CriTrainer: An Adaptive Training Tool for Critical Paper Reading

Kangyu Yuan*
yuanky5@mail2.sysu.edu.cn
Sun Yat-sen University
Zhuhai, China

Hehai Lin*
linhh29@mail2.sysu.edu.cn
Sun Yat-sen University
Zhuhai, China

Shilei Cao*
caoshlei@mail2.sysu.edu.cn
Sun Yat-sen University
Zhuhai, China

Zhenhui Peng†
pengzhh29@mail.sysu.edu.cn
Sun Yat-sen University
Zhuhai, China

Qingyu Guo
qguoag@connect.ust.hk
Hong Kong University of Science and Technology
Hongkong, China

Xiaojuan Ma
mxj@cse.ust.hk
Hong Kong University of Science and Technology
Hongkong, China

ABSTRACT

Learning to read scientific papers critically, which requires first grasping their main ideas and then raising critical thoughts, is important yet challenging for novice researchers. The traditional ways to develop critical paper reading (CPR) skills, e.g., checking general tutorials or taking reading courses, often can not provide individuals with adaptive and accessible support. In this paper, we first derive user requirements of a CPR training tool based on literature and a survey study (N=52). Then, we develop *CriTrainer*, an interactive tool for CPR training. It leverages text summarization techniques to train readers' skills in grasping the paper's main ideas. It further utilizes template-based generated questions to help them learn how to raise critical thoughts. A mixed-design study (N=24) shows that compared to a baseline tool with general CPR guidance, students trained by *CriTrainer* perform better in independently raising critical thinking questions on a new paper. We conclude with design considerations for CPR training tools.

CCS CONCEPTS

• **Human-centered computing**; • **Interactive systems**; • **Natural language processing**; • **Empirical studies**;

KEYWORDS

Human-centered computing; Paper reading; Critical thinking

ACM Reference Format:
Kangyu Yuan, Hehai Lin, Shilei Cao, Zhenhui Peng, Qingyu Guo, and Xiaojuan Ma. 2023. *CriTrainer*: An Adaptive Training Tool for Critical Paper Reading. In *The 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23), October 29-November 1, 2023, San Francisco, CA, USA.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3586183.3606816

1 INTRODUCTION

Reading scientific papers critically is beneficial to researchers [82], *e.g.,* by helping them deepen their understanding and get inspiration in a research domain. To read critically, one needs to not only comprehend the main ideas conveyed through the content and figures but also maintain a skeptical and provisional view of the content [56]. For example, critical readers should be able to summarize the paper's background, motivation, methods, contribution, etc., while also raising critical thinking thoughts like whether the claims are reasonable and why the alternative approaches are not chosen. In this paper, we reflect people's critical reading skills on their performance in summarizing paper content and raising critical thinking questions. These two aspects are also revealed in the review guidelines of top venues (*e.g.,* CHI [1], UIST [2], and CSCW [3]). For instance, a typical review would contain a summary of the introduced ideas, approaches, and contributions, as well as assessments of the strengths and weaknesses from aspects like methodology, analysis, interpretation, significance, and clarity.

However, learning critical paper reading skills is non-trivial for novices who get started in a research domain. From the researchers' perspectives, reading critically is more than answering questions on the texts, *e.g.,* the multiple-choice ones in high-school exams, or open-ended critical thinking questions. It also involves extracting key points from the paper, rephrasing them, and questioning the paper content [60, 76]. This requires readers to actively analyze, synthesize, and evaluate the paper content – the higher-order thinking skills in education suggested in Bloom's taxonomy [10]. These higher-order thinking skills are known to be difficult for students to acquire in various educational scenarios [25, 66].

Traditionally, there are two common ways to learn critical paper reading skills – referring to critical reading guidelines compiled by researchers [37, 60, 76, 79] or seeking support from course instructors, peers, or senior researchers. The former approach suits individual readers at any time but falls short in only providing general guidance, which may not be effective in helping students develop critical reading skills [2, 15, 27, 32, 63]. The latter means, on the other hand, can offer adaptive hints and feedback on specific reading materials yet requires qualified persons who have also read the same materials.

[*] These authors contributed equally to this work.
[†] The Corresponding author

UIST '23, October 29-November 1, 2023, San Francisco, CA, USA

ACM ISBN 979-8-4007-0132-0/23/10…$15.00
https://doi.org/10.1145/3586183.3606816

[1] https://chi2023.acm.org/submission-guides/guide-to-reviewing-papers/
[2] https://uist.acm.org/2023/author-guide/
[3] https://cscw.acm.org/2016/volunteer/DaveRandallReviewingforCSCW.pdf

To mitigate these issues, existing HCI (Human–Computer Interaction) research has explored various in-situ reading support. For example, Peng et al. [65] developed CReBot that asks section-level questions and provides content-independent guidance (*e.g.,* general aspects to answer the questions) to facilitate users in critical paper reading. Chen et al. [19] built Marvista which employs natural language processing models to provide text-specific (content-aware) support for reading online articles. However, CReBot's questions are pre-compiled and do not match well to the specific texts [65], while Marvista focuses on comprehending the online articles instead of facilitating critical thinking [19]. Besides, these tools seldom incorporate educational elements that aim to improve readers' abilities like critical thinking. Prior researchers on HCI and education have proposed a set of intelligent tutors to facilitate skill acquisition, such as QuizBot for learning factual language [70] and ArgueTutor for practicing argumentation writings [83]. Their studies suggest the in-situ support and feedback of the tutors can improve learning engagement and gains [83]. Nevertheless, little is known about what types of support and feedback are required and how to integrate them into critical paper reading training practices. In all, there is a lack of investigations into the design, effectiveness, and user experience of a critical paper reading training tool which has text-specific educational elements.

To this end, we design, develop, and evaluate an adaptive training tool *CriTrainer* for critical paper reading. Based on previous literature on critical thinking (*e.g.,* [56, 79, 82]) and tutoring reading skills [8], we structure a QR2AC critical paper reading training process for each selected paper section (*e.g.,* Introduction). In this framework, learners first should think of comprehension **Q**uestions (*e.g.,* "what, how") and **R**ead the selected section (noted as QR stage). Next, they should conduct the first training task by drafting a summary of the selected section (**A**nswer) and **C**hecking how to improve (noted as AC-1 stage). They then proceed to the second training task in which they should raise relevant critical thinking questions on the selected section and check how to improve (noted as AC-2 stage). We conduct a survey study with 52 university students to identify their difficulties in critical paper reading and the need for support in our structured training process.

Based on the derived findings from the survey study, we develop *CriTrainer* for training critical paper reading skills. *CriTrainer* provides generated comprehension questions in the QR stage. It offers feedback on the drafted summary and highlights key points in the original paper content based on the generated summary in the AC-1 stage. It provides text-specific critical thinking questions generated by our proposed template-based approach in the AC-2 stage. We evaluate *CriTrainer*'s effectiveness and user experience on critical paper reading training through a mixed-design study with 24 undergraduates compared to a baseline tool that provides general guidance in the QR2AC process. We compare participants' performance in independently summarizing the content and raising relevant critical thinking questions on given academic papers before and after the training sessions. We also measure their behaviors and perceptions in the training sessions in which they should read two papers with either *CriTrainer* or baseline tool. Our result shows that compared to those with the baseline tool, participants with *CriTrainer* have significantly more improvement regarding their ability to raise understandable, relevant, and critical questions after the training sessions. After training with either tool, participants can also better express their understanding of the paper content in their drafted summaries. Participants' comments on *CriTrainer* highlight the benefits of its text-specific critical thinking questions. We discuss insights from our findings and offer design considerations for future critical paper reading training tools.

In summary, the main contributions of this work are:

- We propose a critical paper reading training tool *CriTrainer* which offers text-specific guidance and incorporates educational elements in a structured training process.
- We demonstrate the effectiveness of *CriTrainer* in helping participants acquire critical paper reading skills via a mixed-design study.
- We provide insights and design considerations for future critical paper reading training tools.

2 RELATED WORK

In this section, we review the literature on critical paper reading that motivates our work, related paper reading support and intelligent tutoring tools that inspire our design, and recent natural language processing approaches that power *CriTrainer*.

2.1 Critical Paper Reading

Critical thinking is an essential skill in the twenty-first century's education [9, 75]. When applied to paper reading, it requires critical readers to maintain a distance from and keep friendly skepticism towards what authors claim [60, 82]. First, critical readers need to comprehend the paper's content, which means that they should be able to read beyond the literal meaning of the text and grasp the paper's main ideas [56, 77]. The performance of paper summarization can reflect this comprehension ability as learners must take the entire text into consideration and determine its main points when summarizing [29, 68]. After paper comprehension, critical readers should think deeply about the authors' judgments and opinions [56]. One way that reveals this capability is to raise understandable, relevant, and thought-provoking critical thinking questions on the paper content [40, 65], *e.g.,* "Is the motivation of this research clear and strong?". Therefore, in this paper, we measure people's critical reading ability based on their performance in summarizing the paper content and raising critical thinking questions on it.

To develop critical reading skills, people need to learn and practice. One common learning approach is referring to existing guidelines compiled by experienced researchers, which usually contain a set of general critical thinking questions [8, 37, 60, 65, 76, 79, 82, 91]. For instance, the QRAC (**Q**uestion, **R**ead, **A**nswer, **C**heck) [8] guideline has been used in reading courses to train students' skills in reading comprehension. In the QR stage of the guideline, users should read each paper section with comprehension questions [8]. In the AC stage, users should answer the questions and check if their answers could contribute a good summary of the section they read [8]. However, the QRAC guideline does not aid critical thinking during the paper reading process. Peng et al. [65] compiled a guideline for critical paper reading based on the related and publicly available articles, tutorials, and books. The guidelines provide a step-by-step approach to critical paper reading and offer section-level questions from various aspects, such as the motivation aspect of the introduction (*e.g.,* "Why should we care about this research problem?" from the motivation aspect of the introduction). These guidelines and questions serve as a good starting point for practicing critical paper reading. However, they are rather general, which may not be effective for developing critical paper reading skills, especially for novice researchers who would need more specific guidance in the learning process [2, 15, 27, 32, 63]. A more effective learning method would be interacting with lecturers, senior scholars, or peers, *e.g.,* in course activities, paper sharing, and discussions in individual meetings. These experienced persons can offer guidance adaptive to the paper content and timely feedback on their understandings and thoughts. However, they are not always available for individual learners. Our work is motivated by the needs of novice researchers to develop critical paper reading skills. We seek to facilitate them with an adaptive and accessible critical reading training tool. Specifically, we structure a QR2AC critical paper reading training process (Figure 1). It simulates the original QRAC [8] process but differs from it with an additional AC stage for training critical thinking skills and text-specific support.

2.2 Paper Reading Support Tools

HCI researchers have explored a series of tools to support paper reading activities in digital devices [19, 65, 67, 72, 78, 86]. For example, Kim et al. [38] built an interactive document reader that links the text with its corresponding table cells automatically, which can reduce split attention and facilitate reading. Chen et al. [19] developed Marvista that employs various Natural Language Processing (NLP) technology like abstractive summarization to provide text-specific assistance when users are reading online articles. Their main user study showed that Marvista helps them better comprehend the article [19]. Similarly, August et al. [6] created Paper Plain that utilizes NLP techniques to enhance understanding of medical papers. Nevertheless, they focus on comprehending the articles and do not incorporate educational elements for training critical thinking skills.

As for critical paper reading, Tan et al. [78] presented WiREAD, a web-based collaborative platform that supports peer interactions and provides feedback for both students and teachers to engage in critical paper reading together. WiREAD is shown to enhance the critical reading engagement levels of peers and instructors in their user study. [78]. However, it requires other persons to engage in the reading process and therefore lacks scalability. To support individual learners, Peng et al. [65] designed CReBot, a chatbot-style tool that prompts section-level questions and offers general critical thinking guidance during users' paper reading process. Their experiments demonstrated CReBot's engagement and usefulness over static guidelines for routine paper readers but indicated that its benefits are not significant for novice researchers [65]. One possible reason is that for novices, the form of interaction (*e.g.,* static vs. conversational in CReBot) may not be the key for critical reading support, but the text-specific instructions and feedback would be [2, 15, 27, 32, 63].

In line with these tools, our *CriTrainer* can provide individuals with in-situ text-specific paper reading support. However, unlike these tools, we do not aim at improving users' reading performance in each session with *CriTrainer*. We position *CriTrainer* as a training tool whose goal is to enable independent critical paper reading after the training sessions.

2.3 Intelligent Tutoring Tools

To help users acquire knowledge and skills, HCI researchers have explored a variety of intelligent tutoring tools [18, 59, 70, 83, 88, 89, 95, 96]. For example, Ruan et al. [70] developed QuizBot, an AI-powered chatbot that assists students in learning factual knowledge in subjects like science, safety, and English vocabulary. The bot engages users in the learning process by asking questions and providing corrective feedback on users' responses [70]. Similarly, Wambsganss et al. [83] built ArgueTutor, an adaptive dialog-based tutoring system that helps students improve their argumentative writing skills by providing adaptive and instant feedback (*e.g.,* message on what to improve) on their drafted essays. Zhang et al. [96] designed Withyou, an automated adaptive speech shadowing tutoring tool that assists people in learning foreign spoken languages. Withyou automatically adjusts the playback and the difficulty of a speech template through speech recognition and is shown to lead to a larger improvement on spoken language compared to the conventional method [96]. These intelligent tutoring tools commonly structure the learning process in which they offer adaptive support and timely feedback to learners. Our work gets inspired by these tools and extends them with first-hand insights into the design, usefulness, and user experience of an adaptive training tool for critical paper reading.

2.4 Text Summarization and Question Generation Techniques

To provide adaptive training support identified in the later formative study, *CriTrainer* mainly adopt the text summarization and question generation techniques. We briefly review the related ones below.

Text summarization (TS) aims at producing a concise and smooth text which retains the key information and overall meanings of the original content [4, 87]. Basically, there are two main approaches to text summarization: extraction and abstraction. Extraction involves concatenating sentences taken from the original text into the output summary [3, 54], while abstraction involves generating a summary with new and rebuilt sentences using words that might not be in the original text [87, 93]. Both approaches traditionally leverage rule-based [58], sentence-compression [22] or template-based [30] strategies and are now mainly based on deep-learning models [5, 20, 35, 48, 54, 55, 90, 94]. For example, BART uses the standard sequence-to-sequence Transformer architecture and gets high performance on summarization tasks [48]. In this work, we adopt a deep-learning-based abstractive summarization approach to support training on paper comprehension, as its output summary is more closely to the human-written ones [26].

Question generation (QG) aims at creating semantically and syntactically correct questions given various sources like raw text, database, or semantic representation with or without answers [57, 71]. Intelligent tutoring systems have started to exploit QG techniques to support online learning and self-learning, generating natural language questions for users in an online learning system to deepen their understanding of educational material [43, 49]. Rules-based

QG approaches generate questions relying on various rules, like syntactic rules that incorporated linguistic features (e.g., subject-auxiliary inversion) [34], transformation rules based on case grammars (e.g., Agentive, Instrumental, Dative, Factitive, Locative, and Objective) [28], and templates derived from human supervisors [51]. The key to the quality of template-based generated questions is filling the right templates with the right words. To mitigate the scalability issues of rule-based approaches, recent AI researchers commonly train deep neural networks models with large annotated corpus for QG in an end-to-end manner [17, 47, 57, 73, 92]. While these generated questions are more diverse compared to the template-based ones, most of them ask about "what" and "how", which could not satisfy the needs for critical thinking questions about "why" and "how well" [65]. Little work has explored ways to generate critical thinking questions, which would require a large amount of labeled data to develop deep-learning models. In our work, we leverage these neural-network-based models to help users read with generated comprehension questions. To further support learners to raise critical thoughts, we propose a template-based method to generate critical thinking questions filled with AI-detected keywords and validate their appropriateness.

3 FORMATIVE STUDY

To support the acquisition of critical paper reading skills and the need for guiding formative study design, we first structure a QR2AC training process based on literature. We then identify user challenges and needs for support in this process via a survey study.

3.1 QR2AC Critical Reading Training Process

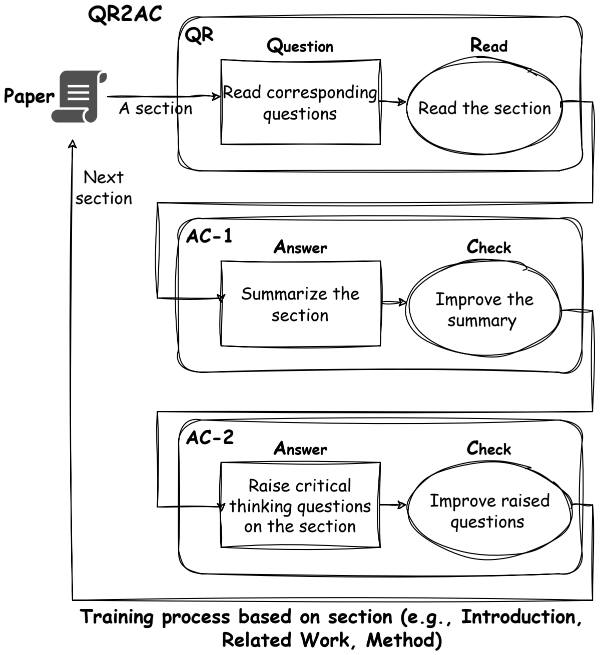


Figure 1: QR2AC Critical Paper Reading Training Process. Participants experience three stages for the selected section.

As reviewed in the Related Work (Section 2.1), the abilities to summarize paper content and raise relevant critical thinking questions are important in critical paper reading. Therefore we focus on training these two skills in this work. We adapt the QRAC ("**Q**uestion, **R**ead, **A**nswer, **C**heck") process [8], which is commonly used in training courses about reading skills [46]. For example, Lee [46] adapt the QRAC framework with a checklist that guides the four steps to organize online collaborative reading activities to train critical thinking skills. We differ from their work by allowing individuals to train critical paper reading skills at any time and providing adaptive support in this process. Specifically, we propose a QR2AC training process (Figure 1): 1) Read each paper section with questions about corresponding content (**QR**); 2) Summarize the reading section and check if it was a good summary that contains the key points of the section (**AC-1**); 3) Raise critical thinking questions in the reading section and check if they were relevant to the paper content and reflect critical thoughts (**AC-2**).

3.2 Survey Study

3.2.1 *Survey Protocal.*

We develop our questionnaire on Microsoft Forms and invite our respondents to fill it online. It first asks for respondents' prior experience and perceived ability in critical paper reading. For those with experience in critical paper reading, we further ask about their encountered difficulties in critical paper reading practices when they were beginners in critical paper reading practices. To elicit novices' needs for support in each training stage, we brainstorm a set of potential features (Table 1) of the training tool, based on their paper reading experience and related works (*e.g.,* CReBot [65], Marvista [19], the theory of scaffolding [33]). We then ask participants to rate their perceived usefulness of these features on a standard five-point Likert scale (1 - not useful at all, 5 - very useful). We also include open-ended questions about other features they would like to have in the training support tool.

3.2.2 *Respondents.*

We recruited 52 students (S1-52; 18 **F**emales, 29 **M**ales, 4 **N**ot to specify) through social networks and word-of-mouth. They all speak English as their second language. Among them were 30 undergraduate, 15 master, and 7 Ph.D. students. Their ages range from 18 to 28 years old with an average age of 23 (*SD* = 2.23). Twenty-five respondents reported having no or little knowledge, 18 respondents had moderate knowledge, and the rest 9 people know a lot about critical paper reading. Forty-nine students indicated prior experience in reading scientific papers.

3.3 Findings

3.3.1 *Challenges of Critical Paper Reading.*

For the 27 respondents with prior knowledge of critical paper reading, we ask about their difficulties in the reading process when they were novices and summarize the results below.

C1: Feel uneasy to capture the key ideas. Eight respondents recalled that they had struggles in identifying the key points of the reading content. "*The paper is commonly lengthy, containing too much information that hinders me from understanding and extracting its main points, e.g., its novelty or contributions*" (S14, **M**ale, age: 21).

C2: Lack of skills in raising critical thoughts. Five students mentioned that they lack skills in questioning the paper content, *e.g.,* what points could be criticized and how to criticize them. "*I*

Table 1: Perceived usefulness of our training tool's potential features in the QR2AC training process; 1/5 - not useful/very useful. We implement those features (bold) with an average score higher than 4 (except the first feature in AC-2 stage).

	Potential Features	M	SD
QR	**(1) Highlight the paragraph where the possible answer locate as a hint**	**4.44**	**0.53**
	(2) Highlight the possible answer for reference	**4.19**	**0.71**
	(3) Provide multiple comprehension questions (e.g., what, how) for users to choose from	**4.15**	**0.53**
	(4) Provide the general guideline of paper reading	3.92	1.07
	(5) Provide a comprehension question one at a time	3.50	0.97
AC-1	**(1) Highlight the possibly key sentences as hints**	**4.54**	**0.50**
	(2) Provide the potential summary for reference	**4.27**	**0.59**
	(3) Provide the general guideline for summarizing	**4.06**	**0.82**
	(4) The tool rates users' summaries for them.	3.81	0.74
	(5) Users assess their own summary.	3.58	0.97
AC-2	(1) Answer users' critical thinking questions	4.29	0.86
	(2) Provide examples of critical thinking questions for reference	**4.23**	**0.72**
	(3) Guide users to raise critical thinking questions	**4.21**	**0.66**
	(4) The tool rates users' critical thinking questions for them.	3.73	0.83
	(5) Users assess their own critical thinking questions.	3.60	0.88

did not have the critical reading mindset and skill set at my early research stage. I normally read the whole paper word by word without assessing whether it is convincing. This is inefficient and ineffective for me to learn from the papers" (S23, **F**emale, age: 25). Instead, they tend to agree with every argument of the authors and seldom think deeply about the paper's content. "*I would easily believe the authors' opinion when entering a new field if there is no other voice that judges it*" (S8, M, age: 23).

3.3.2 Perceived Usefulness of Potential Features. Table 1 shows the respondents' average ratings on the usefulness of the potential tool's features. Using four points as the threshold, we found that in the QR stage, users would find the training tool useful if it highlights the paragraph of the possible answer as a hint, provides the potential answer for reference, and provides various comprehension questions (e.g., "what, how") for users to choose from. In the AC-1 stage, it would be useful to offer the general guideline for summarizing, highlight the possible key sentences as hints and provide the potential summary for reference. Lastly, in the AC-2 stage, the training tool would be useful if it guide users to raise critical thinking questions, provide examples of critical thinking questions for reference, and answer users' critical thinking questions.

3.3.3 Other Expected Features. Our respondents actively indicated their expected features of our training tool in the open-ended questions. In the QR stage, eight participants suggest that if the system provides users with comprehension questions, it can assess how well their thoughts have answered these questions. "*I hope that the tool can indicate the similarities and differences between my answer and the correct answer*" (S48, N, age: 20). Similarly, in the AC-1 stage, seven students would like the tool to indicate if there were redundant or missing points in their drafted summary of the reading content. "*The tool can showcase the points that could be added or deleted from my summary, which could prove guidance on how to improve*" (S2, M, age: 23).

3.4 Design Requirements for *CriTrainer*

Based on the survey findings and related literature on education, we derive the following design requirements (DR) for a critical paper reading training tool.

DR1: In the QR stage, the tool should provide text-specific comprehension (*e.g.,* "what, how") questions and highlight the paragraph of possible answers to these questions to keep the reader's focus on understanding the paper content. Providing questions can increase user engagement by encouraging them to search for answers within the text [8, 31, 62]. Additionally, these questions could help people better understand not only what they read, but also how they read the text [19]. This is also the perceived useful feature rated by our survey respondents (Table 1). Understanding the paper content is the basic requirement for the later summarization and question-asking tasks. Therefore, as expected by the respondents, the tool can further offer features like highlighting the paragraph of possible answers to the questions to consolidate readers' understanding.

DR2: In the AC-1 stage, the tool should provide a referred summary of the reading content and offer feedback on how users' draft summary match with and differ from the referred one. However, the tool should encourage users to spend necessary effort on the summarization task. This could train novices' critical reading ability in terms of capturing the paper's key ideas (C1). The referred summary can serve as a good example of how knowledgeable persons would outline key points [19, 50]. Nevertheless, unlike previous work on reading support tools that aim at easing reading workload, our training tool should encourage users to spend effort in this process. Based on the theory of scaffolding [33] and the respondents' ratings on potential features, it could first let learners have a trial on summarizing paper content, provide hints like possibly important sentences if needed in this process, and feedback on their summaries compared to a referred one.

DR3: In the AC-2 stage, the tool should provide templates of critical thinking questions (*e.g.,* "why, how well") and example text-specific questions based on these templates. However,

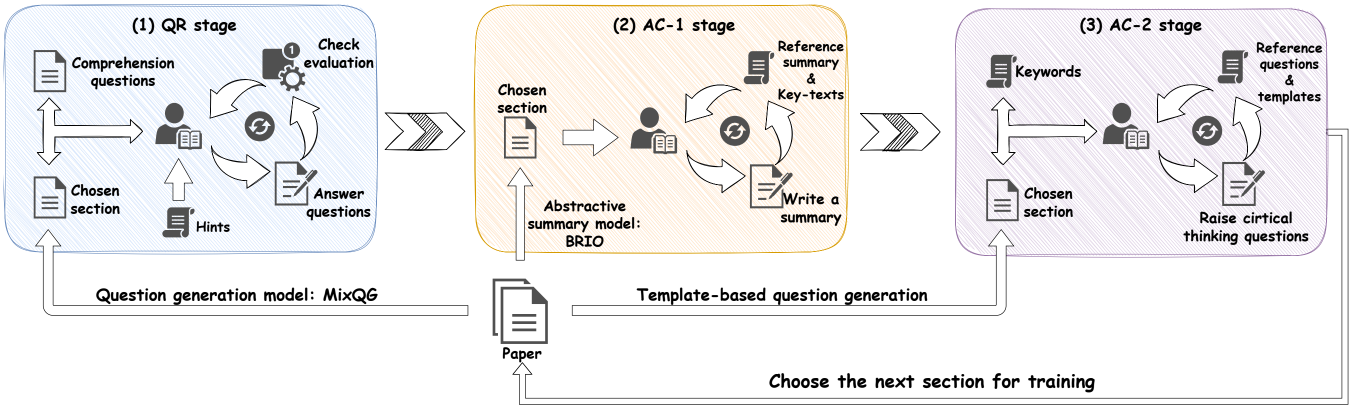

(1) QR stage
Comprehension questions
Check evaluation
Chosen section
Answer questions
Hints
(2) AC-1 stage
Chosen section
Reference summary & Key-texts
Write a summary
Abstractive summary model: BRIO
(3) AC-2 stage
Keywords
Reference questions & templates
Chosen section
Raise cirtical thinking questions
Question generation model: MixQG
Paper
Template-based question generation
Choose the next section for training

Figure 2: QR2AC training process supported by *CriTrainer*: (1) In the QR stage, *CriTrainer* offers comprehension questions about the selected section. (2) In the AC-1 stage, *CriTrainer* offers feedback on the drafted summary and highlights the key points matching the generated summary in the original paper content. (3) In the AC-2 stage, *CriTrainer* offers template-based text-specific critical thinking questions.

the tool should encourage readers to extend their thoughts on these questions and raise their own ones. This could train novices' critical reading ability in terms of raising critical thoughts (C2). The general critical thinking questions (*e.g.,* the ones compiled by Peng et al. [65]) can serve as good starting points for users to learn what these questions look like. However, our training tool should further help users learn how to raise questions that are similar to the referred ones but are more relevant to the reading content. Similar to the AC-1 stage, the tool should encourage them to have a trial on raising their own questions. Besides, while our respondents expect the tool can answer their raised questions, the critical thinking questions normally do not have correct answers. Instead, the tool should encourage users to extend their thoughts on the questions [64, 82, 91]. For example, the tool could highlight the keywords of the questions in the paper that may help re-examine relevant content, which is the expected feature in the survey findings.

4 SYSTEM

Based on the design requirements for the critical paper reading training tool, we develop *CriTrainer* that provides text-specific comprehension questions in the QR stage (DR1), hints and feedback based on the referred summary in the AC-1 stage (DR2), and template-based text-specific critical thinking questions in the AC-2 stage (DR3). We choose to integrate *CriTrainer* as an add-on into Google Docs, a publicly available platform for reading documents (including papers), in the form of a sidebar [36, 44]. The design of *CriTrainer* can be generalized and customized to other reading platforms. The *CriTrainer* add-on is implemented in javascript and connects to the backend Python flask server that processes the paper content and user interaction. *CriTrainer* supports the QR2AC training flow at the section level (*e.g.,* Introduction, Conclusion), as the original QRAC framework [8] does and the previous reading support tool like CReBot suggests [65]. Figure 2 illustrates the section-level training pipeline, *CriTrainer*'s support, and backend models in each stage. In this section, we first present a user scenario to walk through the QR2AC process in *CriTrainer* and then detail the interface design and backend models in each stage.

4.1 User scenario

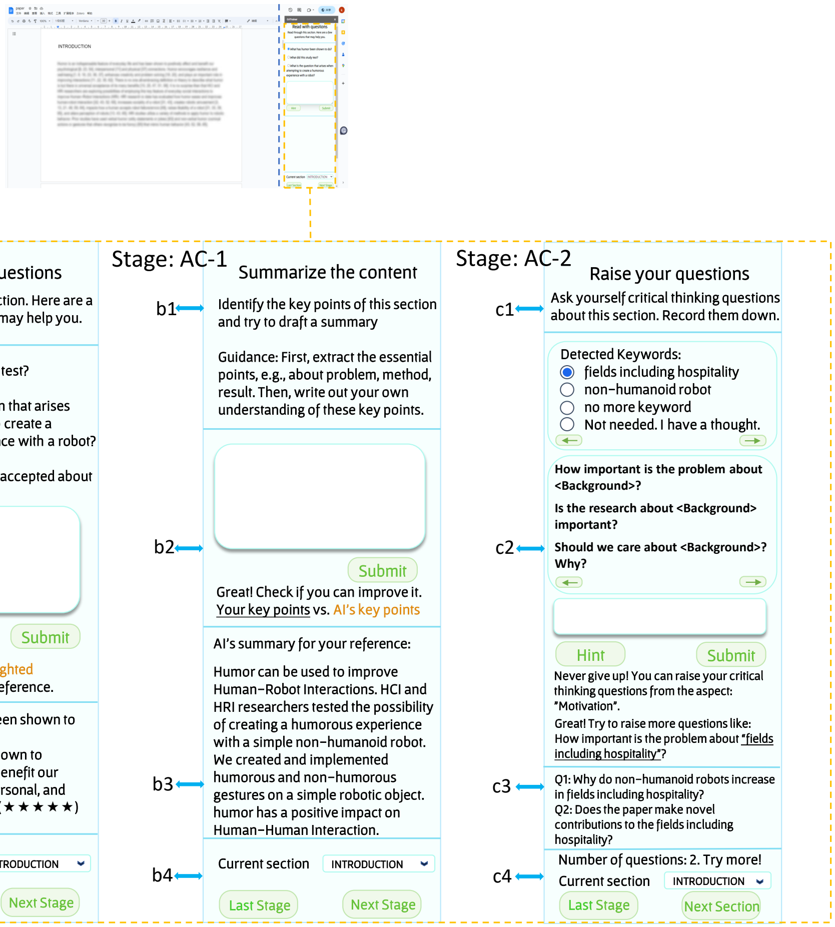
In this scenario, we describe how John, an undergraduate student, practices critical paper reading with *CriTrainer*. His goal is to improve his critical reading skills for the HCI paper-sharing course project which requires him to present and discuss scientific papers.

John first uploads his interested paper to Google Docs and invokes the *CriTrainer* add-on. He selects the Introduction section in *CriTrainer* (Figure 3 a4) and checks the generated comprehension questions (a2). John then reads through the Introduction section. He checks the hint on each question and answers it whenever he wishes. Next, he proceeds to the AC-1 stage (Figure 3 b) to practice paper comprehension skills. John drafts a summary of the Introduction section, compares it with the AI-generated one, refers to the highlighted key-texts in the paper, and makes revisions to improve his summary. After that, John starts to practice critical thinking skills in the AC-2 stage (Figure 3 c). He checks the templates of critical thinking questions and the highlighted keywords in the Introduction section. He is curious about the non-humanoid robot introduced in the paper and raises a question "Why should we care about the non-humanoid robot?". After a deep examination of the sentences about non-humanoid robot in the paper, John has a clearer mind about the motivation of this paper and is more curious about the proposed methods and results. Therefore, he selects other sections of the paper and goes through the QR2AC process again.

In summary, after reading the paper with *CriTrainer*, John is confident that he can present it well and have a deep discussion about it with his classmates. Also, he feels that he is more capable of reading other papers critically.

4.2 QR stage

In this stage, users should read and understand the paper section before proceeding to the two critical reading training tasks.



Read through this section. Here are a few questions that may help you.

● What did this study test?

○ What is the question that arises when attempting to create a humorous experience with a robot?

○ What is universally accepted about humor?

Very Good! The highlighted sentence is for your reference.

Q: What has humor been shown to do?
A: Humor has been shown to positively affect and benefit our psychological, interpersonal, and physical connections.(★★★★★)

Identify the key points of this section and try to draft a summary

Guidance: First, extract the essential points, e.g., about problem, method, result. Then, write out your own understanding of these key points.

Great! Check if you can improve it. Your key points vs. AI's key points

AI's summary for your reference:

Humor can be used to improve Human-Robot Interactions. HCI and HRI researchers tested the possibility of creating a humorous experience with a simple non-humanoid robot. We created and implemented humorous and non-humorous gestures on a simple robotic object. humor has a positive impact on Human-Human Interaction.

Ask yourself critical thinking questions about this section. Record them down.

Detected Keywords:
● fields including hospitality
○ non-humanoid robot
○ no more keyword
○ Not needed. I have a thought.

How important is the problem about <Background>?
Is the research about <Background> important?
Should we care about <Background>? Why?

Never give up! You can raise your critical thinking questions from the aspect: "Motivation".
Great! Try to raise more questions like: How important is the problem about "fields including hospitality"?

Q1: Why do non-humanoid robots increase in fields including hospitality?
Q2: Does the paper make novel contributions to the fields including hospitality?

Number of questions: 2. Try more!

Figure 3: Interface of *CriTrainer* in the QR2AC process. In the QR stage, users can read through a section and answer comprehension questions (a2). In the AC-1 stage, users can draft a summary of the section (b2) and get feedback based on the AI-generated summary (b3). In the AC-2 stage, users can raise critical thinking questions with the support of detected keywords, question templates, and generated questions (c2). In each stage, users can check the task description (a1, b1, c1), track their input content (a3, c3), and navigate to the previous or next stage (a4, b4, c4).

4.2.1 Interaction and interface design. When users select a paper section from the drop-down menu in the bottom part of *CriTrainer* (Figure 3), they will first experience the QR stage. In this stage, *CriTrainer* asks users to read through the selected section and offers a few comprehension questions that may help (a1). Users can select a question, read the section with it, "Submit" an answer to it, or/and check the "Hint" that will highlight the paragraph of referred answer in the paper if they wish (a2). If they submit an answer to the selected question, *CriTrainer* will rate it into five levels (i.e., very good, good, moderate, not too bad, and poor) based on its similarity to the referred answer [4]. Similar to CReBot [65] and Marvista [19], *CriTrainer* will record their submitted answers (a3) so that users can reflect on their understandings when needed. Users can click "Next Stage" to proceed to the first critical paper reading training task in the AC-1 stage.

4.2.2 Backend model. When the user invokes *CriTrainer*, it will access the paper content and send it to the backend server. We utilize the MixQG [57], a state-of-the-art question generation model that is fine-tuned on nine question-answering datasets using T5 pre-trained models [69], to generate the comprehension questions for the reading section. MixQG takes the answer to the intended question and the text about the answer as input and outputs a question. In our practice, we use an extractive summarization model [54] to process each paragraph of the reading section. The output summary and corresponding paragraph serve as the input to MixQG. We present at most five questions [5] in this stage to avoid overwhelming the users. Table 2 provides examples of source texts and comprehension questions generated by *CriTrainer*.

[4] We encode the submitted and referred answers into vectors using bert-base-cased [74] and measure their cosine similarity. 0.8-1: very good, 0.6-0.8: good, 0.4-0.6:moderate, 0.2-0.4:not too bad, and 0-0.2: poor.

[5] The first five questions if there are more than five paragraphs in this paper section.

Table 2: Example source text and generated comprehension questions in the QR stage.

Source text	Comprehension question
To evaluate the effectiveness of our system, we performed a study with six blind users. We observed that the proposed system enabled participants to sense the line movement and to stand in line effectively by themselves. Moreover, participants felt more confident and comfortable to stand in line by themselves after the experiment. Based on our findings and user feedback, we discuss requirements to make the system practical and applicable for other use cases.	How did the researchers evaluate the system?
We developed a smartphone-based system that can detect surrounding people and inform about the distance to the closest person. This system intends to complement blind users' orientation and mobility skills in a social context, allowing them to stand in lines by themselves.	What is the main purpose of the system?

4.3 AC-1 stage

In this stage, users should finish their first training task by drafting a summary of the reading section based on their understanding.

4.3.1 Interaction and interface design. When users proceed to this stage from the QR stage, they can see the task description and general guidance that direct them to identify the key points of the selected section and draft a summary (Figure 3 b1). Users can write down and submit their thoughts (b2), after which they are encouraged to improve their drafts with a machine-generated summary from *CriTrainer* for reference (b3). To help users locate the key points of the drafted and referred summaries, *CriTrainer* will also underline or highlight the sentences in the original paper content that are similar to the drafted or referred summary. In our later experiment, we require participants to finish training tasks about the Introduction section, and the referred summary only shows up after they draft and submit some meaningful content about it (*i.e.,* over 90% meaningful words). This would encourage them to spend the necessary effort on this summarization task (DR2). They can click "Last Stage" to check their QR records or "Next Stage" to proceed to the second critical paper reading training task in the AC-2 stage.

4.3.2 Backend model. To provide a referred summary to learners, *CriTrainer* adopts a state-of-the-art abstractive summarization model BRIO [52], which is trained on three text summarization datasets. It's demonstrated that it can generate fluent, semantically, and syntactically correct summaries [52]. In our practice, *CriTrainer* inputs all paragraphs of the selected section to BRIO and outputs a referred summary to learners. To support highlighted key points in the paper that contributes to the referred summary, we encode each sentence in the selected section and the summary into vectors using bert-base-cased [74] and compute their cosine similarities. The outcome of this process is that for each sentence in the summary, *CriTrainer* can highlight the most similar sentence in the paper, which could serve as the key points that may help users learn how to capture the main ideas of the paper.

4.4 AC-2 stage

In this stage, users should finish their second training task by raising critical thinking questions on the reading section.

4.4.1 Interaction and interface design. When users proceed to this stage from the AC-1 stage, they can first view the task description that requires them to ask critical thinking questions (Figure 3 c1). Before writing down and submitting their critical thoughts, users can check machine-detected keywords about which they may raise relevant questions (c2). To avoid information overload, *CriTrainer* only shows three keywords at a time, and users can click the arrow buttons to check more. Users can select a detected keyword, which will be highlighted in the paper content for easy navigation. *CriTrainer* also offer a "Not needed. I have a thought" option to remind users that they can also ask questions in any form they want. Note that only when users have submitted two critical thinking questions, the question templates (c2) will show up. This interaction design aims to first encourage learners to spend necessary effort and then encourage them to raise more questions with reference, as suggested by our design requirement DR3. Users can click the "Hint" to get the possible aspects for thinking based on the chosen keywords (*e.g.,* Motivation, Novelty, Method). Each time users submit a question, *CriTrainer* will record it (c3, c4) and encourage users to try to raise more with a fixed generated critical thinking question for reference. Users can click "Last Stage" to check their drafted summary again or "Next Section" to proceed to read and get trained on the next paper section.

4.4.2 Backend model. To provide critical thinking question templates, detected keywords, and example text-specific questions, we propose a template-based question generation model. We choose the template-based approach as it does not require a collection of a large context-question labeled dataset that is needed in deep-learning-based methods.

Template Design. We adopt the open-source section-level critical thinking questions from CReBot [65] as the sources of the templates. These questions are compiled by experienced Human-Computer Interaction (HCI) researchers and organized based on "what, how, why, how well", the questioning aspect (*e.g.,* clarity, replicability), and common sections in the critical thinking questions in HCI papers. They are used in our baseline tool in the later experiment. However, these questions are not adapted to the specific text in the paper content and could not satisfy the DR3 for our training tool. In our practice, we select the "why" and "how well" questions ($N = 182$) from CReBot for the template design as they are criticism questions [65]. We get inspired from [13] which suggests that the intention of each sentence in academic papers can be categorized into background, objective, method, result, and others. We examine the selected questions from CReBot and see if they can be converted into templates that ask about those

Table 3: Example text-specific generated critical thinking questions in the AC-2 stage based on our template-based approach. <A> Method, <B> Background, and <C> Result stand for the category of masked keywords.

Section	Template	Critical thinking question
Abstract	Is <B> relevant to the problem I am looking for?	Is **navigating blind pedestrains** relevant to the problem I am looking for?
Introduction	Is the method of this paper, *e.g.,* <A>, novel?	Is the method of this paper, *e.g.,* **infrared depth sensor**, novel?
	Is the research about <B> important?	Is the research about **completely man-made environment** important?
Related work	Do the authors' comments on <B> make sense? Why?	Do the authors' comments on **computer vision-based assistive technologies** make sense? Why?
	Is prior work about <B> adequately reviewed, *e.g.,* in terms of methods and contributions?	Is prior work about **assisting blind people** adequately reviewed, *e.g.,* in terms of methods and contributions?
Method	Whether the descriptions of methods about <A> are clearly presented?	Whether the descriptions of methods about **robotic emotional expressions** are clearly presented?
	Do the authors clarify enough details for me to understand their methods, *e.g.,* about <A>?	Do the authors clarify enough details for me to understand their methods, *e.g.,* about **the greeting opening gesture**?
Discussion	Do the authors point out any potential concerns of their findings about <C>? Can I solve them?	Do the authors point out any potential concerns of their findings about **directly mimicking human**? Can I solve them?
Conclusion	Whether this paper about <A> is useful? Why?	Whether this paper about **built-in RGB camera** is useful? Why?

categories. The results turn out that few questions are about "objective" and "others". For the remaining three sentence categories, our critical thinking question template fills in the keywords of the corresponding questions. Specifically, for each selected question $[w_1, w_2, ..., w_n]$, where w_i denotes a word in the question, its template looks like $[w_1, ..., w_i, < A > / < B > / < C >, w_{i+1}, ..., w_n]$, where $< A > / < B > / < C >$ are keywords about:

<A> Method, which describes the study design;

<B> Background, which motivates the reader to examine the research by setting the general field or topic and stating the shortcomings of the previous study;

<C> Result, which states the major findings and advances the significance of the research by either drawing conclusions or offering recommendations.

Template Development. Three authors of our research team first independently conduct a coding on the twenty randomly sampled questions to derive potential templates. They then meet and discuss with an assistant professor in the HCI domain to refine their templates. For example, the template for the original general question "Why is this problem important?" is "Whether is the <B> important? Why?". We further adopt words like "*e.g.,* " in the templates to mitigate the negative impact of the potentially misdetected keywords. After discussion with the professor, they then code the rest questions and regularly discuss the derived templates for several rounds. In the end, we have 116 templates out of the 182 questions.

Keyword Filling. To fill the templates with proper keywords, we first adopt a pre-trained sequential sentence classification model [21] to classify each sentence of the paper. This model achieves an F1-score of 80.74%, 89.63%, and 80.40% on classifying sentences about method, background, and result respectively. For each sentence that falls into the method, background, or result category, we use a light-weight unsupervised automatic keyword extraction method YAKE [14] to identify its keywords. We tag the detected keywords <A>, <B>, or <C> based on the classified sentence. We fill these keywords in proper templates in the same paper section to generate text-specific questions.

Validation of the Templates and Generated Questions. We apply our template-based question generation approach to four HCI papers published in a top venue CHI, short for The ACM CHI Conference on Human Factors in Computing Systems. These four papers have two common co-authors, who are invited to rate the quality of generated questions and corresponding templates. With the information about the corresponding section, template, and filled keywords, the two co-authors rate each question regarding its understandability, relevance, and criticalness for helping users learn how to raise critical thoughts [45]. For each template, we average the scores of multiple questions that apply the template as its final score. On average, the mean scores for the developed templates are 4.32 (SD = 0.42) for understandability, 3.93 (SD = 1.06) for relevance, and 4.13 (SD = 0.53) for criticalness. We select the templates with scores above 3.5 in all aspects for *CriTrainer* to maintain the quality of the templates. After this process, we have 39 templates asking about "Method" keywords, 13 about "Background", and 20 about "Result". Table 3 presents some examples of our templates along with their generated critical thinking questions.

5 EXPERIMENT

To investigate *CriTrainer*'s impact on novice researchers' critical paper reading training process and outcome, we conduct a mixed-design (tool as between-subjects, time as within-subjects factor) experiment with 24 participants. Our research questions (RQ) are:

RQ1: Compared to the baseline tool that provides general guidance, how would *CriTrainer* help novices improve their critical paper reading skills in (i) summarizing the paper content, and (ii) raising relevant critical thinking questions after training sessions?

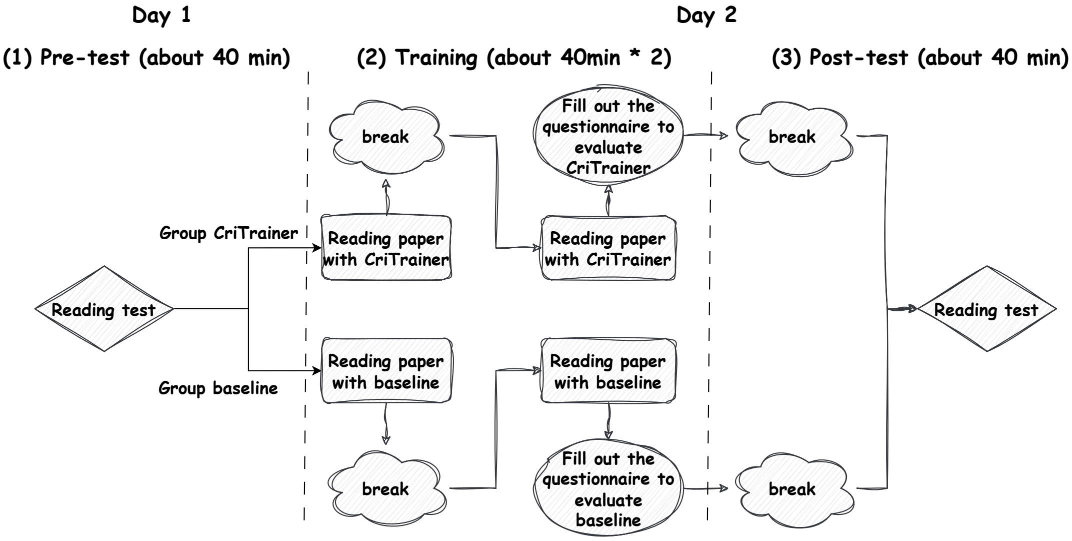

Day 1
Day 2
(1) Pre-test (about 40 min)
(2) Training (about 40min * 2)
(3) Post-test (about 40 min)
Reading test
Group CriTrainer
Group baseline
Reading paper with CriTrainer
break
Reading paper with CriTrainer
Fill out the questionnaire to evaluate CriTrainer
Reading paper with baseline
break
Reading paper with baseline
Fill out the questionnaire to evaluate baseline
break
break
Reading test

Figure 4: Experiment design and procedure. (1) We assess participants' ability in summarzing paper content and raising critical thinking questions in the pre-test. (2) Participants read two papers with either *CriTrainer* or the baseline tool in two training sessions. (3) We assess participants' critical thinking ability in the post-test in a similar way as pre-test.

RQ2: Compared to the baseline tool that provides general guidance, how would *CriTrainer* affect the users' (i) behaviors, and (ii) perceived engagement and workload during the training sessions?

RQ3: Compared to the baseline tool that provides general guidance, how would the novices perceive *CriTrainer* for training critical paper reading skills?

5.1 Baseline

To evaluate the value of our proposed text-specific and interaction features, we choose a baseline tool that does not have these features but has general guidance in the QR2AC process. It sits in the right part of the browser as an add-on with a similar interface as *CriTrainer* (Figure 3). It also structures the training process using the QR2AC structure. Its differences from *CriTrainer* lie in the lack of adaptive text-specific content that we propose. Specifically, the baseline tool does not have the generated comprehension questions and records of user responses in the QR stage (Figure 3 a2 and a3), the highlighted key points and referred summary in the AC-1 stage (b3), and the detected keywords, question templates, and example text-specific questions in the AC-2 stage (c2). Instead, it offers general guidance like those that appeared in b1 and c1. In the QR stage, the baseline tool only informs the users to read through the selected section with their questions, if any. In the AC-1 and AC-2 stages, it offers the same text area for writing down their summary and critical thinking questions but does not have the "Hint" button. In the AC-2 stage, it also provides the section-level general critical thinking questions where the *CriTrainer*'s presented templates come from. In all, the baseline tool simulates how users can learn and exercise critical paper reading skills with general guidance.

5.2 Participants

We recruited 24 undergraduate students through an advertisement posted on a social network in a Chinese university. All participants speak English as their second language, and they are qualified in reading and writing in English with the CET-6 certificate, a national test indicating students' English level of non-English major postgraduates in China. They mainly major in science and engineering backgrounds like Artificial Intelligence and Information and Computing Sciences. Participants generally had little experience in reading academic papers (M = 3.29, SD = 1.10; 1 - No experience at all, 7 - A lot of experience) and self-rated incompetent in critical paper reading (M = 2.75, SD = 0.88; 1/7 - Extremely incompetent/-competent). We randomly assigned participants into a treatment group using *CriTrainer* (P1-12, 7 **F**emales, 5 **M**ales; age: M = 20.58, SD = 0.49) and a controlled group using the baseline tool (P13-24; 9 **F**emales, 3 **M**ales; age: M = 21.08, SD = 0.86).

5.3 Task and Procedure

We conducted the experiment offline. Figure 4 illustrates the procedure for each participant. Since our templates of the generated questions are based on the question pool for reading HCI papers from [65], we sample the training and testing materials from papers published in the HCI venue. Following [65], we choose the CHI late-breaking works (LBWs), which "provide the CHI community with an opportunity to present new and exciting contributions that showcase innovative technologies, extend prior research conversations, detail short self-contained studies, or provide provocations for new work and ideas to emerge" [1]. CHI LBWs are short but complete, which could be suitable for training purposes in a controlled lab study. We sampled four papers with similar word counts (*Mean* = 4251, SD = 77) from CHI2019 and CHI2020 LBWs with few technical details. We left the two LBWs with virtual reality elements as pre- and post-test materials. We used the other two as the training materials; one is about a smartphone-based system for blind people, and the other is about the use of sound for urban runners. Based on a pilot study with two novices, we suggested 40 minutes for the participants to read each paper and informed them that they can spend less or more time if they want. We instructed participants that they need to read the paper's Introduction section in detail and complete the two training tasks on this section and they can read other parts as they want. The focus is on the Introduction section because it is the key part of any academic paper and could be the most valuable and effective content for training critical paper reading skills. All participants read the same four papers in the same order in the procedure, which is described below.

Pre-test. One day before the training sessions, participants took a pre-test on the same given paper for about 40 minutes. They need to write down a summary of the paper's Introduction section and raise relevant critical thinking questions. The purpose of the pre-test is to check participants' critical paper reading skills prior to the training sections, which are used to validate whether they get improved after training.

Training: On the experiment day, participants first viewed a demo from the experimenter on how to use their assigned training tool (*i.e., CriTrainer* or baseline tool) and then started the training sessions. They needed to read two papers with a 10-minute break in between. When reading each paper, we required them to experience a complete QR2AC training process in the Introduction section and recommended that they can use the tool in other sections as they wish. Except for that, we do not restrict whether, when, and how they use the tool. After the two training sessions, participants rated their perceived engagement in the process, workload, and perceptions about the tools. We further asked their opinions on whether and how they learned summarization and question-asking skills with the training tool and their suggestions for improvement.

Post-test: After a 10-minute break from the training session, participants conducted a post-test without any support from the training tools, in which they read another paper, wrote down a summary on its Introduction section, and raise relevant critical thinking questions. We compensated each participant with about $22 USD for around three hours spent in the full experiment.

5.4 Measurement

RQ1. Training outcome. We measure participants' improvements in their critical paper reading skills after the training sessions from two aspects. **i) Summarizing paper content.** In the pre-test and post-test, we measure participants' ability of *S*ummarizing paper content by scoring their written summaries using three items adapted from [42]: *Understanding(S)*: How much does the student seem to understand the main ideas of this section? *Conciseness(S)*: How concise is the written summary? *Overall(S)*: Overall, how good the summary of this section is? **ii) Raising critical thinking questions.** Following [45], for each *Q*uestion participants raise in pre-test and post-test, we score its *Understandability(Q)* (how easy can you understand this question), *Relevance(Q)* (how relevant the question is to this paper section), and *Criticalness(Q)* (how critical do you think this question is). All items are rated on a 5-point Likert Scale (1/5 - very bad/good). We invited the two senior HCI researchers, who co-author the four CHI papers used in validating question templates in subsection 4.4.2, to score the summaries and questions in pre- and post-tests. They are blind to the group information. We average their scores as the final score for each item.

RQ2. Training process. i) Behaviors. To inspect participants' behavior during the training process with *CriTrainer*/Baseline, we log the completion time of each training session, the length of the user-written summary, and the number of user-raised critical thinking questions in each training session. **ii) Perceived engagement and task workload.** We measure participants' perceived engagement in the training process from six aspects adapted from [24, 61], *i.e.,* Concentration, Sense of Ecstasy, Doability, Sense of Serenity, Timelessness Feeling, and Intrinsic Motivation. We also measure their perceived task workload using metrics from NASA Task Load Index [23] regarding *i.e.,* Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration.

RQ3. Perceptions towards the tool. For each tool, we adopt the technology acceptance model from [65, 81, 84] to measure its *usefulness* (four items; Cronbach's α = 0.919); *easy to use* (two items; Cronbach's α = 0.701); and *intention to use* (two items; Cronbach's α = 0.901). We average the scores of multiple items as the final score for each aspect. Besides, we ask for their opinions and suggestions on the tool regarding critical paper reading training support.

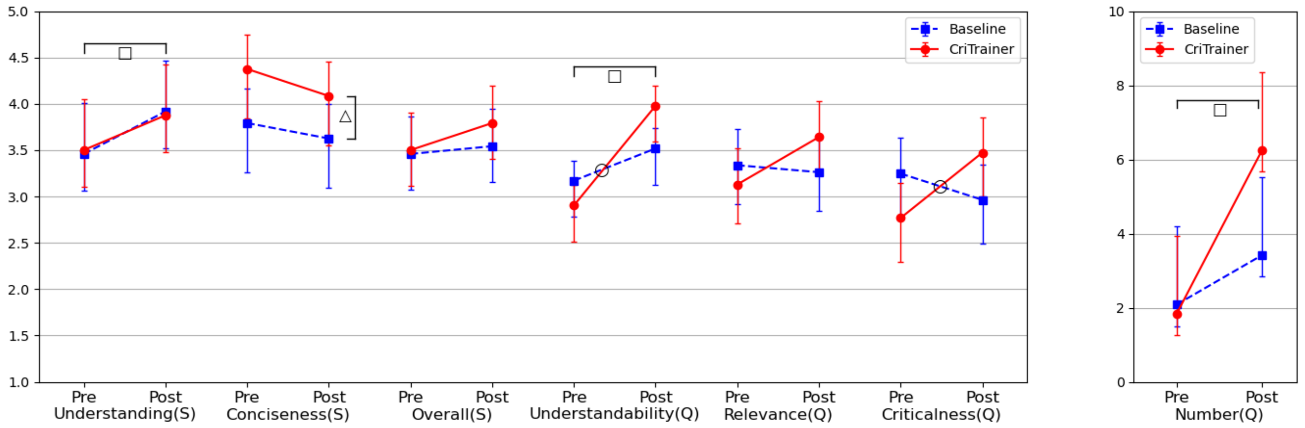
6 ANALYSES AND RESULTS

To evaluate the changes in the ability of the participants in critical paper reading, we conduct a two-way mixed ANOVA to compare the performance of participants in each group (as between-subjects factors) during the pre-test and post-test (as within-subjects). We also perform the Mann-Whitney U test [53] to compare the difference in critical paper reading performance separately in pre- and post-test between the two groups in RQ1. This test confirms that our participants do not have significant differences in their critical paper reading ability before the training sessions. Similarly, for the rated items in RQ2 and RQ3, we use the Mann-Whitney U test [53] to compare the ratings between two user groups. The Mann-Whitney U is a non-parametric test commonly used to compare differences between independent conditions (*e.g.,* in HCI studies (*e.g.,* [16, 19, 39, 80])) especially when the data normality is violated, as confirmed in our cases. In all U tests, we adopt the Bonferroni correction by setting the significance level at 0.05 divided by the number of dimensions within each measurement. For the participants' comments and suggestions on the training tool, we perform a thematic analysis [11]. Two authors first code all the qualitative 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 the pros and cons of each tool (Table 5), which are incorporated into the result presentation below.

6.1 RQ1. Training Outcome

Figure 5 shows participants' performance in pre- and post-tests, which reflect their skills in critical paper reading.

i) Summarizing paper content. Overall, participants have a significant improvement in the Understanding(S) ($p < .05$) of their



(a) The metrics are rated in a 5-point Likert Scale; 1/5: low/high.

(b) The y-axis value represents the number of raised questions.

Figure 5: (RQ1 results) The means and 95% confidence intervals of the expert-rated scores about participants' (a) drafted Summaries and (b) raised critical thinking Questions in the pre- and post-tests; □: p < .05 for time factor (pre- vs. post-test); △: p < .05 for group factor (*CriTrainer* vs. Baseline); ○: p < .05 for interaction between the time and group factors.

written summary with either tool after the training sessions. There are no significant changes in the summaries' conciseness and overall quality between pre- and post-test. As for the group factor, we observe a significant difference in the Conciseness(S) (p < .05) of the summary written by participants using *CriTrainer* compared to those using the baseline tool. There are no significant interaction effects between the time and group factors. In general, both tools could improve users' ability in summarizing paper content regarding showing their understanding on the paper section and overall quality. However, they could perform worse in the summary's conciseness, which could be due to that we did not explicitly require them to pay attention to the length of the summary in the training sessions. Six participants using *CriTrainer* especially value its text-specific guidance in the summarization tasks. "*With the highlighted key points in the paper content and referred summary in CriTrainer, I learned how to grasp the main idea of the section and rephrase it*" (P1, **F**emale, age: 21). "*The referred summary encouraged me to think deeply about the logic among the paragraphs*" (P8, **M**ale, age: 20).

ii) Raising critical thinking questions. After the training sessions, participants in both groups have significant improvements in their skills in raising critical thinking questions regarding the numbers (p < .01) and understandability (p < .01). As for the group factor, there are no significant differences on all measured items of raised questions. However, we have interesting findings on the interaction effects on these items between the time and group factors. Specifically, the understandability and criticalness of questions raised by *CriTrainer* users improve significantly more than Baseline users after training (p < .01). Participants in the *CriTrainer* group also gain improvement on the relevance and criticalness of the raised questions, while those with the baseline tool have a decreasing performance in these two metrics (.05 < p < .1). The follow-up Mann-Whitney U tests on the post-test performance further reveals that participants in *CriTrainer* group raised significantly more (U = 33.5, p < .05/4) critical thinking questions than those in the Baseline group, and these questions are significantly more understandable (U = 29.5, p < .05/4) and critical (U = 37.5, p < .05/4). These results suggest that **compared to the baseline tool, *CriTrainer* led to a better training outcome in terms of their ability in raising more understandable, relevant, and critical questions on the paper content.** Ten participants using *CriTrainer* mention how its detected keywords and templates foster their critical thinking skills in the post-study interviews. "*By referring to the keywords and templates, I gained a deeper understanding of the paper and was able to express my critical thoughts*" (P5, M, age: 21). "*With CriTrainer, I learned to think of the paper's background, method, and results from different aspects*" (P9, F, age: 20). "*I know better how to raise more relevant and reasonable questions after training.*" (P10, M, age: 21). Five participants using the baseline tool desire more text-specific guidance. "*It would be better if it can provide referred critical thinking questions based on the specific paper*" (P13, F, age: 21).

6.2 RQ2. Training Process

i) Behaviors. In general, participants spent significantly more time (U = 37.5, p < .05) in the training process with *CriTrainer* (*M* = 34.58 min, *SD* = 5.25 min) than they did with the baseline tool (*M* = 29.08 min, *SD* = 6.27 min). Meanwhile, we compare the total number of words in user-written summaries during the training process (*CriTrainer*: *M* = 71.58, *SD* = 15.99; Baseline: *M* = 91.88, *SD* = 29.96), but find no significant difference (U = 99.5, p = .12). Furthermore, the number of user-raised critical questions in the *CriTrainer* group (*M* = 7.96, *SD* = 4.74) was significantly greater than the baseline group (*M* = 3.46, *SD* = 1.05) during the training process (U = 21.5, p < .01). This indicates *CriTrainer* motivates participants to devote more effort to raising critical thinking questions, which can be viewed as practicing critical thinking skills. "*The CriTrainer's keywords helped*

Table 4: Users' perceived engagement and workload (RQ2ii) in the training process as well as their perceptions (RQ3) towards the *CriTrainer* or Baseline tool; The significance levels for Engagement, Workload, and Acceptance are .05/6, .05/6, and .05/3 respectively with Bonferroni correction.

Category	Factor	CriTrainer	Baseline	Statistics		
		Mean/S.D.	Mean/S.D.	U	p	Sig.
RQ2 ii) Engagement	Concentration	6.00/0.85	5.92/1.00	70	0.903	
	Sense of Ecstasy	6.42/0.90	5.33/1.37	38	0.038	
	Doability	5.50/1.09	5.58/1.08	67.5	0.783	
	Sense of Serenity	5.58/1.24	5.25/0.62	67	0.759	
	Timelessness Feeling	6.25/0.97	6.00/1.04	59	0.408	
	Intrinsic Motivation	5.58/1.51	5.78/0.97	71	0.952	
RQ2 ii) Workload	Mental Demand	4.50/1.09	5.08/1.00	46	0.118	
	Physical Demand	2.00/0.74	2.92/0.90	31.5	0.014	
	Temporal Demand	2.83/1.11	3.58/0.99	48.5	0.152	
	Performance	4.92/0.67	4.92/0.51	71.5	0.972	
	Effort	4.58/1.17	4.83/0.83	65	0.668	
	Frustration	1.83/0.72	2.08/1.16	67	0.756	
RQ3 Acceptance	Usefulness	6.17/0.72	5.67/0.83	47.5	0.151	
	Easy to Use	5.79/0.96	5.75/0.87	69	0.860	
	Intention to use	5.75/0.92	5.12/1.13	48	0.159	

me think of how to raise questions and its feedback with text-specific questions encouraged me to raise more" (P1, F, age: 21).

ii) Perceived engagement and workload. As for the six items that measure perceived engagement in the training process, there is no significant difference between *CriTrainer* and the baseline tool after Bonferroni correction, which is shown in Table 4. Similarly, we do not observe significant differences between the two tools in the metrics about task workload with Bonferroni correction. We also observe that the workload of the training task is generally lower in *CriTrainer* group regarding mental and temporal demand, effort, and frustration. Ten participants with *CriTrainer* especially favor its highlighting features that help them locate important information more easily, which might reduce their task workload. "*CriTrainer highlights and summarizes the main idea, which makes it easier to practice how to summarize the paper content*" (P12, F, age: 21).

6.3 RQ3. Perceptions Towards the Tool

Table 4 shows the participants' ratings on their technology acceptance of *CriTrainer* and Baseline. We do not observe significant differences in terms of usefulness, ease of use, and intention to use (p > .05). As for their comments on the tools (Table 5), participants with *CriTrainer* generally feel that it can help them learn how to think critically (N = 10), how to identify the key points and summarize them (6). They also like its features on highlighting paper content (9), providing hints and feedback (6), and offering comprehension questions (3). "*After training with CriTrainer, I feel that my perspective on thinking has become more diverse. Additionally, I have started to read and think with a focus*" (P6, F, age: 22) However, *CriTrainer* could be inflexible to use (3) and its templates may be not diverse in some sections (3). Regarding the baseline tool, participants favor its general guidance (6), easy and clear interaction (5), and simple interface (2). "*I like its simple interface and clear interaction*" (P19, M, age: 20). Nevertheless, it can not provide adaptive assistance (5) and could be not enjoyable (3).

7 DISCUSSION

In this work, we propose *CriTrainer* with the goal to help users acquire critical paper reading skills in terms of summarizing paper content and raising relevant critical thinking questions. Our mixed-design study with 24 novice researchers suggested that the amount of improvement in summarization performance with *CriTrainer* is not significantly larger than that with a baseline tool that offers general summarization guidance. One possible reason could be that participants with the baseline tool spent more effort in summarizing the paper content during the AC-1 stage, as suggested by the number of words in their submitted summaries in training sessions; Baseline: M = 91.88, *CriTrainer*: 71.58, p = .12. While participants favor *CriTrainer*'s highlighted key points and referred summary, they may spend less effort in thinking about how to capture them and write them down. This result implies that the AI-powered training tool should balance the amount of support and users' effort during their learning tasks, which we discuss in subsection 7.1.

As for the ability to raise relevant critical thinking questions on the paper, we found that compared to the participants using the baseline tool which provides general section-level questions, those with *CriTrainer* have significantly more improvement in raising understandable, relevant, and critical questions. As commented by our participants, this improvement can be accounted for *CriTrainer*'s detected keywords, question templates, and corresponding text-specific questions. Therefore, our findings support the previous work's [2, 15, 27, 32, 63] implication on the importance of specific guidances in developing critical thinking skills. Our proposed template-based question generation approach can offer such specific guidance to individual learners on a large scale. We acknowledge that the current question templates are derived from a question pool [65] compiled for HCI paper domain and may not be directly

Table 5: (RQ3 results) Summary of users' comments about pros and cons of *CriTrainer*/Baseline

	Pros (the number of participants who mention)	Cons
Critrainer	Facilitate critical thinking (10); Help identify the key points and summarize them (4); Highlight paper content (9); Provide hints and feedback (5); Provided comprehension questions (3)	Inflexible to use (3); Question templates not diverse enough (3)
Baseline	General guidance (6); Easy and clear Interaction (5); Simple interface (2)	Not provide adaptive assistance (5); Lack of fun to use (3)

applied to other research fields. To train the ability to raise critical thinking questions on papers in other domains, we open-source our compiled templates attached in the supplementary materials of this submission and encourage future researchers to further customize them.

Our ideas of *CriTrainer* can also be applied to train users' critical thinking skills in other scenarios such as reading news articles and social media posts. For instance, when reading news articles, a critical thinking training tool like *CriTrainer* can derive question templates from information assessment guidelines [41] to help users learn how to identify the misinformation in the articles. It is also promising to extend the usage scenarios of *CriTrainer* to assist researchers to self-check their paper drafts, help reviewers to identify the submissions' strengths and weaknesses, and facilitate lecturers to prepare critical thinking teaching materials.

7.1 Design Considerations

From our study findings, we derive several design considerations for critical paper reading training tools.

7.1.1 Offer adaptive and interpretable suggestions on how to improve the drafted summary. *CriTrainer* currently provides a referred summary and highlighted key points after users submit a draft in the AC-1 stage. While this information can help participants reflect on their summaries and locate where the key sentences are, they expect more adaptive and interpretable suggestions. First, *CriTrainer* could offer a polished version of their drafted summaries, *e.g.,* using the latest ChatGPT technology [7], which would help them learn how to organize and rephrase the main ideas concisely. Second, *CriTrainer* should explicitly reveal the relationship among the highlighted key points to make them more interpretable. For example, it can identify the fine-grain structure (*e.g.,* previous work, gap, challenge, proposed work, contribution) of the reading section and add the related tags to the highlighted sentence.

7.1.2 Increase the diversity of generated critical thinking questions. While participants generally favor our template-based critical thinking questions, three users of *CriTrainer* suggest that these questions should be more diverse (Table 5). There are two possible ways to address this concern. First, to improve linguistic diversity, *CriTrainer* could leverage pre-trained language models (*e.g.,* GTP4 [12]) to rephrase the templates and generated questions. Second, to diversify the types of questions, we could integrate the template-based and data-driven question generation approaches. We do not suggest a supervised data-driven method that would require a large labeled context-question dataset. Instead, we could collect the existing questions raised in the open reviews of academic papers, group them, and extract templates from the clusters. This could enlarge *CriTrainer*'s template pool for helping users learn how to raise critical thoughts.

7.1.3 Encourage necessary effort spent in the learning tasks. As recommended by the theory of scaffolding [33], *CriTrainer* have adopted a few features to encourage users to first spend effort and then receive support and feedback. For example, *CriTrainer* only shows referred summary if 90% of the submitted words are meaningful, *i.e.,* they are not random letters. The results turn out that during the training sessions, participants did try to raise more questions with *CriTrainer* in the AC-2 stage but tended to write fewer words in the AC-1 stage. Apart from detecting whether users are tricking the training tool, we suggest that future *CriTrainer* should offer hints and feedback step by step rather than presenting the referred summary at once. For example, after the users submit a draft, it can detect the missing key points and prompt a hint on them first. Besides, *CriTrainer* could incorporate a gamification feature (*e.g.,* competition with others) that would encourage users to work hard on the training tasks for better performance.

7.1.4 Balance the tradeoff between flexibility and sufficient support. We do not observe significant differences regarding the user experience between *CriTrainer* and the baseline tool. This result could be accounted for the pros and cons of *CriTrainer* (Table 5). For example, while our proposed features can facilitate critical thinking and comprehension, they also increase the complexity of the interface, making it inflexible to use for some users. This urges further optimizing our interface design to improve user experience. For instance, we could enable customization of the layout, e.g., allowing automatic and manual (un)folding of certain information [65], to simplify the interface based on users' interest.

7.2 Limitations and Future Work

Our work has several limitations. First, as we derive question templates from a pool compiled for HCI papers and primarily target novices in research, we mainly include undergraduate students with science and engineering backgrounds in the experiment. We encourage future work to validate *CriTrainer*'s effectiveness by involving more participants in diverse majors and research expertise. Second, we measure the improvement in participants' critical thinking skills after two training sessions by comparing their performances in summarization and question-asking tasks in pre- and post-test. However, learning critical paper reading is a life-long process for researchers, and the skill set involves other abilities like answering critical thinking questions. Future work could explore how to enhance *CriTrainer* (*e.g.,* with a question-answering module) to support other critical thinking training tasks and evaluate it with a long-term study. Third, we measure novices' critical thinking ability by the number and quality of their raised critical questions [40, 65]. To step forward, we should measure their critical thoughts on these questions, which is an advanced ability of experienced readers. Future work could capture this aspect by rating participants' responses to the pre-compiled critical thinking questions. Fourth, we treat the proposed features as a whole when evaluating the impact of *CriTrainer*. Future work should evaluate it in ablation studies, *e.g.,* the *CriTrainer* tools with and without the hints in the QR2AC process, to explore the value of each feature. Fifth, while the primary target users of *CriTrainer* are novices of all research domains, it is promising that domain experts (*e.g.,* Robotics) could also learn with *CriTrainer* in their unfamiliar areas (*e.g.,* Human-Computer Interaction), which is a direction for future work. Sixth, our experiment treats the training tool as the between-subjects factor to avoid the learning effect on the measured learning outcome. To obtain more robust results, future work could increase the sample size or explore the study design of the tool as the within-subjects factor that mitigates the impact of individual differences. Additionally, all participants in our studies speak English as their second language. People with different cultural backgrounds may have different needs for critical reading support. It would be useful to research how the choice of demographics might have influenced the study results for future work. Lastly, similar to [85], we implement *CriTrainer* as a Google Docs add-on supported by various browsers (*e.g.,* Chrome, Firefox, Edge) to allow easy access and manipulation of paper content. However, some users may be more used to reading academic papers using pdf readers. Future work could seek ways to embed *CriTrainer* into the popular pdf readers if they offer APIs to access and modify the paper content (*e.g.,* color).

8 CONCLUSION

In this paper, we design and build an adaptive training tool *CriTrainer* to help novice researchers develop critical paper reading skills in summarizing paper content and raising critical thinking questions. *CriTrainer* offers a generated summary and highlights the relevant key points in the paper content after users submit a drafted one. It provides text-specific critical thinking questions generated by our proposed template-based approach to help users learn how to raise questions. We compared *CriTrainer* to Baseline which provides general guidance through a mixed-design study with 24 participants. The results show that *CriTrainer* can better improve participants' critical paper reading skills in raising understandable, relevant, and critical questions after the training sessions. Our work offers insights and design considerations for building intelligent tools to train critical thinking skills.

ACKNOWLEDGMENTS

This work is supported by the Young Scientists Fund of the National Natural Science Foundation of China with Grant No. 62202509 and partially supported by the Research Grants Council of the Hong Kong Special Administrative Region, China under General Research Fund (GRF) with Grant No. 16204420.

REFERENCES

[1] Sigchi A. 2023. Late breaking work - CHI 2021. *Retrieved in March 2023 from https://chi2023.acm.org/for-authors/late-breaking-work/.* (2023).

[2] Philip C Abrami, Robert M Bernard, Evgueni Borokhovski, Anne Wade, Michael A Surkes, Rana Tamim, and Dai Zhang. 2008. Instructional interventions affecting critical thinking skills and dispositions: A stage 1 meta-analysis. *Review of educational research* 78, 4 (2008), 1102–1134.

[3] Laith Abualigah, Mohammad Qassem Bashabsheh, Hamzeh Alabool, and Mohammad Shehab. 2020. Text summarization: a brief review. *Recent Advances in NLP: the case of Arabic language* (2020), 1–15.

[4] Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D Trippe, Juan B Gutierrez, and Krys Kochut. 2017. Text summarization techniques: a brief survey. *arXiv preprint arXiv:1707.02268* (2017).

[5] Ayham Alomari, Norisma Idris, Aznul Qalid Md Sabri, and Izzat Alsmadi. 2022. Deep reinforcement and transfer learning for abstractive text summarization: A review. *Computer Speech & Language* 71 (2022), 101276.

[6] Tal August, Lucy Lu Wang, Jonathan Bragg, Marti A Hearst, Andrew Head, and Kyle Lo. 2022. Paper plain: Making medical research papers approachable to healthcare consumers with natural language processing. *ACM Transactions on Computer-Human Interaction* (2022).

[7] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023* (2023).

[8] Sheri Berkeley and Paul J Riccomini. 2013. QRAC-the-code: A comprehension monitoring strategy for middle school social studies textbooks. *Journal of Learning Disabilities* 46, 2 (2013), 154–165.

[9] Marilyn Binkley, Ola Erstad, Joan Herman, Senta Raizen, Martin Ripley, May Miller-Ricci, and Mike Rumble. 2012. Defining twenty-first century skills. In *Assessment and teaching of 21st century skills*. Springer, 17–66.

[10] Benjamin Samuel Bloom. 1956. Taxonomy of educational objectives: The classification of educational goals. (1956).

[11] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[12] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712* (2023).

[13] Jochen WL Cals and Daniel Kotz. 2013. Effective writing and publishing scientific papers, part II: title and abstract. *Journal of clinical epidemiology* 66, 6 (2013), 585.

[14] Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. 2020. YAKE! Keyword extraction from single documents using multiple local features. *Information Sciences* 509 (2020), 257–289.

[15] Roland Case. 2005. Moving critical thinking to the main stage. *Education Canada* 45, 2 (2005), 45–49.

[16] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–22.

[17] Guanliang Chen, Jie Yang, Claudia Hauff, and Geert-Jan Houben. 2018. LearningQ: a large-scale dataset for educational question generation. In *Twelfth International AAAI Conference on Web and Social Media*.

[18] Lijia Chen, Pingping Chen, and Zhijian Lin. 2020. Artificial intelligence in education: A review. *Ieee Access* 8 (2020), 75264–75278.

[19] Xiang'Anthony' Chen, Chien-Sheng Wu, Tong Niu, Wenhao Liu, and Caiming Xiong. 2022. Marvista: A Human-AI Collaborative Reading Tool. *arXiv preprint arXiv:2207.08401* (2022).

[20] Sumit Chopra, Michael Auli, and Alexander M Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*. 93–98.

[21] Arman Cohan, Iz Beltagy, Daniel King, Bhavana Dalvi, and Daniel S Weld. 2019. Pretrained language models for sequential sentence classification. *arXiv preprint arXiv:1909.04054* (2019).

[22] Trevor Anthony Cohn and Mirella Lapata. 2009. Sentence compression as tree transduction. *Journal of Artificial Intelligence Research* 34 (2009), 637–674.

[23] Lacey Colligan, Henry WW Potts, Chelsea T Finn, and Robert A Sinkin. 2015. Cognitive workload changes for nurses transitioning from a legacy system with paper documentation to a commercial electronic health record. *International journal of medical informatics* 84, 7 (2015), 469–476.

[24] Mihaly Csikszentmihalyi. 1990. *Flow: The Psychology of Optimal Experience*.

[25] Sandra Egege and Salah Kutieleh. 2004. Critical Thinking: Teaching Foreign Notions to Foreign Students. *International Education Journal* 4, 4 (2004), 75–85.

[26] Asmaa Elsaid, Ammar Mohammed, Lamiaa Fattouh, and Mohamed Sakre. 2022. A Comprehensive Review of Arabic Text summarization. *IEEE Access* (2022).

[27] Peter Facione. 1990. Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction (The Delphi Report). (1990).

[28] Patrick J Finn. 1975. A question writing algorithm: The value of explicitly specified processes in test construction. *Journal of reading behavior* 7, 4 (1975), 341–367.

[29] Rosalie Friend. 2002. Summing it up. *The Science Teacher* 69, 4 (2002), 40.

[30] Pierre-Etienne Genest and Guy Lapalme. 2011. Framework for abstractive summarization using text-to-text generation. In *Proceedings of the workshop on monolingual text-to-text generation*. 64–73.

[31] Arthur C Graesser, Jennifer Wiley, Susan R Goldman, Tenaha O'Reilly, Moongee Jeon, and Bethany McDaniel. 2007. SEEK Web tutor: Fostering a critical stance while exploring the causes of volcanic eruption. *Metacognition and Learning* 2, 2 (2007), 89–105.

[32] Diane F Halpern. 1998. Teaching critical thinking for transfer across domains: Disposition, skills, structure training, and metacognitive monitoring. *American psychologist* 53, 4 (1998), 449.

[33] Mariane Hedegaard. 2012. The zone of proximal development as basis for instruction. In *An introduction to Vygotsky*. Routledge, 234–258.

[34] Michael Heilman and Noah A Smith. 2010. Good question! statistical ranking for question generation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. 609–617.

[35] Yuuki Iwasaki, Akihiro Yamashita, Yoko Konno, and Katsushi Matsubayashi. 2019. Japanese abstractive text summarization using BERT. In *2019 International Conference on Technologies and Applications of Artificial Intelligence (TAAI)*. IEEE, 1–5.

[36] Hyeonsu Kang, Joseph Chee Chang, Yongsung Kim, and Aniket Kittur. 2022. Threddy: An Interactive System for Personalized Thread-Based Exploration and Organization of Scientific Literature. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) *(UIST '22)*. Association for Computing Machinery, New York, NY, USA, Article 94, 15 pages. https://doi.org/10.1145/3526113.3545660

[37] Srinivasan Keshav. 2007. How to read a paper. *ACM SIGCOMM Computer Communication Review* 37, 3 (2007), 83–84.

[38] Dae Hyun Kim, Enamul Hoque, Juho Kim, and Maneesh Agrawala. 2018. Facilitating Document Reading by Linking Text and Tables. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology* (Berlin, Germany) *(UIST '18)*. Association for Computing Machinery, New York, NY, USA, 423–434. https://doi.org/10.1145/3242587.3242617

[39] Tae Soo Kim, DaEun Choi, Yoonseo Choi, and Juho Kim. 2022. Stylette: Styling the Web with Natural Language. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 5, 17 pages. https://doi.org/10.1145/3491102.3501931

[40] Alison King. 1995. Designing the instructional process to enhance critical thinking across the curriculum. *Teaching of Psychology* 22, 1 (1995), 13–17.

[41] Bill Kovach and Tom Rosenstiel. 2011. *Blur: How to know what's true in the age of information overload*. Bloomsbury Publishing USA.

[42] Geza Kovacs and Robert C Miller. 2014. Smart subtitles for vocabulary learning. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 853–862.

[43] Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler, and Salam Al-Emari. 2020. A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education* 30 (2020), 121–204.

[44] Andrew Kuznetsov, Joseph Chee Chang, Nathan Hahn, Napol Rachatasumrit, Bradley Breneisen, Julina Coupland, and Aniket Kittur. 2022. Fuse: In-Situ Sensemaking Support in the Browser. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) *(UIST '22)*. Association for Computing Machinery, New York, NY, USA, Article 34, 15 pages. https://doi.org/10.1145/3526113.3545693

[45] Nguyen-Thinh Le and Niels Pinkwart. 2015. Evaluation of a question generation approach using semantic web for supporting argumentation. *Research and practice in technology enhanced learning* 10 (2015), 1–19.

[46] Yuan-Hsuan Lee. 2015. Facilitating critical thinking using the C-QRAC collaboration script: Enhancing science reading literacy in a computer-supported collaborative learning environment. *Computers & Education* 88 (2015), 182–191.

[47] Adam D Lelkes, Vinh Q Tran, and Cong Yu. 2021. Quiz-style question generation for news stories. In *Proceedings of the Web Conference 2021*. 2501–2511.

[48] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461* (2019).

[49] David Lindberg, Fred Popowich, John Nesbit, and Phil Winne. 2013. Generating natural language questions to support learning on-line. In *Proceedings of the 14th European Workshop on Natural Language Generation*. 105–114.

[50] Chengzhong Liu, Zeyu Huang, Dingdong Liu, Shixu Zhou, Zhenhui Peng, and Xiaojuan Ma. 2022. PlanHelper: Supporting Activity Plan Construction with Answer Posts in Community-Based QA Platforms. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 454 (nov 2022), 26 pages. https://doi.org/10.1145/3555555

[51] Ming Liu, Rafael A Calvo, and Vasile Rus. 2012. G-Asks: An intelligent automatic question generation system for academic writing support. *Dialogue & Discourse* 3, 2 (2012), 101–124.

[52] Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022. BRIO: Bringing order to abstractive summarization. *arXiv preprint arXiv:2203.16804* (2022).

[53] Henry B Mann and Donald R Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics* (1947), 50–60.

[54] Derek Miller. 2019. Leveraging BERT for extractive text summarization on lectures. *arXiv preprint arXiv:1906.04165* (2019).

[55] Farida Mohsen, Jiayang Wang, and Kamal Al-Sabahi. 2020. A hierarchical self-attentive neural extractive summarizer via reinforcement learning (HSASRL). *Applied Intelligence* 50, 9 (2020), 2633–2646.

[56] Tim Moore. 2013. Critical thinking: Seven definitions in search of a concept. *Studies in Higher Education* 38, 4 (2013), 506–522.

[57] Lidiya Murakhovs' ka, Chien-Sheng Wu, Tong Niu, Wenhao Liu, and Caiming Xiong. 2021. Mixqg: Neural question generation with mixed answer types. *arXiv preprint arXiv:2110.08175* (2021).

[58] Siya Sadashiv Naik and Manisha Naik Gaonkar. 2017. Extractive text summarization by feature-based sentence extraction using rule-based concept. In *2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*. IEEE, 1364–1368.

[59] Benjamin D Nye, Arthur C Graesser, and Xiangen Hu. 2014. AutoTutor and family: A review of 17 years of natural language tutoring. *International Journal of Artificial Intelligence in Education* 24 (2014), 427–469.

[60] University of Toronto. 2020. Reading critically. *Retrieved in February 2023 from https://advice.writing.utoronto.ca/wp-content/uploads/sites/2/critical-reading.pdf*. (2020).

[61] Heather O'Brien. 2016. Theoretical perspectives on user engagement. *Why engagement matters: Cross-disciplinary perspectives of user engagement in digital media* (2016), 1–26.

[62] Annemarie Sullivan Palincsar and Ann L Brown. 1986. Interactive teaching to promote independent learning from text. *The reading teacher* 39, 8 (1986), 771–777.

[63] Richard Paul and Linda Elder. 1992. Critical thinking: What, why, and how. *New directions for community colleges* 77, 2 (1992), 3–24.

[64] Zhenhui Peng, Qingyu Guo, Ka Wing Tsang, and Xiaojuan Ma. 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*. 1–15.

[65] Zhenhui Peng, Yuzhi Liu, Hanqi Zhou, Zuyu Xu, and Xiaojuan Ma. 2022. CReBot: Exploring interactive question prompts for critical paper reading. *International Journal of Human-Computer Studies* 167 (2022), 102898.

[66] Adam M Persky, Melissa S Medina, and Ashley N Castleberry. 2019. Developing critical thinking skills in pharmacy students. *American journal of pharmaceutical education* 83, 2 (2019).

[67] Antoine Ponsard, Francisco Escalona, and Tamara Munzner. 2016. PaperQuest: A visualization tool to support literature review. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. 2264–2271.

[68] Sally A Radmacher and Elizabeth Latosi-Sawin. 1995. Summary writing: A tool to improve student comprehension and writing in psychology. *Teaching of Psychology* 22, 2 (1995), 113–115.

[69] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research* 21, 1 (2020), 5485–5551.

[70] Sherry Ruan, Liwei Jiang, Justin Xu, Bryce Joe-Kun Tham, Zhengneng Qiu, Yeshuang Zhu, Elizabeth L Murnane, Emma Brunskill, and James A Landay. 2019. Quizbot: A dialogue-based adaptive learning system for factual knowledge. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[71] Vasile Rus, Zhiqiang Cai, and Art Graesser. 2008. Question generation: Example of a multi-year evaluation campaign. *Proc WS on the QGSTEC* (2008).

[72] Vinicius Santos, Anderson Iwazaki, Érica Souza, Katia Felizardo, and Nandamudi Vijaykumar. 2021. CrowdSLR: a tool to support the use of crowdsourcing in systematic literature reviews. In *Proceedings of the XXXV Brazilian Symposium on Software Engineering*. 341–346.

[73] Thomas Scialom, Benjamin Piwowarski, and Jacopo Staiano. 2019. Self-attention architectures for answer-agnostic neural question generation. In *Proceedings of the 57th annual meeting of the Association for Computational Linguistics*. 6027–6032.

[74] Peng Shi and Jimmy Lin. 2019. Simple bert models for relation extraction and semantic role labeling. *arXiv preprint arXiv:1904.05255* (2019).

[75] Ibrahim Abu Shihab. 2011. Reading as critical thinking. *Asian Social Science* 7, 8 (2011), 209.

[76] Harry Shum. 2020. You are how you read. *Retrieved in February 2023 from https://www.youtube.com/watch?v=Du7qLsToW-o.* (2020).

[77] E Elona Sochor. 1959. The nature of critical reading. *Elementary English* 36, 1 (1959), 47–58.

[78] Jennifer Pei-Ling Tan, Simon Yang, Elizabeth Koh, and Christin Jonathan. 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.* 430–434.

[79] Terry Tomasek. 2009. Critical reading: Using reading prompts to promote active engagement with text. *International journal of teaching and learning in higher education* 21, 1 (2009), 127–132.

[80] Prasanna Umar, Anna Squicciarini, and Sarah Rajtmajer. 2019. Detection and analysis of self-disclosure in online news commentaries. In *The World Wide Web Conference.* 3272–3278.

[81] Viswanath Venkatesh and Hillol Bala. 2008. Technology acceptance model 3 and a research agenda on interventions. *Decision sciences* 39, 2 (2008), 273–315.

[82] Mike Wallace and Alison Wray. 2021. *Critical reading and writing for postgraduates.* Sage.

[83] Thiemo Wambsganss, Tobias Kueng, Matthias Soellner, and Jan Marco Leimeister. 2021. ArgueTutor: An adaptive dialog-based learning system for argumentation skills. In *Proceedings of the 2021 CHI conference on human factors in computing systems.* 1–13.

[84] Thiemo Wambsganss, Christina Niklaus, Matthias Cetto, Matthias Söllner, Siegfried Handschuh, and Jan Marco Leimeister. 2020. AL: An adaptive learning support system for argumentation skills. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.* 1–14.

[85] Tao Wang and David Redmiles. 2017. Auditory Overview of Web Pages for Screen Reader Users. In *Adjunct Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology* (Québec City, QC, Canada) *(UIST '17 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 193–195. https://doi.org/10.1145/3131785.3131837

[86] Yun Wang, Dongyu Liu, Huamin Qu, Qiong Luo, and Xiaojuan Ma. 2016. A guided tour of literature review: Facilitating academic paper reading with narrative visualization. In *Proceedings of the 9th International Symposium on Visual Information Communication and Interaction.* 17–24.

[87] Adhika Pramita Widyassari, Supriadi Rustad, Guruh Fajar Shidik, Edi Noersasongko, Abdul Syukur, Affandy Affandy, et al. 2020. Review of automatic text summarization techniques & methods. *Journal of King Saud University-Computer and Information Sciences* (2020).

[88] Blake Williford, Matthew Runyon, Wayne Li, Julie Linsey, and Tracy Hammond. 2020. Exploring the potential of an intelligent tutoring system for sketching fundamentals. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.* 1–13.

[89] Rainer Winkler, Sebastian Hobert, Antti Salovaara, Matthias Söllner, and Jan Marco Leimeister. 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.* 1–14.

[90] Weiran Xu, Chenliang Li, Minghao Lee, and Chi Zhang. 2020. Multi-task learning for abstractive text summarization with key information guide network. *EURASIP Journal on Advances in Signal Processing* 2020, 1 (2020), 1–11.

[91] Jinhong Yu. 2015. Analysis of critical reading strategies and its effect on college English reading. *Theory and Practice in Language Studies* 5, 1 (2015), 134.

[92] Xingdi Yuan, Tong Wang, Caglar Gulcehre, Alessandro Sordoni, Philip Bachman, Sandeep Subramanian, Saizheng Zhang, and Adam Trischler. 2017. Machine comprehension by text-to-text neural question generation. *arXiv preprint arXiv:1705.02012* (2017).

[93] Haoyu Zhang, Jianjun Xu, and Ji Wang. 2019. Pretraining-based natural language generation for text summarization. *arXiv preprint arXiv:1902.09243* (2019).

[94] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning.* PMLR, 11328–11339.

[95] Xinlei Zhang, Takashi Miyaki, and Jun Rekimoto. 2016. WithYou: An Interactive Shadowing Coach with Speech Recognition. In *Adjunct Proceedings of the 29th Annual ACM Symposium on User Interface Software and Technology* (Tokyo, Japan) *(UIST '16 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 61–63. https://doi.org/10.1145/2984751.2985704

[96] Xinlei Zhang, Takashi Miyaki, and Jun Rekimoto. 2020. WithYou: automated adaptive speech tutoring with context-dependent speech recognition. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.* 1–12.