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

Sihang Zhao
The Chinese University of Hong Kong, Shenzhen
Shenzhen, China
sihangzhao@link.cuhk.edu.cn

Shoucong Carol Xiong
Zhejiang University
Hangzhou, China
carolhsiung@zju.edu.cn

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

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

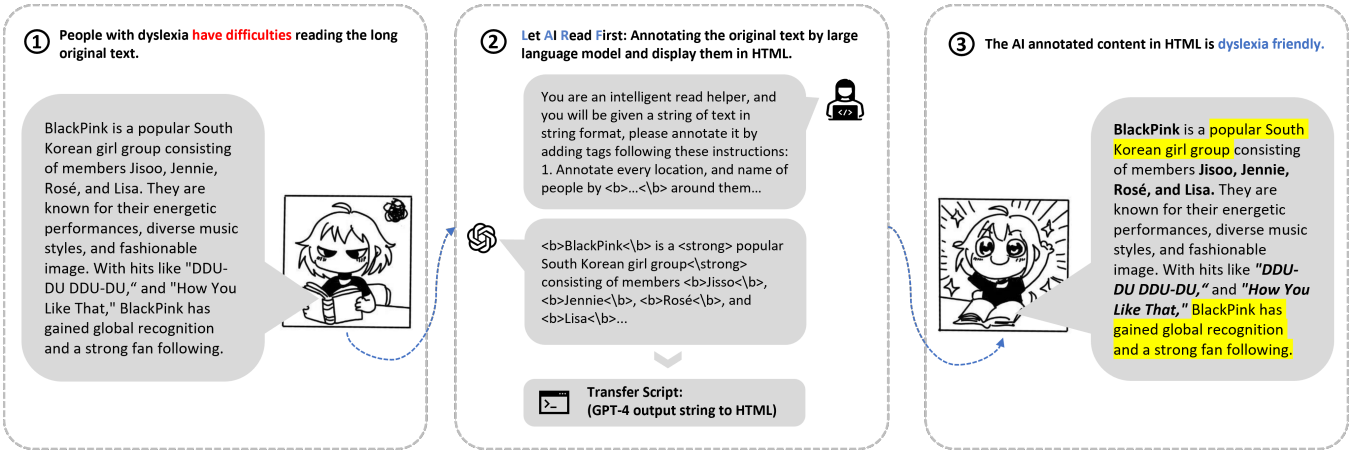

Figure 1: People with dyslexia always have difficulties while reading. We propose a method Let AI Read First (LARF) that uses language models to annotate the original text and display them in HTML format. Our experiment validates that LARF can improve reading performance and improve the reading experience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

Abstract

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 a large-scale between-subjects experiment, involving 150 participants with dyslexia. The results show that LARF significantly improves reading performance and experience for individuals with dyslexia. Results also prove that LARF is particularly helpful for participants with more severe reading difficulties. Furthermore, this work 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**.

Keywords

Dyslexia, accessibility

ACM Reference Format:
Sihang Zhao, Shoucong Carol Xiong, Bo Pang, Xiaoying Tang, and Pinjia He. 2025. Let AI Read First: Enhancing Reading Abilities for Individuals with Dyslexia through Artificial Intelligence. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25), April 26-May 1, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3706599.3720113

CHI EA '25, April 26-May 1, 2025, Yokohama, Japan

ACM ISBN 979-8-4007-1395-8/2025/04
https://doi.org/10.1145/3706599.3720113

1 Introduction

Dyslexia is a neurodevelopmental impairment that affects reading abilities, typically manifested as challenges to reading fluency, speed, and comprehension. 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. Current interventions, mainly in the form of accessible designs, tend to focus on only a few areas: converting text to speech [16], videos or games [17, 23], adjusting text font through electronic readers [26, 29] (e.g., character size, colour, spacing between words), and replacing complex words with simpler synonyms [27]. 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. (2) Converting text descriptions into videos or games manually can be both expensive and non-real-time. (3) Simple synonym substitution and rewriting may alter the original meaning, rhymes or emotions of the original texts. Compared to the available knowledge about reading difficulties and the demonstrated capabilities of AI models, there are relatively few accessible designs that effectively address these challenges [19]. With the rapid development of AI [7], numerous spelling assistance tools for dyslexia have demonstrated considerable capabilities [14, 15, 28]. 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 **AI** **R**ead **F**irst), the first 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 a large-scale experiment (N = 150) to evaluate this hypothesis. Participants self-reported having or likely dyslexia and having English as their mother tongue. They were randomly assigned into three groups: a control group that read the original reading materials directly, a conventional group in which participants read the same materials processed by Bionic Reading [25], and a LARF group that the reading materials annotated by OpenAI's GPT-4 [22]. Using multiple-choice questions, we tested the accuracy in recalling, retrieving details, and comprehension levels. The experimental results show that participants who read GPT-4 annotated texts demonstrate better reading performance compared to those using traditional methods or in 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 GPT-4 annotated texts significantly improve perceived user-friendliness, overall satisfaction, perceived helpfulness, future use, and recommendation tendencies. Users also believed that this method should be applied as a text presentation method for dyslexic populations in more scenarios (e.g., exams, accessible website design).

*Pinjia He is the corresponding author.

2 Related Work and Background

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 [16], 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 [17]. However, these software solutions are often limited by contextual restrictions, as text-to-speech conversion becomes impractical in settings that require silence, such as conferences or exams. Moreover, despite proven effectiveness [11, 31], the high cost of software like Kurzweil3000 limits its widespread adoption [11]. 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 [26, 29]. Santana et al. created Firefixia, which is a browser extension that enables dyslexic readers to tailor websites for enhanced readability [12]. Text4All [33], 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 [29]. 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 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 [27]. 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.

Niklaus et al. evaluated the digital reading rulers and found that digital rulers can help people with dyslexia better focus on the text and improve their reading speed [21]. Li et al. suggested Reader View websites with low visual complexity can benefit the reading performance and user experience of people with and without dyslexia [18]. Despite the wealth of knowledge surrounding reading difficulties, traditional accessible designs addressing these challenges remain limited [19]. 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.

3 Method and Data

3.1 Workflow of LARF

The workflow of LARF is illustrated in Fig. 1, in which LARF takes the original text as input in string format. Guided by preset prompts, GPT-4 processes the original text, incorporating the Hyper Text Markup Language (HTML) tags [4, 5], which can be used to manipulate the display of text, such as "bold," "highlighting," "italics," changing font colour, and adjusting font size. 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 (Fig. 1) shows a segment taken from Wikipedia about BlackPink [6]. GPT-4 was asked to highlight sentences that serve a summarizing role using <mark><\mark>tags and to bold important names 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 the subsequent experiment, we adjusted the prompts by using different labels, thereby modifying the presentation of the text. The detailed default prompts can be found in Appendix A.8.

3.2 Data

In the experiment, we processed the reading materials using GPT-4 API. We also used GPT-4 together with human evaluation to score the participants' short-answer responses in the subsequent experiment. The version of GPT-4 is the ChatGPT July 20 version, with the temperature set to 0 to ensure reproducibility of results. All the specific prompts and generation logs can be found in the supplement material and Appendix A.8. We employed the Bionic Reading [25] 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. Existing research suggests that Bionic Reading can improve students reading proficiency [2]. This tool includes two key parameters: "Fixation," which determines the expression of letter combinations, set to the default value of 3 (ranging from 1 to 5), and "Saccade," which controls the visual jumps between fixations, set to the default value of 10 (ranging from 10 to 50). In this paper, we also apply the default value of 10. The example of Bionic reading can be found in Appendix A.2.

4 Ethic & Transparency

This experiment was approved by the Institutional Review Board (IRB) of our affiliation. All participants were recruited through Prolific, an online research platform [24]. To ensure data protection and confidentiality, participants were informed that their responses would be anonymized, with all identifiable information removed before analysis. Additionally, the survey (including confidentiality information), 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 at https://github.com/LARF2025/LARF-CHI-EA-25.

5 Experiment

5.1 Experiment Setup

Our experiment focused on English language reading. We chose Reading Test 115, Passage 2, "The Step Pyramid of Djoser," a descriptive and factual reading text from the IELTS [10] 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 [1]. The IELTS Academic test is equipped with expertly formulated questions and standardized answers, which further enhance the reliability and validity of our study. Given these qualities, the IELTS Academic test serves as an ideal tool for assessing adult reading performance.

5.2 Method and Experiment Procedure

We recruited 150 participants (M_{age} = 36.8; 33.3% female) from Prolific [24], an online research platform. All participants either had a medical diagnosis of dyslexia, were undergoing a diagnostic process, or strongly suspected they had undiagnosed dyslexia. Additional demographic details are provided in Appendix A.11. Participants were randomly assigned to one of three experimental conditions: control (unmodified text), conventional tool (Bionic Reading), or LARF (GPT-4 annotations). In the LARF condition, participants were not informed that GPT-4 had produced the annotations, in order to minimize any psychological priming effects.

The study began with participants completing a Dyslexia Checklist (refer to Appendix A.4), 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 A.6). Following the recall task, participants were asked to retrieve as many details from the article as possible. Then, the same article was presented again, immediately followed by a reading comprehension assessment on the same page. After reading and finishing the reading comprehension assessment, 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 [8] 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. 2.

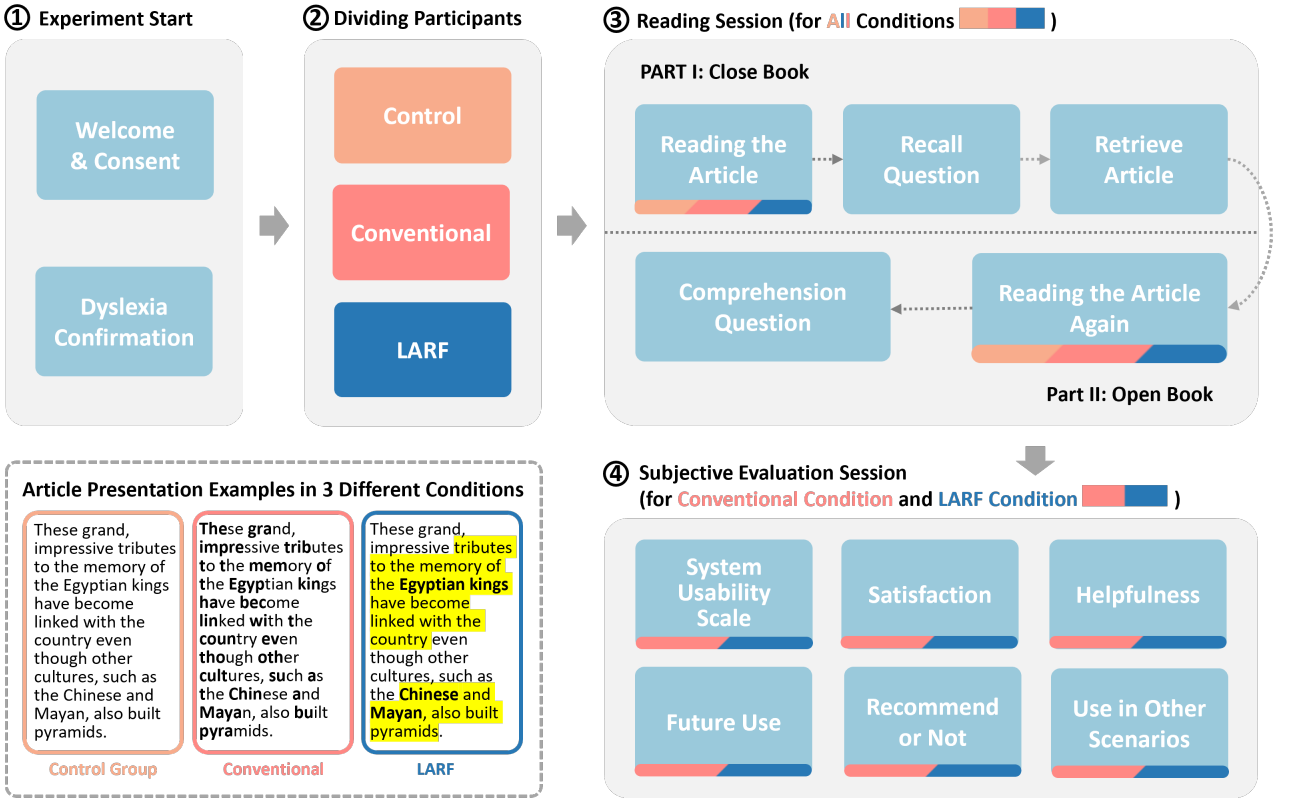

Figure 2: Experiment Procedure. Participants are randomly assigned to three conditions, and then they are asked to finish the reading session. They are required to read the same article but with different presentations. Participants in the conventional condition group and LARF condition group are required to finish a subjective evaluation session after they finish the reading session.

6 Result and Analysis

6.1 Attention Check and Dyslexia Checklist

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.3. Before reading the article, participants assessed their own dyslexia levels using the Dyslexia Checklist (see Appendix A.4 for Dyslexia Checklist). This checklist comprises six items that evaluate 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

6.2 Reading Time

Eight participants are identified as outliers based on their initial reading times, defined as reading times > 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 was conducted with 138 participants. Covariates, including education, age, gender, and dyslexia level, were accounted for in the analysis. We introduced the education level, age, gender, and dyslexia level as covariates in our one-way ANOVA analysis. It shows no significant differences in reading times across conditions ($M_{control} = 117.56, SD = 46.47$; $M_{conventional} = 122.57, SD = 62.70$; $M_{LARF} = 118.26, SD = 50.69$; $F(2, 129) = .160, p = .853$). However, considering the length of the corpus (338 words), and average reading speed of 238 words per minute for English readers [9], those who spent less than 30.2 seconds (0.05 quantile) were considered relatively impatient. Fig. 3(a) shows that participants using LARF did not fall below 30 seconds and were concentrated within a shorter, reasonable range. This suggests that LARF may aid in attracting user attention and enhancing reading patience and confidence.

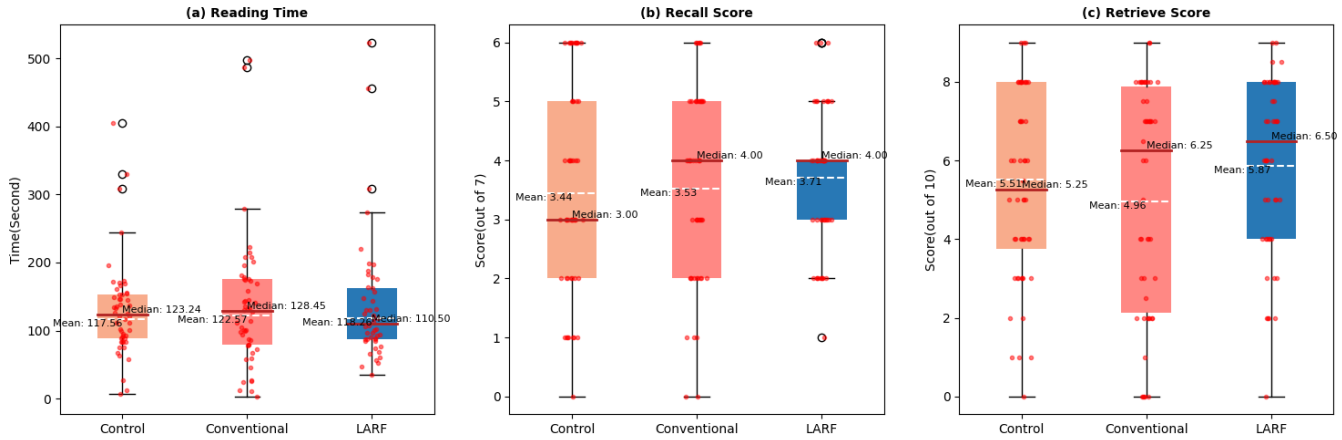

Figure 3: (a) shows the differences in reading time under three different conditions. Though the pattern is not significant, we can observe that users in the LARF group do less "glance over and skip the article." Furthermore, their overall reading time is more concentrated in areas with shorter durations. 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).

6.3 Recall and Retrieve Performance

In the recall section, participants answered six questions, with one point awarded for each correct response (example questions can be found in Appendix A.6). The analysis included education, age, gender, dyslexia level, and reading times as covariates in a one-way ANOVA. As shown in Fig. 3b, **participants in the LARF condition tended 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 ($M_{control} = 3.44, SD = 1.74; M_{conventional} = 3.53, SD = 1.64; M_{LARF} = 3.71, SD = 1.20; F(2, 128) = .303, p = .739$).

In the retrieve section, participants were instructed to retrieve as many details from the article as possible (example question can be found in Appendix A.7.) 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 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 Appendix A.9 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. 3c 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** ($M_{control} = 5.51, SD = 2.38, M_{conventional} = 4.96, SD = 2.89$).

6.4 Comprehension Performance

Comprehension performance was assessed using a similar method 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% of participants in the LARF condition correctly chose the exact two statements. In contrast, this accuracy was observed in 64.71% of participants in the control condition and 67.35% in the conventional condition. Additionally, we evaluated whether participants were able to identify at least one correct statement. In this regard, 100% of participants in the LARF condition succeeded in choosing at least one correct statement, whereas the corresponding figures were 92.16% for the control condition and 87.76% for the conventional condition. We conservatively believe that this indicates **LARF can to some extent enhance the participants' reading comprehension skills.**

6.5 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 Appendix C.1 Fig. 6. **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. The detailed questions and results are shown in Table 2 and Table 3 in Appendix A.10. **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.**

6.6 Post Hoc Evaluation

People with dyslexia often experience different subsets of challenges [20]. 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 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 indicate that LARF is especially helpful for participants with severe dyslexia.** As Fig. 7 in Appendix C 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. 9) and retrieval (Fig. 8) performance.

7 Discussions

Based on our theoretical foundation and software demonstration (refer to Appendix B), as shown in Fig. 10. LARF can be applied to various smart scenarios (e.g., PCs, tablets, and VR), yet high GPU requirements for LLM inference remain a challenge. Exploring smaller models that maintain annotation quality is thus crucial. From an HCI perspective, future work could investigate how best to present AI-generated annotations (e.g., highlight length, colour, or font) for users with reading difficulties, as well as the potential long-term effects on memory and learning. LARF may also be extended to subtitles in videos or live streams, although the impact on neurodiverse populations (such as individuals with ADHD) calls for further exploration. Design solutions should offer customizable annotation settings and integrate seamlessly with existing accessibility features; voice or gesture controls may be essential for VR or compact devices. To ensure privacy, local or end-to-end model inference is preferred, supported by model fine-tuning and well-crafted prompts to enhance annotation quality. Additionally, we have an interesting finding: compared to the control group, Bionic Reading does not appear to improve users' reading performance. In related studies published later than our experiment, they had the similar conclusion [30].

8 Limitation

During our experiments and software development, we faced several limitations. Considering the experiment cost and participants' patience, We kept comprehension and recall tasks relatively simple and limited the number of questions. However, some questions may have been "too easy," resulting in some observed patterns without clear statistical significance. We also decided not to include a "random labelling group" to account for placebo effects, though we believe such effects would be minimal. As mentioned in Section 7, this study does not investigate the interaction between different annotation types nor determine which is most beneficial. We likewise did not explore how to select optimal default prompts for user engagement. While changing font size and colour in HTML is possible, we have not addressed it here. Previous work [13] indicates that letting users set their own preferences can improve reading accuracy, so in our demo users can specify which information they want GPT-4 to annotate and how it should appear. Nevertheless, in our experiment, participants only used the default prompt. For BionicReading, we set the default parameter, and in real-world usage, users can also customise the settings.

9 Conclusion

We introduce LARF, an AI-annotated text approach designed to enhance the reading abilities of individuals with dyslexia. Our Experiment (N=150) validated LARF's effectiveness in improving dyslexic readers' performance and experience, including recall of details, reading comprehension efficiency, and engagement, outperforming the conventional technique.

Acknowledgments

We used LLMs to enhance the linguistic precision and coherence of the content. This work is funded by Guangdong Basic and Applied Basic Research Foundation (No. 2024A1515010145) and Shenzhen Science and Technology Program (No. ZDSYS20230626091302006)

References

[1] H Porter Abbott. 2020. *The Cambridge introduction to narrative.* Cambridge University Press.
[2] Etika Ariyani. 2023. IMPROVING STUDENTS READING PROFICIENCY USING BIONIC METHOD (A CLASSROOM ACTION RESEARCH AT 10th GRADE STUDENTS). *JOEL: Journal of Educational and Language Research* 3, 5 (2023), 215–226.
[3] International Dyslexia Association. [n. d.]. Frequently Asked Questions About Dyslexia. http://www.interdys.org/.
[4] Tim Berners-Lee, Robert Cailliau, Ari Luotonen, Henrik Frystyk Nielsen, and Arthur Secret. 1994. The world-wide web. *Commun. ACM* 37, 8 (1994), 76–82.
[5] Timothy J Berners-Lee and Robert Cailliau. 1990. WorldWideWeb: Proposal for a HyperText project. *World Wide Web Proposal* (1990).
[6] Blackpink. 2024. Blackpink. https://en.wikipedia.org/wiki/Blackpink. https://en.wikipedia.org/wiki/Blackpink Accessed: 2024-09-07.
[7] 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. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).
[8] John Brooke. 1996. Sus: a "quick and dirty'usability. *Usability evaluation in industry* 189, 3 (1996), 189–194.
[9] Marc Brysbaert. 2019. How many words do we read per minute? A review and meta-analysis of reading rate. *Journal of memory and language* 109 (2019), 104047.
[10] British Council. [n. d.]. IELTS. https://www.ielts.org/.
[11] Jennifer Cullen, Sue Keesey, and Sheila R Alber-Morgan. 2013. The effects of computer-assisted instruction using Kurzweil 3000 on sight word acquisition for students with mild disabilities. *Education and Treatment of Children* (2013), 87–103.
[12] Vagner Figueredo de Santana, Rosimeire de Oliveira, Leonelo Dell Anhol Almeida, and Marcia Ito. 2013. Firefixia: An accessibility web browser customization toolbar for people with dyslexia. In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*. 1–4.
[13] Anna Dickinson, Peter Gregor, and Alan F Newell. 2002. 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*. 97–103.
[14] Katharina Galuschka, Ruth Görgen, Julia Kalmar, Stefan Haberstroh, Xenia Schmalz, and Gerd Schulte-Körne. 2020. Effectiveness of spelling interventions for learners with dyslexia: A meta-analysis and systematic review. *Educational Psychologist* 55, 1 (2020), 1–20.
[15] Steven M Goodman, Erin Buehler, Patrick Clary, Andy Coenen, Aaron Donsbach, Tiffanie N Horne, Michal Lahav, Robert MacDonald, Rain Breaw Michaels, Ajit Narayanan, et al. 2022. Lampost: Design and evaluation of an ai-assisted email writing prototype for adults with dyslexia. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility*. 1–18.
[16] Kristen Laga, Daniel Steere, and Domenico Cavaiuolo. 2006. Kurzweil 3000. *Journal of Special Education Technology* 21, 2 (2006), 79.

[17] Andres Larco, Jorge Carrillo, Nelson Chicaiza, Cesar Yanez, and Sergio Luján-Mora. 2021. Moving beyond limitations: Designing the helpdys app for children with dyslexia in rural areas. *Sustainability* 13, 13 (2021), 7081.
[18] Qisheng Li, Meredith Ringel Morris, Adam Fourney, Kevin Larson, and Katharina Reinecke. 2019. The impact of web browser reader views on reading speed and user experience. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–12.
[19] Jacob E McCarthy and Sarah J Swierenga. 2010. What we know about dyslexia and web accessibility: a research review. *Universal Access in the Information Society* 9 (2010), 147–152.
[20] Meredith Ringel Morris, Adam Fourney, Abdullah Ali, and Laura Vonessen. 2018. Understanding the needs of searchers with dyslexia. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–12.
[21] Aleena Gertrudes Niklaus, Tianyuan Cai, Zoya Bylinskii, and Shaun Wallace. 2023. Digital Reading Rulers: Evaluating Inclusively Designed Rulers for Readers With Dyslexia and Without. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.
[22] OpenAI. [n. d.]. ChatGPT. https://openai.com/.
[23] Mikel Ostiz-Blanco, Javier Bernacer, Irati Garcia-Arbizu, Patricia Diaz-Sanchez, Luz Rello, Marie Lallier, and Gonzalo Arrondo. 2021. Improving reading through videogames and digital apps: A systematic review. *Frontiers in psychology* 12 (2021), 652948.
[24] Prolific. 2023. Prolific - Online Participant Recruitment for Surveys and Market Research. https://www.prolific.com/
[25] Bionic Reading. [n. d.]. Bionic Reading. https://bionic-reading.com/.
[26] Luz Rello and Ricardo Baeza-Yates. 2013. Good fonts for dyslexia. In *Proceedings of the 15th international ACM SIGACCESS conference on computers and accessibility*. 1–8.
[27] Luz Rello and Ricardo Baeza-Yates. 2014. Evaluation of DysWebxia: a reading app designed for people with dyslexia. In *Proceedings of the 11th Web for All Conference*. 1–10.
[28] Luz Rello, Clara Bayarri, Yolanda Otal, and Martin Pielot. 2014. A computer-based method to improve the spelling of children with dyslexia. In *Proceedings of the 16th international ACM SIGACCESS conference on Computers & accessibility*. 153–160.
[29] Luz Rello, Gaurang Kanvinde, and Ricardo Baeza-Yates. 2012. 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*. 1–9.
[30] Joshua Snell. 2024. No, Bionic Reading does not work. *Acta Psychologica* 247 (2024), 104304.
[31] Robert A Stodden, Kelly D Roberts, Kiriko Takahashi, Hye Jin Park, and Norma Jean Stodden. 2012. Use of text-to-speech software to improve reading skills of high school struggling readers. *Procedia Computer Science* 14 (2012), 359–362.
[32] Tate. n.d.. Carnation, Lily, Lily, Rose by John Singer Sargent. https://www.tate.org.uk/art/artworks/sargent-carnation-lily-lily-rose-n01615. https://www.tate.org.uk/art/artworks/sargent-carnation-lily-lily-rose-n01615 Accessed: 2024-09-11.
[33] V Topac. 2012. The development of a text customization tool for existing web sites. In *Text Customization for Readability Symposium*.

A Experiment Details

A.1 Bionic Reading

Bionic Reading is an application that present the first one or few character with bold effect. The Example of Bionic Reading is shown in Figure 5.

A.2 Bionic Reading Data

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

A.3 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.4 Dyslexia Checklist

We use the Dyslexia checklist 1 to ask participants to evaluate their extent of different difficulties in reading.

A.5 Reasons Using Prolific

Prolific encourages participants to disclose any health-related conditions, including dyslexia, allowing researchers to recruit specific individuals with relevant health conditions. Second, to control for factors such as time of day and time zone, which could potentially impact participants' cognitive function, Prolific allows us to limit recruitment to participants within the same time zone.

A.6 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.7 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.8 Default Prompt for GPT-4

This prompt is used for our experiment, 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 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>, , 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.9 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.10 Subjective Evaluation

The subjective evaluation consists of a system usability scale (see Table 2) and a general subjective evaluation scale (see Table 3).

A.11 Participants

Participants' demographics in the experiment can be found in Table 4.

B Software Demo

The LARF software demo shown in Figure 4 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.

C Supplemental Figures

C.1 Subjective Evaluation Result in Experiment

The subjective evaluation includes 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 ($M_conventional = 3.69, SD = 2.10; M_{LARF} = 4.96, SD = 1.96; F(1, 95) = 9.388, p = .003$). 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).

C.2 Post Hoc Evaluation

The Post hoc evaluation for the experiment: 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.3 More Scenarios

Besides LARF, there are lots of potential applications in different modalities and scenarios under the same idea: let AI decide and tell people what is worth attention. Some example is given in Figure 10.

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

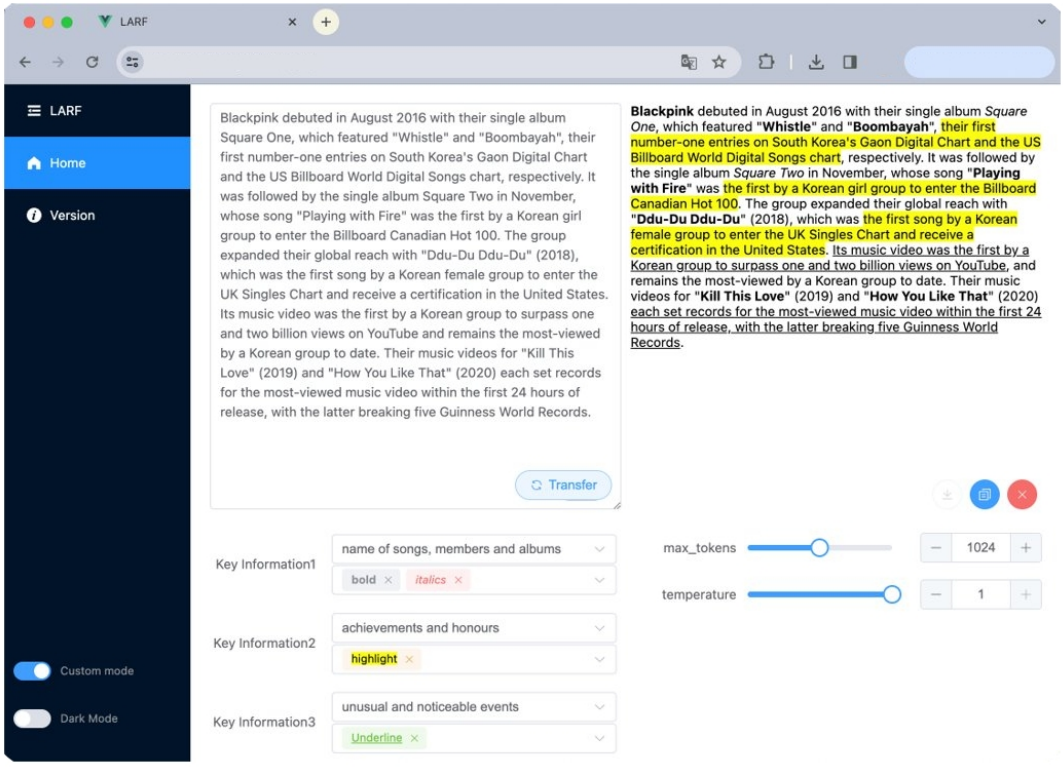

Figure 4: The demo of the custom mode of LARF software application on PC.

Table 1: Dyslexia Checklist for the Experiment

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 2: Subjective Evaluation - System Usability Scales

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 < 0.01
(3) SUS-3 is a reversed-scale question

Table 3: 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-readers, 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

Table 4: Participants Demographic in the Experiment

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		

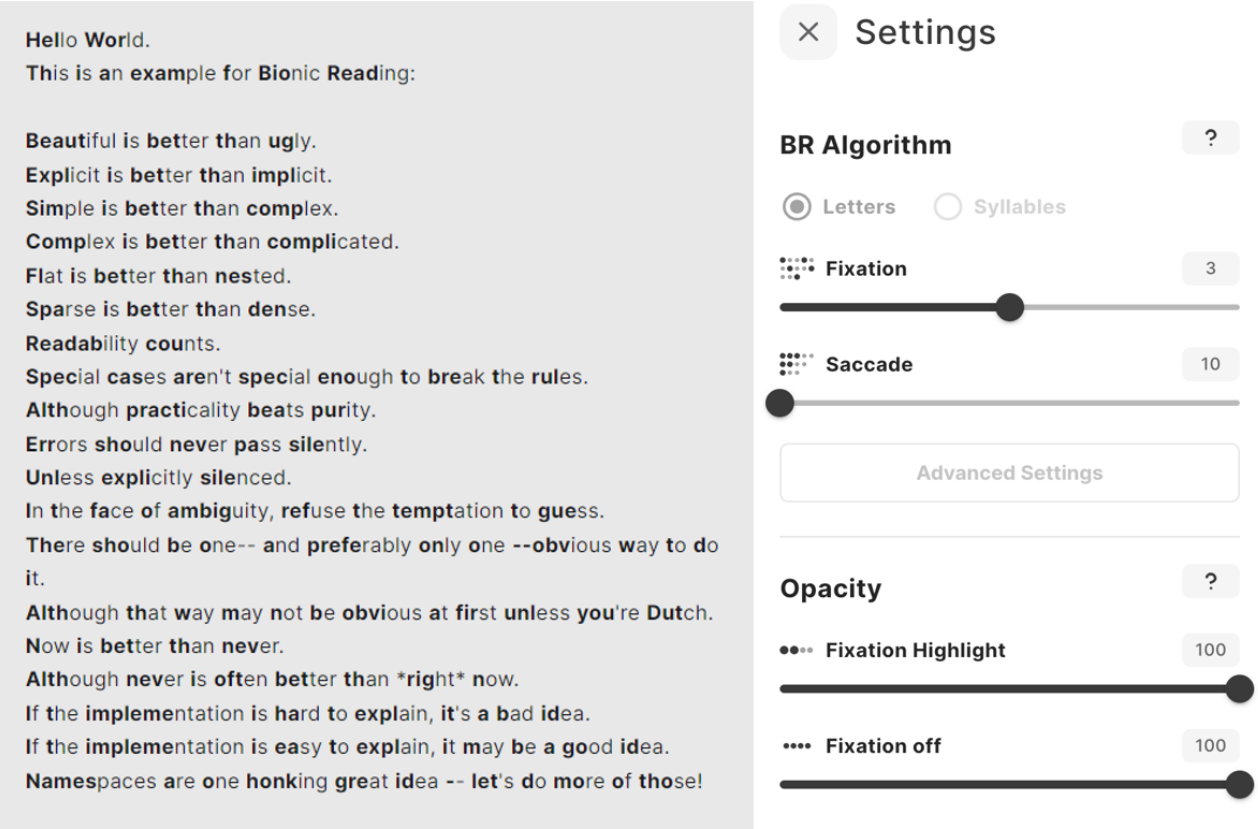

Figure 5: An example of the result and parameters of Bionic Reading

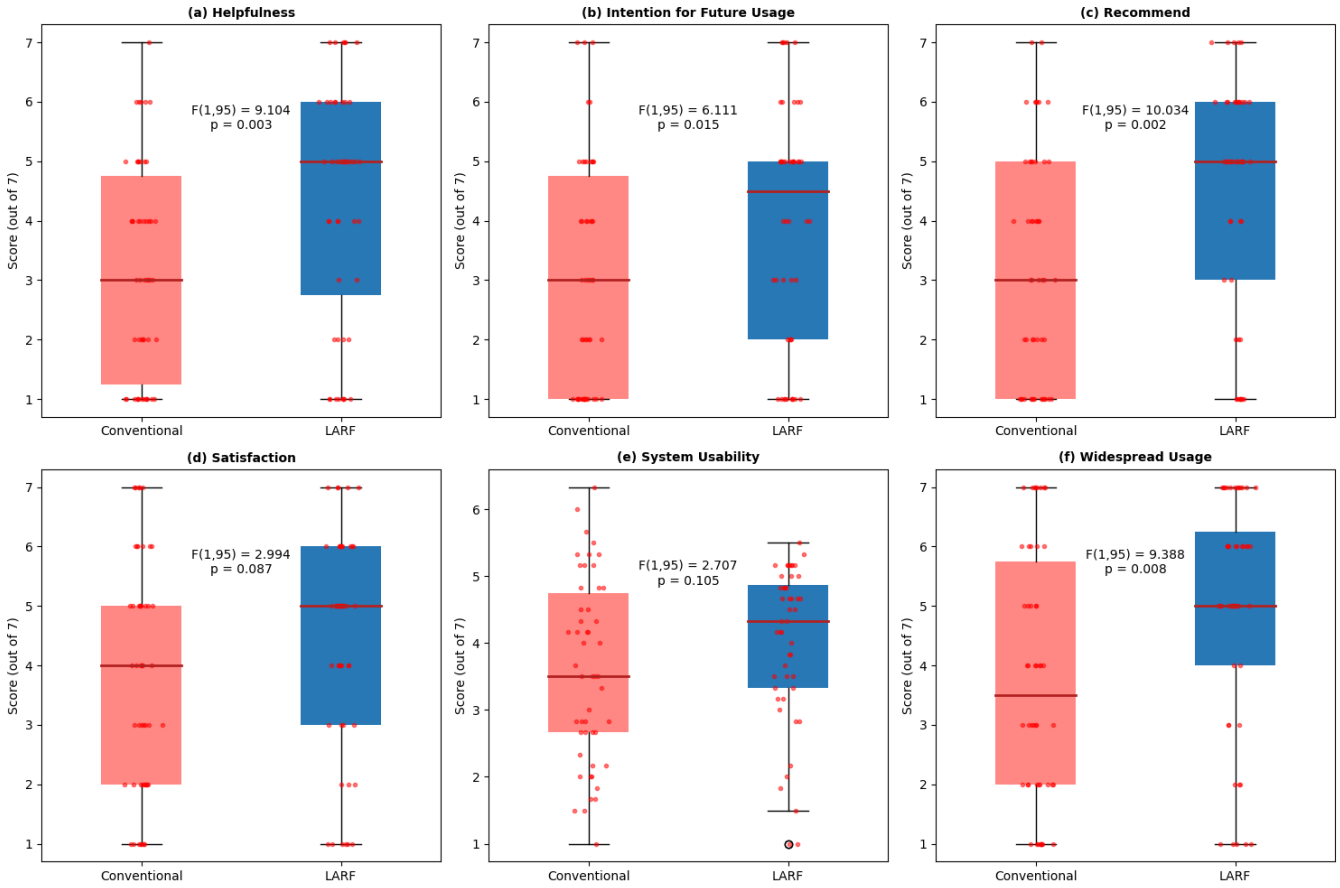

Figure 6: The subjective evaluation result. Participants with dyslexia exhibited a clear preference for LARF, considering text annotated with LARF to be effective, user-friendly, and worthy of broader adoption in various contexts.

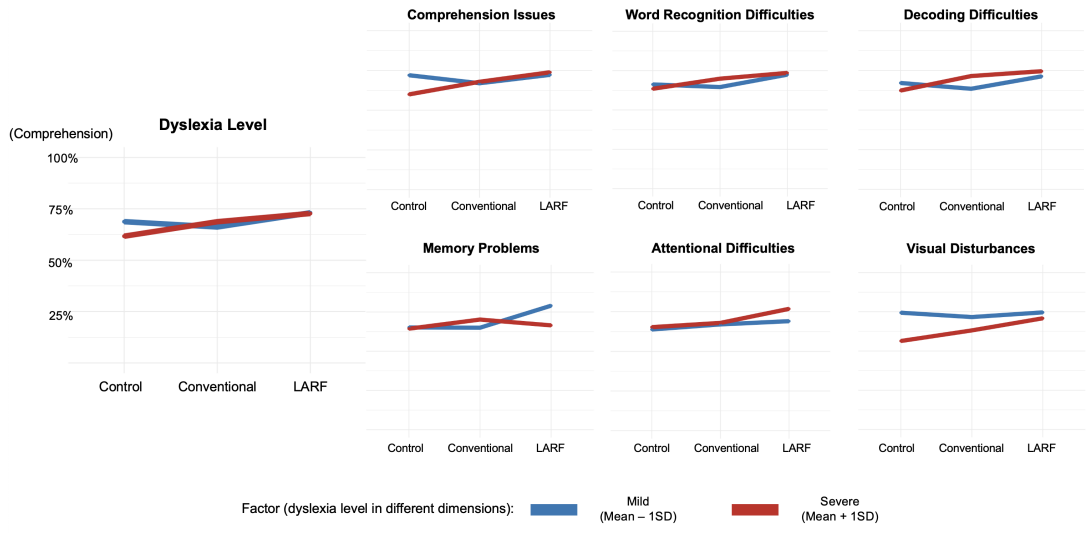

Figure 7: Post hoc evaluation for comprehension performance. The y-axis represents the accuracy of reading comprehension. In the group with severe symptoms, LARF exhibited significant improvement compare to the conventional group and control group.

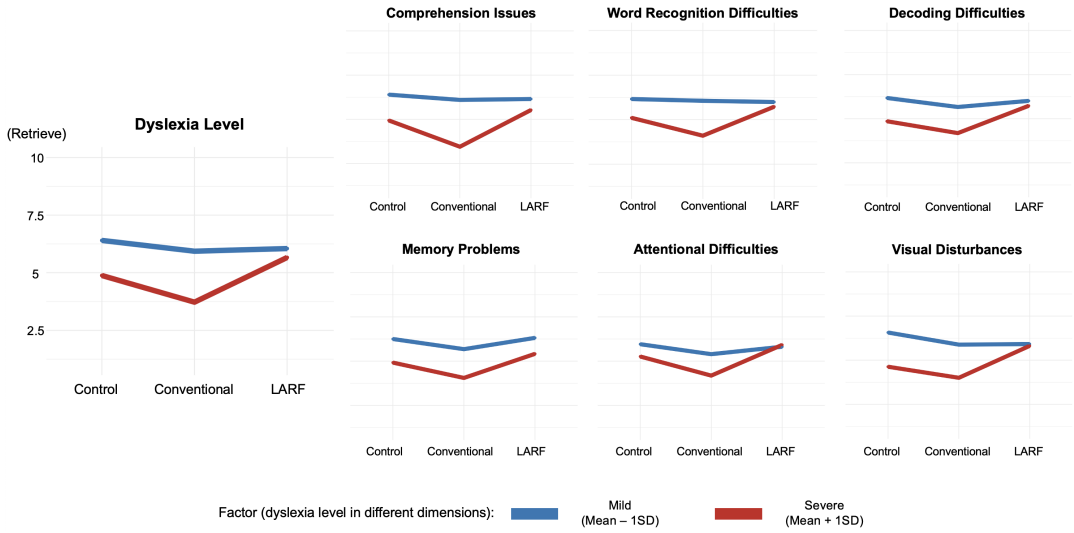

Figure 8: Post hoc evaluation for retrieving performance. The y-axis represents the scores for retrieve, with a maximum score of 10. While in the group with mild symptoms, LARF did not exhibit improvement, it significantly enhanced users' retrieval abilities in the group facing more severe reading challenges, whereas conventional tools had almost entirely negative impacts.

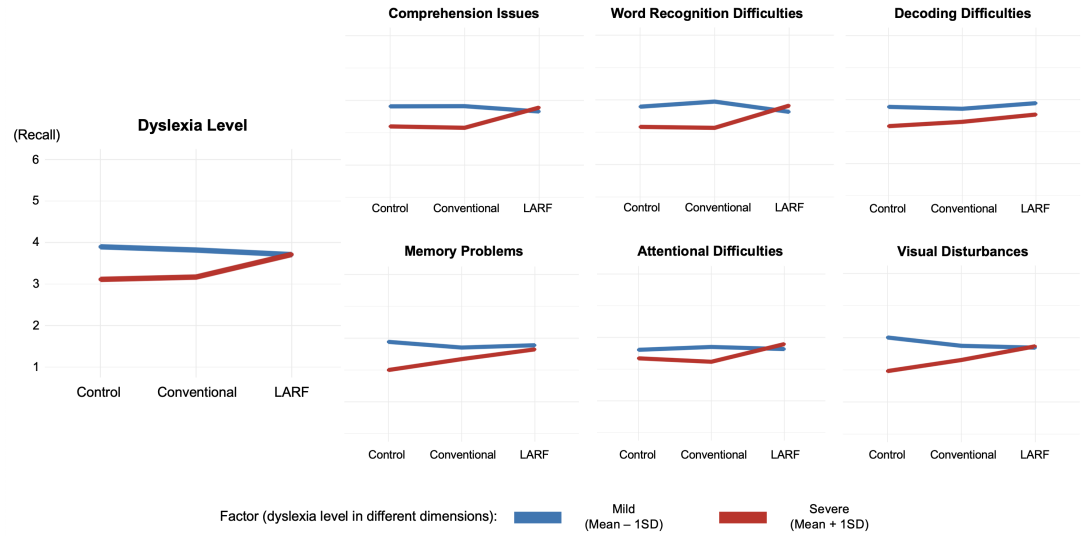

Figure 9: Post hoc evaluation for recall performance. The y-axis represents the scores for recall, with a maximum score of 6. LARF similarly provided substantial assistance to the group with more severe symptoms, even surpassing the group with mild symptoms who also used LARF.

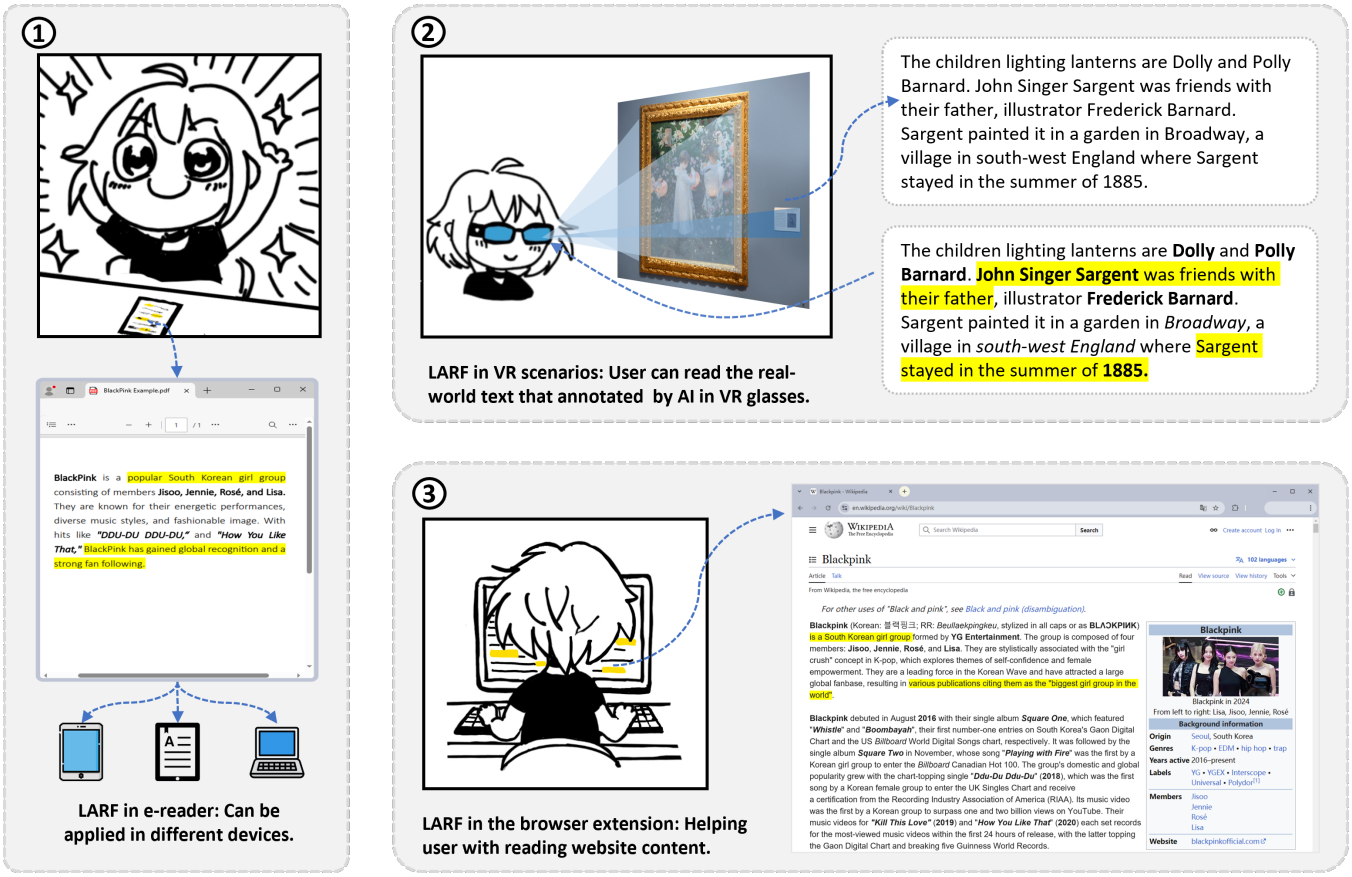

Figure 10: Real-world application scenarios that can apply LARF. In the first subplot, the user is using an e-reader which is integrated with LARF, this device can be a tablet, a smartphone or a laptop. In the second subplot, the user is wearing VR glasses, looking at the "Carnation, Lily, Lily, Rose" by John Singer Sargent [32]. 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[6].)