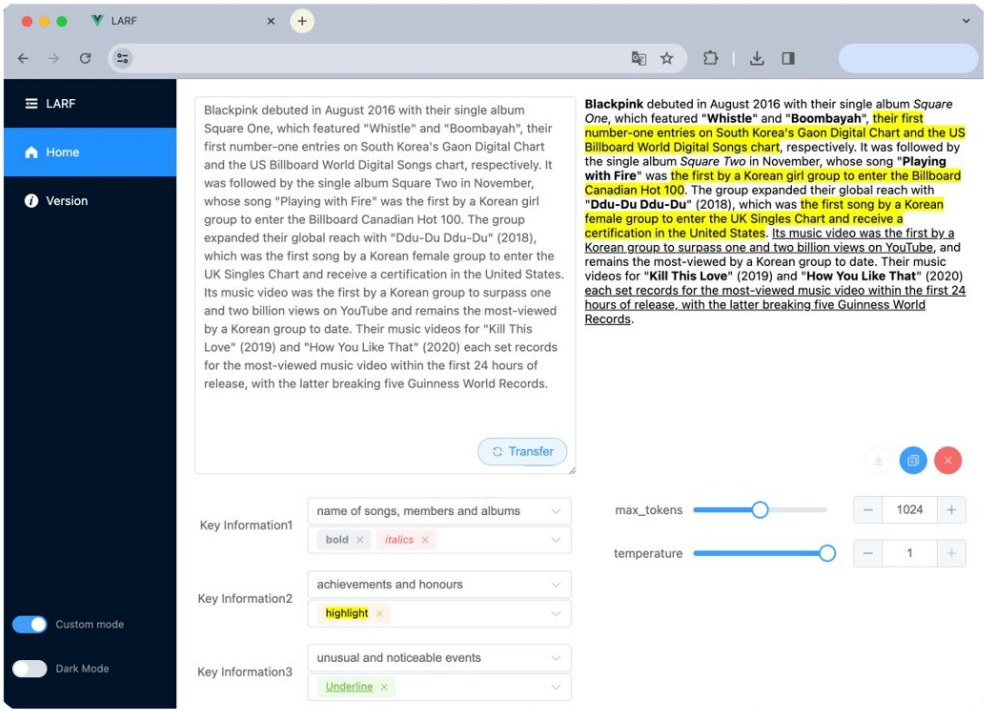


Figure 6: The demo of the custom mode of LARF software application on PC.

In subsequent experiments and practical applications, we adjust the prompts by using different labels, thereby modifying the presentation of the text.

3.2 GPT-4 Data

In Experiment 1 and Experiment 2, we processed the corpus using the default prompts of LARF based on the GPT-4 API. We also used prompt GPT-4 as judges to score the short-answer questions in subsequent experiments. The version of GPT-4: ChatGPT July 20 version. Specific prompts and generation logs can be found in the supplementary material.

3.3 BionicReading Data

We employed the BionicReading as a representative of conventional tools to process the corpora in subsequent experiments like fig. 7, as it is one of the most widely used reading performance improvement solutions. “Fixation” defines the expression of the letter combinations. It can be set as a value from 1 to 5. We use 3, which is also defined as the base value. With “Saccade” we can define the visual jumps from Fixation to Fixation. It can be set as a value from 10 to 50. In this paper, we also apply the default value of 10.

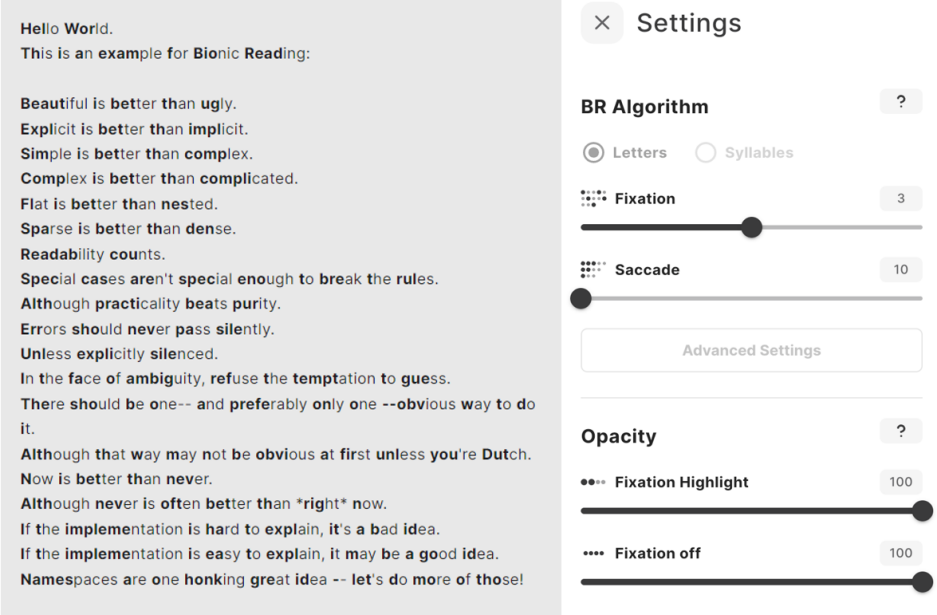


Figure 7: An example of the result and parameters of Bionic Reading

4 Ethic & Transparency

4.1 Ethics

The experiments involved in this work have been approved by the Institutional Review Board (IRB) of our affiliation. All participants in Experiment 1 and Experiment 2 completed informed consent forms and were recruited through the Prolific platform. All of the participants in Experiment 3 were from an ADHD community in China, who voluntarily participated in the software experience activity.

4.2 Transparency

All raw experimental data, GPT-4 processing steps, and data analysis code have been uploaded to an open-source GitHub repository. Criteria and results for the subjective scoring portion can also be found in the GitHub repository. The demo of LARF is also open to the public for free trials.

5 Experiment 1: LARF's Effectiveness among Dyslexia Participants

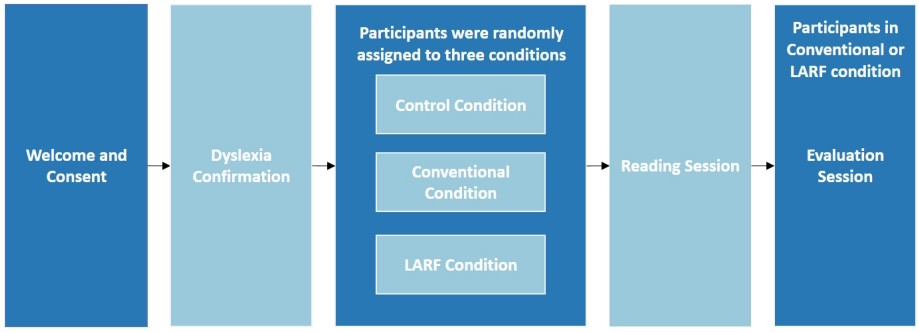
The primary objective of the first experiment is to validate our main hypothesis, namely, whether LARF compared to traditional methods (i.e., BionicReading) and control groups can improve and to what extent affect the reading performance and reading experience of individuals with dyslexia, on both objective measures and subjective evaluations.

5.1 Experiment Setup

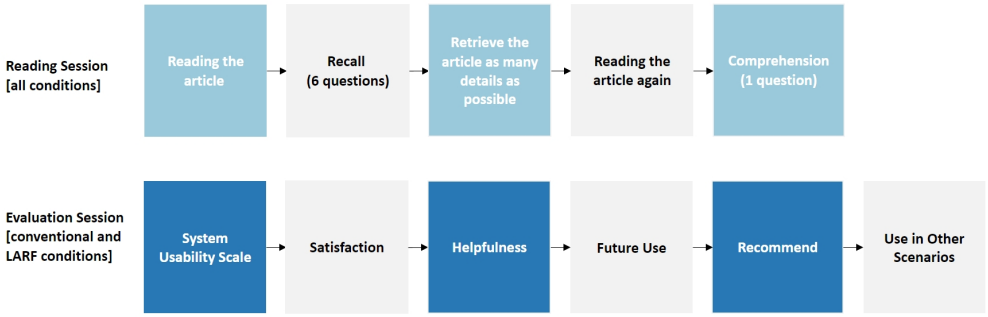
In this experiment, our primary focus lies on studying English, as it is the world's most spoken language and the third most spoken native language[38]. Large language models such as GPT-4 predominantly utilize English in their training data, and consequently, their performance is most proficient when processing English text[39]. Investigating the effectiveness of our proposed GPT-4-based annotation methods (i.e., LARF) within an English-language framework can streamline the selection process for both participants and corpora. This focused approach can potentially enhance the robustness and generalizability of the method under consideration.

We selected a descriptive and factual reading text from the IELTS Academic as the corpus for our study. This decision was motivated by the comprehensive nature of the IELTS Academic reading test, which employs a long-form format featuring texts sourced from books, journals, magazines, and newspapers[40]. We chose this test to scrutinize participants' reading efficiency for several reasons. Firstly, the descriptive and factual text provides comprehensive and

verifiable information about a subject[41]. Secondly, the IELTS Academic test is equipped with expertly formulated questions and standardized answers. These questions' careful design and standardization add another layer of reliability and validity to our study, making the IELTS Academic test an optimal tool for assessing adult reading performance. In particular, we chose Reading Test 115, Passage 2, "The Step Pyramid of Djoser," as our corpus.



(a) Experiment Procedure



(b) Session Details

Figure 8: The session and procedure of Experiment 1

5.2 Method and Experiment Procedure

We recruited our participants on Prolific[42]. Prolific is an online research platform that connects researchers with a diverse pool of participants for academic studies, surveys, and experiments. 150 individuals who have been medically diagnosed with dyslexia, are in the process of being diagnosed, or strongly suspect they have undiagnosed dyslexia from Prolific participated in our study ($M_{age} = 36.8$, 33.3% Female). Participants were randomly assigned to one of three experimental conditions: the control condition, the conventional tool condition, or the LARF condition. In the control condition, the article was presented without any modifications or annotations. In the conventional tool condition, modifications of the article were based on the Bionic Reading Method. Lastly, in the LARF condition, the article was annotated by LARF. Participants are not pre-informed that texts in the LARF group are annotated by GPT-4, nor is the rationale behind annotations (e.g., bolding of significant characters' names) disclosed. While this approach may potentially attenuate the effects of LARF, it mitigates the influence of psychological priming on the experimental outcomes.

The study began with participants completing a Dyslexia Checklist, designed to assess the severity of various reading-related challenges they face based on personal experiences. Afterwards, they read an article and answered a series of recall questions to evaluate their retention of key details, such as the main character's name and aspects of a described pyramid. Our design included six recall questions, alongside an attention check question (refer to Appendix, Table 2). Following the recall task, participants were asked to retrieve as many details from the article as possible. The article was then presented, immediately followed by a reading comprehension assessment on the same page.

After reading, participants in the control condition provided demographic information (age, gender, educational background) and concluded the experiment. Participants in the conventional tool and LARF conditions, however, also evaluated the modifications and annotations made by the respective tools. We initially used an adapted version of the

System Usability Scale[43] to assess tool usability. Participants then rated the perceived helpfulness and satisfaction with the tool, their intention to continue using it, and the likelihood of recommending it to others. Those in the conventional tool condition ended their participation after providing demographic information. In contrast, participants in the LARF condition were additionally queried about their preference for a personalized LARF tool before providing their demographic details. The experiment procedure is shown in Fig. 8a and session details are shown in Fig. 8a.

5.3 Result and Analysis

Attention Check

Of the initial 150 participants, 2 failed to pass the attention check and were consequently excluded from further analysis. The remaining 148 participants were included in subsequent analyses. There are 51 participants in the control condition, 49 in the conventional tool condition, and 48 in the LARF tool condition.

Dyslexia Checklist: Before reading the article, participants assessed their own dyslexia levels using the Dyslexia Checklist (see supplementary material for Dyslexia Checklist). This checklist comprises six items that evaluate various aspects: comprehension issues, word recognition difficulties, decoding difficulties, memory problems, attentional difficulties, and visual disturbance. We calculated the average scores from these items to determine each participant's overall dyslexia level (Cronbach's alpha = 0.91). Statistical analysis revealed no significant differences in dyslexia levels across the three conditions ($M_{control} = 3.80$, $SD = 1.65$, $M_{conventional} = 3.48$, $SD = 1.41$; $M_{LARF} = 3.49$, $SD = 1.29$; $F(2, 145) = .755$, $p = .472$), which indicated that participants are balanced among three conditions

Reading Time: We initially focused on the time participants took to read the passage before the recall section, using this as the primary measure of reading time. Ten participants were identified as outliers based on their initial reading times (defined as initial reading time $>$ Q3 + 1.5$\times$IQR or $<$ Q1 - 1.5$\times$IQR) and were thus excluded from this part of the analysis. Consequently, the final analysis on reading time was conducted with 138 participants. Covariates including education, age, gender, and dyslexia level were considered in the analysis. The one-way ANOVA showed no significant differences in reading times across conditions ($M_{control} = 117.56, SD = 46.47; M_{conventional} = 122.57, SD = 62.70; M_{LARF} = 118.26, SD = 50.69; F(2, 135) = .160, p = .853$). However, considering the first corpus' length of 388 words, and average reading speeds of 238 words per minute for English readers[44], those who spent less than 30.2 seconds (0.05 quantile) are considered relatively impatient. Fig. 9(a) shows that although there were no significant differences in reading times, participants using LARF did not fall below 30 seconds and was concentrated within a shorter, reasonable range. This suggests that LARF may aid in attracting user attention and enhancing reading patience and confidence.

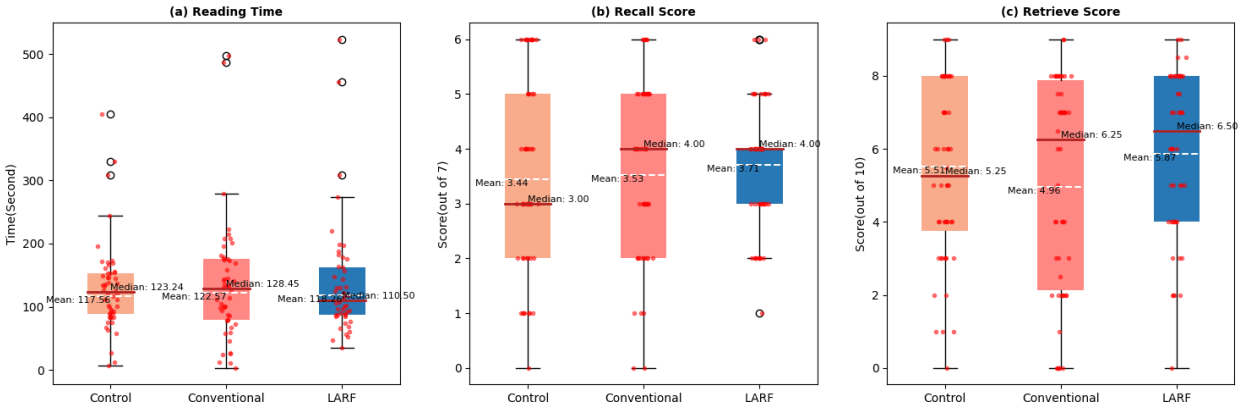


Figure 9: (a) shows the differences in reading time under three different conditions. Though the pattern is not significant, we can observe that users in the LARF group do less 'glance over and skip the article.' Furthermore, their overall reading time is more concentrated in areas with shorter durations. (b) and (c) respectively represent the scores of users in the retrieve and recall phases. It can be observed that compared to other groups, participants reading the LARF-marked texts exhibit better recall ability (marginally significant) and a superior capability to remember the details of the articles (significant).

Recall Performance: In the recall section, participants were asked six questions, earning one point for each correct answer. For this part of our analysis, we included education, age, gender, dyslexia level, and reading time as covariates in the one-way ANOVA analysis. The result Fig. 9b reveal that participants in the LARF condition tended to score

higher than the other two conditions, though the difference was not statistically significant ($M_{control} = 3.44, SD = 1.74, M_{conventional} = 3.53, SD = 1.64; M_{LARF} = 3.71, SD = 1.20; F(2, 128) = .215, p = .807$).

Retrieve Performance: Following the recall section, participants were instructed to retrieve as many details from the article as possible. We employ GPT-4 to evaluate the quality of participants' retrieval performance, utilizing a scoring range of 0 to 10. The assessment scores of GPT-4 for 148 participants underwent verification by two human reviewers, each of whom independently cross-checked the scores. The reviewers made only one significant correction to the scores, which was clearly erroneous. The scoring criteria and records employed by GPT-4 are available for reference in the supplementary materials and the GitHub repository. Additionally, GPT-4 provided the reason for the assigned retrieval performance scores. A similar one-way ANOVA analysis was conducted. The results in Fig. 9c clearly show a significant difference across three conditions ($F(2, 128) = 3.465, p = .034$). Participants in the LARF condition ($M_{LARF} = 5.87$, $SD = 2.30$) scored higher than the other two conditions ($M_{control} = 5.51$, $SD = 2.38$, $M_{conventional} = 4.96$, $SD = 2.89$).

Comprehension Performance: In our study, the final objective measure was reading comprehension, assessed using a method analogous to that employed in the IELTS examination. Participants were required to identify the correct two statements out of six that were presented in the article. To ensure accuracy in scoring, participants selecting more than two statements were automatically assigned a score of zero, as per our predefined criteria that only two statements were correct.

Our analysis revealed that 72.92% (35/48) of participants in the LARF condition correctly chose the exact two statements. In contrast, this accuracy was observed in 64.71% (33/51) of participants in the control condition and 67.35% (33/49) in the conventional condition. Additionally, we evaluated whether participants were able to identify at least one correct statement. In this regard, 100% (48/48) of participants in the LARF condition succeeded in choosing at least one correct statement, whereas the corresponding figures were 92.16% (47/51) for the control condition and 87.76% (43/59) for the conventional condition. We conservatively believe that this indicates LARF can to some extent enhance the participants' reading comprehension skills.

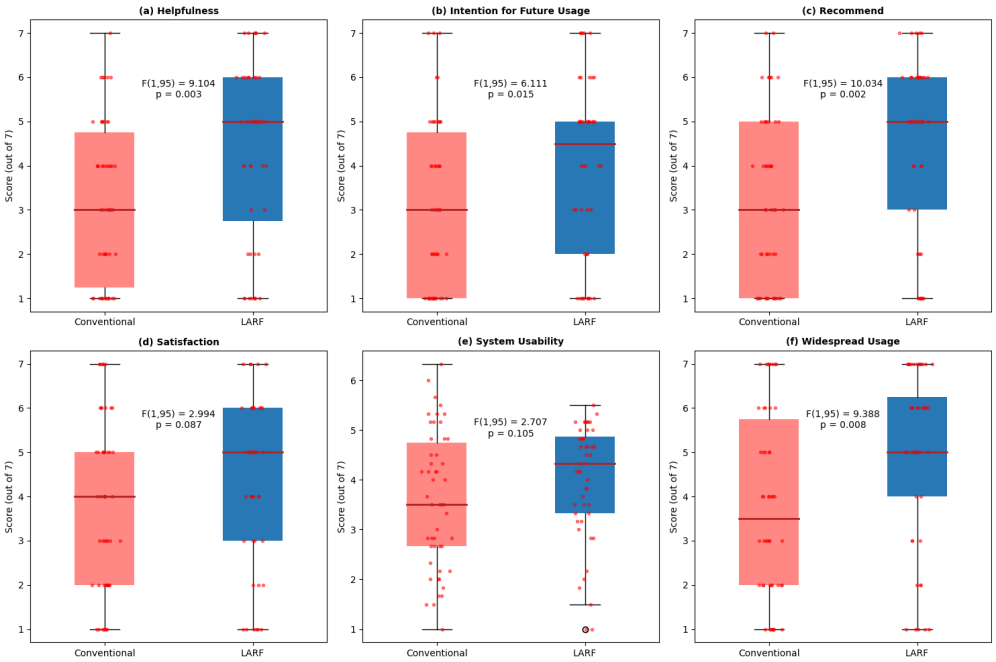


Figure 10: The subjective evaluation result. Participants with dyslexia exhibited a clear preference for LARF, considering text annotated with LARF to be effective, user-friendly, and worthy of broader adoption in various contexts.

Subjective Evaluation: We conducted a separate analysis to compare the subjective evaluations of the annotation tools between the conventional and LARF conditions. The questionnaire items and corresponding results are presented in Fig. 10. Overall, participants in the LARF condition rendered more favourable evaluations than those in the conventional condition. Notably, participants exposed to LARF-generated annotations reported more positive perceptions and future behaviour tendencies across multiple dimensions. These included system usability

($M_{conventional} = 4.09, SD = 1.42; M_{LARF} = 4.43, SD = 1.36; F(1, 95) = 1.469, p = .229$), satisfaction of the tool ($M_{conventional} = 3.76, SD = 1.92; M_{LARF} = 4.42, SD = 1.84; F(1, 95) = 2.994, p = .087$), perceived helpfulness ($M_{conventional} = 3.14, SD = 1.76; M_{LARF} = 4.29, SD = 1.99; F(1, 95) = 9.104, p = .003$), intention for future usage ($M_{conventional} = 2.94, SD = 1.89; M_{LARF} = 3.92, SD = 2.01; F(1, 95) = 6.111, p = .015$), recommend ($M_{conventional} = 3.18, SD = 1.87; M_{LARF} = 4.42, SD = 1.97; F(1, 95) = 10.034, p = .002$), and widespread usage ($M_conventional = 3.69, SD = 2.10; M_{LARF} = 4.96, SD = 1.96; F(1, 95) = 9.388, p = .003$).The detailed questions and results are shown in Table **??** and Table 4 in Appendix.

Specifically, participants in the LARF condition expressed a favourable inclination towards customizing the LARF tool. This preference was quantitatively reflected, with the mean score for the desire to customize LARF being 5.04 (SD = 1.41). The specific subjective scale questions and data can be found in supplementary material.

5.4 Post Hoc Evaluation

Given the diverse severity of dyslexia experienced by participants, which results in distinct challenges in reading, we conducted a post hoc evaluation to assess LARF's efficacy across varying degrees and categories of reading difficulties. Based on previous research, we calculated the Mean 1SD for each dyslexia item. Participants whose self-reported dyslexia scores are higher than M + 1SD were identified as having severe dyslexia, while those less than M – 1SD were identified as having mild dyslexia. Fig. 17 in Appendix show that LARF tended to improve the recall, retrieval, and comprehension performance for those with severe dyslexia.

6 Experiment 2: LARF Effectiveness among General Population

Building on the insights from Experiment 1, which focused on dyslexic participants, Experiment 2 expands the scope of our investigation to the general population. This experiment aims to evaluate the impact of text modifications and annotations generated by either conventional tools (i.e., the Bionic Reading Method) or the LARF tool on participants without dyslexia. We aim to understand how the effectiveness of these tools varies across different reader profiles. Similar to the previous experiment, we assess both objective measures (e.g., reading speed, efficiency, and completion rates) and subjective evaluations (e.g., perceived tool helpfulness, and future usage). We recruited 160 participants from Prolific, ensuring a diverse sample representative of the general population.

6.1 Method and Experiment Procedure

The procedure in Experiment 2 mirrors that of Experiment 1 to maintain methodological consistency. 160 individuals on Prolific participated in our study ($M\_age = 40.6, 48.13\%$ Female). Participants were randomly assigned to one of three experimental conditions: control condition, conventional tool condition, or LARF annotation condition. We used the same article in Experiment 1. Initially, participants were presented with a Dyslexia Tendency Adult Checklist [7], a tool designed to assess dyslexia tendencies based on personal experiences. Subsequently, they were in similar reading sessions and evaluation sessions as participants in Experiment 1.

6.2 Result and Analysis

Dyslexia Adult Checklist: The mean dyslexia score for the sample was 32.68, with a standard deviation of 8.17. Given the relatively small sample size, which means only 14 participants exhibited symptoms aligned with mild, moderate, or severe dyslexia according to the criteria outlined in the Adult Dyslexia Test. We opted to treat the dyslexia score as a continuous variable in subsequent analyses.

Reading Time: In a manner akin to Experiment 1, we conducted a statistical analysis of the time expended by users during their initial reading phase. Out of 148 participants, 10 were excluded based on this criterion, leaving 138 for subsequent analysis. Mirroring the approach in Experiment 1, we controlled for variables including education, age, gender, and level of dyslexia. A one-way ANOVA was performed to compare reading times across conditions. The result indicate no significant differences ($M_{control} = 104.33, SD = 58.38, M_{conventional} = 125.94, SD = 70.25; M_{LARF} = 109.22, SD = 62.82; F(2, 129) = 1.510, p = .225$), between conditions. However, the same pattern can be observed that participants in the LARF group were concentrated within the shorter time intervals and also with a lower median reading time.

Recall Performance: Participants were asked seven questions in the recall section this time, earning one point for each correct answer. For this part of our analysis, we included education, age, gender, dyslexia level, and reading time as covariates in the one-way ANOVA analysis. The result revealed that participants in the LARF condition scored

marginally higher than the other two conditions. However, the difference was not statistically significant ($M_{control} = 5.26, SD = 1.47, M_{conventional} = 5.15, SD = 1.89; M_{LARF} = 5.37, SD = 1.74; F(2, 128) = 1.016, p = .365$).

Retrieve Performance: We used the same method as Experiment 1 to judge participants' retrieve performance by GPT-4. Based on GPT-4's generated explanations for each score, we were able to identify and exclude participants who appeared to have cheated or provided random responses in the retrieval section. Out of the initial 148 participants, 10 were consequently disqualified, resulting in a final sample of 138 participants for the performance analysis.

A similar one-way ANOVA analysis was conducted. The results showed that participants in the LARF condition ($M_{LARF} = 7.67, SD = 1.64$) scored higher than the other two conditions ($M_{control} = 7.10, SD = 2.17; M_{conventional} = 7.20, SD = 2.31$). But the differences are not significant, either in the base model ($F(1, 135) = .985, p = .376$) and control model (i.e., controlling for dyslexia score, education level, age, and gender; $F(1, 128) = .480, p = .620$).

Comprehension Performance: The final measure for objective outcomes in our study was reading comprehension, evaluated using questions analogous to those in the IELTS examination. Our analysis did not reveal statistically significant differences in reading comprehension performance across three conditions ($M_{control} = 4.20, SD = 2.03; M_{conventional} = 4.02, SD = 1.84; M_{LARF} = 4.12, SD = 1.51$), neither in the base model ($F(2, 145) = .127, p = .881$) nor in the control model ($F(2, 138) = .288, p = .750$).

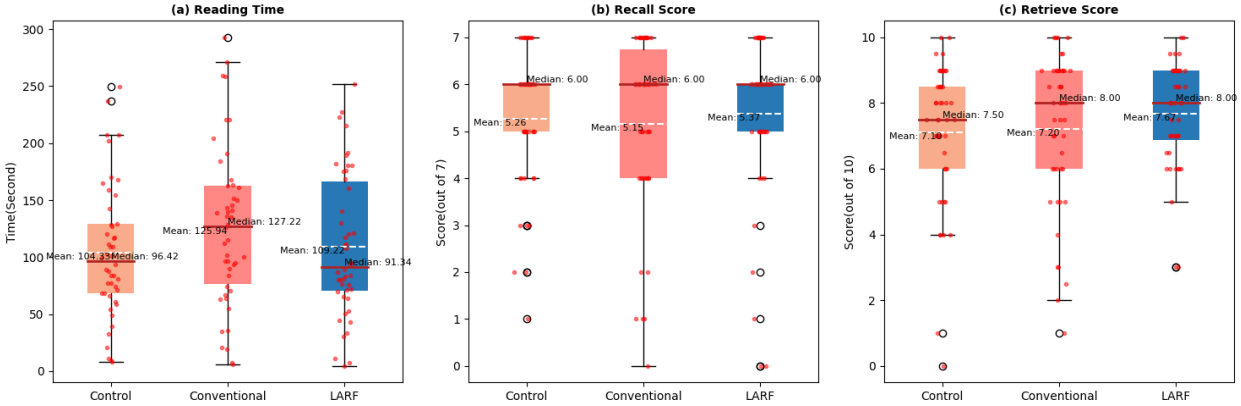


Figure 11: Fig. 12 (a) illustrates that there is no significant difference in performance between the LARF group and the control group in typical readers, despite the median time spent by the LARF group being notably lower compared to the other two groups. (b) and (c) respectively reflect the performance of the general population in the recall and retrieve tasks. Overall, in typical readers, the improvement in the LARF group is not as significant as the improvement observed in individuals with dyslexia in Experiment 1.

Subjective Evaluation: Similar to Experiment 1, we conducted a separate analysis to compare the subjective evaluations of the annotation tools between the conventional and LARF conditions. The questionnaire items and completed corresponding results are presented in the supplementary material. A selection of some prominent phenomena is shown in Fig. 12. Overall, participants in the LARF condition rendered more favourable evaluations than those in the conventional condition, replicating the findings in Experiment 1. Notably, participants exposed to LARF annotations reported significantly more positive perceptions and future behaviour tendencies across multiple dimensions. These included system usability ($Mconventional = 4.33, SD = 1.45; MLARF = 4.87, SD = 1.39; F(1, 97) = 3.619, p = .060$), perceived helpfulness of the annotations ($Mconventional = 3.29, SD = 2.31; MLARF = 4.70, SD = 1.83; F(1, 97) = 11.419, p = .001$), outcomes of the annotation integration ($Mconventional = 3.74, SD = 1.95; MLARF = 4.58, SD = 1.90; F(1, 97) = 4.795, p = .031$), confidence in reading ($Mconventional = 4.59, SD = 2.19; MLARF = 4.86, SD = 1.98; F(1, 97) = 4.527, p = .036$), and intention for future usage ($Mconventional = 2.98, SD = 2.13; MLARF = 3.86, SD = 1.99; F(1, 97) = 4.527, p = .036$).

In summary, we did not observe a direct significant improvement in reading performance with LARF in the general population from the result of the objective evaluation. However, in the subjective evaluation, participants shows a positive perception of LARF, suggesting a favourable reading experience for LARF users.

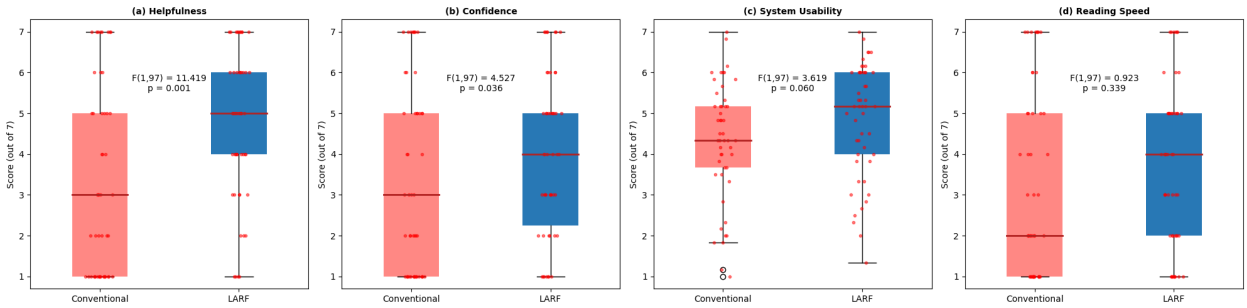


Figure 12: The subjective evaluation shows typical readers also hold a significant positive attitude to LARF (e.g. participants believe LARF can help them read faster and is easy to use).

6.3 Post Hoc Evaluation

We conducted similar post hoc evaluations as in Experiment 1. The results in Fig. 13 and Table. **??** show that LARF was more effective for participants with higher dyslexia tendencies when they were doing retrieval tasks and reading comprehension tasks. We coded the control condition as a baseline and centred the dyslexia scores around their mean. We controlled for demographic variables and time spent on the retrieval task to isolate the impact of our experimental conditions. Results also show that the interaction effect between the LARF condition and dyslexia is both positive and statistically significant (b = .111, SE = .051, t = 2.182, p = .031). This demonstrates that the negative impact of dyslexia on retrieve performance is mitigated in the LARF condition relative to the control condition. In simpler terms, when dyslexia scores are high, being in the LARF condition (vs. the control condition) exerts a positive influence on retrieve performance.

We didn't replicate this effect when we considered recall performance, which may be because the recall questions from IELTS are too easy for typical readers whose first language is English. However, when we compare users with recall scores above the average score, we observe that participants in the LARF group spent significantly less time than those in the Control and Conventional groups, $F(2, 70) = 3.914, p = .024$).

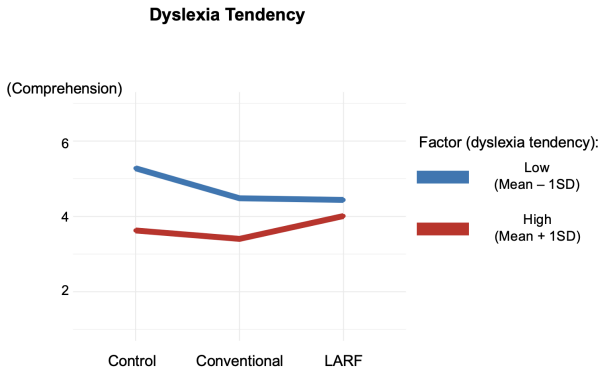


Figure 13: LARF significantly improved reading comprehension in participants self-reporting higher dyslexia tendencies.

Another interesting finding is Participants in the LARF condition demonstrated a markedly higher completion rate ($P_{LARF} = 98\%$) compared to those in the conventional condition ($P_{Conventional} = 85.7\%$) and the control condition ($P_{control} = 91.8\%$). A Fisher's exact test indicated that the difference in completion rates among the three conditions approached statistically significant, with a p-value of 0.069. This phenomenon illustrates that despite no additional requirement to treat this survey seriously, users in the LARF group still exhibited greater patience, aligning with the results of subjective evaluation where users perceived that LARF could enhance their reading confidence and patience.

Table 1:
UNSTANDARDIZED COEFFICIENTS OF
THE MODEL FOR THE RESULT OF EXPERIMENT

	Retrieve Performance
(Intercept)	6.005*** (1.937)
Conventional	-.283 (.393)
LARF	.145 (.403)
Dyslexia	-.084** (.033)
Conventional × Dyslexia	-.011 (.050)
LARF × Dyslexia	.111**(.051)
Age	.008 (.013)
Gender	-.585* (.326)
Education Level 2	1.370 (1.947)
Education Level 3	1.330 (1.919)
Education Level 4	1.234 (1.965)
Education Level 5	.888 (2.098)
Retrieve Time	.005*** (.001)

Notes:
(1) Standard errors are in parentheses;
(2) *p <0.1, **p <0.05, ***p <0.01

7 Follow-up Experiment

Based on the results of the first and second experiments, we found that LARF is particularly beneficial for individuals who struggle to maintain attention during reading and face difficulties in comprehension and memory retention. Given the high comorbidity of Attention Deficit Hyperactivity Disorder (ADHD) with dyslexia [45] and most dyslexia-related inclusive designs have primarily focused on English[12], we were curious about whether GPT-4's language transfer capabilities could enable LARF to be effective across languages. For these reasons, we conducted a small-scale follow-up study in a real-world, non-English-speaking context. We openly invited participants from an ADHD community in China ($N = 14, meanage = 23, 35.7\% female$) who self-reported reading difficulties to use LARF. They were asked to provide feedback on their experience with LARF through a questionnaire.

7.1 Method and Experiment Procedure

After completing the Dyslexia Checklist, identical to that used in Experiment 1, participants freely utilized LARF with both the Chinese translation and the original English version of the text "The Step Pyramid of Djoser" (used in Experiments 1 and 2) available for selection. Subsequently, participants were involved in a series of questionnaires and discussions, including scales for feedback on LARF usage and open-ended suggestions.

7.2 Result and Analysis

Fig. 14 shows all participants report difficulty in maintaining focus while reading for extended periods, leading to easy distractions. And Fig. 15A significant proportion (64.3%, 71.4%, and 78.6%, respectively) noted substantial improvements in reading speed, efficiency, and ability to review key information with LARF, while 57% felt it enhanced their reading confidence.

Fig. 16 (b) shows 92.9% of the participants agreed that LARF-annotated texts could serve as an accessible presentation format in special education contexts, such as exams for learning disabilities like dyslexia and ADHD, and in accessible web design. This aligns with the findings from Experiment 1. Fig. 16 (c) shows that after informing LARF that the annotations were generated by GPT-4, 14.3% of users expressed distrust in AI(Fig 12. c). Although the proportion of users exhibiting distrust in AI technology is relatively small, this finding underscores the importance of addressing user trust in AI. This aspect could represent a significant area for investigation in future research.

In the qualitative study phase, we first gathered overall perceptions of LARF. Commonly cited views included the usefulness of customization features for focusing on desired information and diverse annotation methods enhancing reading patience and reducing re-reading. However, some negative feedback included the excessive brightness of yellow highlights, low accuracy in custom annotations, and a preference for modifications in font and spacing over annotations.

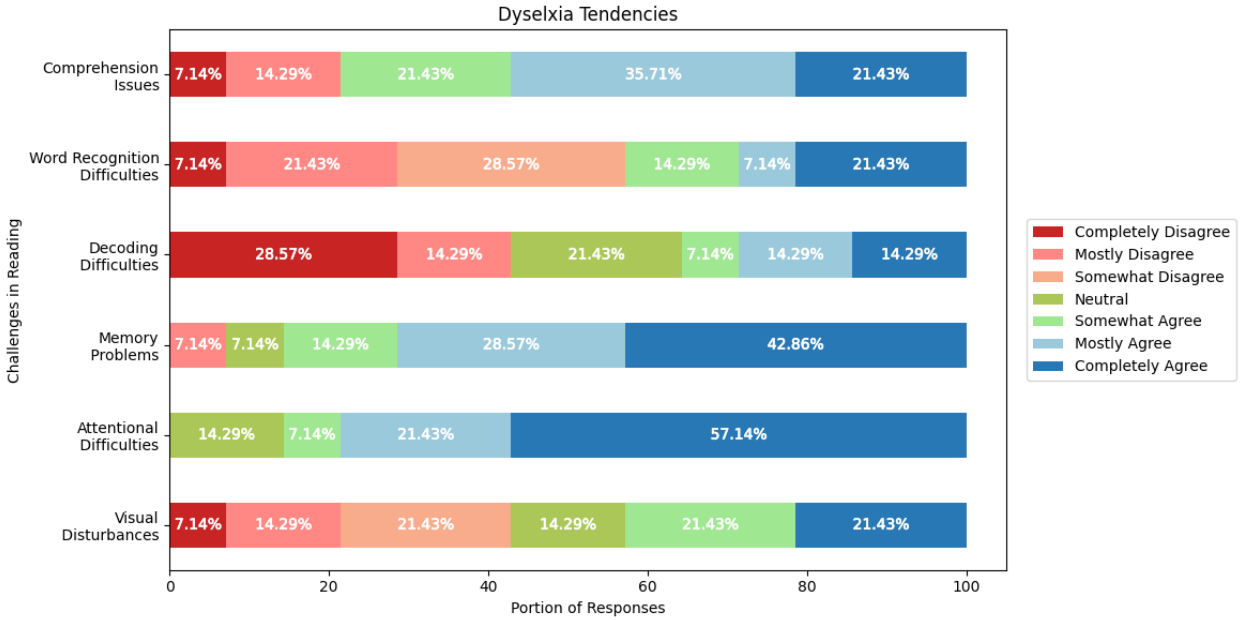


Figure 14: The Dyslexia Tendency Scale includes participants' self-reported tendencies towards reading difficulties. The scale shows that 100% of the participants reported difficulties with attention deficit during reading, while only 35.71% of participants reported struggles with decoding

Participants also provided development suggestions for LARF and expectations for future reading assistance software. Key recommendations included: 1. Clarifying the logic behind default annotations. 2. Expanding customization options (e.g., changing font colour, adding markings, adjusting reading background and line spacing). 3. Integrating AI-powered summarization of articles and paragraphs. 4. Extending functionality to PDF reading. These insights will be further discussed in the Discussion section, offering valuable directions for future development.

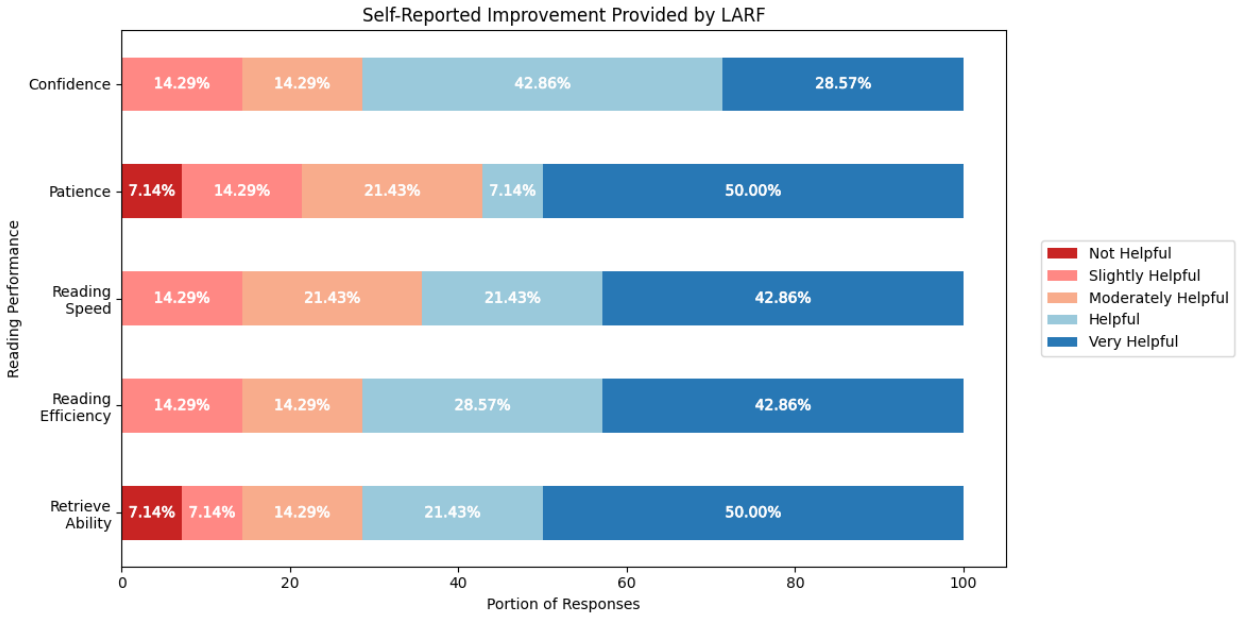


Figure 15: Follow upstudy on users' experiences and performance improvements in reading with LARF. Most users believe that LARF enhances their reading speed, patience, efficiency, and ability to recall details.

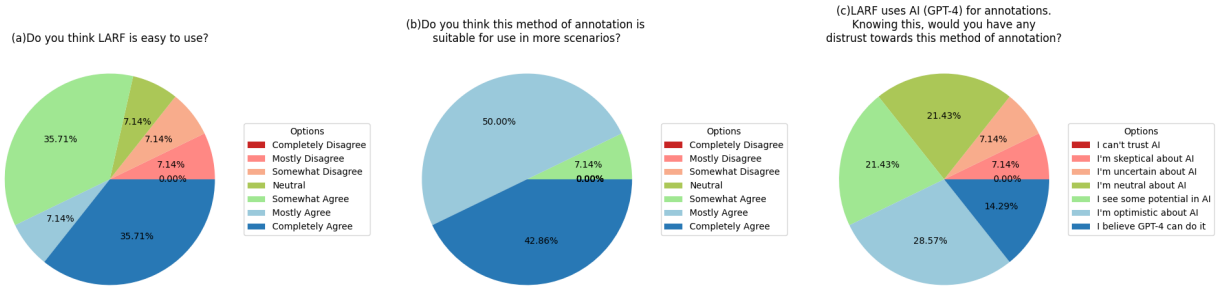


Figure 16: (a)Surveying the ease of use of LARF, reveals that 50% of users think LARF is easy to use. (b) Regarding participants' support for the extension of LARF to other scenarios, shows that 100% of the participants expressed a positive view. (c) is about the trust level in AI annotation tools among users after being informed that LARF is generated by GPT-4, with 14.3% of the users expressing scepticism towards AI

8 Discussions

8.1 Limitation

The focal point of this article is providing LARF, an AI-annotated presentation method for dyslexia, and evaluating it through a series of evaluations and user studies. During the experimental process and software development, the study encountered several limitations:

Limitation 1: Due to the limitation of participant numbers in online questionnaires and potential concentration issues in online surveys, the reading comprehension and recall tasks in the evaluation were not set to a high level of difficulty. This resulted in some evaluations observing only patterns but without significant outcomes. We did not construct a 'random labelling group' in this article to exclude the influence of the placebo effect, although intuitively such an effect should be small.

Limitation 2: This work primarily focuses on proposing and evaluating the method of using AI-annotated text in accessible design. To simplify software development, LARF's demo backend algorithm was based on prompts designed for the large predictive model GPT-4. The current version of LARF, functioning as a software application, still relies on the GPT-4 API and is thus subject to constraints related to network conditions and certain costs. In practical application, training or fine-tuning a smaller model specifically for this use case would be a more economical and effective approach.

Limitation 3: The study did not explore the interaction among different types of annotations (e.g., bolding, highlighting), nor did it investigate which annotation form would be most beneficial for users. Identifying optimal prompts for user engagement was also not examined. Adjusting the font size and colour is also achievable in HTML format, which is not discussed in this paper.

Limitation 4: In the field of accessible design for reading disabilities, more attention has been given to Synthetic languages like English and Spanish, while less focus has been placed on Analytic languages such as Chinese and Vietnamese, where the connection between script and pronunciation or meaning differs. Although the follow-up study tests LARF's performance in Chinese, the scale is limited due to the lower prevalence and diagnosis rates of dyslexia in China compared to the first two experiments.

8.2 Future Work

Building on the theoretical foundation provided in this paper, future research could be particularly valuable in the following areas:

In contemporary research, we believe Extract QA[46] can be an appropriate NLP downstream task for our AI annotations target. Reflecting on the effectiveness evidenced in this study, the development of NLP models capable of precisely annotating specific information in extensive texts emerges as a topic of potential significance. Furthermore, establishing a framework for assessing the accuracy, recall rates, inference speed, and other measurable metrics related to such models, accompanied by the formulation of pertinent benchmarks, might also constitute an area meriting academic exploration. For Limitation 3: Integrate more kinds of HTML formats and user-requested features such as 'help summarizing highlighted content', as mentioned in follow-up studies, presents considerable research potential for enhancing existing reading assistance methods. For limitation 4: The linguistic transfer capabilities of LLMs like GPT-4

offer novel potential and research value for the design of reading assistance solutions in regions with lesser-known languages. In regions like China, where dyslexia is less prevalent, collecting sufficient samples is still challenging. However, this also highlights the significant potential for AI and LLMs in helping to screen dyslexia and other neurodiversity conditions.

9 Conclusion

In this article, we introduce LARF, a pioneering approach employing AI-annotated text to enhance the reading abilities of individuals with dyslexia, addressing the limitations of traditional methods. Experiment 1 involved 150 participants with reported dyslexia, engaging in a series of subjective and objective evaluations. The study validated LARF's effectiveness in improving dyslexic readers' performance and experience, including recall of details, reading comprehension efficiency, and patience, while also outperforming the conventional technique (i.e. BionicReading). Subsequent post hoc evaluations and Experiment 2 extended the investigation to typical readers. Although the enhancement in reading skills was less pronounced in the mildly or non-dyslexic population, LARF still improved reading experience across the general population, indicating its broader application potential. In the discussion, we explore future directions for AI-annotated text-assisted reading within LARF, including the integration with NLP's Extractive QA tasks to develop smaller, faster, and more precise proprietary models. These insights contribute to the HCI community, delineating pathways for advanced assistive reading solutions.

References

[1] Zhichao Xia, Fumiko Hoeft, Linjun Zhang, and Hua Shu. Neuroanatomical anomalies of dyslexia: Disambiguating the effects of disorder, performance, and maturation. *Neuropsychologia*, 81:68–78, 2016.

[2] Robin L Peterson and Bruce F Pennington. Developmental dyslexia. *The lancet*, 379(9830):1997–2007, 2012.

[3] International Dyslexia Association. Frequently asked questions about dyslexia. `http://www.interdys.org/`.

[4] S Gunnel Ingesson. Growing up with dyslexia: Interviews with teenagers and young adults. *School psychology international*, 28(5):574–591, 2007.

[5] James H Smith-Spark, Lucy A Henry, David J Messer, Elisa Edvardsdottir, and Adam P Zięcik. Executive functions in adults with developmental dyslexia. *Research in developmental disabilities*, 53:323–341, 2016.

[6] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[7] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

[8] Katharina Galuschka, Ruth Görgen, Julia Kalmar, Stefan Haberstroh, Xenia Schmalz, and Gerd Schulte-Körne. Effectiveness of spelling interventions for learners with dyslexia: A meta-analysis and systematic review. *Educational Psychologist*, 55(1):1–20, 2020.

[9] Luz Rello, Clara Bayarri, Yolanda Otal, and Martin Pielot. A computer-based method to improve the spelling of children with dyslexia. In *Proceedings of the 16th international ACM SIGACCESS conference on Computers & accessibility*, pages 153–160, 2014.

[10] Kristen Laga, Daniel Steere, and Domenico Cavaiuolo. Kurzweil 3000. *Journal of Special Education Technology*, 21(2):79, 2006.

[11] Andres Larco, Jorge Carrillo, Nelson Chicaiza, Cesar Yanez, and Sergio Luján-Mora. Moving beyond limitations: Designing the helpdys app for children with dyslexia in rural areas. *Sustainability*, 13(13):7081, 2021.

[12] Mikel Ostiz-Blanco, Javier Bernacer, Irati Garcia-Arbizu, Patricia Diaz-Sanchez, Luz Rello, Marie Lallier, and Gonzalo Arrondo. Improving reading through videogames and digital apps: A systematic review. *Frontiers in psychology*, 12:652948, 2021.

[13] Luz Rello, Gaurang Kanvinde, and Ricardo Baeza-Yates. Layout guidelines for web text and a web service to improve accessibility for dyslexics. In *Proceedings of the international cross-disciplinary conference on web accessibility*, pages 1–9, 2012.

[14] Luz Rello and Ricardo Baeza-Yates. Good fonts for dyslexia. In *Proceedings of the 15th international ACM SIGACCESS conference on computers and accessibility*, pages 1–8, 2013.

[15] Luz Rello and Ricardo Baeza-Yates. Evaluation of dyswebxia: a reading app designed for people with dyslexia. In *Proceedings of the 11th Web for All Conference*, pages 1–10, 2014.

[16] Jacob E McCarthy and Sarah J Swierenga. What we know about dyslexia and web accessibility: a research review. *Universal Access in the Information Society*, 9:147–152, 2010.

[17] OpenAI. Chatgpt. `https://openai.com/`.

[18] Meredith Ringel Morris, Adam Fourney, Abdullah Ali, and Laura Vonessen. Understanding the needs of searchers with dyslexia. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2018.

[19] Jennifer Cullen, Sue Keesey, and Sheila R Alber-Morgan. The effects of computer-assisted instruction using kurzweil 3000 on sight word acquisition for students with mild disabilities. *Education and Treatment of Children*, pages 87–103, 2013.

[20] Robert A Stodden, Kelly D Roberts, Kiriko Takahashi, Hye Jin Park, and Norma Jean Stodden. Use of text-to-speech software to improve reading skills of high school struggling readers. *Procedia Computer Science*, 14:359–362, 2012.

[21] Vagner Figueredo de Santana, Rosimeire de Oliveira, Leonelo Dell Anhol Almeida, and Marcia Ito. Firefixia: An accessibility web browser customization toolbar for people with dyslexia. In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*, pages 1–4, 2013.

[22] Anna Dickinson, Peter Gregor, and Alan F Newell. Ongoing investigation of the ways in which some of the problems encountered by some dyslexics can be alleviated using computer techniques. In *Proceedings of the fifth international ACM conference on Assistive technologies*, pages 97–103, 2002.

[23] Peter Gregor, Anna Dickinson, Alison Macaffer, and Peter Andreasen. Seeword—a personal word processing environment for dyslexic computer users. *British Journal of Educational Technology*, 34(3):341–355, 2003.

[24] V Topac. The development of a text customization tool for existing web sites. In *Text Customization for Readability Symposium*, 2012.

[25] Bionic Reading. Bionic reading. `https://bionic-reading.com/`.

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

[27] Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021.

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

[29] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[30] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.

[31] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

[32] Anna Rogers, Matt Gardner, and Isabelle Augenstein. Qa dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension. *ACM Computing Surveys*, 55(10):1–45, 2023.

[33] Matt Gardner, Jonathan Berant, Hannaneh Hajishirzi, Alon Talmor, and Sewon Min. Question answering is a format; when is it useful? *arXiv preprint arXiv:1909.11291*, 2019.

[34] Maxime Peyrard. A simple theoretical model of importance for summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1059–1073, Florence, Italy, July 2019. Association for Computational Linguistics.

[35] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[36] Timothy J Berners-Lee and Robert Cailliau. Worldwideweb: Proposal for a hypertext project. 1990.

[37] Tim Berners-Lee, Robert Cailliau, Ari Luotonen, Henrik Frystyk Nielsen, and Arthur Secret. The world-wide web. *Communications of the ACM*, 37(8):76–82, 1994.

[38] Ethnologue. Languages of the world. (2022). `https://www.ethnologue.com/`.

[39] Queenie Luo, Michael J Puett, and Michael D Smith. A perspectival mirror of the elephant: Investigating language bias on google, chatgpt, wikipedia, and youtube. *arXiv preprint arXiv:2303.16281*, 2023.

[40] H Porter Abbott. *The Cambridge introduction to narrative*. Cambridge University Press, 2020.

[41] Michael Alley. The craft of scientific writing. Technical report, Springer, 1996.

[42] Prolific. Prolific - online participant recruitment for surveys and market research, 2023.

[43] John Brooke. Sus: a "quick and dirty'usability. *Usability evaluation in industry*, 189(3):189–194, 1996.

[44] Marc Brysbaert. How many words do we read per minute? a review and meta-analysis of reading rate. *Journal of memory and language*, 109:104047, 2019.

[45] Lauren M McGrath and Catherine J Stoodley. Are there shared neural correlates between dyslexia and adhd? a meta-analysis of voxel-based morphometry studies. *Journal of Neurodevelopmental Disorders*, 11(1):1–20, 2019.

[46] Ali Mohamed Nabil Allam and Mohamed Hassan Haggag. The question answering systems: A survey. *International Journal of Research and Reviews in Information Sciences (IJRRIS)*, 2(3), 2012.

A Tables

Table 2: Dyslexia Checklist

Term	**Scale**	**Description**
Understanding	1–7	To what extent do you have difficulty understanding the meaning of sentences or paragraphs, even if individual words can be recognized?
Recognition	1–7	To what extent do you struggle to correctly and fluently recognize letters and words, which can lead to slow reading speed and misinterpretation of words?
Memory	1–7	To what extent do you struggle to remember what has been read, especially understanding longer texts or story plots?
Decoding	1–7	To what extent do you have difficulty blending letters into words and understanding word pronunciation rules, affecting reading fluency and comprehension?
Attention	1–7	To what extent do you have difficulty maintaining focus while reading for an extended period, leading to easy distractions?
Visual Disturbance	1–7	How frequently do you encounter visual disturbances during reading, such as letters or words appearing distorted, jumbled, or overlapping?

Table 3:
SUBJECTIVE EVALUATION - 1

System Usability Scales		Mean (SD)		Statistics (F(1, 95))	p-value
Conventional	**LARF**	**Conventional**	**LARF**		
I believe that I would frequently like to read articles with these types of bold labels on certain occasions.	I believe that I would frequently like to read articles with these types of highlights, underlines, or bold labels on certain occasions.	3.35 (2.07)	3.77 (1.96)	1.073	p =.303
I think understanding these bold labels was not difficult for me.	I think understanding these highlights, underlines, or bold labels was not difficult for me.	3.96 (1.78)	4.31 (1.84)	.927	p =.338
I believe I would need the support of a technical person to read an article with these bold labels.[reversed-scale]	I believe I would need the support of a technical person to read an article with these highlights, underlines, or bold labels.[reversed-scale]	5.55 (1.62)	5.29 (1.86)	.538	p =.465
I found that the bold labels were well-integrated.	I found that the highlights, underlines, or bold labels were well-integrated.	3.63 (2.02)	4.23 (1.68)	2.500	p =.117
I would imagine that most people would learn to read with these bold labels very quickly.	I would imagine that most people would learn to read with these highlights, underlines, or bold labels very quickly.	3.96 (1.84)	4.65 (1.89)	2.543	p =.114
I felt very confident reading with the bold labels.	I felt very confident reading with the highlights, underlines, or bold labels.	4.06 (1.73)	4.40 (1.83)	.859	p =.356

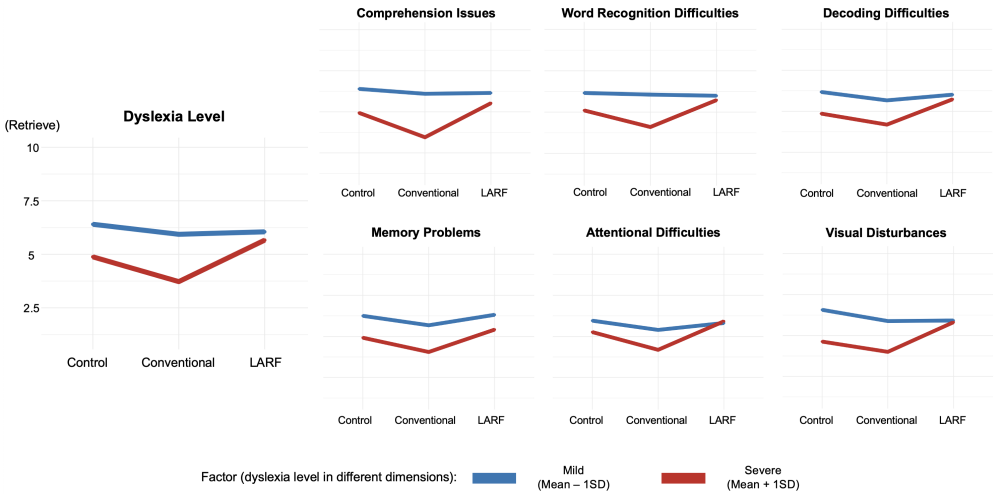
Notes:
(1) Standard errors are in parentheses;
(2) *p <0.1, **p <0.05, ***p
(3) SUS-3 is a reversed-scale question <0.01

Table 4:
SUBJECTIVE EVALUATION - 2

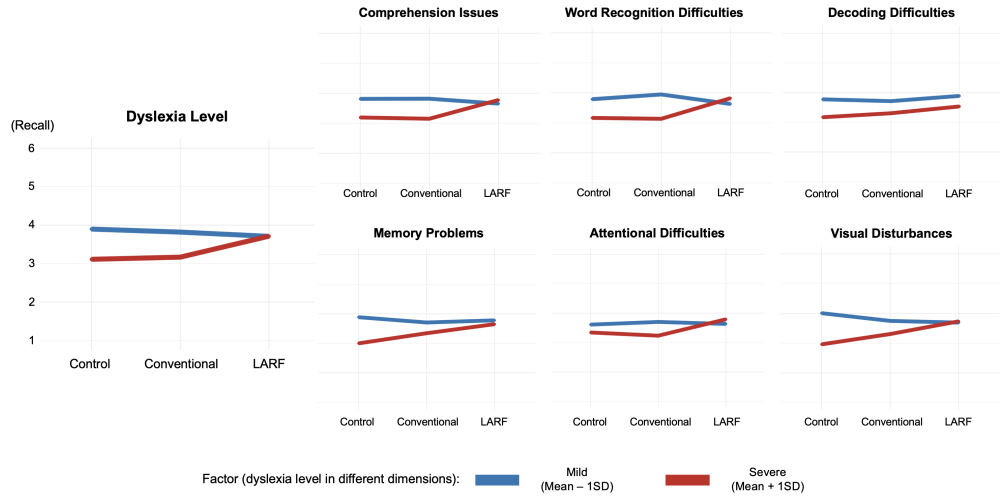
Metrics	Question	Mean (SD)		Statistics (F(1, 95))	p-value
		Conventional	LARF		
Satisfaction	What is your overall satisfaction with this kind of presentation (highlights, underlines, or bold labels annotations/bold labels annotations) when you read articles?	3.76 (1.92)	4.42 (1.84)	2.994	.087*
Helpfulness	To what extent do you think you will continue to use this kind of presentation (highlights, underlines, or bold labels annotations/bold labels annotations) in future reading?	3.14 (1.76)	4.29 (1.99)	9.104	.003**
Intention for Future Use	To what extent do you believe the marks in the articles helped you concentrate on the key information?	2.94 (1.89)	3.92 (2.01)	6.111	.015*
Recommendation	To what extent will you recommend this kind of presentation (highlights, underlines, or bold labels annotations/bold labels annotations) to others?	3.18 (1.87)	4.42 (1.97)	10.034	.002**
Intention for Widespread Usage	Do you think this kind of presentation (highlights, underlines, or bold labels annotations/bold labels annotations) is suitable for widespread use in other contexts? For example, in special exam papers for people with reading disabilities, integrated into e-reader, or for online academic paper reading?	3.69 (2.10)	4.96 (1.96)	9.388	.003**

Notes:
(1) Standard errors are in parentheses;
(2) *p <0.1, **p <0.05, ***p <0.01

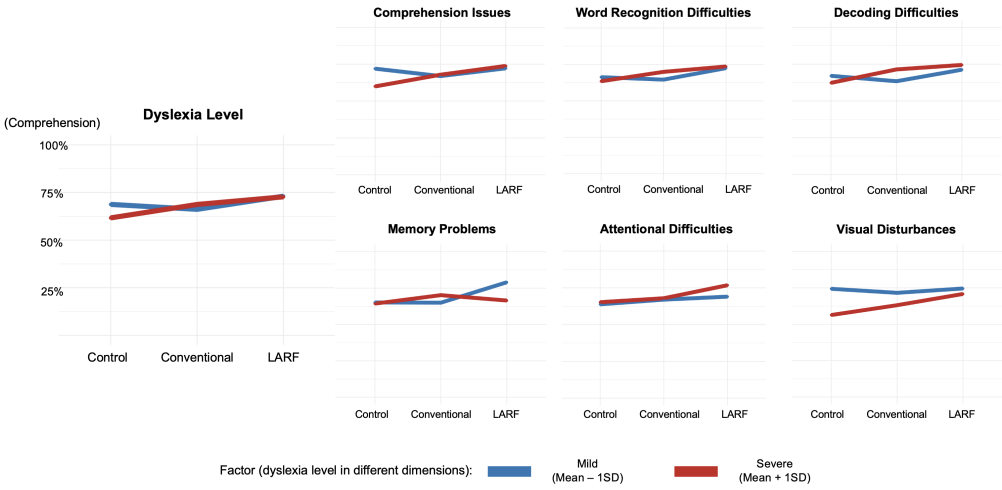
B Figures



(a) Post hoc evaluation for retrieving performance. The y-axis represents the scores for recall, with a maximum score of 10. While in the group with mild symptoms, LARF did not exhibit improvement, it significantly enhanced users' retrieval abilities in the group facing more severe reading challenges, whereas conventional tools had almost entirely negative impacts.



(b) Post hoc evaluation for recall performance. The y-axis represents the scores for recall, with a maximum score of 6. LARF similarly provided substantial assistance to the group with more severe symptoms, even surpassing the group with mild symptoms who also used LARF.



(c) Post hoc evaluation for recall performance. The y-axis represents the accuracy for reading comprehension. In the group with severe symptoms, LARF exhibited a consistent improvement compared to the control group and the conventional group.

Figure 17: The Post hoc evaluation for experiment 1. Given that individuals with dyslexia may encounter varying types and degrees of reading challenges, we categorized each symptom in the dyslexia checklist into "severe" and "mild". The red line depicted in the figure represents the performance of users facing more significant challenges in that specific