



## CReBot: Exploring interactive question prompts for critical paper reading

Zhenhui Peng <sup>a,\*</sup>, Yuzhi Liu <sup>b</sup>, Hanqi Zhou <sup>d</sup>, Zuyu Xu <sup>c</sup>, Xiaojuan Ma <sup>c</sup>

<sup>a</sup> Sun Yat-sen University, School of Artificial Intelligence, Xiangzhou District, Tangji Road, Zhuhai, 519082, Guangdong Province, China

<sup>b</sup> Hong Kong University of Science and Technology, School of Science, Clear Water Bay, Kowloon, Hong Kong, China

<sup>c</sup> Hong Kong University of Science and Technology, Department of Computer Science and Engineering, Clear Water Bay, Kowloon, Hong Kong, China

<sup>d</sup> University of Tübingen, Tübingen, Germany



### ARTICLE INFO

#### Keywords:

Paper reading  
Interactive tools  
Pedagogical conversational agent  
Question prompts  
Critical thinking

### ABSTRACT

Pre-compiled guidelines with a static question list can stimulate critical thinking while reading a scientific paper. However, they could be less engaging than taking live question prompts from others. In this paper, we develop CReBot that interactively asks section-level critical thinking questions and customize it for routine paper readers with prior research experience and novices new to research. Our first within-subjects study with 24 routine readers demonstrates CReBot's engagement and usefulness over static guidelines. Then, with more teacher-like question-specific hints prepared for CReBot, we conduct another within-subjects study with 20 novices. The results, however, indicate that CReBot might not be better than static guidelines for beginners. Nevertheless, both user groups favor CReBot's contextualized questions and interaction flexibility. We conclude with design implications for interactive tools to facilitate critical reading.

### 1. Introduction

Critical thinking is one of the higher-order skills that university students need to learn (Fadel et al., 2015; OECD, 2018). When applied to reading scientific publications, it requires readers to question the texts as they study a claim, a method, a result, or other content (Wallace and Wray, 2016). For example, a critical reader should ask whether the evidence convincingly supports a paper's claims and why these claims matter to the audience (Wallace and Wray, 2016). Critical reading guidelines or prompts that consist of a checklist of questions compiled by experienced researchers provide great material to encourage critical thinking (University of Toronto, 2020; Shum, 2020; Keshav, 2007; Tomasek, 2009), e.g., "What are the authors trying to do in writing this? How convincing is what they are saying? And in conclusion, what use can I make of this?" (Wallace and Wray, 2016).

While these checklist-like question prompts serve as a good starting point for practicing critical paper reading, they are rather static and thus tend to be less engaging and useful than receiving interactive questions and opinions about the paper from others (e.g., peers, teachers, supervisors) (Wilson et al., 2004; Wiles et al., 2016). For example, in a group meeting, colleagues can usually raise critical thinking questions about the presented paper content, stimulating the presenter to think of the paper from different aspects. For paper reading on computer screens, existing works have explored possible means to facilitate critical thinking, such as online collaborative reading platforms (Tan

et al., 2016) and the projection of teachers' reading patterns on digital papers (Cheng et al., 2015). However, these methods require human peers or teachers to read the same materials beforehand or at the same time to raise suitable questions. They hence fall short in scalability as such qualified social questioners are not always available.

A bot is a possible alternative to supporting critical paper reading by imitating how humans interactively ask questions. As summarized by Weber et al. (2021), the pedagogical conversational agents (the bots in educational domains) have been demonstrated to be effective and engaging in a variety of tasks. They act as tutors or peers and interact with users of different ages in multimodal ways (Ruan et al., 2019; Winkler et al., 2020). For example, Wambsganss et al. built ArgueTutor that helps students to write convincing arguments by offering adaptive feedback on persuasiveness, and they showed that it is more enjoyable than a baseline tool with general and documented recommendations (Wambsganss et al., 2021). Winkler et al. developed a bot named Sara that asks students questions about their understanding of programming concepts during video lectures. They also demonstrated that it improved students' learning in programming tasks (Winkler et al., 2020). Such interactive bots often follow the theory of scaffolding (Mariane, 2002) by prompting questions and providing support during users' task completion processes. Nevertheless, little work has looked into the design, usefulness, and user experience of an interactive question-prompts

\* Corresponding author.

E-mail addresses: pengzh29@mail.sysu.edu.cn (Z. Peng), yliugt@connect.ust.hk (Y. Liu), hanqi.zhou@uni-tuebingen.de (H. Zhou), zxubo@connect.ust.hk (Z. Xu), mxj@cse.ust.hk (X. Ma).

bot for critical paper reading. Unlike the knowledge acquisition tasks in which there is usually a correct answer for each question, an interactive critical thinking facilitator should encourage readers to develop their own understanding and interpretation (Wallace and Wray, 2016; Yu, 2015). Also, it should offer users sufficient critical thinking support while mitigating any possible interference in the paper reading process (Head et al., 2021). Furthermore, the design and usefulness of such an interactive facilitation may vary across the users' levels of reading and research experience with the paper topics, as previous research suggests that novices generally desire more assistance and guidance and perform differently in their tasks compared with seniors (Miller and Bailey, 2014). Therefore, despite the success of interactive bots in other educational scenarios, gaps exist in exploring the design and usage of a question-asking bot to engage users with different levels of research experiences in critical paper reading.

To this end, we develop CReBot (Critical Reading Bot) that asks section-level questions and provides guidance in real-time to facilitate users in critical paper reading and explore its usage in reading HCI (Human–Computer Interaction) publications. Specifically, our design and evaluation of CReBot are customized for two types of potential users in two stages. In the first stage, we target the CReBot as a peer-like agent for assisting routine paper readers who generally have prior experience in research and paper reading, e.g., postgraduate students. As they read each section of a paper in a web browser, CReBot, located in the sidebar, prompts associated questions from the critical paper reading guidelines pre-compiled by our design team. Users can respond to the bot's question, switch to another question of the same or a different section, and add their own critical thoughts any time they want. We evaluate this CReBot against the conventional checklist of instructions and questions (denoted as "Guideline") via a mixed-methods, within-subjects study with 24 routine paper readers, including 19 graduates and five undergraduates with prior research experience. The results show that CReBot better engages these routine paper readers in the reading process by encouraging more critical reading behaviors, and it is perceived significantly more useful and easier to use. Participants suggest that such interactive facilitation from CReBot might also be useful for novices who are new to scientific research or seldom read academic papers before.

Inspired by the first-stage findings, we refine CReBot as a teacher-like agent specifically for novices in the second stage. First, we verify the novices' needs and identify their requirements of CReBot via a Speed Dating study (Zimmerman and Forlizzi, 2017; Ma et al., 2015), where we present eight novices with storyboards of CReBot usage scenarios and with low-fidelity prototypes for user enactments. Then, based on the Speed Dating results, we mainly augment the previous CReBot design by offering hints from experts that help novices look for possible answers to each prompted critical thinking question. Next, we evaluate the refined CReBot for novices via a within-subjects study with 20 undergraduate students who have little or no research experience. Experimental results show that CReBot achieves comparable performance as the Guideline in engaging novices in the critical reading process and helping them comprehend and critique the paper content. Both tools are deemed equally useful and easy to use by these participants. Novices generally favor CReBot's interactive questions and hints for critical paper reading but feel that the CReBot's interactions sometimes cause interruptions as it does not match their traditional reading habits.

In summary, we contribute a bot prototype to explore interactive question prompts for critical paper reading support. Our two-stage investigation adds to the understanding of how people with different levels of paper reading experience perceive and read scientific publications with such a bot. We also discuss the generality of CReBot and propose design considerations for improving the usefulness and user experience of interactive technologies for paper reading and critical thinking support. Meanwhile, we contribute a question bank that encourages critical thinking of research publications.

## 2. Related work

### 2.1. Critical paper reading and its guidelines

The ability to think critically is widely regarded as an essential skill people should gain through high-level education (Fadel et al., 2015; OECD, 2018). Critical thinking requires not only understanding the problem, text, or answer (i.e., *comprehension*) but also "making clear, reasoned judgments" (*criticism*) to them (Beyer and Kappa, 1995). For those who need to read scientific papers to learn from existing literature and get insights for their works, it is crucial to think critically in their reading practices (Wallace and Wray, 2016). To do that, readers should actively listen to what a text says and respond in their own voice, e.g., they ask questions, think of examples to challenge the text, and relate it to their purposes or experiences (Wilson et al., 2004; Bakhtin et al., 2010). To learn and exercise critical reading skills, people can get guidance from senior scholars, who usually give tutorials and organize classroom reading activities (Wilson et al., 2004; Wilson, 2016). In a more accessible way, they can also refer to online critical thinking questions and guidance compiled by experienced researchers (Wallace and Wray, 2016; Keshav, 2007; University of Toronto, 2020; Yu, 2015; Shum, 2020). For example, to read critically, a handout from the University of Toronto recommends that people should analyze "what the patterns of the text are", interpret "what the patterns of the argument mean", and evaluate "how well the text does what it does and what its value is" (University of Toronto, 2020). Such guidelines could be useful for both novices and those who already have some experience in critical paper reading, as a written set of general rules, principles, or pieces of advice from others are generally helpful for users to complete given tasks (Miniukovich et al., 2019; Bharadwaj et al., 2019; da Rocha Tomé Filho et al., 2019; Agapie et al., 2018; Lai et al., 2020).

While these critical reading guidelines (denoted as "Guideline" condition in the experiments) traditionally contain a checklist of question prompts that users can ask about the text, they are static and generic, which could be less engaging and useful than the interactive questions from others. As argued by Lev Vygotsky in the theory of scaffolding (Kim, 2001), we learn best when we socially interact with others, who offer us questions and hints that are adaptive to our status and help us improve. Ngoon et al. provided empirical evidence to this insight by showing that the interactive techniques – adaptively presented suggestions and guidance – improve the quality of feedback from novice reviewers on creative work, compared to the static ones (Ngoon et al., 2018). In this paper, we explore the design of interactive tools for reading papers critically, and we compare our tool against the Guideline to evaluate its usefulness and user experience with both routine paper readers and novices.

### 2.2. Interactive tools to facilitate paper reading

Researchers in HCI (Human–Computer Interaction) have explored various interactive tools to facilitate people's reading experience. These tools can be categorized into two groups regarding how they manipulate the reading materials. The first group of approaches does a content "reduction" usually for helping people skim a text and grab its main idea (Lee et al., 2016; Kobayashi and Kawashima, 2019; Kim et al., 2018; Graham, 1999). For example, Kobayashi et al. proposed to sequentially fade-out characters sentence-by-sentence for each paragraph and demonstrated its effectiveness in improving readers' text comprehension (Kobayashi and Kawashima, 2019). Kim et al. presented an interactive reading system that automatically links document text with corresponding table cells and showed that it increased users' accuracy and speed in navigating the matched sentences of table cells (Kim et al., 2018). The other way of reading assistance "enriches" the reading materials with extra resources and guidance (Wang et al., 2016; Khan et al., 2020; Subramonyam et al., 2020; Romat et al., 2019; Collins et al., 2009; Peng, 2021). For example, Khan et al. designed an audio

skimming app that blends auditory and visual reading for situational impairments (e.g., walking) (Khan et al., 2020). Head et al. introduced an augmented reading interface that can provide users with easy access to the definitions of technical terminology and mathematical symbols (Head et al., 2021). To help users quickly track and glance related work, Wang et al. developed a visualization system that presents the literature review as interactive slides (Wang et al., 2016).

Existing ideas of interactive facilitation for paper reading fall into the content enrichment type of methods (McCartney et al., 2018; Cheng et al., 2015). For example, Cheng et al. developed a SocialReading system that shares teachers' gaze data for an academic paper and validated that it can improve students' reading comprehension of that paper (Cheng et al., 2015). Tan et al. designed and implemented WiREAD, which offers a collaborative environment with practices, questions, and discussions for peers and instructor to engage in critical paper reading together (Tan et al., 2016). Unlike the tools mentioned above, CReBot can facilitate critical reading via question prompts without the need to have teachers or peers read the same materials beforehand or at the same time.

### 2.3. Interactive bots in educational domains

According to the Google's English dictionary provided by Oxford Languages, a 'bot' is an autonomous program on the internet or another network that can interact with systems or users. The bot's accessibility and ability to imitate human behaviors make it increasingly popular in educational domain (Kerry et al., 2009; Heffernan and Koedinger, 2002). Pedagogical conversational agents (PCA) are representatives of educational bots (Weber et al., 2021). PCAs often leverage the theory of scaffolding (Mariane, 2002) and serve as the person who better understands the material and scaffolds the material in smaller chunks that expand the learners' knowledge. For example, the tutoring system AutoTutor has been used to teach college students in multiple domains, such as computer literacy and critical thinking (Nye et al., 2014). AutoTutor provides explanations, feedback, scaffolding, deep reasoning questions, and subject content in online courses, and multiple studies have demonstrated its effectiveness in improving learning gains (Nye et al., 2014). Similarly, Wambsganss et al. designed ArgueTutor that judges argumentative writing performance of users' essay and suggests how to improve (Wambsganss et al., 2021). Educational bots commonly use questions as a start to engage users in learning. For example, Winkler et al. developed Sara that acts like a teacher to asks students questions during an online video lecture (Winkler et al., 2020). They demonstrated in a lab experiment that Sara could significantly improve learning than the without-Sara condition in a programming learning task (Winkler et al., 2020). Ruan et al. created QuizBot, an interactive agent that asks questions and provides feedback to users' answers in learning factual knowledge about science, safety, and English vocabulary (Ruan et al., 2019). They showed that QuizBot engaged users better in the learning process than the traditional flashcard tool, and users preferred the bot strongly for casual learning (Ruan et al., 2019).

Inspired by these successful educational bots, we explore the possibility of an interactive bot for critical paper reading support. Our CReBot uses questions as the main interactive materials because they are generally effective in encouraging thinking (LW et al., 2001; Liao et al., 2020; Syed et al., 2020). CReBot does not give corrective feedback to readers' responses as Sara and QuizBot do in their knowledge acquisition tasks with standard answers. Instead, it focuses on prompting suitable critical thinking questions and encouraging users with hints to find or think of the answers (Wallace and Wray, 2016; University of Toronto, 2020). We seek to provide empirical evidence for whether the general benefits of PCAs, e.g., improved engagement and efficiency in learning (Weber et al., 2021; Ruan et al., 2019; Winkler et al., 2020), also exists in such a question-asking bot in critical reading support tasks. In all, to the best of our knowledge, our work is the first to probe the design, usefulness, and user experience of interactive question prompts from a bot for critical paper reading.

### 3. Stage 1: Developing CReBot for routine paper readers

Our exploration consists of two stages. In the first stageIn the first stage, we develop CReBot that acts as a peer-like assistant for those who generally have paper reading experience and need to frequently read scientific papers. Critical paper reading is an important requirement for them (Wallace and Wray, 2016), and CReBot that imitates the colleagues' question-asking behaviors could be a good facilitator in this process. In this paper, we treat them as "routine paper readers" to differentiate this user group from Stage 2's novices who have little paper reading experience or are new to research. These routine paper readers can be full-time researchers in companies, postgraduate students, or undergraduates that have deep engagement in paper reading and research. As the first work to explore the design and usage of an interactive question-asking bot for critical reading support, we evaluate CReBot to address the following research questions:

**RQ1:** Compared to the conventional guidelines that list the questions (denoted as Guideline in this paper), how would an interactive question-asking bot (CReBot) affect the *routine paper readers'* (i) behaviors, (ii) perceived engagement and difficulty in the process, and (iii) perceived performance in the outcome of critical paper reading?

**RQ2:** (i) How would the *routine paper readers* use the CReBot and Guideline in practice, and (ii) what is their acceptance towards such tools for critical paper reading?

To support the critical thinking during users' paper reading process, we present our system: CReBot. Based on the surveyed Related Work, the existing paper reading support tools (Duggan and Payne, 2011; McCartney et al., 2018; Cheng et al., 2015) do not provide interactive critical thinking questions and guidance, which are helpful for critical reading (Wallace and Wray, 2016) and scaffold users' learning process (Mariane, 2002). Therefore, to provide such critical reading support for general users, we derived the following design requirements (DR) and criteria (C) for CReBot:

**DR1:** To engage users in the critical paper reading process, CReBot should provide guidance on critical thinking and ask related questions (Wallace and Wray, 2016; of Leeds, 2021).

**DR2:** To be easy to generalize (C1: reproducibility) (Xiao et al., 2020), easy to access (C2: practicality) (Peng et al., 2020; Xiao et al., 2020), and easy to learn (C3: familiarity) (Wambsganss et al., 2020), CReBot should adopt publicly available technologies that do not require installation of extra hardware/software and use an intuitive chatbot-like interface design.

In this section, we first present how we fulfill DR1 with the critical reading guidelines and questions customized in the HCI (Human-Computer Interaction) paper domain as a case demonstration. We then describe how we follow DR2 to develop CReBot, a proof-of-concept prototype that facilitates users to read papers critically by interactively prompting section-level questions from the guidelines on the fly. Next, we report an empirical evaluation of CReBot with our Stage 1 target users and present the results.

#### 3.1. Critical reading guidelines with questions customized in HCI domain

To build up the knowledge base of CReBot, we compile critical paper reading guidelines based on the related and publicly available articles, tutorials, and books (University of Toronto, 2020; Wallace and Wray, 2016; Shum, 2020; Mitzenmacher, 2020; Keshav, 2007; Alake, 2020). The guidelines start with (0) *setting a reading goal* that can help readers establish a constructive purpose for reading the paper critically (Wallace and Wray, 2016) and (1) *understanding the general idea of the paper* from the title, abstract, introduction, heading of each section, and conclusions (Wallace and Wray, 2016; Keshav, 2007). When (2) *digging into each section*, a critical reader should first *comprehend* its content (i.e., identify **what** the authors are trying to do) and then *criticize* it by examining **how** the authors achieve their purposes, understanding **why** the authors do it in this way, and assessing **how**

well the authors' writing and logic are (Wallace and Wray, 2016; University of Toronto, 2020; LW et al., 2001). Finally, readers can reflect on the paper by summarizing its strengths, weaknesses, and how it contributes to their reading goals (Wallace and Wray, 2016; University of Toronto, 2020; Mitzenmacher, 2020).

Next, we customize the prototypical questions of guidelines into the HCI paper domain as a demonstrated case of CReBot, since the structure of a paper and the contextual critical thinking questions for each paper section could vary across different research fields. Using papers published on CHI2019 (The ACM CHI Conference on Human Factors in Computing Systems<sup>1</sup>), a top HCI (Human–Computer Interaction) venue, as an example, we showcase our three-phases question customization process as below:

**1. Identify common paper sections.** We first collect all CHI2019 proceeding papers and late-breaking works (LBW) from its website and extract their outlines using the PyPDF2 package (Stamy, 2020), which captures the papers' section and subsection titles. We sort these titles by their numbers of occurrences and include the frequently used ones for further analysis (in our case: > 3 for proceeding papers, 168 items; > 1 for LBW, 93 items). Then, two authors independently go through these headings and inductively group them into potential categories. After determining the categories of common (sub)sections through two rounds of comparisons and discussions, they assign the headings to each category independently. The intraclass correlation coefficient (ICC), a standard measure for quantifying the degree to which a fixed number of raters have consistent judgments (Bartko, 1966), is 0.761, suggesting a good consistency in the coding (Cicchetti, 1994). Finally, they resolve the disagreement via discussion and output 19 categories of sections with commonly used titles (Fig. 1).

**2. Collect section-level questions via interviews with domain researchers.** To collect a question bank for critical reading in the HCI domain, we conduct semi-structured online interviews with four HCI researchers (three males, one female; age:  $M = 28.8$ ,  $SD = 1.7$ ) using the video conference tool Zoom (Zoom Video Communications, 2020). They all have over three years of experience in reading HCI papers (90+ papers per year), submitting CHI papers (at least one accepted), and reviewing CHI papers. In the interview with each participant, we first present the general guidelines and ask for their relevance to personal critical reading experiences. Then, we invite the interviewee to recall what critical thinking questions are in mind when reading paper (sub)sections in each of the 19 categories, with example titles, contents, and prototypical questions on the shared materials.<sup>2</sup> In the end, we ask for their research experience, common means to exercise critical paper reading, and feedback to the potential question-prompting bot for critical reading support. We compensate each participant \$10 for about 50 min spent in the study. All interviewees agree that the general guidelines are consistent with their critical reading experiences and are positive about the potential bot's usefulness. They usually read HCI papers about creativity support, virtual reality, gamification, education, visualization, human-AI interaction, and voice interface. They mention three ways for learning critical paper reading: experience of and instructions received for reviewing HCI papers; reviewers' comments on their own submissions; and questions from lab mates or advisors. Therefore, apart from the 276 section-level questions from the interviewees, we also add two other sources to enrich the question bank: (1) guidelines for reviewers in CHI papers (e.g., Sigchi (2020)); and (2) audiences' questions about papers shared in weekly HCI group meetings for three weeks at a local university. In total, we collect 363 critical thinking questions for common (sub)sections in CHI publications. They also suggest that the questions should be further categorized, as readers can think critically about a (sub)section from different aspects, e.g., the *motivation* and *results* in the Introduction.

**3. Assign critical thinking aspects and levels to each question.** Two authors who majored in HCI further contextualize the questions by first identifying the aspects in each question category via two rounds of coding and discussions on the 363 questions. They then independently label the questions in each aspect into levels of *what*, *how*, *why*, or *how well*. They reach an excellent level of agreement over the labels (ICC = 0.920) and discuss and resolve the conflicts. Fig. 1 shows the example section categories and corresponding questions with critical thinking aspects and levels.<sup>3</sup>

### 3.2. CReBot System design and implementation

#### 3.2.1. User interaction with CReBot

We build CReBot as a responsive web-based application (Fig. 2a) embedded in a paper reading web page powered by react-pdf-highlighter (Tyurin, 2020) and pdf.js (Mozilla, 2020) (C1, C2). The right part of the interface (i.e., part i in Fig. 2) displays the paper content which users can zoom in/out, highlight, and comment on (C3). CReBot stays in the left sidebar to set up users' expectation of where to interact with (C3). Following the interface design of educational bots (Winkler et al., 2020; Ruan et al., 2019; Xiao et al., 2020), CReBot displays the interaction history (ii) and its current message (iii) in a chat window (C3). Users can receive from CReBot the critical thinking questions with context indicated by “category ⇒ aspect ⇒ level” on top and the general guidance for this level of questions by hovering the mouse on “Tip”. To interact with CReBot, users can (1) type down and “send” their critical thoughts in reply to the current question; (2) see “next question” of the same section; and (3) switch to questions of a specific section or subsection (iv). When ready, users can click “Finish reading” to get questions and guidance to reflect on the paper (v). To allow users to record their own critiques of the papers, if any, CReBot also offers an “add their critical question” option. Further, to help users record and recover their critical thinking flow, CReBot has an “export all chat logs” feature. Both features can also help the CReBot to be expanded as a community-based critical reading tool by encouraging users' contributions on its question pool in the future.

#### 3.2.2. Backend server and interaction logic

After the web app loads in a paper, it extracts the paper's outline and sends it to a python flask server, which uses the fuzzywuzzy package (SeatGeek, 2020) to compute similarity scores between each (sub)section title and our collection of heading items in the 19 categories (e.g., item “Design implication” in category “Implication”). It returns the (sub)sections of the loaded paper that have similarity scores over a threshold (90% in our case) as the candidates for CReBot to ask questions about (Fig. 2a part iv). After a self-introduction, CReBot follows the steps of guidelines by first asking users for reading goals and encouraging them to grasp the paper's general idea. It then stimulates the readers to dig into each section of the paper and prompt relevant critical questions. CReBot will prompt a new question every time readers answer the current one or click “Next Question” using a weighted-chance strategy. Specifically, we manually put more weights on questions in the same critical thinking level (50%), followed by the next level (30%) and the next aspect (20%). This strategy could encourage enough thinking on the current level's questions before users go into the next level or aspect. To achieve this logic, CReBot will generate a random variable between 0 and 1, with the number  $\in [0,0.5]$ ,  $\in [0.5,0.8]$ ,  $\in [0.8,1.0]$  for determining the next question in the same level, next level, and next aspect. When it runs out the questions of all aspects, it will circle back with a random question about the current (sub)section. This strategy can increase the coverage of possible question alternatives (Wilson et al., 2004) and follow the

<sup>1</sup> <https://chi2019.acm.org/>

<sup>2</sup> The interview materials can be found in Appendix A or <http://zhenhuipeng.com/>.

<sup>3</sup> The full set of CHI section categories and customized critical thinking questions can be found in Appendix A or <http://zhenhuipeng.com/>.

Categories	Aspects	Example questions	Levels	Categories	Aspects	Example questions	Levels
Abstract		<ul style="list-style-type: none"> <li>• <b>Background:</b> What is the general background of this paper? (1)</li> <li>• <b>Target:</b> Who is it in dialogue with? (1)</li> <li>• <b>Proposed work:</b> How did the authors solve the problems? (2)</li> <li>• <b>Findings:</b> What is the conclusion of the paper (1)</li> <li>• <b>Relevance:</b> Is this paper related to your research interests? (4)</li> </ul>		Participant		<ul style="list-style-type: none"> <li>• <b>Sample size:</b> Is the sample size of the participants enough? (4)</li> <li>• <b>Background / Demographics:</b> What is the background of the participants? (1)</li> <li>• <b>Recruitment:</b> Are the participants required to have domain knowledge? (2)</li> <li>• <b>Validity:</b> Are the participants representative as targeted users? (4)</li> </ul>	
Introduction		<ul style="list-style-type: none"> <li>• <b>Background / Problem:</b> What is the background of the problem? (1)</li> <li>• <b>Motivation:</b> Why should we care about this research problem? (3)</li> <li>• <b>Proposed work:</b> Is the method of this work novel? (4)</li> <li>• <b>Findings:</b> What is the benefit of the findings? (4)</li> <li>• <b>Contribution:</b> Why do the contribution and benefit matter? (3)</li> </ul>		Task		<ul style="list-style-type: none"> <li>• <b>Purpose:</b> How do the authors decide this task? (2)</li> <li>• <b>Design:</b> Is this task well-designed? (4)</li> <li>• <b>Clarity / Replicability:</b> Do the authors clearly explain the task? (4)</li> </ul>	
Related Work		<ul style="list-style-type: none"> <li>• <b>Logic / Structure:</b> What categories of related works are presented? (1)</li> <li>• <b>Previous work:</b> Why are people doing this research? (3)</li> <li>• <b>Need of this work:</b> How is this problem situated in the literature? (3)</li> </ul>		Procedure		<ul style="list-style-type: none"> <li>• <b>Details:</b> Why are the authors running the procedure in this way? (2)</li> <li>• <b>Payment:</b> Do the participants get reasonable payment for their time? (4)</li> <li>• <b>Clarity:</b> Does the procedure include any necessary detail? (4)</li> </ul>	
				Measures		<ul style="list-style-type: none"> <li>• <b>Purpose:</b> What is the purpose of each measurement? (1)</li> <li>• <b>Details:</b> Do the measures reflect the variables properly? (4)</li> <li>• <b>Validity:</b> Have these measures previously used in literature or not? (2)</li> </ul>	

Fig. 1. Example categories of sections/subsections in CHI papers, corresponding questions with critical thinking aspects and levels (1 - what, 2 - how, 3 - why, 4 - how well).

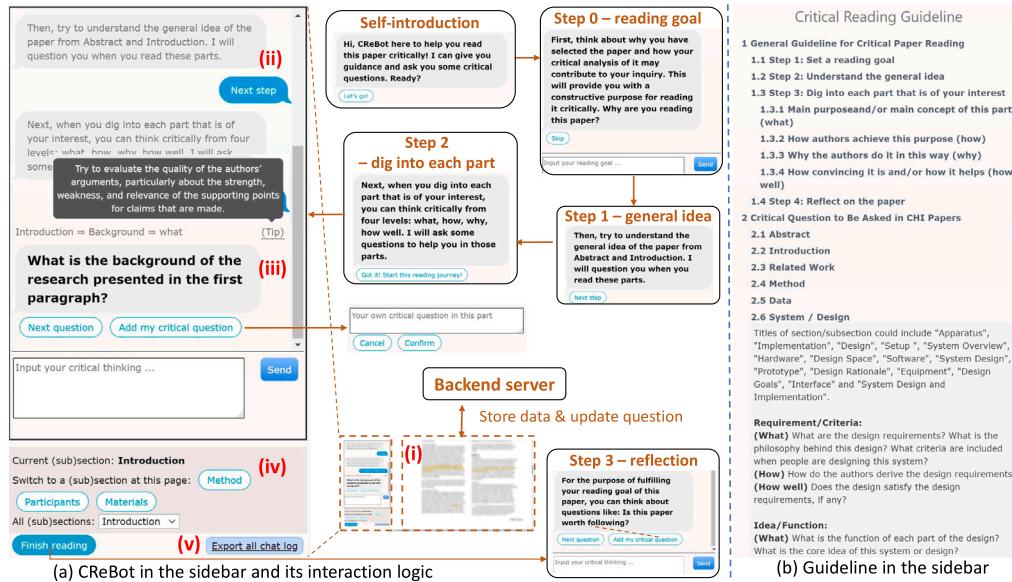


Fig. 2. (a) CReBot and interaction logic following the compiled critical paper reading guidelines: (i) Interface for the reading material in both CReBot and Guideline conditions; (ii) Interaction history; (iii) Current message; (iv) A panel for switching a (sub)section; (v) Buttons to proceed and export the reading notes. (b) The Guideline tool that documents the critical paper reading guidelines.

“select a topic randomly” approach to show the bot’s proactivity — one social characteristic of chatbots. Further, when users scroll across pages, CReBot will proactively switch to an unseen section of the new page and ask questions as the high-proactivity agents could provide sufficient facilitation to users (Peng et al., 2019; Chaves and Gerosa, 2021). At the current stage, CReBot cannot respond to users’ questions but mainly serve as a critical thinking questioner. We discuss this limitation and future work in Section 5.4.

### 3.2.3. Dialogue design

We position CReBot more as an interactive reading support tool than an anthropomorphized actor. Specifically, CReBot’s messages in step 2 of the critical paper reading process (Fig. 2) are pure questions, such that users can focus on critical thinking without the additional need to apply human social rules and expectations during the interaction (Rapp et al., 2021). Nevertheless, CReBot’s dialogue design matches to certain conversational intelligence of common chatbots (Chaves and Gerosa, 2021), such as conscientiousness (e.g., show buttons and chat log to maintain conversation flow) and communicability (e.g., advertise the functionality and suggest the next step).

## 3.3. Experiment

To investigate CReBot’s usefulness for facilitating routine paper readers in critical reading and its user experience compared with that of the static guidelines, we conduct a mixed-methods, within-subjects study with 24 participants (P1-24; 9 Females, 15 Males; age range

19 – 61,  $M = 25.5, SD = 7.9$ ). We adopt the within-subjects rather than the between-subjects study design to minimize the random noise from the participants, e.g., their mood or stress levels. To mitigate the transfer of knowledge between two conditions, we set a 24-hours interval between two tasks and counter-balance the tools’ order as described in Section 3.3.2. We recruit participants via social networks, word of mouth, and snowballing, and they have proven certificates of fluent English skills or are native English speakers. Fourteen participants are postgraduate students, five are undergraduates, and the rest five are researchers in the industries. They all have a routine need to read scientific papers. Twenty of them major(ed) in Computer Science (CS), and the others major(ed) in Cognitive Science, Design, or Electronic Engineering. In general, our participants have moderate experiences in reading CS papers ( $M = 5.0, SD = 1.5$ ) and HCI papers ( $M = 4.6, SD = 1.6$ ; 1 - No experience at all, 7 - A lot of experience), and they are interested in HCI research ( $M = 6.1, SD = 1.0$ ; 1 - No interest at all, 7 - A great deal of interest). On average, their self-reported skills in critical reading is 4.1 ( $SD = 1.2$ ; 1/7 - Extremely incompetent/competent).

### 3.3.1. Baseline: The guideline tool

We denote the baseline tool as Guideline (Fig. 2b), which takes the popular form of existing critical reading assistance (e.g., University of Toronto (2020), Keshav (2007)) with a documented list of questions that can be asked about the paper. We embed it in the left sidebar of the web app to minimize unfair comparison with CReBot due to other usability issues irrelevant to critical reading. Users can check the Guideline conveniently without switching to another webpage. A click

on a title or subtitle will unfold/fold its corresponding content. The Guideline could serve as a stronger baseline than a condition without any guidance, because participants can decide whether to use the tools or not while reading the papers. Future work can involve a control group without either the CReBot or Guideline to study the effects of critical thinking guidance on critical readers.

### 3.3.2. Task

Participants read two CHI late-breaking works (LBW; six pages excluding reference) critically with either the CReBot or Guideline as if they were going to share the papers in an HCI group meeting, and the audience would ask them questions. LBW “represents work that has not reached a level of completion or maturity that would warrant the full refereed selection process” and “should have potential to, with time, make a contribution to the body of HCI knowledge” (Sigchi, 2019). Therefore, they are suitable for CReBot evaluations in the lab study. We randomly sample two papers with little technical content (e.g., math equation), equal length, and different topics from CHI2019 LBWs to lower the technical bar for participants and mitigate the learning effects across two conditions. Paper 1 is about user perception with strategies of artificial intelligence in offering suggestions (3151 words), and paper 2 is about the importance of visualizing the human body in a virtual reality application (3279 words). Both papers are perceived with moderate difficulties to read in the post-survey (item: “The paper itself is difficult to read”; 1/7 - Strongly disagree/agree; paper 1:  $M = 4.3$ ; paper 2:  $M = 3.5$ ). None of the participants had read these papers prior to our study. We counterbalance the order of two tools/papers conditions using Latin Square to minimize the potential order effect. To reduce learning effects and minimize fatigue, we schedule the two reading tasks of each participant at least 24 h apart.

### 3.3.3. Measurements

**RQ1. i) Reading behaviors.** We log the completion time of each reading task, the number of highlights and comments, and the number of times scrolling into the previous/next page. The task completion time includes the time interacting with CReBot/Guideline since when checking and responding to its questions, users’ attention would switch back and forth between the chat box and the paper to locate the relevant content. **ii) Perceived engagement and difficulty in the critical reading process.** We adapt three 7-point Likert scale questions (Cronbach’s  $\alpha = 0.852$ ; Table 1) about focused attention from the perceived engagement scale (Wu et al., 2020). We also measure participants’ perceived difficulty in reading the assigned paper critically (one item adapted from (Hart and Staveland, 1988)) to estimate how CReBot/Guideline affects their cognitive load. **iii) Perceived performance in the critical reading outcome.** We initially collected users’ responses to four questions (e.g., what are the potential weaknesses of the paper) about the paper in each post-survey. However, we found that the responses are quite open-ended and hard to quantify the quality in a consistent manner. For example, if two users point out the same number of weaknesses of the given paper but cover different points at different levels, it may be difficult to compare their critical thinking outcomes. We instead obtain the participants’ ratings of their perceived critical reading performance to understand how CReBot might help from the users’ point of view. We measure it from three aspects regarding deep examination of the arguments, identifications of possible flaws, and reinterpretation of the paper for improved clarity (Althusser and Balibar, 1970) (Table 1).

**RQ2. i) Interaction with CReBot/Guideline.** To understand how participants use CReBot, we log the number of their textual responses to the prompted questions, the number of questions added by users, as well as the number of times users click “Next question” and switch a (sub)section to be prompted questions about. For the usage of Guideline, we record the amount of clicking behaviors on the headings of each section category. We also ask for their frequency of using the

tool in the questionnaire (Table 1) and inquire under what circumstance they use it in the interview after each task. **ii) Interruption and technology acceptance.** To evaluate user perceptions towards CReBot/Guideline, we first ask participants to rate the levels of reading interruptions by the tools (adapted from (Peng et al., 2019)). Then we adapt the technology acceptance model (Wambsganss et al., 2020; Venkatesh and Bala, 2008) to measure the following in each condition: **usefulness** (four items, Cronbach’s  $\alpha = 0.890$ ); **easy to use** (four items,  $\alpha = 0.841$ ); and **intention to use** (two items,  $\alpha = 0.862$ ). With the high  $\alpha$  scores ( $> 0.8$ ) of the items, we average the ratings of multiple questions as the final score for each factor in the acceptance model, which tend to correlate almost perfectly with “real” factor scores but are easier to understand<sup>4</sup>.

Previous work about PCAs suggests that a scaffolding-based bot could improve the learning experience and outcome compared to static, non-scaffold tools (Ruan et al., 2019; Winkler et al., 2020). Considering that CReBot can interactively prompt questions like a more knowledgeable human to scaffold the critical reading process, we hypothesize that:

**(H1A)** Compared to Guideline, CReBot significantly engages participants more in the critical reading process.

**(H1B)** Compared to Guideline, CReBot can significantly improve users’ perceived critical reading performance;

In comparison to Guideline, users perceive CReBot to be **(H2 A)** significantly more useful, **(H2B)** easier to use, and **(H2C)** of a stronger intention to be used in the future.

### 3.3.4. Procedure

Fig. 3 illustrates the study procedure conducted online via Zoom. At the beginning of each task, we first present task instructions and demonstrate the web app interface using a third example paper via a shared screen on a demo website. Then participants proceed to the main task and start reading the given paper with CReBot/Guideline. We allocate 35 min for each reading session based on a pilot study with two users and tell the participants that they can finish early or take more time if needed. We also inform them that we do not restrict whether, when, and how they use the tool. After reading each paper, participants rate their engagement in the reading process, perceived performance, and perceptions about the tool in a post-survey. We further conduct a semi-structured interview with them to make sense of the ratings and collect feedback on when they use the tool. Upon completion of two tasks, we ask which tool they prefer and why as well as suggestions to improve it. After debriefing, each participant receives USD \$20 as compensation.

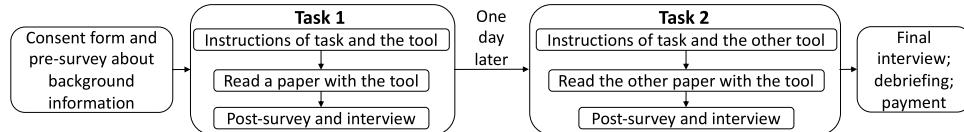
## 3.4. Analysis and results

We perform Wilcoxon signed-rank test (Woolson, 2008) to assess the difference in the participants’ ratings regarding various measurements of the CReBot and Guidelines. The Wilcoxon signed-rank test is commonly used to compare two sets of scores that come from the same participants (e.g., in HCI studies (Kang et al., 2021; Yan et al., 2021; Weinman et al., 2021)). The test affirms that the quantitative results do not suffer from the tool/paper order or tool-paper assignment. For the interview recordings, two of the authors transcribe them into text and conduct a thematic analysis Braun and Clarke (2006) subsequently. They first familiarize themselves by reviewing all the interview data independently, and after discussion, they form a list of initial codes. After several rounds of coding with comparison and discussion, they consolidate different codes into potential overarching themes, which are CReBot/Guideline’s pros and cons (Table 2), usage patterns of the tools, and suggestions for improvement. Lastly, they independently

<sup>4</sup> Suggested by the SPSS Factor Analysis — Beginners Tutorial. <https://www.spss-tutorials.com/spss-factor-analysis-tutorial/>

**Table 1**

Measurement	Items
RQ1/3 (ii) Perceived engagement (Wu et al., 2020)	<b>Engagement</b> (Cronbach's $\alpha = 0.852$ ) • I was absorbed in this critical reading process. • I was so involved in reading this paper that I lost track of time. • I was really engaged in the critical reading process.
RQ1/3 (ii) Perceived difficulty (Hart and Staveland, 1988)	• How difficult it was for you to read this paper critically? (1 – Very easy, 7 – Very Difficult)
RQ1/3 (iii) Perceived performance (Althusser and Balibar, 1970)	• I had a <i>deep examination</i> of some claims, their supporting points and/or possible counterarguments in this paper. • I <i>identified some possible ambiguities and flaws</i> in the author's reasoning, and even thought some ways to address them comprehensively in this paper. • I <i>reinterpret</i> and reconstructed some points of the paper for improved clarity and readability in this paper.
RQ2/4 (i) Interaction	• How frequently did you refer to CReBot's guidance for critical reading? (1 – Never, 7 - Always)
RQ2/4 (ii) Interruption (Peng et al., 2019)	• I found CReBot interrupting my reading process.
RQ2/4 (ii) Technology acceptance (Wambsganss et al., 2020; Venkatesh and Bala, 2008)	<b>Usefulness</b> (Cronbach's $\alpha = 0.890$ ) • The use of the CReBot enables me to read papers in a more critical manner. • Using CReBot improves my performance in digesting this paper. • The use of CReBot enhances my effectiveness in my critical reading task. • I find the CReBot useful in my critical reading process. <b>Ease of use</b> (Cronbach's $\alpha = 0.841$ ) • I would find the CReBot to be flexible to interact with. • My interaction with the CReBot is clear and understandable. • Interacting with the CReBot does not require a lot of my mental effort. • I find it easy to get what I want from the CReBot.
	<b>Intention to use</b> (Cronbach's $\alpha = 0.862$ ) • If the CReBot is available there to help me read my interested papers critically, I would use it. • I intend to be a heavy user of the CReBot when I want to have a critical reading on the papers.

**Fig. 3.** Procedure of the within-subjects (tool: CReBot vs. Guideline) experiment. In each task, participants read a given paper with either CReBot or Guideline based on their assigned order.**Table 2**

Summary of routine paper readers' interview responses to pros and cons of CReBot/Guideline in Stage 1.

	Pros (# of participants mentioned)	Cons (#)
CreBot	Conversational style of interaction (10); Checklist for examining papers deeply (8); Help understand the paper's idea (6); Contextual questions (10); Auto question update + selection panel (4); "Export all chat log" feature (10)	Bring distraction (2); Some questions not matched to the reading context enough (7); Malfunctions of auto question update (9)
Guideline	Checklist for examining papers deeply (8); Systematic and detailed tutorial for critical reading (10); More freedom for exploration (3)	Bring distraction (4); lack of interactive features (8); Too much content on a page (16)

assign the final codes to the interview data and reach substantial agreement (Cohen's  $\kappa = 0.628$ ). They resolve the disagreement with discussions. We count the occurrences of codes and incorporate these qualitative findings in the following presentation of our results.

### 3.4.1. RQ1 results

**Table 3** summarizes the RQ1 results about CReBot's impact on user's critical reading process and outcome in comparison to the Guideline. **i) Reading behaviors.** In general, participants spend significantly more time on reading a paper with CReBot ( $Mdn = 36.00\text{mins}$ ) than

**Table 3**

RQ1 results in the stage 1's experiment with 24 routine paper readers. 7-point Likert scale (1/7 - Strongly disagree/agree) for (ii) and (iii).

Mean (SD). \*\*:  $p < .01$ , \*\*\*:  $p < .001$ . #: number of occurrences.

(i) Reading behaviors			(ii) Critical reading process		
	Time (min) ***	Highlight #	Comment #	Scroll pages #	Engagement ***
CReBot	<b>36.4 (8.4)</b>	11.8 (8.2)	2.4 (3.6)	58.0 (33.6)	<b>5.49 (1.07)</b>
Guideline	30.9 (8.4)	13.3 (10.3)	3.5 (4.7)	58.0 (33.2)	4.54 (1.34)
(iii) Perceived performance					
	Examine arguments	Identify flaws		Reinterpret the paper **	
CReBot	5.17 (1.27)	4.63 (1.41)		<b>5.08 (1.41)</b>	
Guideline	4.79 (1.06)	4.88 (1.26)		4.13 (1.26)	

they do with the Guideline ( $Mdn = 33.00\text{mins}$ );  $T = 226.50$ ,  $z = -3.25$ ,  $p < .001$ ,  $\eta^2 = .66$ . They actively read papers in their tasks, with no significant differences ( $p > .05$ ) between the two conditions regarding the number of highlights, comments, and scrolling across pages (Table 3i). However, we observe in the CReBot condition that participants additionally respond to a total of 129 what, 33 how, 13 why, and 85 how well questions from the bot in the chat window. This indicates that participants devote extra efforts to digesting and addressing the critical points raised by CReBot, which can be viewed as practicing critical thinking skills. It would be interesting for future work to study whether and whether these reading behaviors impact users' critical thinking skills.

**ii) Perceived engagement and difficulty in critical reading.** In general, the participants feel that they are significantly more engaged with the reading material in the CReBot condition ( $Mdn = 5.67$ ) than in the Guideline condition ( $Mdn = 4.67$ );  $T = 232.00$ ,  $z = -3.43$ ,  $p < .001$ ,  $\eta^2 = .70$ ; H1A accepted. Ten participants explicitly mention that CReBot's conversational style of interaction (Table 2) helps them focus on critical reading. "I like the bot. It asks me questions like my advisors to keep me thinking during the reading process" (P24, Male, age: 22). No significant difference is found regarding the perceived difficulty in reading the paper critically with both tools ( $p > .05$ ).

**iii) Perceived performance in critical reading outcome.** On average, the participants feel that they do equally well in both conditions in terms of having a deep examination of the arguments as well as identifying possible flaws in the authors' reasoning ( $p > .05$ ). They comment that both CReBot (eight users) and Guideline (also eight users) can serve as a checklist for examining the papers more deeply. "After I took a pass on the paper, I chatted with CReBot. It asked me about the strengths and weaknesses of this paper, which reminded me to examine the paper's weak and novel points" (P21, M, 23). "If I read the paper by myself, I would have blind spots. The Guideline reminded me to think of my goals for reading this paper and how it can contribute to my future research. I checked the discussion part and made highlights accordingly" (P10, Female, age: 27). However, participants find that they do better in reinterpreting the paper for improved clarity under CReBot's ( $Mdn = 5.00$ ) than Guideline's ( $Mdn = 4.00$ ) assistance;  $T = 104.00$ ,  $z = -2.56$ ,  $p < .01$ ,  $\eta^2 = .52$ . H1B is thus partially accepted. By answering CReBot's questions, users can understand the paper's ideas better, as suggested by six participants. "CReBot helped me sort out the pipeline of the paper in my own way" (P8, M, 22).

#### 3.4.2. RQ2 results

**i) Interaction with CReBot/Guideline.** As shown in Table 4i, overall, participants rate that they refer to CReBot ( $Mdn = 6.00$ ) for critical reading assistance significantly more frequently than to Guideline ( $Mdn = 3.00$ ) in the sidebar;  $T = 231.00$ ,  $z = -4.04$ ,  $p < .001$ ,  $\eta^2 = .82$ . In the CReBot condition, participants frequently click the "Next question" button ( $M = 17.79$ ,  $SD = 16.74$ ) and write down their critical thoughts in response to a question in the text box ( $M = 11.04$ ,  $SD = 8.35$ ). Four participants add their own critical questions and consider it a practical function to record their thoughts. "I can

add questions to CReBot, which is similar to taking notes of my personal critical thoughts about the paper" (P21, M, 23). On average, participants switch to a different (sub)section on the current page 2.42( $SD = 2.22$ ) times and select a specific part of the paper from the dropdown menu 3.21( $SD = 3.64$ ) times. When using Guideline, they click the headings in the outline to unfold the content of interests for around 11.96( $SD = 6.84$ ) times.

Regarding the **timings** for using the tools, our thematic analysis on participants' interview responses to "when did you usually use the tool" reveals quite a difference between two conditions (Table 4i). Participants commonly use CReBot as they are reading the paper while using Guideline before they start to read the paper content or after they go through the full paper. They interact with CReBot either right before digging into each section (nine users) so as to "bring the questions to start reading" (P5, M, 27), in the middle of reading a section (six) in cases of encountering "some difficulties in understanding the paper in depth" (P6, F, 23), or after going over a section (ten) to "check if they have processed and understood this part thoroughly" (P15, F, 25). This shows that CReBot encourages users to exercise critical thinking as they read each part of the paper.

#### ii) Interruption and technology acceptance.

Table 4ii shows the participants' ratings on their perceptions of CReBot and Guideline. Both tools generally make little **interruption** to the users' reading processes. Yet, a few participants report that CReBot (two users) and the Guideline (four) in the sidebar may occasionally distract them from their reading materials. "The interaction with CReBot shifted my attention away from the paper" (P19, M, 21). In terms of **technological acceptance**, participants rate CReBot ( $Mdn = 5.88$ ) to be significantly more **useful** for critical reading than the Guideline ( $Mdn = 4.50$ ) in the sidebar;  $T = 272.50$ ,  $z = -4.10$ ,  $p < .001$ ,  $\eta^2 = .84$ ; H2A accepted. While several participants acknowledge both tools' usefulness as a checklist for in-depth paper examination (Table 2), ten participants particularly value the contextualized questions prompted by CReBot. "CReBot's questions were connected to my pace of reading and really helped me digest the corresponding section" (P16, F, 23). However, seven people using CReBot find that "some questions are still not well customized to the given paper" (P12, M, 23). For the Guideline, the lack of interaction features prevents it from realizing its full value, as suggested by eight participants. "The Guideline is not useful as I cannot write down my thoughts to the questions" (P21, M, 23). Yet, 12 users consider the Guideline as a "systematic" and "detailed" tutorial for critical reading. "The Guideline looked like a manual which provides a detailed approach for reading papers critically" (P20, M, 25).

Apart from the usefulness, participants feel that CReBot ( $Mdn = 5.63$ ) is significantly **easier to use** than Guideline ( $Mdn = 4.88$ ) (Table 4b);  $T = 234.50$ ,  $z = -2.94$ ,  $p < .01$ ,  $\eta^2 = .60$ ; H2B accepted. Four participants comment that CReBot's feature of updating questions automatically when they scroll onto a new page is "convenient" (P15, F, 25), and the panel for switching/selecting (sub)sections to be asked questions about is "flexible" (P18, F, 25). Nevertheless, nine users report a glitch in the current design of this question updating feature when CReBot fails to recognize their intention behind the scrolling action, e.g., "when I scrolled back and forth to find answers for its question,

**Table 4**

RQ2 results. Note that one participant may use the tools at multiple timings (*before/during/after* reading the full paper or each section). 7-point Likert scale (1/7 - Strongly disagree/agree) for subjective measures. Mean (SD). \*\*:  $p < .01$ , \*\*\*:  $p < .001$ .

(i) Interaction with CReBot/Guideline						
Frequency of usage ***	Timing of usage #					
	Before read	During reading the paper			After read	Rare usage
		Bef sec	In sec	Aft sec		
CReBot	<b>5.67 (1.46)</b>	3	9	6	10	2
Guideline	3.13 (1.42)	9	2	7	1	5
(ii) Perception towards CReBot/Guideline						
Interrupt reading	Acceptance		Usefulness ***	Easy to use **	Intention to use ***	
	3.00 (1.50)	<b>5.77 (0.71)</b>		<b>5.58 (0.79)</b>	<b>5.92 (0.99)</b>	4.64 (1.40)
CReBot	3.04 (1.88)	4.49 (1.07)				4.52 (1.65)

*the bot popped up another one*" (P24, M, 22). For the Guideline, while three users feel that it provides "*more freedom for exploration*" (P4, M, 30), 16 participants complain that it conveys too much content at the same time. "*It is too wordy. Sometimes I want to find the needed questions, but I have to go through all questions of that part to get those key points*" (P15, F, 25). All in all, participants have a significantly stronger **intention to use** CReBot ( $Mdn = 6.00$ ) than the Guideline ( $Mdn = 4.25$ ) in their future critical reading practices;  $T = 220.50$ ,  $z = -3.67$ ,  $p < .001$ ,  $\eta^2 = .75$ ; **H2C** accepted.

In addition to the strengths of CReBot mentioned above, ten users particularly favor its "Export all chat log" feature, saying that it "*really helps me if I need to present this paper by reminding my critical thoughts about the paper*" (P13, M, 61). Besides, 14 participants are strongly positive that CReBot could improve their critical reading skills after a long-time usage. "*CReBot is useful. If I use it for several weeks, I believe that I can better grasp and examine the key points of the papers and form a consistent critical reading habit.*" (P23, F, 23). Moreover, readers may gradually reduce the use of both tools in the long run, as anticipated by five participants. "*Regardless of which tool [is employed], users will construct their own critical reading patterns after using it multiple times, and they will rely less and less on it*" (P17, M, 27). This is actually one of the goals of CReBot and matches the theory of scaffolding (Kim, 2001), which states that as learners become more independent in their thinking, the support from others can gradually fade away. For the interaction design, eight participants suggest that CReBot could be more dynamically integrated into the reading interface. For example, its chat window can "*pop up when users make a highlight*" (P20, M, 25) or can be "*attached to related paragraphs when asking questions*" (P24, M, 22).

**In summary, for users who often read research publications and have general research experiences,** CReBot with interactive question prompts can significantly better engage them in the critical reading process and enhance their perceived performance in paper reinterpretation compared to the Guideline. Participants frequently interact with CReBot as they go over each section. They consider it significantly more useful and easier to use, and they favor its chat-like interaction that maintains their attention and its contextualized questions that stimulate thinking. Four participants mention that both tools could be particularly useful for novices who are new to paper reading by "*teaching them how to read a paper critically and efficiently*" (P11, M, 22). This motivates us to further investigate whether and how the interactive support from CReBot would help novices read scientific papers critically in Stage 2, as there is an essential need to teach critical thinking skills to students (Fadel et al., 2015) through the exercise of paper reading (Bhattacharyya et al., 2018).

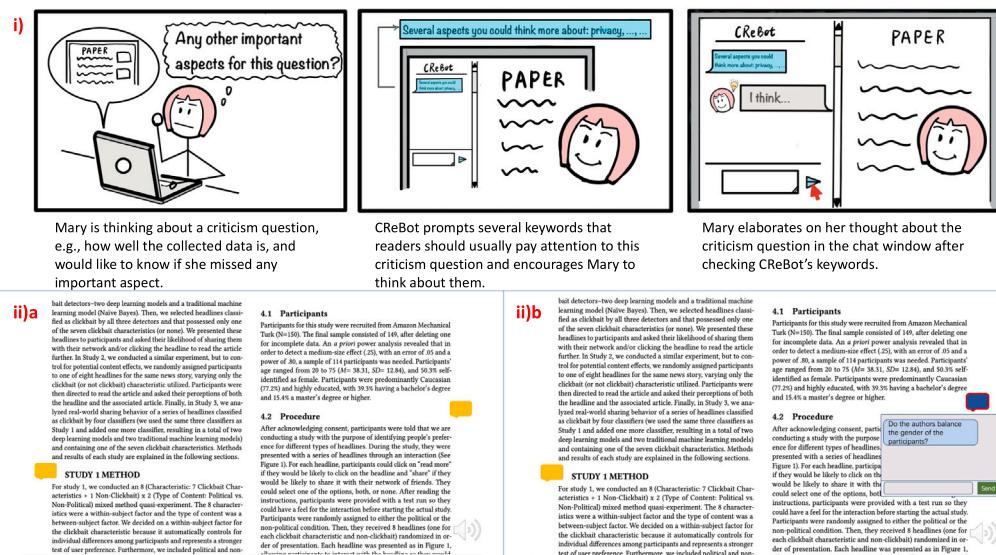
#### 4. Stage 2: Customizing CReBot for novices

Encouraged by the findings in Stage 1 and the need to teach critical thinking skills to students (Fadel et al., 2015; Kilgo et al., 2015), we

proceed to our second-stage exploration on CReBot's usefulness and user experience for novices of scientific research. The representatives are junior undergraduates who have little or no paper reading experience but have the desire to learn critical thinking and interests in going for advanced academic studies. The evaluation study of CReBot in stage 2 is identical to that of stage 1 except for the target user groups and the design of the bot having been refined to accommodate the new user needs and preferences. Previous work suggests that novices' needs for technological support in their tasks may be different from the experts' needs (Miller and Bailey, 2014). Therefore, we start by verifying the needs and identifying the refined requirements of CReBot specifically for novices. In this section, we first present a speed dating study for this purpose and then describe our refinement on CReBot as well as the user experiments and results.

#### 4.1. Refined design requirements of CReBot for novices

To explore novices' potential needs and requirements of CReBot, we follow a user-centered design "speed dating" approach developed by Zimmerman and Forlizzi (Zimmerman and Forlizzi, 2017). This approach can reveal user needs or desires for new products or services with a small sample of participants and has been used by design teams either online or offline (Ma et al., 2015; Zimmerman and Forlizzi, 2017). In our speed dating sessions, we seek to investigate the concepts of a bot with interactive questions and guidance to support novices in critical paper reading. Specifically, we explore their needs for associated question prompts when digging into each section (step 2 in our compiled guidelines) since the results of our Stage 1's experiment imply that this step took most of the time for a critical reader. For a higher-level exploration of the CReBot's concept, we group the questions from four critical thinking levels to two sets based on their goals. **Comprehension questions** ask "what" is the purpose, claim, or premise in the text (e.g., motivation of the paper) – people usually can find direct answers to these questions from the paper. **Criticism questions** ask "how", "why", and "how well" (e.g., how well does the paper link to the related work) – readers usually need to analyze and evaluate the text to compose their own answers to these questions (University of Kent, 2021; of Leeds, 2021). We recruit eight undergraduate students (four females, four males, P1-8) from a local university with English as the teaching language. They are 18–20 years old ( $M = 19.00$ ,  $SD = 0.76$ ) and major in either computer science (CS), physics, math, or life science. They report limited experience in reading papers ( $M = 3.38$ ,  $SD = 0.92$ ; 1 - no experience at all, 7 - a large amount of experience) and almost no experience in reading HCI papers ( $M = 1.25$ ,  $SD = 0.46$ ), which are the reading materials in the speed dating study.



**Fig. 4.** Speed dating materials used in Stage 2. (i) An example storyboard used for examining novices' need for criticism question prompts from CReBot. (ii) An example slide animation for user enactment in the design case of CReBot in a pop-up window — slide *a* is first presented, and then slide *b* is automatically played after a pre-set timing.

#### 4.1.1. Speed dating materials and procedure

Following the common practice of speed dating (Zimmerman and Forlizzi, 2017), we prepare 1) storyboards to verify the user needs and the usefulness of the potential CReBot's functions and 2) a user enactment (UE) environment to explore novices' preferred interaction designs of CReBot.

**Storyboards.** We present the potential usage scenarios of CReBot to participants via storyboards with abstract bot components rather than the workable version in the first stage to keep the exploration more open. Participants view five storyboards one by one in a fixed order that explore the following needs brainstormed by our design team: comprehension questions, response to readers' answers to the comprehension questions, hints on where to find possible answers to comprehension questions, criticism questions, and hints to think of the criticism questions. Each storyboard consists of three figures with associated text explanation about a reader's problem during paper reading, the CReBot's assistance, and the reader's reaction to the bot (Fig. 4i).<sup>5</sup>

**User enactment.** Results of Stage 1's study suggest that users may use CReBot at different timings and favor different forms of integration into paper content. We thus explore through user enactment novices' preferred interaction design in 2 (location: sidebar vs. pop-up window on paper content)  $\times$  3 (timing: before vs. during vs. after reading related parts) conditions illustrated as separate animated slides (Fig. 4ii). In each condition, participants first see a full-screen slide that depicts the reading interface before the CReBot takes any action. After a given timing (i.e., 2, 24, 46 s<sup>6</sup>), the animation depicting the assigned bot action, that is, asking a question in the sidebar or the pop-up window, will be automatically played. We randomize the order of the five slide animations and prepare each animation with a different paragraph from CHI papers to read and a CReBot's question to ask. We conduct the speed dating online via Zoom with participants' consent and start by the five storyboard tasks, in which they are asked to 1) empathize with

<sup>5</sup> The storyboards and enactment slides can be found in Appendix A or <http://zhenhuipeng.com/>.

<sup>6</sup> The 46 s for the "after reading related parts" conditions is determined based on the average time that two of the authors spend reading each given paper paragraph. We choose 2 s for the "before" conditions to allow users to glance at the interface first before reading and select the mean of 46 s and 2 s, i.e., 24 s, for the "during" conditions.

#### 4.1. Participants

Participants for this study were recruited from Amazon Mechanical Turk (N=150). The final sample consisted of 149 after deleting one for being a duplicate entry. We conducted a power analysis that in order to detect a medium-size effect (.25), with an error of .05 and a power of .80, a sample of 114 participants was needed. Participants age ranged from 18 to 65 years old ( $M = 30.3$ ,  $SD = 12.84$ ), and 50.3% self-identified as female. Participants were predominantly Caucasian (77.2%) and highly educated, with 39.3% having a bachelor's degree and 15.4% a master's degree or higher.

#### 4.2. Procedure

After obtaining consent, participants were told that we are conducting a study with the purpose of identifying people's preference for different types of headlines. During the study, they were presented with a series of headlines and asked to read them. They could select one of the options, both or none. After reading the instructions, participants were provided with a test run so that they could have a better understanding before starting the actual study. Participants were randomly assigned to either the political or the non-political condition as well as a within-subject factor for the clickbait characteristic because it automatically controls for individual differences among participants and represents a stronger test of user preference. Furthermore, we included political and non-

#### 4.1. Participants

Participants for this study were recruited from Amazon Mechanical Turk (N=150). The final sample consisted of 149 after deleting one for being a duplicate entry. We conducted a power analysis that in order to detect a medium-size effect (.25), with an error of .05 and a power of .80, a sample of 114 participants was needed. Participants age ranged from 18 to 65 years old ( $M = 30.3$ ,  $SD = 12.84$ ), and 50.3% self-identified as female. Participants were predominantly Caucasian (77.2%) and highly educated, with 39.3% having a bachelor's degree and 15.4% a master's degree or higher.

#### 4.2. Procedure

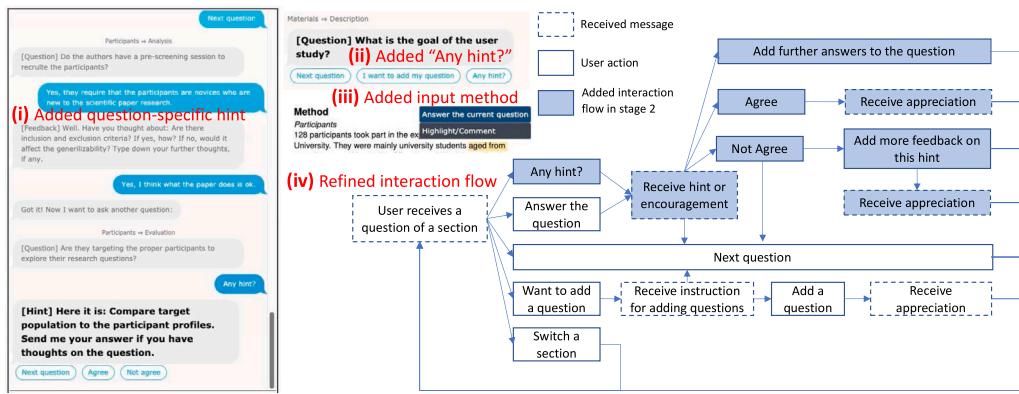
After obtaining consent, participants were told that we are conducting a study with the purpose of identifying people's preference for different types of headlines. During the study, they were presented with a series of headlines and asked to read them. They could select one of the seven clickbait characteristics (clickbait, clickbait + headline, clickbait + headline + type of content, political vs. Non-Political mixed method quasi-experiment). The 8 characteristics were a within-subject factor and the type of content was a between-subject factor. Participants were randomly assigned to either the clickbait characteristic because it automatically controls for individual differences among participants and represents a stronger test of user preference. Furthermore, we included political and non-

the given persona; 2) report the frequency of facing a similar problem and the perceived usefulness of the bot's support in a 7-point Likert scale; and 3) share how they usually deal with the problem and what other functions they need for the bot. Next, participants enact with CReBot via six slide animations in a random order, during which they are required to 1) read the assigned paragraph with the bot; 2) indicate likeability (1–7 points) of the bot's interaction design and why; and 3) explain how they would answer the bot's question. Each participant spends around 40 min for the whole study and gets USD 6.2 as a token of appreciation.

#### 4.1.2. Findings and design requirements for CReBot refinement

The user needs for CReBot that asks comprehension and criticism questions and gives corresponding guidance are confirmed via the storyboards, as indicated by the participant ratings across the five scenarios: (1) frequency ( $M = 4.00 - 5.25$ ; 1/7 - Never/Always) of various problems encountered in paper reading and (2) usefulness of CReBot ( $M = 4.88 - 5.63$ ; 1/7 - Not useful at all/Very useful) for problem mitigation.<sup>7</sup> Participants comment that they normally get questions and hints from "professors or online forums" (P1) or "senior students" (P5, 6, 7) when facing issues depicted in the storyboards, suggesting that the CReBot could be a "natural alternative for offering similar support" (P2). For the interaction design of CReBot, results from user enactment show that participants generally prefer the bot in the sidebar ( $M = 5.17$ ) rather than in the pop-up window ( $M = 4.46$ ; 7 - like it very much). The pop-up design has lower ratings as it would "unexpectedly occlude the paper content" (P1, 3, 4, 5). On average, participants express higher likeability to the CReBot when it proactively asks questions 'after' ( $M = 5.38$ ) users finish reading related parts, compared to the 'before' ( $M = 4.69$ ) and 'during' ( $M = 4.38$ ) timings. However, it is very difficult to track people's reading focus on the computer screen without access to webcams or additional eye-tracking devices (Navalpakkam et al., 2011a), and proactive intervention at an unexpected timing would "be distracting" (P5, 7), similar to the negative feedback for CReBot's proactive questions in the first-stage study (Table 2). Instead, all participants suggest that CReBot should offer options for users to access its support at any time. They also expect various means to answer the bot's questions, including typing

<sup>7</sup> The ratings for each storyboard can be found in Appendix A or <http://zhenhuipeng.com/>.



**Fig. 5.** (a) Refined CReBot for assisting novices in critical paper reading in stage 2. Compared to previous version in stage 1, it mainly adds specific hints/feedback (similar content but triggered differently, inspired by the scaffolding interaction logic of Sara (Winkler et al., 2020) and QuizBot (Ruan et al., 2019)) that encourage thinking of each question (i, ii), an “Answer the current question” option for selecting the text from the paper as an answer (iii), and interaction flow for the question-specific hints (iv).

(mentioned by 8 participants) or copying and pasting (6) in the message box, as well as highlighting the corresponding paper content as a quick answer (6). Based on these findings, we derive two additional design requirements apart from DR1 and DR2 for further customizing CReBot for novices:

**DR3:** To offer sufficient support for novices’ critical paper reading, CReBot should provide detailed hints for each comprehension or criticism question.

**DR4:** To mitigate the interruption to novices’ reading process, CReBot should stay in the sidebar, offer multiple ways for user input, and proactively ask questions only when it can accurately track the user’s reading pace; otherwise, it should let users initiate the interaction.

#### 4.2. Refined CReBot for assisting novices

We refine CReBot for novices based on the aforementioned requirements. Compared with the previous version depicted in Fig. 2, the refined CReBot has the following improvements (Fig. 5). First, it removes the high-level guidance and concentrates on providing question prompts to help novices dig into each paper section. This is to keep its functions focused on the most time-consuming step of critical reading and make them easy to understand for novices. Second, it provides teacher-like hints pre-compiled by HCI researchers for each critical thinking question (DR1). Correspondingly, CReBot extends its interaction flow to handle the user interaction with its hints (Fig. 5i, ii, iv). For example, it offers options for users to agree or disagree with the hints, which can serve as a means to encourage further critical thinking and as user feedback for improving the CReBot. Third, it offers an additional input option for users to select the text of the paper and directly answer the current question as shown in Fig. 5ii (DR2). Fourth, it deducts the proactive question update feature when users scroll across pages as it could not accurately capture user’s reading pace and might cause disturbance (DR2). We leave this proactive feature to future work when there is a better reading pace inference solution with minimum sensing load.

To implement the second change, we need a high-quality dataset of critical thinking question-hint pairs. We tried to create such a dataset using public reviews posted in OpenReview<sup>8</sup>. However, the collected reviews are unstructured, making it difficult to automatically generate the question-hint pairs. Moreover, these reviews are mostly for papers submitted to AI conferences (e.g., NeurIPS, ICML), which could differ from the reviews of HCI papers in our demonstrated CReBot use case. To this end, to build up the knowledge base of our CReBot prototype, we refer to domain experts’ knowledge similar to the methods

in Huang et al. (2021), Cheng et al. (2015). We invite three HCI researchers (males, age: 26, 25, 23) who got a Ph.D. degree or are Ph.D. students who majored in HCI to pre-compile question-specific hints in a brainstorming workshop. The first author, as one of the three participants, prepares the workshop materials, including common CHI section categories with titles and example critical thinking questions from the first-stage study (Section 3.1) as well as templates of hints (e.g., from (Bangor University, 2021; RMIT University, 2021)). Inspired by the online critical thinking guidelines (of Leeds, 2021; University of Kent, 2021), we ask brainstorms to contribute questions from three perspectives – “comprehension – description” (i.e., what), “criticism – analysis” (how, why), and “criticism – evaluation” (how well) (DR1) – which come to mind while reading the paper contents under each section category. They also need to produce ideas about where one may look for possible answers or how to approach each proposed question. The online brainstorming session lasts for about two hours and results in a total of 103 question-hint pairs, which are further reviewed and refined by a professor who has published over 20 CHI papers. Fig. 5i shows some example pairs, and the full dataset can be found in Appendix A or <http://zhenhuipeng.com/>.

#### 4.3. Experiment

Our evaluation of the refined CReBot for assisting novices in critical paper reading is guided by the following two research questions:

**RQ3:** Compared to the conventional guidelines that lists the questions (Guideline), how would an interactive question-asking bot (CReBot) affect novices’ i) behaviors, ii) perceived engagement and difficulty in the process, and iii) perceived performance in the outcome of critical paper reading?

**RQ4:** i) How would novices use CReBot and Guideline in practice, and ii) what would be their acceptance of such tools for critical paper reading?

We conduct a within-subjects study with 20 participants (P1–20; 11 Females, 9 Males; age range 18 – 22,  $M = 20.1$ ,  $SD = 1.0$ ) recruited via word of mouth in a local university with English as the teaching language. Seven major in computer science (CS), and the rest major in life science, math, finance, law, and so on. Different from those who already have general research experience and moderate experience in reading CS/HCI papers in stage 1, the targeted participants in stage 2 are novices with little or no experience in reading CS/HCI papers ( $M = 2.4/1.5$ ; 1 - No experience at all, 7 - A lot of experience). The user study design is similar to that in stage 1 (Section 3.3). The baseline tool Guideline is similar to the one as shown in Fig. 2b but replaces the 363 critical thinking questions with the 103 question-hint pairs used by the refined CReBot. Participants have two CHI LBW reading tasks the same as those used in stage 1. From post-surveys, the perceived

<sup>8</sup> <https://openreview.net/>

**Table 5**

RQ3 results in the stage 2's experiment with 20 novices. 7-point Likert scale (1/7 - Strongly disagree/agree) for (ii) and (iii). Mean (SD).

\*:  $p < .05$ , + :  $.05 < p < .1$ . #: number of occurrences.

	(i) Reading behaviors				(ii) Critical reading process	
	Time (min)	Highlight #	Comment #	Scroll pages #	Engagement	Difficulty *
CReBot	33.8 (5.8)	8.3 (7.5)	4.9 (3.9)	66.7 (51.6)	4.44 (0.88)	<b>4.50 (1.32)</b>
Guideline	32.9 (4.8)	9.7 (11.5)	3.1 (4.8)	62.5 (25.6)	4.58 (1.17)	3.90 (1.33)
(iii) Perceived performance						
	Examine arguments	Identify flaws		Reinterpret the paper +		
CReBot	4.45 (1.00)	4.20 (1.32)		<b>4.35 (1.42)</b>		
Guideline	4.55 (1.15)	4.45 (1.19)		3.80 (1.64)		

reading difficulties of paper 1 and 2 are 3.85 and 3.75; item “The paper itself is difficult to read”, 1/7 - Strongly disagree/agree. We follow the same experimental **procedure** in stage 1 as depicted in Fig. 3. For RQ3, the **measures** include the logged reading behaviors, rated perceived engagement and difficulty in the critical reading process, and rated perceived performance in the critical reading outcome, using the same post-survey as that for RQ1 (see Section 3.3.3 and Table 1). As for RQ4, we also adopt the same post-survey as used for RQ2 to assess participants’ interaction behaviors with CReBot/Guideline, perceived interruption, and technology acceptance of each tool. We additionally log timestamped user interaction with the question-specific hints and the “Answer the current question” feature (Fig. 5iii) of the refined CReBot.

#### 4.4. Analysis and results

Note that in this experiment, we do not intend to make a direct comparison about CReBot’s usefulness for routine paper readers and novices; instead, we explore the refined CReBot’s interactive question prompts for novices compared to the documented question list. We run similar Wilcoxon signed-rank tests used in stage 1 (Section 3.4) on the quantitative measurements. For the qualitative data from the transcribed interview recordings, we go through a similar thematic analysis as that in stage 1 and identify novices’ positive and negative feelings to CReBot/Guideline, their usage patterns of the tools, and suggestions for improvement. The summative check of inter-rater reliability regarding the final codes indicates a substantial agreement between the two annotators (Cohen’s  $\kappa = 0.735$ ). The disagreement on the final coding results is resolved by a final discussion. We incorporate these codes into the presentation of the results below.

##### 4.4.1. RQ3 results

Table 5 summarizes the RQ3 results about CReBot’s impact on novices’ critical reading process and outcome in comparison to the Guideline. **i) Reading behaviors.** All participants actively read and annotate (i.e., reading time, highlight, comment, scroll pages; all with  $p > .05$ ) the papers with both tools (Table 5i). **ii) Perceived engagement and difficulty in critical reading.** In general, participants feel that they engage in the critical reading process equally well with both tools ( $p > .05$ ). However, participants consider it significantly more difficult to read the assigned papers critically with CReBot than with Guideline;  $T = 96.00$ ,  $z = -2.14$ ,  $p < .05$ ,  $\eta^2 = .48$ . “*The bot somehow made me feel stressed and reduced my reading speed, making the critical reading a more difficult task for me*” (P10, M, 20). This suggests that CReBot’s might create a more cognitively demanding reading environment than Guideline for novices. **iii) Perceived performance in critical reading outcome.** On average, participants in both conditions have similar perceived performance regarding argument examination and flaw identification ( $p > .05$ ), though they tend to feel that they perform better with CReBot ( $Mdn = 5.00$ ) than with Guideline ( $Mdn = 4.50$ ) in paper reinterpretation ( $T = 101.50$ ,  $z = -1.80$ ,  $0.05 < p < .1$ ,  $\eta^2 = .40$ ).

#### 4.4.2. RQ4 results

**i) Interaction with CReBot/Guideline.** As shown in Table 6i), participants reflect that they use CReBot ( $Mdn = 4.00$ ) for critical thinking significantly more frequently than Guideline ( $Mdn = 3.00$ ) in the sidebar;  $T = 74.50$ ,  $z = -2.06$ ,  $p < .05$ ,  $\eta^2 = .46$ . Participants frequently check the “Next question” ( $M = 17.15$ ,  $SD = 13.10$ ) and answer the bot’s questions ( $M = 7.20$ ,  $SD = 6.29$ ). Half of the answers are inputted by selecting the paper content and clicking the pop-up “Answer the current question” ( $M = 3.70$ ,  $SD = 3.80$ ) – a new feature in the refined CReBot (Fig. 5iii). Our participants also ask for hints (i.e., click “Any hint?”) quite often ( $M = 4.75$ ,  $SD = 6.70$ ) and actively express their feelings to the received question-specific guidance. In total, they “agree” to the bot’s guidance 78 times and do “not agree” to it ten times, suggesting that they carefully check the hints and that the pre-compiled hints generally are of high quality. Only one participant added his own critical thinking question, and this observation is acceptable as novices may still learn to read papers critically. It could also be difficult to contribute new questions to CReBot in our lab study. For the usage of Guideline, participants click the headings in the outline to unfold the content of interest around 10.60( $SD = 5.64$ ) times. Regarding the **timings** at which users checked the questions from the two tools, we have an interesting finding (Table 6i). With the Guideline, novices usually turn to the questions after reading each paper section (number of occurrences: 11) or the entire paper (7), while with CReBot, some of them refer to the critical questions before (6) or while reading each section (6). This implies that CReBot could encourage some novices to adjust their reading habits from “read first” to “think first” or “think in situ”. Three participants explicitly comment that this reading behavior change is helpful. “*CReBot’s questions caught my attention as I read each paper section, and I started to think about them and highlighted potential answers from the section as I proceeded. This makes my critical thinking on this paper more efficient*” (P3, F, 19). However, another three users mention that they are not used to such a change. “*Taking the questions in mind before or during reading each section is not my familiar reading style. It is like being challenged by the bot, and I do not like this feeling*” (P1, F, 20). **ii) Interruption and technology acceptance.** As shown in Table 6ii), both tools generally make little **interruption** (e.g., both  $M < 4.00$ ) to the users’ reading processes ( $p > 0.1$ ). In terms of **technological acceptance**, participants deem CReBot and Guideline equally useful and easy to use. They also have similar levels of intention to use both tools in the future ( $p > .05$  in all three aspects).

We further dig into the qualitative data to probe what aspects of CReBot/Guideline work or do not work for novices (Table 7) as well as their expectations for future interactive critical paper reading support. In general, both tools are valuable for novices by providing good critical thinking questions, as suggested by eight/six users with CReBot/Guideline. “*The bot asked me some insightful questions that help me grab key points of the paper to think of*” (P19, F, 21). Eight participants also appreciate the question-specific hints from both tools, saying that “*they are inspiring for locating answers or getting an angle to criticize the paper*” (P11, M, 22). Similar to the qualitative results (Table 2) in stage 1, novices favor the flexible interaction features (mentioned by 7 users) and contextual questions (3) of CReBot and the clear and systematic

**Table 6**

RQ4 results. Note that one participant may use the tools at multiple timings (before/during/after reading full paper or section). 7-point Likert scale (1/7 - Strongly disagree/agree) for subjective measures. Mean (SD). \*\*:  $p < .01$ , + :  $.05 < p < .1$ .

		(i) Interaction with CReBot/Guideline					
		Timing of usage #					
		Before read		During reading the paper			
		Bef sec	In sec	Aft sec	After read	Rare usage	
CReBot	<b>4.35 (1.46)</b>	2	6	6	2	0	
Guideline	3.30 (1.17)	3	1	11	7	1	

		(ii) Perception towards CReBot/Guideline						
		Interrupt reading	Acceptance					
		Usefulness	Easy to use					
CReBot	3.60 (1.76)	4.76 (1.38)	4.13 (1.21)	4.27 (1.52)				
Guideline	2.75 (1.37)	4.95 (1.36)	4.73 (1.39)	4.43 (1.40)				

**Table 7**

Summary of novices' interview responses to pros and cons of CReBot/Guideline in Stage 2.

	Pros (# of participants mentioned)	Cons (#)
CReBot	Good critical thinking questions (8); Inspiring hints (8); Flexible interaction (7); Contextual questions (3)	Some questions not matched to the reading context enough (3); Some hints are general (3)
Guideline	Good critical thinking questions (6); Inspiring hints (8); Clear structure (4)	Lack of interactive features (4); Too much content on a page (7)

structure (4) of Guideline. “*I like the conversational style of interaction with CReBot, which is flexible and interesting.*” (P4, F, 19). “*The Guideline interface is simpler and more concise for me*” (P15, F, 20). Nevertheless, the Guideline is criticized by seven users for that “*it provides too much content at the same time, making it hard to read and use*” (P6, F, 21). Six participants mention that some of the CReBot’s questions and hints still do not match the section content very well, and they desire “*more questions and hints that are associated with the specific details of this paper and even some follow-up questions to the previous user responses*” (P19, F, 21). Regarding the envisioned design of a future interactive critical paper reading support tool, participants suggest some additional features: 1) certain questions could be presented with “yes/no” or “multiple choices” option buttons to enable quick answers (commented by 3 users) and 2) it could have a panel in which readers can store their questions and hints for reflection and future usage (3).

**In summary, for novices who have rarely read scientific papers before**, the current design of CReBot with interactive questions and associated hints may not be a better solution than the static checklist of questions and guidance (Guideline) for assisting critical paper reading. The pre-compiled critical thinking questions and guidance presented by either tool are generally valuable for novices. While CReBot is still favored for its flexible interaction and contextual support, it could interfere novices’ traditional reading habits and make them uncomfortable. We discuss possible reasons why CReBot is not deemed more helpful for novices than Guideline for routine paper readers in the next section.

## 5. Discussion

### 5.1. Reflection of CReBot for assisting routine paper readers and novices

From our first-stage findings, CReBot with interactive section-level question prompts is more useful for engaging routine paper readers in critical thinking than Guideline. These people can be postgraduates and researchers who have experience in scientific research and are already familiar with the scenarios (e.g., group meeting, rebuttal) of responding to questions and/or criticisms regarding the papers they read or write. CReBot simulates the active Q & A activity in these scenarios to help this group of users think critically before, during, and after reading each section of a paper (Table 4). This is proven beneficial in our stage 1 study. This is in line with the findings in Liao et al. (2020) that the questions users may ask for understanding artificial

intelligent (AI) systems can guide design practitioners to think of and create user-centered explainable AI applications.

In the second-stage exploration, we refine CReBot with additional teacher-like hints that helps novices approach each critical thinking question. While the questions and hints are appreciated by the novices in the user study, the interactive way of presenting them by CReBot does not outperform the traditional method which documents all information as a static list (Table 6ii). One possible reason is that CReBot’s interactive questions lead some novices to think before or while reading each paper section (Table 6i), which might not be a familiar or comfortable ways of reading. Furthermore, as the novices in our study use CReBot more frequently than Guideline, they possibly face questions that are hard to answer more often due to a lack of prior experience in scientific research. This may make the critical reading task more difficult for them (Table 5ii). For example, the question “How do the authors interpret the numerical results statistically?” may be difficult for novices who are unfamiliar with common statistical analysis methods used in (HCI) research (Shepherd and Sande, 2014). Theoretically, this could be also explained by the Yerkes-Dodson law (Yerkes and Dodson, 1908), which states that people’s task performance increases with physiological or mental arousal, but only up to a point. The critical paper reading tasks would be more intellectually demanding for the beginners who lack prior research experience and thus they may require a lower level of arousal for optimal performance (Diamond et al., 2007). Compared to the static Guideline, the interactive questions from CReBot might cause higher arousal than the optimal level, leading to a potential decrease in the bot’s usefulness and user experience for novices in critical paper reading.

Our two-stage study offers empirical evidence that a question prompting PCA is helpful for routine paper readers. They also imply the need to customize the interactive reading support tools for users with different research experience. For routine paper readers, the reading support tools can imitate the human interaction in offline co-located group paper discussions (Haller et al., 2010), such as question prompts in our case, commenting, and association of related works. For novices, a more complete picture of critical reading knowledge space (e.g., the organized question list in Fig. 5b) rather than the fragment questions prompted by CReBot is needed (Mombini et al., 2020). The reading support tools can act as a teacher to present this type of structured knowledge space to novices via a scaffolding manner (Mariane, 2002),

e.g., raising questions, checking users' answers, and giving corrective feedback if applicable. To personalize the reading support based on users' levels of critical thinking skills, the interactive tools can invite the readers to ask questions about the papers like what our CReBot currently supports. These questions raised by readers can be further compared with the high-quality critical thinking questions to assess their competency in paper reading.

## 5.2. Extension of CReBot to technology-enhanced learning communities

**For other paper domains.** In the development process of CReBot, we create a question bank for paper reading. It includes 363 critical thinking questions from interviews with domain researchers in stage 1 (Section 3.1) and 103 question-hint pairs from a brainstorming workshop in stage 2 (Section 4.2). As a proof-of-concept, these questions are customized by HCI researchers and mapped to the common CHI paper sections in the scope of this work. Following the same pipeline, researchers could easily adapt this initial question pool to other research domains, enabling CReBot to provide critical reading assistance specific to the nature of these domains.

**For critical writing and review.** Given CReBot's benefits for routine paper readers, we anticipate that CReBot can also facilitate them in scientific writing and paper reviewing, which are two other common research activities that involve critical thinking. Critical writing skills lie in the capability of convincing readers to accept the authors' claims (Wallace and Wray, 2016) and are reciprocal to critical reading (Bazerman, 1995). Therefore, it is promising that authors can leverage our CReBot to self-check if their writings sufficiently address potential readers' concerns. Reviewers for conferences and journals can also employ our CReBot to identify and articulate the strengths and weaknesses of a submission, and provide constructive comments in terms of how the submission could be improved. Future research can explore the potential extensions of CReBot to these contexts. For example, to assist critical writing, designers can add features such as the persuasiveness of the arguments (Wambsgaass et al., 2020) and recommended similar arguments from published papers (Peng et al., 2020; Hui et al., 2018).

**For collaborative reading.** Beyond the usage for individual paper reading, CReBot can also involve a group of people to read collaboratively, as hoped by five participants in the first-stage experiment. “CReBot can record all users' questions and answers so that people can check others' thoughts”. (P4 in the first CReBot evaluation study, M, 30). This idea is similar to the shared highlights and notes of Amazon Kindle (Reader, 2020), the collaborative critical reading platform (Tan et al., 2016), and cooperative note-taking system (Kam et al., 2005), which requires human peers to form a reading group. We suggest that CReBot can be integrated into online groups for academic paper reading (e.g., PeerLibrary (Anon, 2021b), Fermat's Library (Anon, 2021a), and Reddit r/MachineLearning (u/valetudoo, 2021)). In this case, CReBot not only acts as a facilitator of critical reading but also a “social organizer” (Seering et al., 2019) who connects users with other community members based on their critical thoughts on the same paper and encourages them to share ideas and hold discussions. At the same time, CReBot will be able to evolve itself iteratively with users' question-answer pairs.

**For critical thinking support on social networks and teamwork.** CReBot also has the potential to assist in the critical assessment of information on social media (Machete and Turpin, 2020; MacAvaney et al., 2019). For example, CReBot can ask questions that guide readers to stay critical and identify fake news (Machete and Turpin, 2020). This is beneficial for people suffering from the massive amount of misinformation online, e.g., about COVID-19 (Puig et al., 2021). Moreover, CReBot can facilitate members of a team by acting as an active questioner that helps them to think critically in group discussions and decision-making.

## 5.3. Design considerations for interactive critical reading support

### 5.3.1. Providing proactive assistance when users get stuck

During the critical paper reading process, users especially those new to scientific research (i.e., novices) may get stuck answering critical thinking questions from time to time. To offer sufficient facilitation in this case, CReBot could add more proactive mechanisms (Peng et al., 2019) such as auto-highlights of keywords, sentences, or paragraphs that might be of interests to the readers (e.g., by tracking discourse markers like “we found that” (Khan et al., 2020)) when they get entangled in a question. Common ways of detecting such struggles to initiate proactive support include but are not limited to detecting extensively long pause with certain paper content through interface interaction logs or eye-tracking (Navalpakkam et al., 2011b; Cheng et al., 2015); the latter is more accurate but has the trade-offs of privacy concerns and the possibility of involving additional hardware.

### 5.3.2. Supporting user customization of prompted questions and guidance

As suggested by three participants in stage 2, CReBot can offer another panel in which users can easily store their questions and hints. For example, it can offer a “save” or “bookmark” checkbox next to every question prompt and guidance (Fig. 5i). A check on this box can automatically copy this message to the designated panel so that users can easily reflect on their preferred critical thoughts on the questions later. This is like the “idea checking area” in MetaMap (Kang et al., 2021) that stores searched images for visual metaphor ideation and the “Saved Designs” view in the GRIDS (Dayama et al., 2020) tool that lets designers record their design solutions.

### 5.3.3. Tutoring novices via an Ad-Hoc critical thinking quiz

Two novices in the second-stage experiment suggest that CReBot could better help them learn critical thinking if it can set up an ad-hoc quiz to test how well they have mastered the paper content and the critical reading skills. “*For a given text, CReBot can ask us to identify its weak points and give us corrective feedback, such that we can learn how our critical thoughts can be improved*” (P11 in the second CReBot evaluation study, M, 22). We thus suggest that with the known weak points in the paper, CReBot can offer such a quiz throughout novices' paper reading exercises. For example, we can leverage the reviewers' comments on some papers (e.g., those in the OpenReview (OpenReview, 2021)) and extract the points that mention the papers' weaknesses (e.g., using keywords like “not convincing” and “need to improve”). CReBot can post a quiz about flaw identification when readers start to read related sections of paper and provide corrective feedback with the pre-set answers. This is an instructional scaffolding strategy (Mariane, 2002) commonly used in the classroom and recently in educational chatbots like Sara (Winkler et al., 2020) and QuizBot (Ruan et al., 2019) which are shown to be effective in helping students to learn.

## 5.4. Limitations and future work

Our work has several limitations. First, we show in stage 1 that routine paper readers find CReBot more useful for stimulating them to reinterpret the reading materials for improved clarity than the Guideline (Table 3iii). However, given that our focus is primarily on how users interact with the system, we do not carry out pre-post tests to measure changes in their critical understanding of the paper (e.g., the strengths and weaknesses of the paper) after using the CReBot/Guideline. We thus do not get to evaluate CReBot's effectiveness objectively. Future work can derive metrics of critical thinking performance and examine how well CReBot can help users improve their mastery of scholarly publications. Second, our results are from a short-term study, in which participants read our selected CHI late-breaking works in a simulated scenario where they need to present the papers in a group meeting. In real-world contexts, people might read other styles of papers (e.g., CHI full papers with various topics

and/or different types of contributions) and have different purposes for reading papers critically (e.g., homework for research seminar course and literature review (Shum, 2020; Wallace and Wray, 2016)). In the future, we need to explore the usage of CReBot in a long-term field study in which users can read papers of their interests to see if people's critical reading skill improves and eventually they could read papers critically without the bot. Third, we invite HCI researchers to contribute to the section-level questions and guidance for CReBot, which are perceived helpful but sometimes do not match very well to specific papers as noted by the participants in two stages. Fourth, CReBot cannot currently respond to users' questions with adaptive answers. Future work can explore automatic question generation and question answering based on the paper's content to provide more adaptive questions and hints at scale. For example, it can use template-based methods that fill phrases (e.g., semantic role labels (Lindberg et al., 2013)) from the reading materials into some templates or use machine-learning models that are trained on labeled question-answer-context data (Du et al., 2017) when such a dataset is available in the scientific paper domain. Fifth, we recruit 24 and 20 participants for evaluation of CRebot in studies 1 and 2, respectively. The sample sizes were acceptable for a Wilcoxon ranked test for a within-subjects study (Dwivedi et al., 2017). They were comparable to the numbers of participants in system-based HCI works (e.g., Kang et al. (2021), Yan et al. (2021), Weinman et al. (2021)) with similar study designs. We acknowledge that involving more participants with highly diverse research backgrounds can further deepen our understandings of the user experience and effectiveness of CReBot, and this will be part of our future work. Lastly, certain user experience problems with CReBot (e.g., reported in Tables 2 and 7) need to be addressed in the later design iterations with more detailed scenarios (e.g., critical reading exercises) and target users (e.g., undergraduate students in a course).

## 6. Conclusion

This paper probes the design, usefulness, and user experience of the interactive question prompts for critical paper reading support. Our first-stage study aims to offer such facilitation to routine paper readers who have prior research experience and need to read papers critically. We build a CReBot system that interactively asks section-level questions from the pre-compile critical paper reading guidelines when users are reading each paper section. Our within-subjects experiment with 24 routine paper readers shows that CReBot can better engage them in the critical reading process and is perceived significantly more useful and easier to use than the static checklist-like Guideline. In our second-stage study, we further customize CReBot for novices who are new to scientific research by adding more question-specific hints, based on the identified needs and requirements in a speed dating study with novices. However, the results from another within-subjects study with 20 novices imply that CReBot is not perceived as more useful and easier to use than the static Guideline. Novices appreciate the critical thinking questions and associated guidance from either tool but the CReBot sometimes unexpectedly interferes with their conventional reading habits. The main takeaways of our works include (1) the CReBot prototype to exploring interactive question prompts for critical paper reading support, (2) demonstrated value of CReBot for facilitating routine paper readers, (3) requirements of interactive critical reading support for novices and how they work with CReBot, and (4) design implications for interactive tools to enhance learning. Meanwhile, we contribute a question bank with guidance that encourages critical thinking of scientific research publications.

## CRediT authorship contribution statement

**Zhenhui Peng:** Conceptualization, Methodology, System development, User study, Data analysis, Writing – original draft, Writing – review & editing. **Yuzhi Liu:** System development, User study, Writing

– review & editing. **Hanqi Zhou:** System development, User study, Writing – review & editing. **Zuyu Xu:** System development, User study, Writing – review & editing. **Xiaojuan Ma:** Conceptualization, Writing – review & editing, Data analysis, Supervision, Funding acquisition.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgment

This work is partially supported by the partially supported by the Fundamental Research Funds for the Central Universities, Sun Yat-sen University, No. 22qntd110 and the Hong Kong General Research Fund (GRF) with grant No. 16204819.

## Appendix A. Supplementary data

Supplementary material related to this article can be found online at <https://doi.org/10.1016/j.ijhcs.2022.102898>.

## References

- Agapie, E., Chinh, B., Pina, L.R., Oviedo, D., Welsh, M.C., Hsieh, G., Munson, S., 2018. Crowdsourcing exercise plans aligned with expert guidelines and everyday constraints. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. In: CHI '18, Association for Computing Machinery, New York, NY, USA, pp. 1–13. <http://dx.doi.org/10.1145/3173574.3173898>.
- Alake, R., 2020. How you should read research papers according to andrew ng (stanford deep learning lectures). Retrieved in July 2020 from <https://towardsdatascience.com/how-you-should-read-research-papers-according-to-andrew-ng-stanford-deep-learning-lectures-98ecbd3ccfb3>.
- Althusser, L., Balibar, E., 1970. Reading Capital.
- Anon, 2021a. Fermat's library. Retrieved in September 2021 from <https://fermatslibrary.com/>.
- Anon, 2021b. Peerlibrary. Retrieved in September 2021 from <https://peerlibrary.org/>.
- Bakhtin, M., Holquist, M., Emerson, C., 2010. The Dialogic Imagination: Four Essays. In: University of Texas Press Slavic Series, University of Texas Press, URL <https://books.google.com.tw/books?id=JKZztxqd1Pc>.
- Bangor University, 2021. What is critical writing? Retrieved in August 2021 from <https://www.bangor.ac.uk/studyskills/study-guides/critical-writing.php.en>.
- Bartko, J.J., 1966. The intraclass correlation coefficient as a measure of reliability. Psychol. Rep. 19 (1), 3–11. <http://dx.doi.org/10.2466/pr0.1966.19.1.3>.
- Bazerman, C., 1995. The Informed Writer: Using Sources in the Disciplines. In: Practice & pedagogy, Houghton Mifflin, URL <https://books.google.com.tw/books?id=Ui9DN0a0JKoC>.
- Beyer, B.K., Kappa, P.D., 1995. Critical thinking. Educ. Found.
- Bharadwaj, A., Siangliuue, P., Marcus, A., Luther, K., 2019. Critter: Augmenting creative work with dynamic checklists, automated quality assurance, and contextual reviewer feedback. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. In: CHI '19, Association for Computing Machinery, New York, NY, USA, pp. 1–12. <http://dx.doi.org/10.1145/3290605.3300769>.
- Bhattacharyya, P., Chan, C.W., Waraczynski, M., 2018. How novice researchers see themselves grow. Int. J. Scholarship Teach. Learn. 12, 3.
- Braun, V., Clarke, V., 2006. Using thematic analysis in psychology. Qual. Res. Psychol. 3 (2), 77–101. <http://dx.doi.org/10.1191/1478088706qp063oa>.
- Chaves, A.P., Gerosa, M.A., 2021. How should my chatbot interact? A survey on social characteristics in human–chatbot interaction design. Int. J. Hum. Comput. Interact. 37, 729–758.
- Cheng, S., Sun, Z., Sun, L., Yee, K., Dey, A.K., 2015. Gaze-based annotations for reading comprehension. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. In: CHI '15, Association for Computing Machinery, New York, NY, USA, pp. 1569–1572. <http://dx.doi.org/10.1145/2702123.2702271>.
- Cicchetti, D., 1994. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instrument in psychology. Psychol. Assess. 6, 284–290. <http://dx.doi.org/10.1037/1040-3590.6.4.284>.
- Collins, C., Carpendale, S., Penn, G., 2009. DocuBurst: Visualizing document content using language structure. Comput. Graph. Forum (Proc. of the Eurographics/IEEE-VGTC Symposium on Visualization (EuroVis)) 28 (3), 1039–1046. <http://dx.doi.org/10.1111/j.1467-8659.2009.01439.x>.

- da Rocha Tomé Filho, F., Mirza-Babaei, P., Kapralos, B., Moreira Mendonça Junior, G., 2019. Let's play together: Adaptation guidelines of board games for players with visual impairment. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. In: CHI '19, Association for Computing Machinery, New York, NY, USA, pp. 1–15. <http://dx.doi.org/10.1145/3290605.3300861>.
- Dayama, N.R., Todi, K., Saarelainen, T., Oulasvirta, A., 2020. GRIDS: Interactive layout design with integer programming. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, pp. 1–13.
- Diamond, D., Campbell, A., Park, C.R., Halonen, J.D., Zoladz, P., 2007. The temporal dynamics model of emotional memory processing: A synthesis on the neurobiological basis of stress-induced amnesia, flashbulb and traumatic memories, and the Yerkes-Dodson law. *Neural Plast.* 2007.
- Du, X., Shao, J., Cardie, C., 2017. Learning to ask: Neural question generation for reading comprehension. pp. 1342–1352. <http://dx.doi.org/10.18653/v1/P17-1123>.
- Duggan, G.B., Payne, S.J., 2011. Skim reading by satisficing: Evidence from eye tracking. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. In: CHI '11, Association for Computing Machinery, New York, NY, USA, pp. 1141–1150. <http://dx.doi.org/10.1145/1978942.1979114>.
- Dwivedi, A.K., Mallawaarachchi, I., a. Alvarado, L., 2017. Analysis of small sample size studies using nonparametric bootstrap test with pooled resampling method. *Stat. Med.* 36, 2187–2205.
- Fadel, C., Bialik, M., Trilling, B., 2015. Four-Dimensional Education: The Competencies Learners Need to Succeed. p. 177.
- Graham, J., 1999. The reader's helper: A personalized document reading environment. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. In: CHI '99, Association for Computing Machinery, New York, NY, USA, pp. 481–488. <http://dx.doi.org/10.1145/302979.303139>.
- Haller, M., Leitner, J., Seifried, T., Wallace, J.R., Scott, S.D., Richter, C., Brandl, P., Gokcezade, A., Hunter, S., 2010. The NiCE discussion room: Integrating paper and digital media to support co-located group meetings. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. In: CHI '10, Association for Computing Machinery, New York, NY, USA, pp. 609–618. <http://dx.doi.org/10.1145/175326.1753418>.
- Hart, S.G., Staveland, L.E., 1988. Development of NASA-TLX (task load index): Results of empirical and theoretical research. In: Hancock, P.A., Meshkati, N. (Eds.), *Human Mental Workload*. In: Advances in Psychology, vol. 52, North-Holland, pp. 139–183. [http://dx.doi.org/10.1016/S0166-4115\(08\)62386-9](http://dx.doi.org/10.1016/S0166-4115(08)62386-9), URL <https://www.sciencedirect.com/science/article/pii/S0166411508623869>.
- Head, A., Lo, K., Kang, D., Fok, R., Skjonsberg, S., Weld, D.S., Hearst, M.A., 2021. Augmenting scientific papers with just-in-time, position-sensitive definitions of terms and symbols. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA.
- Heffernan, N., Koedinger, K., 2002. An intelligent tutoring system incorporating a model of an experienced human tutor. In: Intelligent Tutoring Systems.
- Huang, G., Qian, X., Wang, T., Patel, F., Sreram, M., Cao, Y., Ramani, K., Quinn, A.J., 2021. AdapTutAR: An adaptive tutoring system for machine tasks in augmented reality. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. In: CHI '21, Association for Computing Machinery, New York, NY, USA, <http://dx.doi.org/10.1145/3411764.3445283>.
- Hui, J.S., Gergle, D., Gerber, E.M., 2018. IntroAssist: A tool to support writing introductory help requests. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. In: CHI '18, Association for Computing Machinery, New York, NY, USA, pp. 1–13. <http://dx.doi.org/10.1145/3173574.3173596>.
- Kam, M., Wang, J., Iles, A., Tse, E., Chiu, J., Glaser, D., Tarshish, O., Cannay, J., 2005. Liveneotes: A system for cooperative and augmented note-taking in lectures. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. In: CHI '05, Association for Computing Machinery, New York, NY, USA, pp. 531–540. <http://dx.doi.org/10.1145/1054972.1055046>.
- Kang, Y., Sun, Z., Wang, S., Huang, Z., Wu, Z., Ma, X., 2021. MetaMap: Supporting visual metaphor ideation through multi-dimensional example-based exploration. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. In: CHI '21, Association for Computing Machinery, New York, NY, USA, <http://dx.doi.org/10.1145/3411764.3445325>.
- Kerry, A., Ellis, R., Bull, S., 2009. Conversational Agents in E-Learning. In: Applications and Innovations in Intelligent Systems XVI, Springer London, pp. 169–182.
- Keshav, S., 2007. How to read a paper. *Comput. Commun. Rev.* 37, 83–84. <http://dx.doi.org/10.1145/1273445.1273458>.
- Khan, T.A., Yoon, D., McGrenere, J., 2020. Designing an eyes-reduced document skimming app for situational impairments. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. In: CHI '20, Association for Computing Machinery, New York, NY, USA, pp. 1–14. <http://dx.doi.org/10.1145/3313831.3376641>.
- Kilgo, C.A., Sheets, J.K.E., Pascarella, E., 2015. The link between high-impact practices and student learning: some longitudinal evidence. *Higher Educ.* 69, 509–525.
- Kim, B., 2001. Social Constructivism. Emerging perspectives on learning, teaching, and technology.
- Kim, D.H., Hoque, E., Kim, J., Agrawala, M., 2018. Facilitating document reading by linking text and tables. In: Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology. In: UIST '18, Association for Computing Machinery, New York, NY, USA, pp. 423–434. <http://dx.doi.org/10.1145/3242587.3242617>.
- Kobayashi, J., Kawashima, T., 2019. Paragraph-based faded text facilitates reading comprehension. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. In: CHI '19, Association for Computing Machinery, New York, NY, USA, pp. 1–12. <http://dx.doi.org/10.1145/3290605.3300392>.
- Lai, V., Liu, H., Tan, C., 2020. "Why is 'Chicago' deceptive?" towards building model-driven tutorials for humans. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. In: CHI '20, Association for Computing Machinery, New York, NY, USA, pp. 1–13. <http://dx.doi.org/10.1145/3313831.3376873>.
- Lee, B., Savisaari, O., Oulasvirta, A., 2016. Spotlights: Attention-optimized highlights for skim reading. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. In: CHI '16, Association for Computing Machinery, New York, NY, USA, pp. 5203–5214. <http://dx.doi.org/10.1145/2858036.2858299>.
- of Leeds, U., 2021. Critical thinking - a model for critical thinking. Retrieved in August 2021 from [https://library.leeds.ac.uk/info/1401/academic\\_skills/105/critical\\_thinking/2](https://library.leeds.ac.uk/info/1401/academic_skills/105/critical_thinking/2).
- Liao, Q.V., Gruen, D., Miller, S., 2020. Questioning the AI: Informing design practices for explainable AI user experiences. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, pp. 1–15.
- Lindberg, D., Popowich, F., Nesbit, J., Winne, P., 2013. Generating natural language questions to support learning on-line. In: Proceedings of the 14th European Workshop on Natural Language Generation. Association for Computational Linguistics, Sofia, Bulgaria, pp. 105–114, URL <https://aclanthology.org/W13-2114>.
- LW, A., DR, K., PW, A., KA, C., Mayer, R., PR, P., Rath, J., MC, W., 2001. A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives, 1st edition 080131903X, Pearson.
- Ma, X., Yu, L., Forlizzi, J.L., Dow, S.P., 2015. Exiting the design studio: Leveraging online participants for early-stage design feedback. In: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work and Social Computing. In: CSCW '15, Association for Computing Machinery, New York, NY, USA, pp. 676–685. <http://dx.doi.org/10.1145/2675133.2675174>.
- MacAvaney, S., Yao, H.-R., Yang, E., Russell, K., Goharian, N., Frieder, O., 2019. Hate speech detection: Challenges and solutions. *PLoS One* 14.
- Machete, P., Turpin, M., 2020. The use of critical thinking to identify fake news: A systematic literature review. *Responsible Des. Implement. Use Inf. Commun. Technol.* 12067, 235–246.
- Mariane, H., 2002. The Zone of Proximal Development as Basis for Instruction. In: An introduction to Vygotsky, Routledge, pp. 183–207.
- McCartney, M., Childers, C., Baiduc, R.R., Barnicle, K., 2018. Annotated primary literature: A professional development opportunity in science communication for graduate students and postdocs. *J. Microbiol. Biol. Educ.* 19.
- Miller, S.R., Bailey, B., 2014. Searching for inspiration: an in-depth look at designers example finding practices.
- Miniukovich, A., Scaltritti, M., Sulpizio, S., De Angeli, A., 2019. Guideline-based evaluation of web readability. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. In: CHI '19, Association for Computing Machinery, New York, NY, USA, pp. 1–12. <http://dx.doi.org/10.1145/3290605.3300738>.
- Mitzenmacher, M., 2020. How to read a research paper. Retrieved in June 2020 from <https://www.eecs.harvard.edu/~michaelm/postscripts/ReadPaper.pdf>.
- Mombini, H., Tulu, B., Strong, D., Agu, E., Lindsay, C., Loretz, L., Pedersen, P., Dunn, R., 2020. Do novice and expert users of clinical decision support tools need different explanations? In: Proceedings of the ... Americas Conference on Information Systems Americas Conference on Information Systems. 2020.
- Mozilla, 2020. Pdf.js. Retrieved in September 2020 from <https://github.com/mozilla/pdf.js>.
- Navalpakkam, V., Rao, J., Slaney, M., 2011a. Using gaze patterns to study and predict reading struggles due to distraction. In: CHI EA '11, Association for Computing Machinery, New York, NY, USA, pp. 1705–1710. <http://dx.doi.org/10.1145/1979742.1979832>.
- Navalpakkam, V., Rao, J., Slaney, M., 2011b. Using gaze patterns to study and predict reading struggles due to distraction. In: CHI EA '11 Extended Abstracts on Human Factors in Computing Systems. In: CHI EA '11, Association for Computing Machinery, New York, NY, USA, pp. 1705–1710. <http://dx.doi.org/10.1145/1979742.1979832>.
- Ngoon, T.J., Fraser, C.A., Weingarten, A.S., Dontcheva, M., Klemmer, S., 2018. Interactive guidance techniques for improving creative feedback. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, pp. 1–11.
- Nye, B., Graesser, A., Hu, X., 2014. Autotutor and family: A review of 17 years of natural language tutoring. *Int. J. Artif. Intell. Educ.* 24, <http://dx.doi.org/10.1007/s40593-014-0029-5>.
- OECD, 2018. The future of education and skills: Education 2030. Retrieved in January 2022 from <http://www.oecd.org/education/2030/oecd-education-2030-position-paper.pdf>.

- OpenReview, 2021. Openreview.net. Retrieved in September 2021 from <https://openreview.net/>.
- Peng, Z., 2021. Designing and evaluating intelligent agents' interaction mechanisms for assisting human in high-level thinking tasks. In: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA.
- Peng, Z., Guo, Q., Tsang, K.W., Ma, X., 2020. Exploring the effects of technological writing assistance for support providers in online mental health community. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. In: CHI '20, Association for Computing Machinery, New York, NY, USA, pp. 1–15. <http://dx.doi.org/10.1145/3313831.3376695>.
- Peng, Z., Kwon, Y., Lu, J., Wu, Z., Ma, X., 2019. Design and evaluation of service robot's proactivity in decision-making support process. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. In: CHI '19, Association for Computing Machinery, New York, NY, USA, pp. 1–13. <http://dx.doi.org/10.1145/3290605.3300328>.
- Puig, B., Blanco-Anaya, P., Pérez-Maceira, J.J., 2021. "Fake news" or real science? Critical thinking to assess information on COVID-19. In: Frontiers in Education.
- Rapp, A., Curti, L., Boldi, A., 2021. The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots. *Int. J. Hum. Comput. Stud.* 151, 102630.
- Reader, T.D., 2020. How to download your kindle notes and highlights and export them. Retrieved in August 2020 from <https://the-digital-reader.com/2020/06/28/how-to-download-your-kindle-notes-and-highlights-and-export-them/>.
- RMIT University, 2021. Critical essay: Landscape architecture. Retrieved in August 2021 from <https://www.dlsweb.rmit.edu.au/slc/LandscapeArchitectureCriticalEssay/assets/Landscape%20architecture-PDF.pdf>.
- Romat, H., Henry Riche, N., Hinckley, K., Lee, B., Appert, C., Pietriga, E., Collins, C., 2019. Activeink: (th)inking with data. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. In: CHI '19, Association for Computing Machinery, New York, NY, USA, pp. 1–13. <http://dx.doi.org/10.1145/3290605.3300272>.
- Ruan, S., Jiang, L., Xu, J., Tham, B.J.-K., Qiu, Z., Zhu, Y., Murnane, E.L., Brunskill, E., Landay, J.A., 2019. QuizBot: A dialogue-based adaptive learning system for factual knowledge. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. In: CHI '19, Association for Computing Machinery, New York, NY, USA, pp. 1–13. <http://dx.doi.org/10.1145/3290605.3300587>.
- SeatGeek, 2020. Fuzzywuzzy. Retrieved in September 2020 from <https://github.com/seatgeek/fuzzywuzzy>.
- Seering, J., Luria, M., Kaufman, G., Hammer, J., 2019. Beyond dyadic interactions: Considering chatbots as community members. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. In: CHI '19, Association for Computing Machinery, New York, NY, USA, pp. 1–13. <http://dx.doi.org/10.1145/3290605.3300680>.
- Shepherd, M., Sande, C., 2014. Reading mathematics for understanding—From novice to expert. *J. Math Behav.* 35, 74–86. <http://dx.doi.org/10.1016/j.jmathb.2014.06.003>.
- Shum, H., 2020. You are how you read. Retrieved in June 2020 from <https://www.youtube.com/watch?v=Du7qLsToW-o>.
- Sigchi, A., 2019. Late breaking work - CHI 2019. Retrieved in September 2020 from <https://chi2019.acm.org/authors/late-breaking-work/>.
- Sigchi, A., 2020. Guide to reviewing papers - CHI 2020. Retrieved in September 2020 from <https://chi2020.acm.org/guide-to-reviewing-papers/>.
- Stamy, M., 2020. PyPDF2. Retrieved in September 2020 from <https://github.com/mstamy2/PyPDF2>.
- Subramonyam, H., Seifert, C., Shah, P., Adar, E., 2020. TexSketch: Active diagramming through pen-and-ink annotations. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. In: CHI '20, Association for Computing Machinery, New York, NY, USA, pp. 1–13. <http://dx.doi.org/10.1145/3313831.3376155>, URL <https://doi.org.lib.ezproxy.ust.hk/10.1145/3313831.3376155>.
- Syed, R., Collins-Thompson, K., Bennett, P., Tang, M., Williams, S., Iqbal, S., Tay, W., 2020. Improving learning outcomes with gaze tracking and automatic question generation. In: The Web Conference 2020 (Formerly WWW Conference). URL <https://www.microsoft.com/en-us/research/publication/improving-learning-outcomes-with-gaze-tracking-and-automatic-question-generation/>.
- Tan, J.P.-L., Yang, S., Koh, E., Jonathan, C., 2016. Fostering 21st century literacies through a collaborative critical reading and learning analytics environment: User-perceived benefits and problematics. In: Proceedings of the Sixth International Conference on Learning Analytics & Knowledge. In: IAK '16, Association for Computing Machinery, New York, NY, USA, pp. 430–434. <http://dx.doi.org/10.1145/2883851.2883965>.
- Tomasek, T., 2009. Critical reading: Using reading prompts to promote active engagement with text. *Int. J. Teaching Learn. Higher Educ.* 21, 127–132.
- Tyurin, A., 2020. React-pdf-highlighter. Retrieved in September 2020 from <https://github.com/agentcooper/react-pdf-highlighter>.
- University of Kent, 2021. Critical thinking and writing. Retrieved in August 2021 from <https://www.kent.ac.uk/learning/documents/student-support/value-map/valuemap1516/criticalthinkingandwriting171015alg.pdf>.
- University of Toronto, 2020. Reading critically. Retrieved in June 2020 from <https://www.utsc.utoronto.ca/twc/sites/utsc.utoronto.ca.twc/files/resource-files/CriticalReading.pdf>.
- u/valetudoo, 2021. ML/DL paper reading group. Retrieved in September 2021 from [https://www.reddit.com/r/MachineLearning/comments/oonijn/r\\_mdl\\_paper\\_reading\\_group/](https://www.reddit.com/r/MachineLearning/comments/oonijn/r_mdl_paper_reading_group/).
- Venkatesh, V., Bala, H., 2008. Technology acceptance model 3 and a research agenda on interventions. *Decis. Sci.* 39 (2), 273–315. <http://dx.doi.org/10.1111/j.1540-5915.2008.00192.x>, arXiv:<https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-5915.2008.00192.x>.
- Wallace, M., Wray, A., 2016. Critical Reading and Writing for Postgraduates. In: Student Success, SAGE Publications, pp. 29–43, URL <https://books.google.com.hk/books?id=2bGzCwAAQBAJ>.
- Wambsganss, T., Kueng, T., Soellner, M., Leimeister, J.M., 2021. ArgueTutor: An adaptive dialog-based learning system for argumentation skills. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA.
- Wambsganss, T., Niklaus, C., Cetto, M., Söllner, M., Handschuh, S., Leimeister, J.M., 2020. AL: An adaptive learning support system for argumentation skills. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. In: CHI '20, Association for Computing Machinery, New York, NY, USA, pp. 1–14. <http://dx.doi.org/10.1145/3313831.3376732>.
- Wang, Y., Liu, D., Qu, H., Luo, Q., Ma, X., 2016. A guided tour of literature review: Facilitating academic paper reading with narrative visualization. pp. 17–24. <http://dx.doi.org/10.1145/2968220.2968242>.
- Weber, F., Wambsganss, T., Rüttimann, D., Söllner, M., 2021. Pedagogical agents for interactive learning: A taxonomy of conversational agents in education completed research paper.
- Weinman, N., Drucker, S.M., Barik, T., DeLine, R., 2021. Fork it: Supporting stateful alternatives in computational notebooks. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. In: CHI '21, Association for Computing Machinery, New York, NY, USA, <http://dx.doi.org/10.1145/3411764.3445527>.
- Wiles, J., Allen, R., Butler, R., 2016. Owning my thoughts was difficult: Encouraging students to read and write critically in a tertiary qualitative research methods course. *J. Univ. Teach. Learn. Practice* 13, <http://dx.doi.org/10.53761/1.13.1.8>.
- Wilson, K., 2016. Critical reading, critical thinking: Delicate scaffolding in english for academic purposes (EAP). *Thinking Skills Creat.* 22, 256–265. <http://dx.doi.org/10.1016/j.tsc.2016.10.002>, URL <https://www.sciencedirect.com/science/article/pii/S1871187116301365>.
- Wilson, K., Devereux, L., Macken-Horarik, M., Tringham-Jack, C., 2004. Reading readings: How students learn to (dis)engage with critical reading.
- Winkler, R., Hobert, S., Salovaara, A., Söllner, M., Leimeister, J.M., 2020. Sara, the lecturer: Improving learning in online education with a scaffolding-based conversational agent. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. In: CHI '20, Association for Computing Machinery, New York, NY, USA, pp. 1–14. <http://dx.doi.org/10.1145/3313831.3376781>.
- Woolson, R., 2008. Wilcoxon signed-rank test. ISBN: 047146242X, <http://dx.doi.org/10.1002/9780471462422.eoct979>.
- Wu, Z., Jiang, Y., Liu, Y., Ma, X., 2020. Predicting and diagnosing user engagement with mobile UI animation via a data-driven approach. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. In: CHI '20, Association for Computing Machinery, New York, NY, USA, pp. 1–13. <http://dx.doi.org/10.1145/3313831.3376324>.
- Xiao, Z., Zhou, M.X., Chen, W., Yang, H., Chi, C., 2020. If I hear you correctly: Building and evaluating interview chatbots with active listening skills. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. In: CHI '20, Association for Computing Machinery, New York, NY, USA, pp. 1–14. <http://dx.doi.org/10.1145/3313831.3376131>.
- Yan, L., Glassman, E.L., Zhang, T., 2021. Visualizing examples of deep neural networks at scale. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. In: CHI '21, Association for Computing Machinery, New York, NY, USA, <http://dx.doi.org/10.1145/3411764.3445654>.
- Yerkes, R.M., Dodson, J.D., 1908. The relation of strength of stimulus to rapidity of habit-formation. *J. Compar. Neurol. Psychol.* 18 (5), 459–482. <http://dx.doi.org/10.1002/cne.920180503>, arXiv:<https://onlinelibrary.wiley.com/doi/pdf/10.1002/cne.920180503>.
- Yu, J., 2015. Analysis of critical reading strategies and its effect on college english reading. *Theory Pract. Lang. Stud.* 5, 134. <http://dx.doi.org/10.17507/tpls.0501.18>.
- Zimmerman, J., Forlizzi, J., 2017. Speed dating: Providing a menu of possible futures. *She Ji Des. Econ. Innov.* 3, 30–50.
- Zoom Video Communications, I., 2020. Zoom. Retrieved in September 2020 from <https://zoom.us/>.