Let AI Read First: Enhancing Reading Abilities for Individuals with Dyslexia through Artificial Intelligence

SIHANG ZHAO, The Chinese University of Hong Kong, Shenzhen, China
SHOUCONG, XIONG, Zhejiang University, China
BO PANG, Chinese Academy of Science, China
XIAOYING TANG, The Chinese University of Hong Kong, Shenzhen, China
YUHANG ZHAO, University of Wisconsin-Madison, USA
PINJIA HE, The Chinese University of Hong Kong, Shenzhen, China

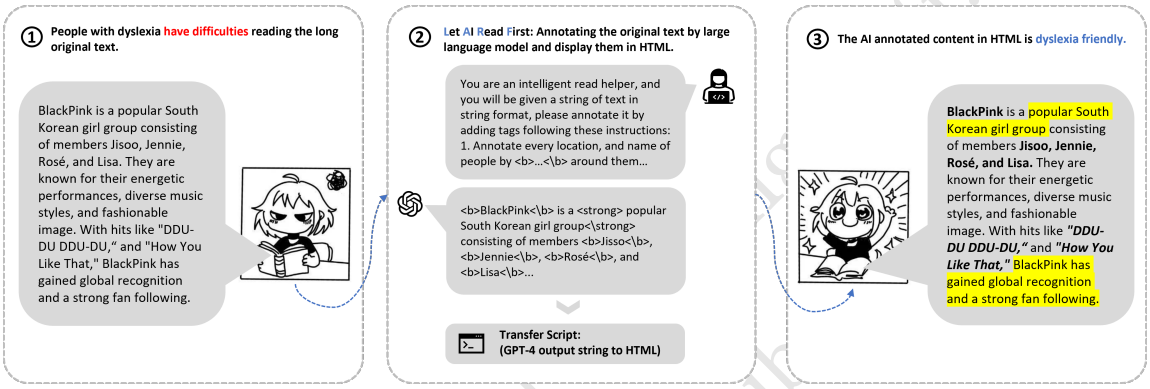

Fig. 1. People with dyslexia always have difficulties while reading. We propose a method Let AI Read First (LARF) that uses language models to annotate the original text and display them in HTML format. Our experiment validates that LARF can improve reading performance and improve the reading experience for individuals with dyslexia.

Dyslexia, a neurological condition affecting approximately 12% of the global population, presents significant challenges to reading ability and quality of life. Existing assistive technologies are limited by factors such as unsuitability for quiet environments, high costs, and the risk of distorting meaning or failing to provide real-time support. To address these issues, we introduce LARF (Let AI Read First), the first strategy that employs large language models to annotate text and enhance readability while preserving the original content. We evaluated LARF in two large-scale between-subjects experiments, involving 150 participants with dyslexia and 160 participants from the general population. The results show that LARF significantly improves reading performance and experience for individuals

Authors' Contact Information: Sihang Zhao, sihangzhao@link.cuhk.edu.cn, The Chinese University of Hong Kong, Shenzhen, Shenzhen, China; Shoucong, Xiong, Zhejiang University, Hangzhou, China, carolhsiung@163.com; Bo Pang, Chinese Academy of Science, Beijing, China, bopang@cnic.cn; Xiaoying Tang, The Chinese University of Hong Kong, Shenzhen, Shenzhen, China, tangxiaoying@cuhk.edu.cn; Yuhang Zhao, University of Wisconsin-Madison, Madison, WI, USA, yuhang.zhao@cs.wisc.edu; Pinjia He, The Chinese University of Hong Kong, Shenzhen, Shenzhen, China, hepinjia@cuhk.edu.cn.

Manuscript submitted to ACM

with dyslexia. Results also prove LARF is particularly helpful for participants with more severe reading difficulties. Furthermore, our work proposes design guidelines and discusses potential research directions opened up by LARF for the HCI community.

CCS Concepts: • **Human-centered computing → Human computer interaction (HCI)**; **Accessibility**; **Accessibility design and evaluation methods**.

Additional Key Words and Phrases: Dyslexia, accessibility

ACM Reference Format:

Sihang Zhao, Shoucong, Xiong, Bo Pang, Xiaoying Tang, Yuhang Zhao, and Pinjia He. 2018. Let AI Read First: Enhancing Reading Abilities for Individuals with Dyslexia through Artificial Intelligence. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 30 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Dyslexia constitutes a category of neurodevelopmental impairments that affect reading abilities, typically manifested as challenges to 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 affects their comprehension, fluency, and vocabulary. This condition leads to academic difficulties and increases the risk of bullying in children and adolescents [4]. Additionally, adults with dyslexia commonly experience considerable deficits in executive functioning [5]. Current interventions, mainly in the form of accessible designs, tend to focus on only a few areas: converting text to speech [6], videos or games [7, 8], adjusting text font through electronic readers [9, 10] (e.g., character size, colour, spacing between words), and replacing complex words with simpler synonyms [11]. Nevertheless, these efforts often exhibit one or more of the following limitations:

(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) Manual annotation relies heavily on human effort. 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).

Compared to the available knowledge about reading difficulties and the demonstrated capabilities of Artificial intelligence (AI) models, there are relatively few accessible designs that effectively address these challenges [12]. With the rapid development of AI [13], especially in natural language processing(NLP) [14], numerous spelling assistance tools for dyslexia have demonstrated considerable capabilities [15, 16]. However, we have not yet discovered any existing reading assistance tools or research that has utilised or discussed how to integrate state-of-the-art AI techniques to address these issues in assistive reading tools for people with dyslexia.

Therefore, to fill these gaps, we propose an AI-based presentation strategy to assist people with dyslexia in reading. We introduce LARF (**L**et **A**I **R**ead **F**irst), an AI-based method that annotates "important" information in texts with highlights, bolding, underlining, and other marks. This approach aims to help readers focus more easily on the key content of the original text, thereby enhancing 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.

Our main hypothesis is that LARF can improve the overall reading performance and experience of people with reading difficulties. Consequently, we conducted two large-scale experiments to evaluate this hypothesis.

We first conducted Experiment One (N = 150) in which participants were from the United Kingdom and the United States and self-reported as having dyslexia, all of whom had English as their mother tongue. Participants were divided into three groups. The groups included: a control group that read the original reading materials directly, a conventional group in which participants read the same materials processed by Bionic Reading [17], and a LARF group that the reading materials are annotated by OpenAI's GPT-4 [18]. We tested the accuracy in recalling and retrieving details, and comprehension levels, using multiple-choice questions. The experimental results show that participants who read GPT-4-annotated texts demonstrate better recall and retrieval performance compared to those using traditional methods and the control groups. Participants were also asked to complete a series of subjective evaluations to assess their user experience with LARF or the conventional tool. The results indicate that LLM-annotated texts significantly improve perceived user-friendliness, overall satisfaction, perceived helpfulness, future use, and recommendation tendencies. Users also believe that this method should be applied as a text presentation method for dyslexic populations in more scenarios (e.g., exams, accessible website design).

Given that dyslexia is a spectrum disorder, we designed Experiment Two (N = 160) to examine whether LARF is also helpful for individuals in the general population who may experience varying degrees of reading difficulties (from none to severe), but have not necessarily self-identified as dyslexic. We conducted a similar experiment to that in Experiment One, involving 160 participants from the United States. We observed similar results and found that LARF is more helpful for individuals with more severe reading difficulties.

Furthermore, previous research suggests that users may employ different strategies in experimental settings versus real-world environments (e.g., in exams [19]). Additionally, most dyslexia-related inclusive designs have focused primarily on English and Spanish [8]. To this end, we conducted a follow-up study. We recruited 14 self-reported dyslexic individuals from a neurodivergent community in China, all of whom were bilingual in Chinese and English. We provided them with a software demo (Appendix B) integrating LARF powered by GPT-4. We allowed them to use our software demo freely and collected their feedback on how AI-annotated text affected their reading experience. The results show that all users reported that the GPT-4 annotated text improved their reading experience. Based on their feedback on LARF; their hopes for future assistive reading software, and the results from Experiments One and Two, we proposed design guidelines and discussed potential research directions opened up by LARF for the HCI community.

In summary, our main contributions are:

- We propose LARF, the first method to leverage LLMs for text annotation, improving text accessibility for individuals with reading difficulties.
- We conducted two large-scale between-subjects experiments, involving participants with reading disabilities(N = 150) and from the general population(N = 160), to validate the effectiveness of LARF. Additionally, we found this tool to be particularly beneficial for individuals with severe reading disabilities.
- We developed a LARF-based software demo and conducted a follow-up study (N = 14) to explore how users engage with AI-annotated text in real-world scenarios.
- We suggest future research directions based on LARF for the HCI community.

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 [6], 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 [7]. 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 [20, 21], the high cost of software like Kurzweil3000 limits its widespread adoption [20]. 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 [9, 10]. Santana et al. [22] created Firefixia, which is a browser extension that enables dyslexic readers to tailor websites for enhanced readability. Text4All [23], an online service for web pages, and the Android IDEAL eBook reader4 for e-books are customisation tools informed by previous research in dyslexic individuals [9]. Text4All extends its offerings to include medical language adaptation, terminology annotation, and language analysis. Currently, a popular method called Bionic Reading [17] revises texts so that the most concise parts of words are highlighted. This guides the eye over the text, and the brain remembers previously learnt words more quickly. Although these methods can be applied in a broader range of contexts, they treat all text as a uniform entity, lacking a targeted emphasis on key segments such as definitions or summary sentences. This results in substantial room for improvement to improve reading performance and experience.

"In another approach, complex words are replaced with simpler synonyms to aid comprehension [11]. However, such an approach not only fails to guarantee accuracy in the context of substitution (i.e., it may completely distort the original intent of the text) but may also affect the literary attributes of the text, such as emotional intensity and rhythm.

Despite the wealth of knowledge surrounding reading difficulties, traditional accessible designs addressing these challenges remain limited [12]. Considering the rapid advancements in AI, the incorporation of AI models with superior reading comprehension and creativity into accessible design offers a promising area for further exploration. As these models become increasingly versatile and powerful, their intersection with 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 [13], with the advancement of natural language processing (NLP) tasks, specifically language models (LMs) [24–27], being particularly prominent. Despite the intricate structures and profound mathematical foundations underlying language models, they can essentially function as a "simple" black box, processing input text to generate appropriate output for the given task [14].

ChatGPT [18], published by OpenAI, is the successor to InstructGPT [28], fine-tuned using Reinforcement Learning with Human Feedback (RLHF) [29], and has gained phenomenal attention and widespread discussion not only within the NLP community but also globally. In this paper, we used GPT-4 to generate re-formatted content that is more friendly for individuals with reading challenges. However, we do not discuss the theoretical background or technical details.

2.3 Hyper Text Markup Language (HTML)

In Hyper Text Markup Language (HTML) [30, 31], 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

HTML tag	Description	Example
<b>	Bold the text	**Example**
<i>	Italicize the text	*Example*
<u>	Underline the text	Example
	Highlight the text	Example
...	...	...

Fig. 2. 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, etc. The figure provides examples of only a few common default tags and their display results.

display text in bold as shown in Figure 2. We can modify the appearance of text without changing its content by using HTML tags. In real-world scenarios, HTML can produce diverse visual effects; however, in this work, we focus only on fundamental HTML tags.

3 Method and Data

3.1 Workflow of LARF

The workflow of LARF is illustrated in Fig. 1, in which the original text is input as a string. Guided by preset prompts, GPT-4 processes the original text, incorporating the HTML tags. The resulting HTML-tagged text is output as a new string. Subsequently, a Python script compiles this HTML-tagged string into an HTML file, serving as the final output. Consequently, users receive a presentation where specific information has been modified with bold formatting or highlighting, while the textual content remains entirely unchanged. The simple example in Fig. 1 shows a segment taken from Wikipedia about BlackPink [32]. GPT-4 was asked to highlight sentences that serve a summarizing role using <mark><\mark>tags. It also asked to bold important names of people and items using <b><\b>tags. After processing the output of GPT-4 with the transfer scripts, the user gets the GPT-4-annotated content shown on the right-hand side.

In subsequent experiments and practical applications, we adjusted the prompts by using different labels, thereby modifying the presentation of the text. The detailed default prompts can be found in the Appendix A.6.

3.2 GPT-4 Data

In Experiment One and Experiment Two, we processed the reading materials using GPT-4 API. We also used GPT-4 together with human evaluation to score the participants' short-answer questions in subsequent experiments. The version of GPT-4: ChatGPT July 20 version. We set the temperature as 0 to ensure the result is reproducible. All the specific prompts and generation logs can be found in the supplement material and Appendix.

3.3 Bionic Reading Data

We employed the Bionic Reading [17] as a representative of conventional tools to process the corpora in subsequent experiments, 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 the default 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.

An Example of Bionic Reading
BlackPink **is a popu**lar **Sou**th **Kor**ean **gi**rl **gro**up **consi**sting **o**f members **Jisoo**, **Jennie**, Rosé, and **Li**sa. **The**y **a**re **kno**wn **f**or **the**ir **energ**etic **performanc**es, diverse **mus**ic **style**s, and **fashio**nable **ima**ge. **Wi**th **hi**ts **li**ke "**D**DU-**D**U **D**DU-**D**U," **a**nd "**H**ow **Y**ou **Li**ke **Th**at," **Black**Pink **h**as **gai**ned **glo**bal **recogn**ition **a**nd **a str**ong **f**an **follo**wing.

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 the participants in Experiment 1 and Experiment 2 were recruited through the Prolific platform and completed informed consent forms. All of the participants in Experiment 3 were from a neurodivergent community in China, who voluntarily participated in the software experience activity and they were also informed consent forms.

4.2 Transparency

All raw experimental data, GPT-4 processing history (including evaluations and annotations), and data analysis code are available in the supplementary materials. Examples of our prompts and questions in the questionnaires are provided in Appendix A. The LARF demo is publicly available for free trials.

5 Experiment One: 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., Bionic Reading) and control groups, can improve reading performance and to what extent it affects the reading experience of individuals with dyslexia. We evaluated both reading performance and reading experience using objective measures and subjective evaluations.

5.1 Experiment Setup

In this experiment, we focus on English speakers, as English is the world's most spoken language and the third most spoken native language[33]. 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[34].

We chose Reading Test 115, Passage 2, "The Step Pyramid of Djoser," a descriptive and factual reading text from the IELTS[35] Academic as the corpus in this 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[36]. 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[37]. 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 a proper tool for assessing adult reading performance.

5.2 Method and Experiment Procedure

We recruited our participants from Prolific[38]. Prolific is an online research platform that connects researchers with a diverse participant pool for academic studies, surveys, and experiments. 150 individuals who have been medically

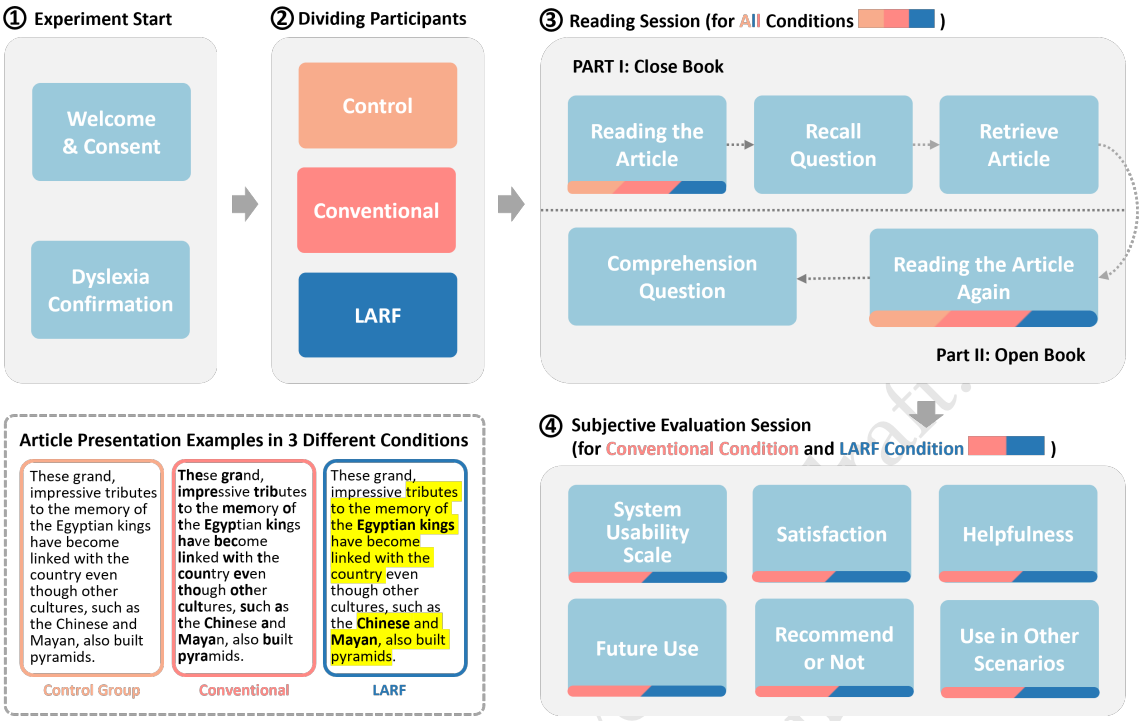

Fig. 3. Experiment Procedure. Participants are randomly assigned to three conditions, and then they are asked to finish the reading session. They are required to read the same article but with different presentations. Participants in the conventional condition group and LARF condition group are required to finish a subjective evaluation session after they finish the reading session.

diagnosed with dyslexia, are in the process of being diagnosed, or strongly suspect they have undiagnosed dyslexia participated in our study (M_{age} = 36.8, 33.3% Female). Details about participants' demographics can be found in Appendix A.1. 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 GPT-4. Participants were not pre-informed that texts in the LARF group are annotated by GPT-4, nor was the rationale behind annotations (e.g., bolding of significant characters' names) disclosed. While this approach may attenuate the effects of LARF, it mitigates the influence of psychological priming on the results.

The study began with participants completing a Dyslexia Checklist (refer to Appendix A.3), 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 ??,). 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 completed the experiment. Participants in the conventional tool and LARF conditions also evaluated the modifications and annotations made by these tools. We used an adapted version of the System Usability Scale[39] to assess tool usability. Participants then rated the tool's perceived helpfulness, satisfaction, intention to continue using it, and likelihood of recommending it to others. Participants in the conventional tool condition completed the experiment after providing demographic information. In contrast, participants in the LARF condition were asked about their preference for a personalized LARF tool before providing demographic information. The experimental procedure and session details are shown in Fig. 3.

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. The detail of the attention check is given in the Appendix A.2.

Dyslexia Checklist: Before reading the article, participants assessed their own dyslexia levels using the Dyslexia Checklist (see Appendix A.3 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 reveals 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 indicates 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 are identified as outliers based on their initial reading times (defined as initial reading time > Q3 + 1.5 × IQR or < Q1 - 1.5 × IQR) and were thus excluded from this part of the analysis. Consequently, the final analysis on reading time is conducted with 138 participants. Covariates including education, age, gender, and dyslexia level are considered in the analysis. The one-way ANOVA shows 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[40], those who spend less than 30.2 seconds (0.05 quantile) are considered relatively impatient. Fig. 4(a) shows that participants using LARF do not fall below 30 seconds and are concentrated within a shorter, reasonable range. This suggests that LARF may aid in attracting user attention and enhancing reading patience and confidence.

Recall Performance: In the recall section, participants are asked six questions (example questions can be found in Appendix A.4), earning one point for each correct answer. For this part of our analysis, we include education, age, gender, dyslexia level, and reading time as covariates in the one-way ANOVA analysis. The result Fig. 4b reveals that participants in the LARF condition tend to score higher (7.8% higher than the control group and 5.1% higher than conventional group) 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 (example question can be found in Appendix A.5.) 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

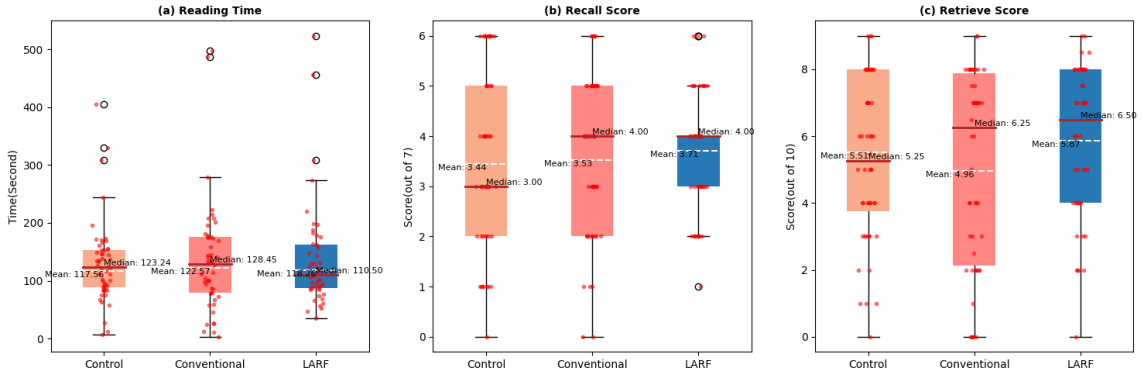

Fig. 4. (a) shows the differences in reading time under three different conditions. Though the pattern is not significant, we can observe that users in the LARF group do less "glance over and skip the article." Furthermore, their overall reading time is more concentrated in areas with shorter durations. Subfigure (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).

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 can be found in AppendixA.7 and the GPT-4 score logs are available for reference in the supplementary materials. A similar one-way ANOVA analysis was conducted. The results in Fig. 4c 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 (6.5% higher than the control group and 18.3% higher than the conventional group) 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.

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. 5. **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

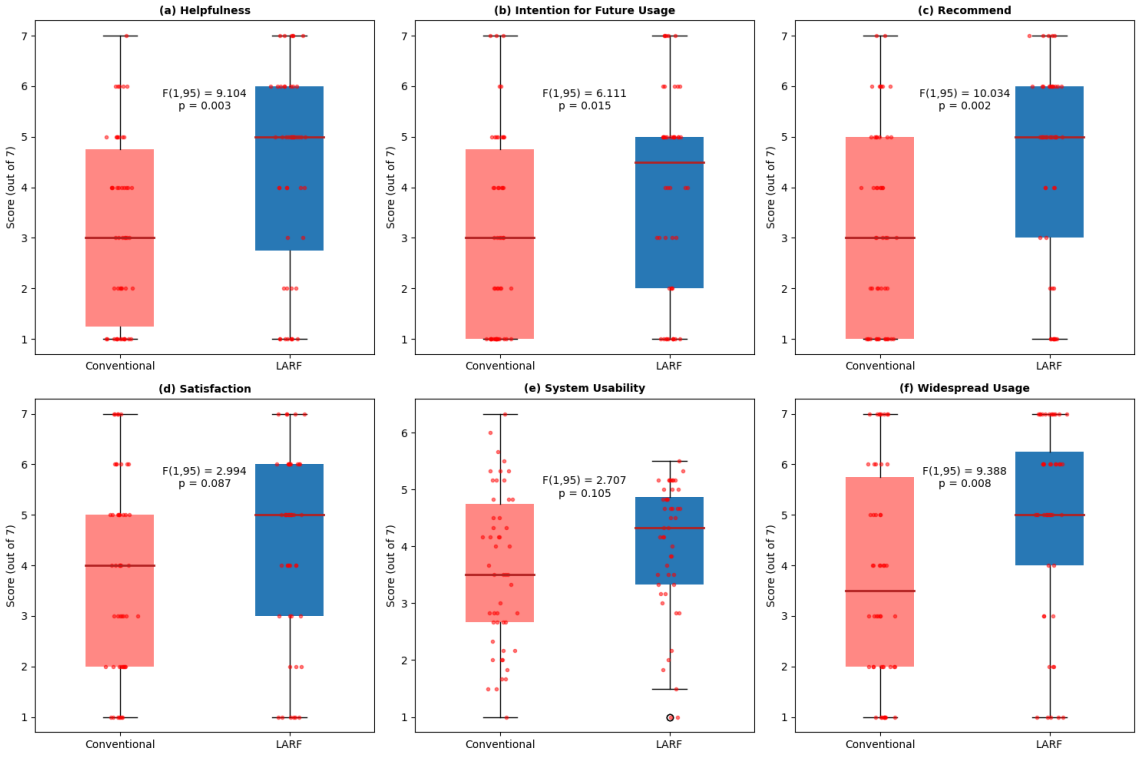

Fig. 5. The subjective evaluation result. Participants with dyslexia exhibited a clear preference for LARF, considering text annotated with LARF to be effective, user-friendly, and worthy of broader adoption in various contexts.

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 5 and Table 6 in Appendix A.8. **The result suggests that participants in the LARF group show more overall satisfaction, they also reported that LARF is more helpful and easier to use compared to Bionic Reading.**

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).

5.4 Post Hoc Evaluation

People with dyslexia often experience different subsets of challenges [41]. Given the varying severity of dyslexia among participants, resulting in distinct reading challenges, we conducted a post hoc evaluation to assess LARF's efficacy across different degrees and categories of reading difficulties, focusing on its effects on various symptoms of reading

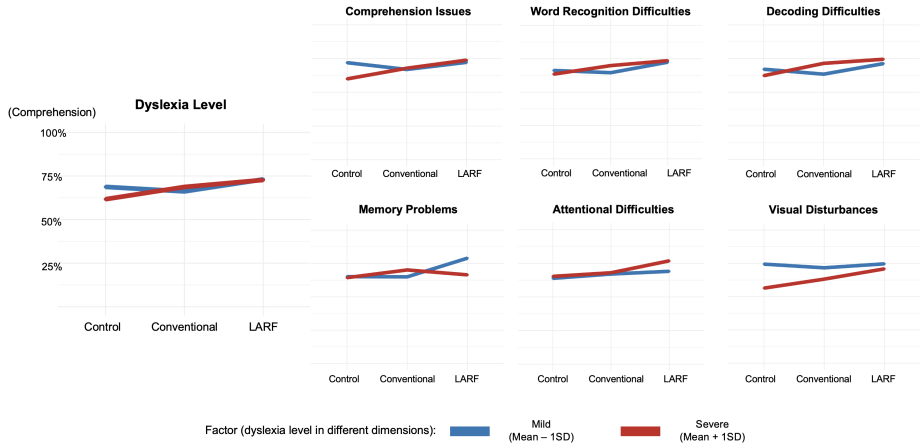

Fig. 6. Post hoc evaluation for comprehension performance. The y-axis represents the accuracy of reading comprehension. In the group with severe symptoms, LARF exhibited significant improvement compare to the conventional group and control group.

disabilities. Based on previous research, we calculated the mean ± 1 standard deviation (SD) for each dyslexia item. Participants whose self-reported dyslexia scores were higher than M + 1 SD were classified as having severe dyslexia, while those with scores lower than M − 1 SD were classified as having mild dyslexia. **Results show that LARF is especially helpful for participants with severe dyslexia.** As Fig. 6 shows, the improvement in participants' reading performance is more pronounced in those with severe dyslexia. Similar results can also be observed in their recall (Fig. 13) and retrieval (Fig. 12) performance.

6 Experiment Two: LARF Effectiveness among General Population

As dyslexia is a spectrum disorder, people in the general population may also face challenges during reading. Building on the insights from Experiment One, which focused on dyslexic participants, Experiment Two expands the scope of our investigation to the general population. 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 without filtering any information about dyslexia, ensuring a diverse sample representative of the general population.

6.1 Method and Experiment Procedure

The procedure in Experiment Two mirrors that of Experiment One to maintain methodological consistency. 160 individuals on Prolific participated in our study (M_age = 40.6, 48.13% Female). Detailed information about participants can be found in Appendix A.1. 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 One. Initially, participants were presented with a Dyslexia Tendency Adult Checklist [42], 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 One. Based on the attention check and the participants' answers, we were able to identify and exclude participants who appeared to have cheated or provided random responses (e.g., some participants copied

the original text of our reading materials and directly pasted it in the retrieval section and some of them use less than 10 seconds to finish all the reading and questions.) Out of the initial 160 participants, 22 were consequently disqualified, resulting in a final sample of 138 participants for the performance analysis.

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, which means only 14 participants exhibited symptoms aligned with 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 One, 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 One, 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 are 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 reveals that participants in the LARF condition scored marginally higher than the other two conditions ($M_{control} = 5.26, SD = 1.47, M_{conventional} = 5.15, SD = 1.89; M_{LARF} = 5.37, SD = 1.74; F(2, 128) = 1.016, p = .365$).

Retrieve Performance: We used the same method as Experiment 1 to evaluate participants' retrieve performance by GPT-4. 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 ($F(2, 138) = .288, p = .750$).

Subjective Evaluation: Similar to Experiment One, 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. 8. **Overall, participants in the LARF condition rendered more favourable evaluations than those in the conventional condition, replicating the findings in Experiment One**. 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 ($Mconventional = 4.59, SD =$

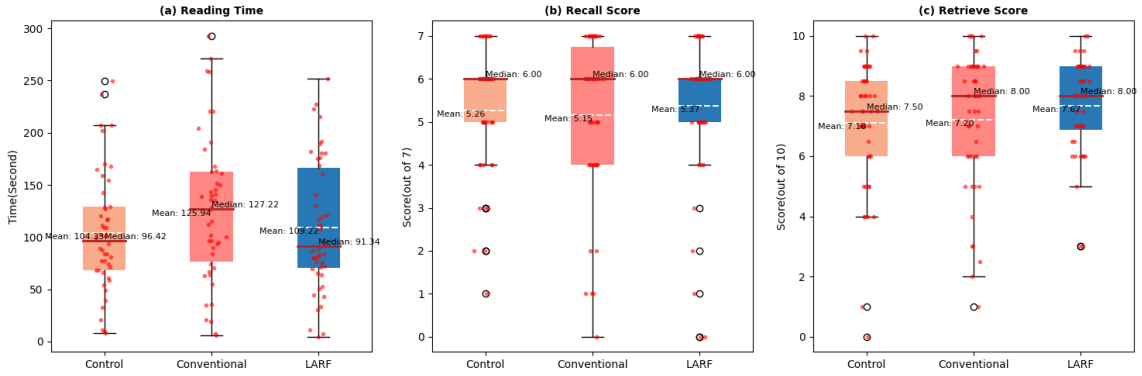

Fig. 7. illustrates that there is no significant difference in performance between the LARF group and the control group in typical readers, despite the median time spent by the LARF group being notably lower compared to the other two groups. (b) and (c) respectively reflect the performance of the general population in the recall and retrieval 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.

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 show a positive perception of LARF, suggesting a favourable reading experience for LARF users.

6.3 Post Hoc Evaluation

The results of Experiment Two on the general population showed that although the LARF group exhibited a trend of better reading performance compared to the other two groups, the difference was not statistically significant. Given that the results of Experiment One indicate that LARF is more effective for individuals with more severe reading difficulties, we conducted a post hoc evaluation similar to that in Experiment One. The results in Fig. 9 and Table. **??** also shows that LARF is more effective for participants with higher dyslexia tendencies when they are doing retrieval tasks and reading comprehension tasks in the general population. 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 higher, being in the LARF condition (vs. the control condition) exerts a more 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 and the difference is also statistically significant ($F(2, 70) = 3.914, p = .024$.)

Another interesting finding is Participants in the **LARF condition demonstrates a markedly higher completion rate** ($P_{LARF} = 98\%$) compared to those in the conventional condition ($P_{Conventional} = 85.7\%$) and the control condition

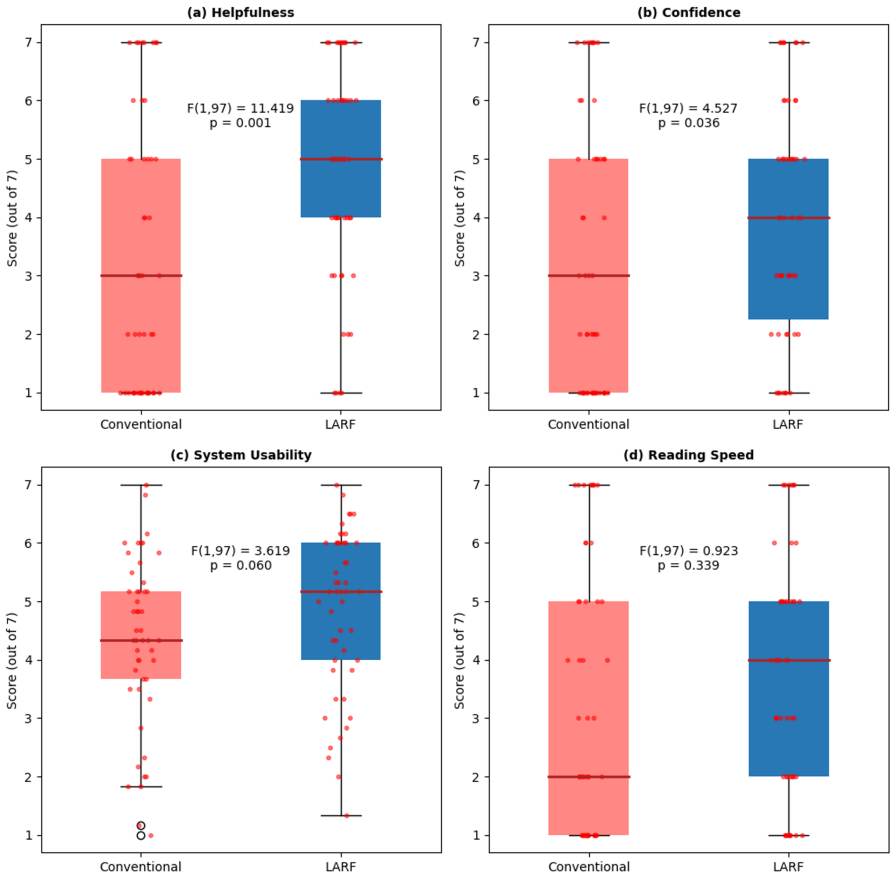

Fig. 8. The subjective evaluation shows typical readers also hold a significant positive attitude toward LARF (e.g. participants believe LARF can help them read faster and is easy to use).

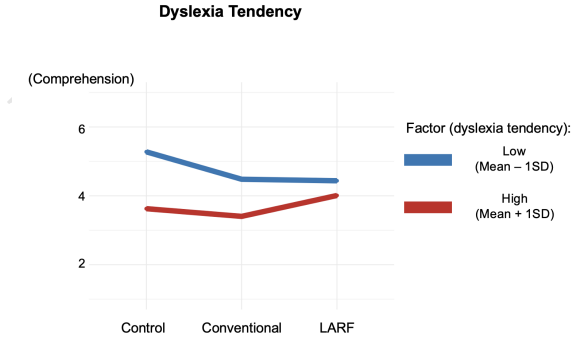

Fig. 9. LARF significantly improved reading comprehension in participants self-reporting higher dyslexia tendencies.

($P_{control}$ = 91.8%). A Fisher's exact test indicates that the difference in completion rates among the three conditions approaches is 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 exhibite greater patience, aligning with the results of subjective evaluation where users perceive that LARF could enhance their reading confidence and patience.

Table 1. Unstandardized Coefficient 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

Since participants in Experiment One and Experiment Two were only allowed to read pre-annotated text generated by our default prompts, we wanted to know whether our strategy would still work in real-world use. We also wanted to know whether GPT-4's language transfer capabilities could enable LARF to be effective across languages. Therefore, we conducted a small-scale follow-up study. We invited participants from a neurodivergent community in China ($N = 14, meanage = 23, 35.7\% female$) who self-reported reading difficulties. We provided them with a software demo integrated with LARF (The detailed introduction of the demo can be found in Appendix B.) They were asked to freely use our demo and then provide feedback on their experience through a questionnaire.

7.1 Method and Experiment Procedure

After completing the Dyslexia Checklist, identical to that used in Experiment One, participants were provided with the same reading materials "The Step Pyramid of Djoser" (used in Experiments One and Two) together with its Chinese translation version. They are also allowed to process their own textual content with our demo. 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

All participants reported difficulty in maintaining focus while daily reading for extended periods, leading to easy distractions (Fig. 15.) And a 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.)

There are 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 (Fig. 17 (b).) This aligns with the findings from Experiment 1. After informing LARF that the annotations were generated by GPT-4, 14.3% of users expressed distrust in AI (Fig. 17 (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. None of the users reported differences between reading Chinese content and English content.

We 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, we also received some negative feedback including the excessive brightness of yellow highlights, low accuracy in custom annotations, and a preference for modifications in font and spacing over annotations.

Participants also provided development suggestions for our software demo and 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. We have a detailed discussion on the design guidelines and future research topics about assistive reading tools for HCI community in Section 8.1 based on these valuable suggestions.

8 Discussions

Our work shows that annotating text by LLMs can improve the reading performance and reading experience of people with reading difficulties. The results also show that the LARF strategy is more helpful for users with more severe reading challenges. In this section, we discuss the related research directions in the HCI community and promising real-world application scenarios that this work opens up. We also discuss the limitations of our work.

8.1 Future Work for HCI

Our work establishes the theoretical foundation and provides a software demo, but there are additional potential application scenarios for integrating LARF. As shown in Fig. 10, LARF can be incorporated into smart devices such as PCs, laptops, tablets, and even VR devices. The annotations in our demo and experiments are generated using OpenAI's GPT-4 API, however, LLM inference requires substantial GPU resources. Currently, the devices that can use LLMs for text annotation locally remain limited. Therefore, exploring how smaller models can accomplish this task and achieve optimal annotation quality is a meaningful topic for both the HCI and NLP communities. In existing research, we consider the NLP downstream topic Extract QA[44] aligns with our objectives.

While our work demonstrates the effectiveness of AI-based text annotation, how to make AI-generated annotations more accessible to users with reading difficulties is another direction the HCI community could further investigate. For example, prior studies have explored how the length of the highlighted text affects reading comprehension[45], and Tran et al. have examined the impact of data visualization colours on users with ADHD [46]. HCI researchers could

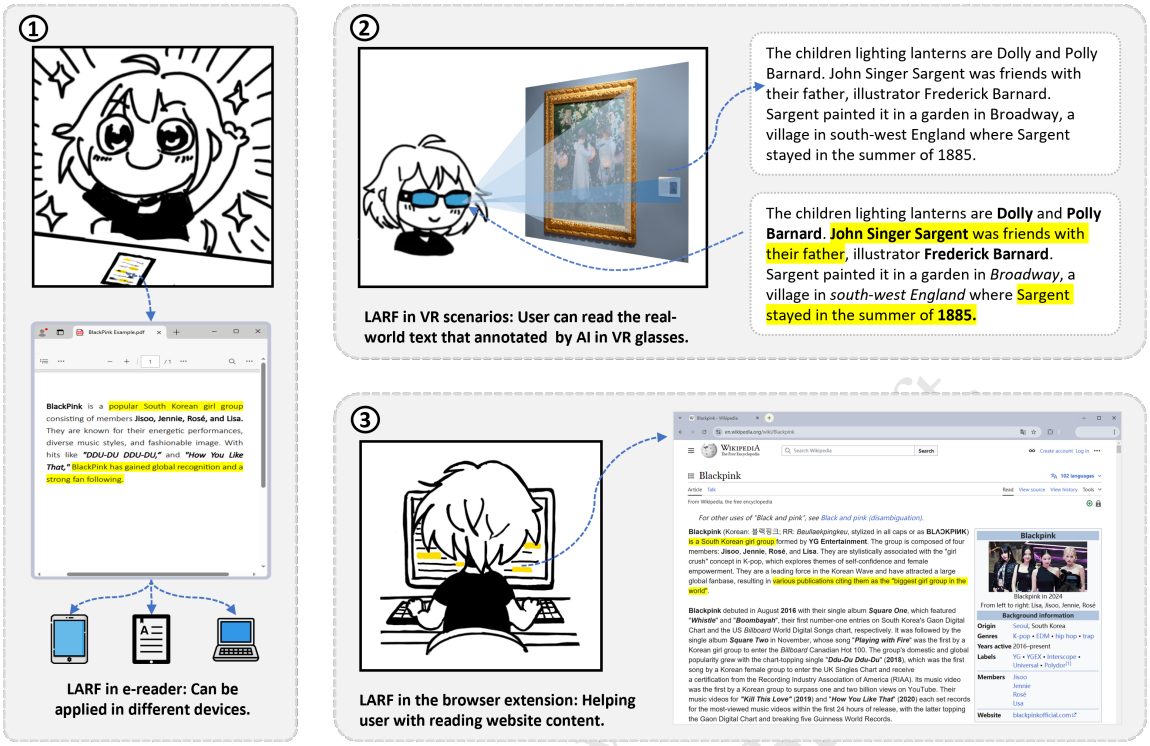

Fig. 10. Real-world application scenarios that can apply LARF. In the first subplot, the user is using an e-reader which is integrated with LARF, this device can be a tablet, a smartphone or a laptop. In the second subplot, the user is wearing VR glasses, looking at the "Carnation, Lily, Lily, Rose" by John Singer Sargent [43]. The VR headset with built-in LARF functionality helped annotate the description next to the painting, making it easier to read. In the Third subplot, LARF is integrated into a browser extension and helps the user reading the online content (the web page in the figure is the BlackPink item in Wikipedia[32].)

explore ways to optimize the display of AI-generated annotations, such as the colour, brightness, or font of highlighted text. Additionally, we can investigate how to tailor annotations for different types of content, which may involve more complex prompts and task-specific fine-tuning.

LARF could also potentially be applied to subtitles in videos, live streams, and online courses. It is worth noting that research on such assistive annotations is still limited. For example, some studies suggest that the improvement in reading ability may result from the annotation process rather than the annotations themselves [**?**]. This raises an important question: could long-term reliance on AI-generated annotations, such as those in the LARF strategy, lead to a decline in memory or learning abilities over time? This is a topic worth further investigation. Additionally, dyslexia often co-occurs with other neurodevelopmental disorders, such as Attention Deficit Hyperactivity Disorder (ADHD) [47]. Future work could focus on whether AI-annotated text might introduce additional distractions for neurodiverse populations and how these distractions could be minimized.

8.2 Design Guidelines

In this section, we provide recommendations for future accessibility designs based on the results of our experiments using LARF. Users expressed a desire for customizable annotation settings (in Section 5.3 and Section 7). For example,

in our software demo (refer to Appendix B), we implemented a text input box and a dropdown menu for selecting annotation types. This functionality could be integrated into e-readers or browser extensions. However, for applications in VR or on smaller screens (e.g., smartwatches), voice commands and gesture-based controls may be required. Users also wanted the ability to adjust the font size, spacing, and other text properties (Section 7), which is consistent with previous research [9]. Therefore, future designs should focus on integrating LARF together with existing accessibility features but not independently.

Additionally, in the follow-up experiment, two users expressed concerns about AI annotations, particularly regarding privacy and annotation quality. To address these concerns, future designs should consider implementing local or end-to-end models for inference to ensure user privacy. Furthermore, fine-tuning the models and using more sophisticated, well-crafted prompts could improve the quality of AI-generated annotations.

8.3 Limitation

During the experimental process and software development, the study encountered several limitations:

To ensure that participants were not impatient with the questions (especially when the participants have dyslexia, reading long content may cause cognitive load), our reading comprehension and recall tasks in the evaluation were not set at high difficulty and we limited the number of questions. However, some of the questions have the potential risk of being "too easy" for participants. This resulted in some evaluations observing only patterns but without significant outcomes. We did not conduct a "random labelling group" in this article to exclude the influence of the placebo effect, although intuitively such an effect should be small. As mentioned in Section 8.1, this study does not explore the interaction among different types of annotations (e.g., bolding, highlighting), nor does it investigate which annotation form would be most beneficial for users. Identifying optimal default prompts for user engagement is also not examined. Adjusting the font size and colour is also achievable in HTML format, which is not discussed in this paper.

Previous work also shows that allowing users to set their own preferences can improve their reading accuracy compared to the default setting [48]. So in our software demo, users can input what kinds of information and in what types of presentation they want GPT-4 to annotate. However, in Experiment One and Experiment Two, the participants only used the default prompt.

9 Conclusion

In this article, we introduce LARF, an AI-annotated text approach designed to enhance the reading abilities of individuals with reading difficulties and to address the limitations of traditional methods. Experiment One involved 150 participants with reported dyslexia, who completed 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 engagement, outperforming the conventional technique (i.e., Bionic Reading). Subsequent post hoc evaluation and Experiment Two further suggest that LARF is particularly beneficial for individuals with severe reading difficulties. The follow-up experiment shows that LARF is also useful and satisfying in real-world settings. In the discussion, we suggest future directions for AI-annotated text-assisted reading within LARF. 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] Kristen Laga, Daniel Steere, and Domenico Cavaiuolo. Kurzweil 3000. *Journal of Special Education Technology*, 21(2):79, 2006.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

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

[18] OpenAI. Chatgpt. https://openai.com/.

[19] Kirsti Lonka, Sari Lindblom-Ylänne, and Sini Maury. The effect of study strategies on learning from text. *Learning and Instruction*, 4(3):253–271, 1994.

[20] 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.

[21] 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.

[22] 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.

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

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

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

[31] 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.

[32] Blackpink. Blackpink. https://en.wikipedia.org/wiki/Blackpink, 2024. Accessed: 2024-09-07.

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

[34] 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.

[35] British Council. Ielts. https://www.ielts.org/.
[36] H Porter Abbott. *The Cambridge introduction to narrative*. Cambridge University Press, 2020.
[37] Michael Alley. The craft of scientific writing. Technical report, Springer, 1996.
[38] Prolific. Prolific - online participant recruitment for surveys and market research, 2023.
[39] John Brooke. Sus: a "quick and dirty'usability. *Usability evaluation in industry*, 189(3):189–194, 1996.
[40] 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.
[41] 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.
[42] Lefly Pennington. Adult reading history questionnaire. https://dyslexiaida.org/screening-for-dyslexia/dyslexia-screener-for-adults/.
[43] Tate. Carnation, Lily, Lily, Rose by John Singer Sargent. https://www.tate.org.uk/art/artworks/sargent-carnation-lily-lily-rose-n01615, n.d. Accessed: 2024-09-11.
[44] 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.
[45] Nikhita Joshi and Daniel Vogel. Constrained highlighting in a document reader can improve reading comprehension. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–10, 2024.
[46] Tien Tran, Hae-Na Lee, and Ji Hwan Park. Discovering accessible data visualizations for people with adhd. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2024.
[47] Hani Zohra Muhamad. Dyslexia with attention deficit hyperactivity disorder: A case study. *Asia Pacific Journal of Developmental Differences*, 1(2):238–252, 2014. Correspondence to: Hani Zohra Muhamad, Dyslexia Association of Singapore, 1 Jurong West Central 2, #05-01, Jurong Point, Singapore 648886.
[48] 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.

A Experiments Details

A.1 Participants

Table 2. **Participants Demographic in Experiment One**

Gender		Age		Education	
Male	95	18-24	11	Less than high school	2
Female	50	25-34	56	High School graduate	58
Non-binary/Unknown	5	35-44	44	Bachelor degree (or currently in processing)	56
		45-54	26	Master degree (or currently in processing)	26
		55-64	12	Doctor degree (or currently in processing)	6
		65-74	0		
		75+	0		

Table 3. **Participants Demographic in Experiment Two**

Gender		Age	
Male	83	18-24	12
Female	77	25-34	42
Non-binary/Unknown	0	35-44	49
		45-54	28
		55-64	19
		65-74	7
		75+	3

A.2 Attention Check

In our attention check, participants are asked to answer the question where including the instruction that the correct answer is "Water". Participants who fail in this question will be marked as not focused.

In the modern era, explorers and archaeologists uncovered the secrets of the pyramid's chambers. The stories of Djoser, Imhotep, and the countless hands that had shaped the monument were revealed, shedding light on the ancient world's mysteries. **To show that you have read the instructions carefully, please ignore the items below about the explorers' findings and instead choose "Water".** Based on the information in the preceding paragraph, which of these objects did explorers find?

- o Gold
- o Diamond
- o Rosewood
- o Water
- o Stele

A.3 Dyslexia Checklist

Table 4. **Dyslexia Checklist for Experiment One**

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?

A.4 Recall Question

Two examples of our recall questions are given below:

Where is the Step Pyramid of Djoser at?
o Saqquira o Saqqara o Saqqura o Saqqarua

Which king in ancient Egypt does this article discuss?
Please input your answer:

A.5 Retrieve Question

An example of our retrieve question is given below:

Please retrieve the article and provide as many details as possible (such as what specific data the article presents, what names appear, and the relationships between the characters and events, etc.)
Please input your answer:

A.6 Default Prompt for GPT-4

This prompt is used for Experiment One and Experiment Two, as well as the default prompt in our software demo (default model). It is designed to be used in general situations but not for particular articles. The detailed prompt is displayed below:

You are an intelligent reader helper and you will be given a string of text in string format, please annotate it by adding tags following these instructions:
1. Please annotate every date, number, location, and name of people or events in the paragraph by adding `<strong>` tags around them.
2. Please highlight sentences and phrases in the paragraph that can summarize the core content of the paragraph or serve as a conclusion to the description by adding `<mark>` tags around them.
3. Please underline sentences and phrases in the paragraph that are unusual or need to be particularly noted by adding `<u>` tags around them.
4. You can add as many `<mark>`, `<strong>`, or `<u>` tags in one paragraph as necessary to highlight or bold important text.
5. Please make sure to use and only use the 3 types of annotations above to annotate each paragraph of the text.
6. Don't make the highlights or underlines too long or too often if it is not necessary.
7. You are allowed to add only the above previously mentioned HTML tags, and that's the only change you can make to the text. YOUR OUTPUT MUST KEEP THE CONTENT OF THE ARTICLE THE SAME AS THE ORIGINAL ONE.
8. Your output should only contain the marked text with added tags, which can be directly presented in HTML. Don't add anything else like "Here is your output" and so on.
9. Keep the original language; i.e., if the context was given in Chinese, your output should be Chinese as well.

A.7 GPT-4 Evaluation Criteria

Here is the prompt for GPT-4 to give the score, we use a one-shot learning method to give GPT an example of a 6-point answer:

Please play the role of a rater and help me rate some answers. you will be given an article. Please read it, and you will be given some information about this article. I need you to score each item by their completeness and accuracy from 0 to 10.
A 0-point represents the entrance is very poor and basically contains no correct or important information and a 10 means the entrance is almost perfect.
A 5-point answer should have some details correct but misses or get some key information wrong, and the overall understanding of the article is partially correct.
A 7-point entrance should contain some correct details, such as the correct name, time, data, etc., or provide a not-bad summary of the overall article. However, it may be a lack of coherent logic or could miss some important information.
A 9-point entrance should contain most of the correct details, such as the correct name, time, data, etc., and it should also contain a logically coherent and accurate summary of the full text.

Here is the original article
*****
ORIGINAL ARTICLE
*****
Now you should directly give a score and the reason you give that score, and here is an example of 6-point entrance: The entrance is: 10.5 m high, with 13 false doors, there were tombs made of mud and clay before stone pyramids, the third Egyptian dynasty was the first to build of stone.
And the answer is:
Score: 6
The entrance provides important details such as the height of the wall (10.5 meters) and the number of false doors (13). It also correctly mentions that tombs were made of mud and clay before the construction of stone pyramids and that the Third Dynasty of Egypt was the first to build with stone. However, it could have provided more information about the Step Pyramid itself, such as its final dimensions or its significance in Egyptian history. And its logic is not very coherent.

A.8 Subjective Evaluation

Table 5. **Subjective Evaluation - System Usability Scales**

System Usability Scales		Mean (SD)		Statistics	p-value
Conventional	**LARF**	**Conventional**	**LARF**	**(F(1, 95))**	
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 <0.01
(3) SUS-3 is a reversed-scale question

Table 6. **Subjective Evaluation - All**

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 Software Demo

The LARF software demo shown in Figure 11 is an interactive interface that users can open in a browser via a link. Users can copy and paste the text into the text box on the left, and by clicking the "Transfer" button, they can obtain the annotated text in the text box on the right. By checking the "Custom mode" option on the left, users can activate the custom prompt feature. When Custom mode is off, LARF will process the text using the same prompt as in the previous experiments. When Custom mode is on, users can enter the information they want to be annotated (such as the names of songs, members, and albums shown in the figure) and specify how they want this information to be annotated in the Key Information section below.

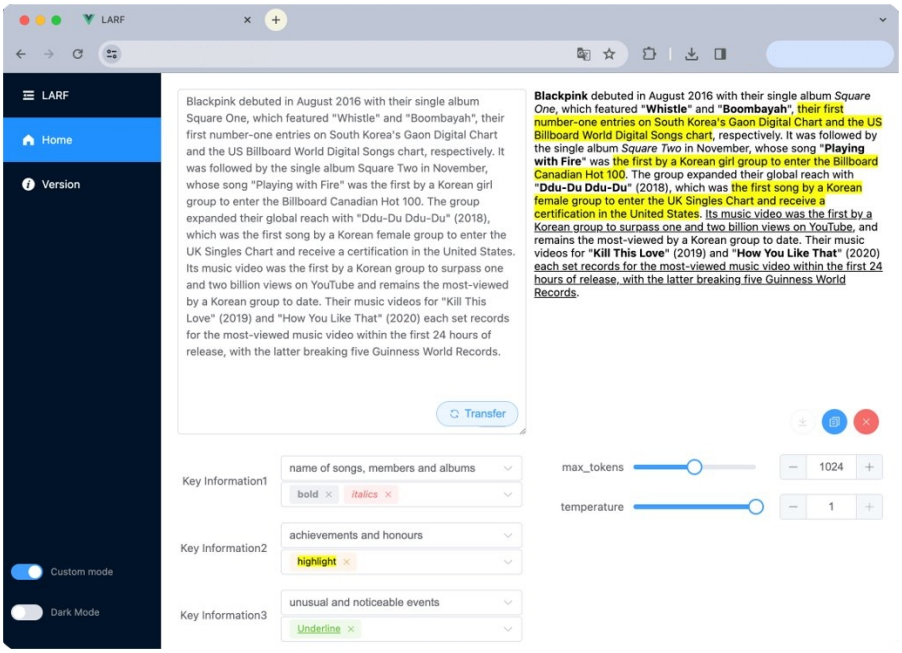

Fig. 11. The demo of the custom mode of LARF software application on PC.

C Supplemental Figures

C.1 Post Hoc Evaluation

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 item. The plot shows that LARF significantly improved recall, retrieval, and comprehension performance in individuals with more severe symptoms.

C.2 Bionic Reading

The Example of Bionic Reading is shown in Figure 14.

C.3 Follow-up Study

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009

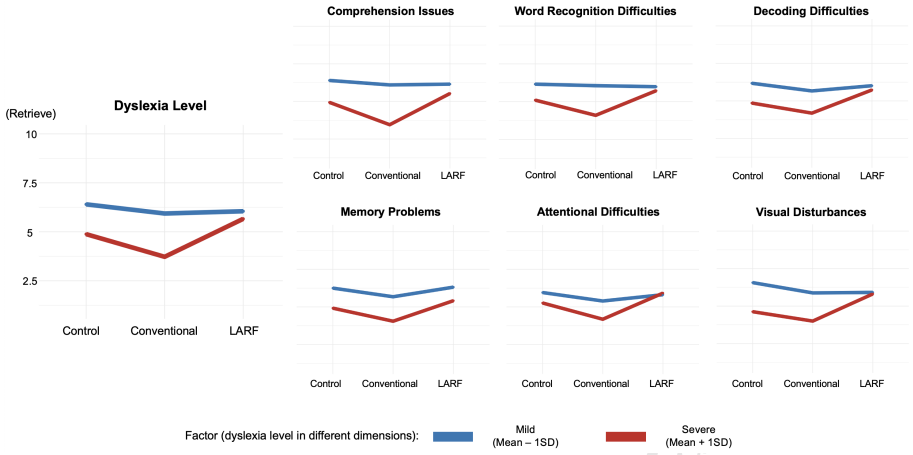

Fig. 12. Post hoc evaluation for retrieving performance. The y-axis represents the scores for retrieve, with a maximum score of 10. While in the group with mild symptoms, LARF did not exhibit improvement, it significantly enhanced users' retrieval abilities in the group facing more severe reading challenges, whereas conventional tools had almost entirely negative impacts.

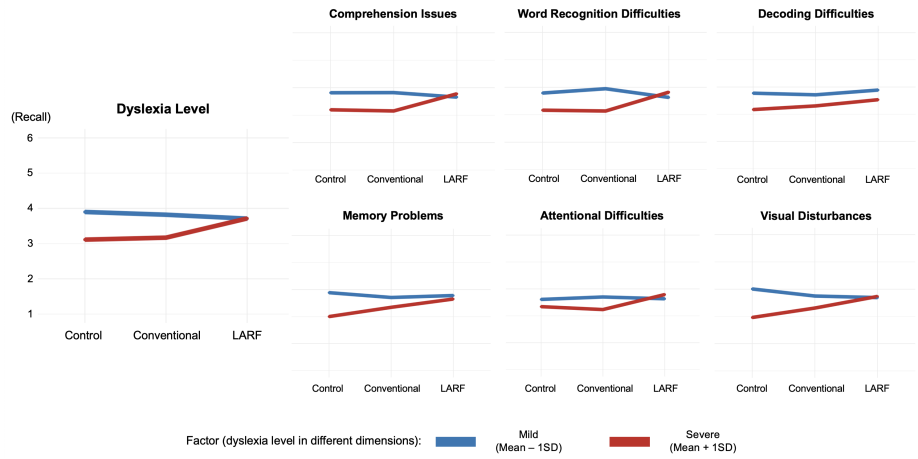

Fig. 13. Post hoc evaluation for recall performance. The y-axis represents the scores for recall, with a maximum score of 6. LARF similarly provided substantial assistance to the group with more severe symptoms, even surpassing the group with mild symptoms who also used LARF.

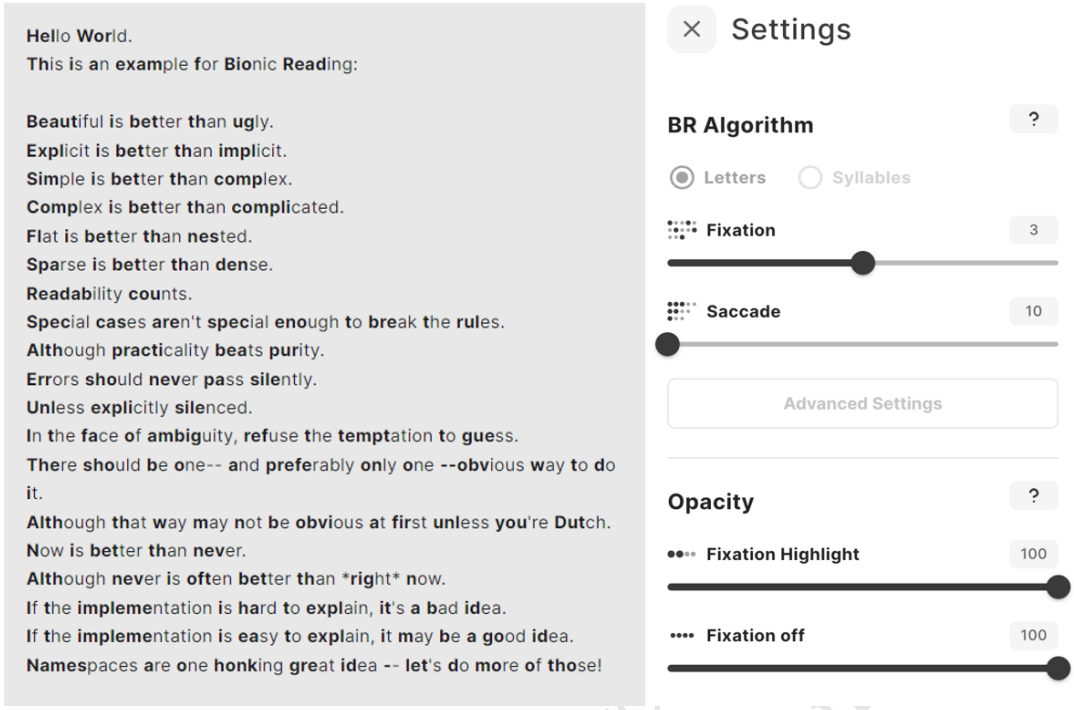

Fig. 14. An example of the result and parameters of Bionic Reading

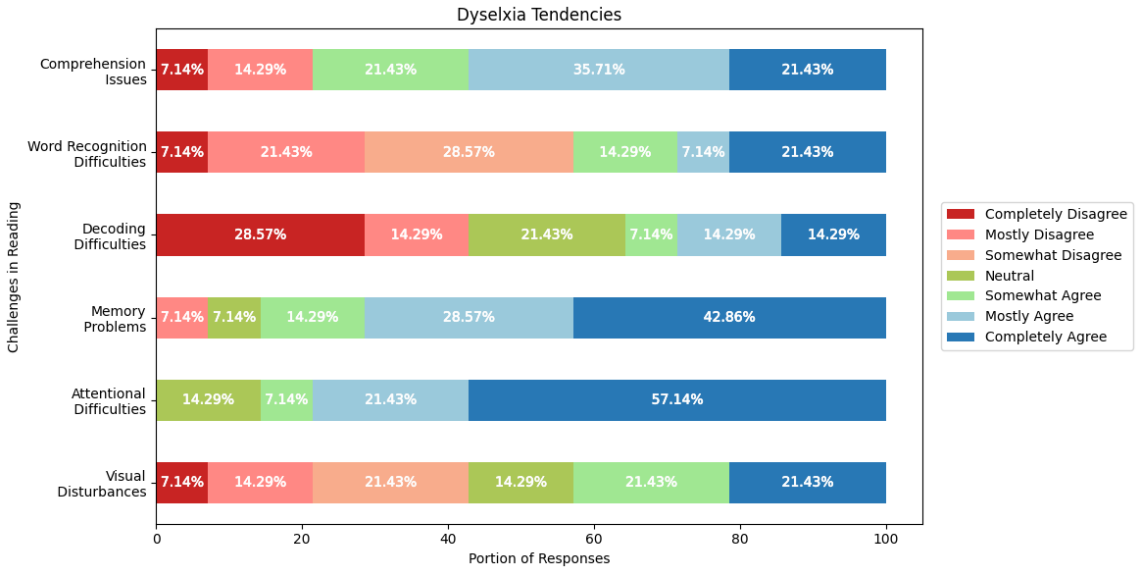

Fig. 15. The Dyslexia Tendency Scale includes participants' self-reported tendencies towards reading difficulties. The scale shows that 100% of the participants reported difficulties with attention deficit during reading, while only 35.71% of participants reported struggles with decoding

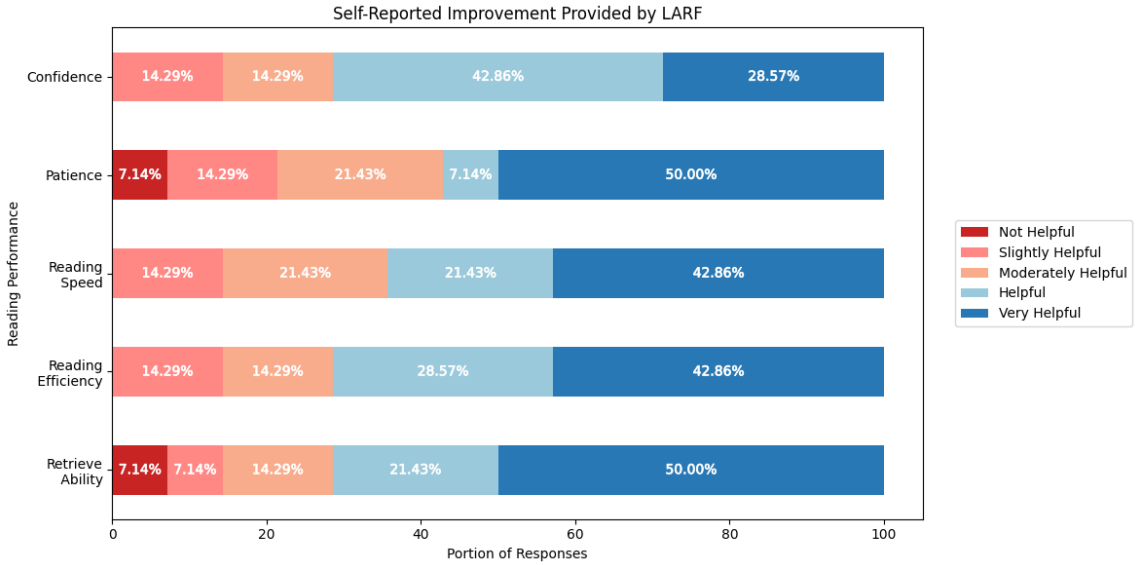

Fig. 16. Follow upstudy on users' experiences and performance improvements in reading with LARF. Most users believe that LARF enhances their reading speed, patience, efficiency, and ability to recall details.

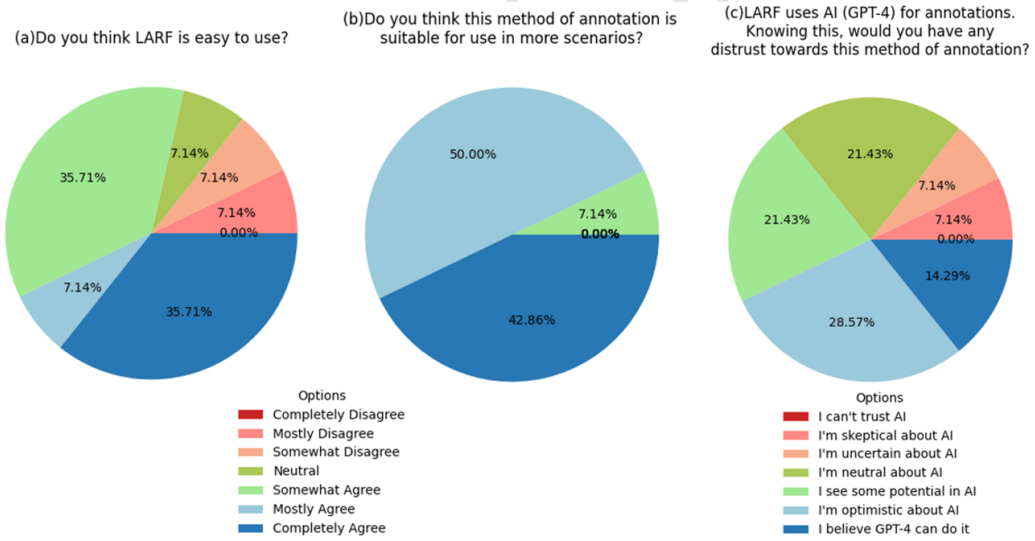

Fig. 17. (a)Surveying the ease of use of LARF, reveals that 50% of users think LARF is easy to use. (b) Regarding participants' support for the extension of LARF to other scenarios, shows that 100% of the participants expressed a positive view. (c) is about the trust level in AI annotation tools among users after being informed that LARF is generated by GPT-4, with 14.3% of the users expressing scepticism towards AI