ReadingQuizMaker: A Human-NLP Collaborative System that Supports Instructors to Design High-Quality Reading Quiz Questions

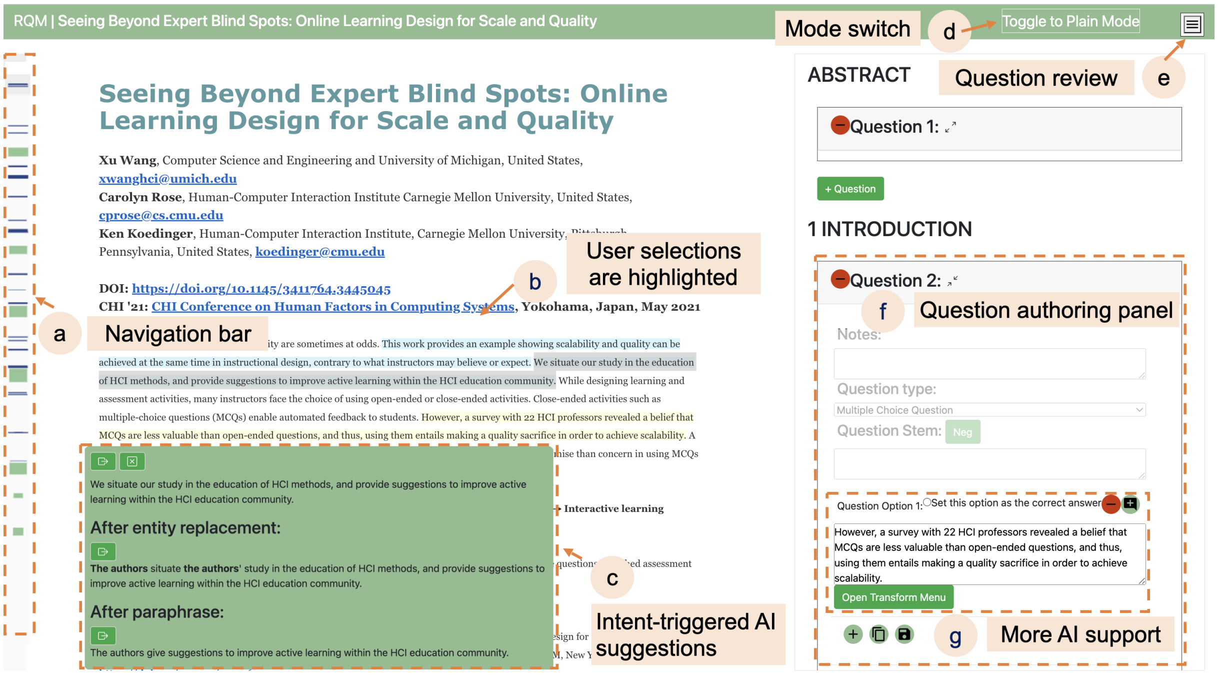
Xinyi Lu
University of Michigan
Ann Arbor, United States
lwlxy@umich.edu

Simin Fan
University of Michigan
Ann Arbor, United States
oliviaaa@umich.edu

Jessica Houghton
University of Michigan
Ann Arbor, United States
houghj@umich.edu

Lu Wang
University of Michigan
Ann Arbor, United States
wangluxy@umich.edu

Xu Wang
University of Michigan
Ann Arbor, United States
xwanghci@umich.edu



RQM | Seeing Beyond Expert Blind Spots: Online Learning Design for Scale and Quality

Seeing Beyond Expert Blind Spots: Online Learning Design for Scale and Quality

Xu Wang, Computer Science and Engineering and University of Michigan, United States, xwanghci@umich.edu
Carolyn Rose, Human-Computer Interaction Institute Carnegie Mellon University, United States, cprose@cs.cmu.edu
Ken Koedinger, Human-Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania, United States, koedinger@cmu.edu

DOI: https://doi.org/10.1145/3411764.3445045
CHI '21: CHI Conference on Human Factors in Computing Systems, **Yokohama, Japan, May 2021**

ity are sometimes at odds. This work provides an example showing scalability and quality can be achieved at the same time in instructional design, contrary to what instructors may believe or expect. We situate our study in the education of HCI methods, and provide suggestions to improve active learning within the HCI education community. While designing learning and assessment activities, many instructors face the choice of using open-ended or close-ended activities. Close-ended activities such as multiple-choice questions (MCQs) enable automated feedback to students. However, a survey with 22 HCI professors revealed a belief that MCQs are less valuable than open-ended questions, and thus, using them entails making a quality sacrifice in order to achieve scalability. A

We situate our study in the education of HCI methods, and provide suggestions to improve active learning within the HCI education community.

After entity replacement:

The authors situate **the authors**' study in the education of HCI methods, and provide suggestions to improve active learning within the HCI education community.

After paraphrase:

The authors give suggestions to improve active learning within the HCI education community.

ABSTRACT

Question 1:

+ Question

1 INTRODUCTION

Question 2:

Notes:

Question type:

Multiple Choice Question

Question Stem: Neg

Question Option 1: ○ Set this option as the correct answer

However, a survey with 22 HCI professors revealed a belief that MCQs are less valuable than open-ended questions, and thus, using them entails making a quality sacrifice in order to achieve scalability.

Open Transform Menu

Figure 1: The ReadingQuizMaker interface contains a Navigation Bar (a), a Paper Panel (middle), and a Question Authoring Panel (right). Users can navigate a paper using the Navigation Bar. The green blocks indicate tables and figures. Users can select content in the Paper Panel and transfer them to the Question Authoring Panel to create questions. ReadingQuizMaker provides AI suggestions (e.g., paraphrase) based on the user selection of text (c). User selections are highlighted (b). On the Question Authoring Panel, users can edit question stem and options (f) and preview more AI suggestions (g). Users can switch between a section-based mode and a plain state (d), and review all the questions they have created (e).

ABSTRACT

Despite that reading assignments are prevalent, methods to encourage students to actively read are limited. We propose a system ReadingQuizMaker that supports instructors to conveniently design high-quality questions to help students comprehend readings. ReadingQuizMaker adapts to instructors' natural workflows of creating questions, while providing NLP-based process-oriented support. ReadingQuizMaker enables instructors to decide when and which NLP models to use, select the input to the models, and

CHI '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9421-5/23/04…$15.00
https://doi.org/10.1145/3544548.3580957

edit the outcomes. In an evaluation study, instructors found the resulting questions to be comparable to their previously designed quizzes. Instructors praised ReadingQuizMaker for its ease of use, and considered the NLP suggestions to be satisfying and helpful. We compared ReadingQuizMaker with a control condition where instructors were given automatically generated questions to edit. Instructors showed a strong preference for the human-AI teaming approach provided by ReadingQuizMaker. Our findings suggest the importance of giving users control and showing an immediate preview of AI outcomes when providing AI support.

CCS CONCEPTS

• **Applied computing** → **Computer-assisted instruction**; **Interactive learning environments**.

KEYWORDS

Reading Quiz, Active Learning, Human-AI Teaming, Automatic Question Generation

ACM Reference Format:
Xinyi Lu, Simin Fan, Jessica Houghton, Lu Wang, and Xu Wang. 2023. ReadingQuizMaker: A Human-NLP Collaborative System that Supports Instructors to Design High-Quality Reading Quiz Questions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3544548.3580957

1 INTRODUCTION

Assigned readings are an integral part of almost any college class. Instructors believe that class readings are important learning activities and can enhance class discussion [17, 46]. However, research has shown that it has been a nationwide problem in higher education that students do not complete reading assignments [13, 15, 18, 24, 26]. It is estimated consistently across studies in different subject domains that just 20-30% of undergraduate students read the materials that they are assigned for classes [13, 15, 26]. With the prevalence of social media and short, fast-paced snippets of information, there is a further decline in college reading over the past decade [29, 76]. The reasons for the low compliance are multifold, it could be that students lack motivation [15, 51], are deficient in reading skills [72], have constraints on their time [63, 67], and undervalue the importance of reading [43, 62].

How can instructors better support students' academic reading practices? Pedagogical strategies and digital tools have been proposed to assist students' reading experience. In recent years, many instructors choose to use social annotation tools. For instance, Perusall [68] has been one of such popular tools and is widely used for pre-class reading assignments. Research has shown that students spend more time reading and have better performance on in-class exams after using Perusall [60]. However, research has also shown that social annotation tools only work better for students who have self-regulated learning skills [22]. One weakness of such tools resides in the fact that students do not get feedback on their understanding of the content [9]. Over the years, reading questions is a common strategy employed by instructors to actively engage students. On the one hand, question answering provides an active learning experience compared to passively reading a text [23, 34], which is demonstrated through decades of educational research to be effective at improving comprehension and learning outcomes [9, 23, 48, 77]. On the other hand, students get immediate feedback [77] on quiz questions that help them self-evaluate their understanding or go back to read certain parts of the text more carefully [46]. For students who are less proficient in academic reading, carefully-designed reading questions aid them in focusing and extracting essential information [28, 72] that may otherwise get lost [72].

However, high-quality and thought-provoking questions take a significant amount of effort to design. Both instructors and students do not favor detail-oriented quiz questions for readings, since they simply check whether the student has read specific content or not [71, 72]. As suggested in prior work, good reading questions should guide the students as they read, help them identify what is important, underline what they should understand by the end, prompt thoughts about the main issues and implications in the content, and prepare the students to come to class ready to talk about the readings [32]. Although there are many existing NLP-based automatic question generation systems, the adoption of them in classrooms is low [8, 49], mainly because those models are only suitable for specific domains, such as language learning and math teaching, and the generated questions are often of low quality and limited in types and difficulty levels [8, 49].

In this work, we propose a human-NLP (Natural Language Processing) collaborative system *ReadingQuizMaker*, to support instructors to create high-quality reading questions. ReadingQuizMaker adapts to instructors' natural workflows of the question construction procedure, while providing NLP-based process-oriented support. The design of ReadingQuizMaker is informed by a need-finding study with 11 instructors from 7 different universities, which suggests that instructor input is critical in the question creation process since they rely on domain expertise and external resources when creating questions. When providing AI assistance on question creation, instructors want to make sure they have full control and flexibility on when and how to use AI [10].

An overview of ReadingQuizMaker is shown in Figure 1. Users can select texts on the Paper Panel and send them to the Question Authoring Panel as question options. ReadingQuizMaker provides an immediate preview of AI suggestions, such as entity replacement and paraphrasing using the user-selected texts as input. On the Question Authoring Panel, users get suggestions on question stems, and can use the NLP Toolbox to improve question options and create distractors powered by a negation model. All the texts instructors have used are highlighted in the Paper Panel and visualized through the Navigation Bar, allowing users to check content coverage.

We ran an evaluation study with 13 instructors from 10 different universities to evaluate ReadingQuizMaker. All participants successfully created questions that they were satisfied with using ReadingQuizMaker. Participants commented that the interface was easy to use, helped them create better questions and could help save time compared to their usual instructional design practice. Participants also found the AI suggestions in ReadingQuizMaker to be useful and desirable. About 60% of the AI paraphrase suggestions were adopted by the users after they read them. About 60% of the AI negation suggestions were adopted by the users. Paraphrase suggestions were more often read and adopted by users compared to the summarization suggestions, because paraphrase suggestions are more discoverable without requiring the user to make an explicit request. Users shared that they got inspiration from AI and would make sure they read the AI results before using them to avoid errors.

In the study, we compared ReadingQuizMaker with a baseline condition where instructors received automatically generated questions and had an opportunity to edit and improve them. Participants strongly preferred the human-AI teaming approach offered by ReadingQuizMaker, and found the questions that were automatically generated to be of lower quality. Instructors also found editing automatically generated questions to be more challenging compared to creating questions from scratch. Participants considered that having a sense of control was critical in their question creation process.

This paper makes the following three contributions:

- A formative study revealing instructors' challenges in creating reading quiz questions, and design requirements for developing tools to support the process.
- ReadingQuizMaker, a novel system that provides NLP-based process-oriented support to instructors while they create questions. The design of ReadingQuizMaker adapts to instructors' natural workflows of creating questions, expedites the question creation process through novel interaction designs, and provides instructors with AI suggestions to augment their question creation experience.
- An evaluation of ReadingQuizMaker that demonstrates the usability and utility of the system. The study shows the promises of the human-AI teaming approach in supporting creative instructional design work, and suggests the importance of giving user control when providing AI support and showing an immediate preview of AI outcomes to increase adoption.

2 RELATED WORK

2.1 Low Compliance in College Reading Assignments

There has been low compliance with reading assignments among college students. It is found that only 20% - 30% undergraduate students do readings for class [13], which leads to undesirable academic performance [24, 73]. One major reason for low compliance is the lack of motivation [17, 50, 51]. Many students underestimate the importance of reading [73]. Studies found that students view readings as a complement to lectures [16], and tend not to spend extra time on reading [13, 16, 62]. Deficiency in reading skills also leads to students' low compliance to read [17, 72], especially when they encounter increasing complexity of the reading assignments and visualizations [72]. Time restriction is ranked as the "number one constraint" that prevents students from completing reading assignments [13, 46]. This suggests better practices are needed to help students read, by reducing the difficulty of reading assignments, giving students who are deficient in reading scaffolding and practice, and providing feedback to students' reading processes to increase engagement.

2.2 Strategies to Support Reading Practices

Prior work has proposed a large variety of strategies to help students actively read. One line of work encourages students to take active notes, e.g., learning log [20] or index cards [19]. Other work supports collaborative reading [44], e.g., enabling students to share notes [59], make podcasts [12], share posts on discussion forums or social media [29, 44]. Social annotation tools such as Perusall [68] was found to be effective in increasing students' reading time. These strategies emphasize the importance of active reading where activities are designed to sustain students' active attention during reading. However, research has also shown that many of these strategies only work better for students who have self-regulated learning skills [22]. One weakness of such tools is that students do not get feedback on their understanding of the content [9]. Another frequently used approach is reading quiz questions [13, 35], in the form of multiple-choice questions which enable immediate feedback. However, designing high-quality and thought-provoking reading quiz questions can be difficult [81, 82].

2.3 Interfaces to Support Active Reading

One line of research focuses on mimicking the physical paper reading experience in digital environments [64], with a focus on navigation and note taking. For example, Pearson et. al developed digital stickers that can be used as bookmarks for digital reading [66]. LiquidText introduced a workspace for users to interact with their comments. [80]. More recent work includes novel UI designs to augment scientific paper reading experiences [36, 37] with support on understanding formula notations and math. In our work, we focus on developing technologies to support reading comprehension in higher education contexts. We aim to support instructors to create high quality reading questions that would in turn help students' conceptual understanding of the text.

2.4 Question Generation Techniques for Educational Purposes

With the increasing awareness of the importance of active reading and active learning, researchers working at the intersection of AI and education have developed techniques to support question creation. One line of work uses crowdsourcing techniques to produce new questions. For example, UpGrade creates questions based on prior student solutions [83] and QMAps encourage students to generate questions for each other [91]. Another line of work develops end-to-end NLP models for question creation. Existing automatic question generation techniques are good at creating factual questions [27, 49], while not being able to generate questions that target higher Bloom goals [14]. On question stem and open-ended question generation, Willis et. al used KPE-Gen to extract key phrases and generate question stem. BERT [78] and PLM [85] models are used for open-ended question creation as well. QG-Net trained a recurrent neural network structure to incorporate the context information and generate wh-format question word-by-word [84]. On multiple-choice question generation, prior approaches used name entity recognition and topic modeling to identify salient sentences and extract keywords for question options [54, 55]. However, the main drawback of existing question generation systems is that they often work for a single domain [79, 92] and the generated questions are often of low quality and limited in types and difficulty levels [21, 39, 49, 65]. In this work, we introduce a Human-AI teaming approach, where AI provides process-oriented to human instructors when they design multiple-choice questions. The goal is to make the question creation process more robust and flexible, and produce higher quality questions.

2.5 Human-AI Systems for Education

Since pure AI systems tend to have high uncertainty of model capability and high complexity of model output [89], human-AI collaboration has been explored in a variety of domains [25, 86]. This concept was introduced and studied in education in recent years. Human-AI educational systems often have human instructors lead the instructional decision making process, while AI provides support along the way. Previous studies have explored a diverse set of human-AI approaches, mostly focusing on supporting in-classroom teaching, e.g., visualizing student progress and struggle to teachers through dashboards [38, 57, 58], smart wearable devices [42, 70], and ambient awareness tools [7], helping teachers assign students to teams [87], and improving classroom orchestration [56, 88], The effectiveness of AI support in other stages of teaching and instructional design is under explored, for example, helping teachers prepare materials and questions [74]. In this work, we investigate the capability of a human-AI teaming approach to support teachers in their quiz question creation.

2.6 Human-AI System Design Guidelines

In recent years, researchers have proposed guidelines for designing and developing human-AI systems [10]. In one of the mostly adopted guidelines, Amershi et al. proposed 18 guidelines that are applicable to different interaction scenarios to improve user experiences. Several guidelines that are in particular relevant to this work include "Show contextually relevant information", "Support efficient invocation", "Support efficient dismissal", and "Support efficient correction" [10]. The design of ReadingQuizMaker was inspired by these principles and we aim to examine how users interact with a human-AI system on a creative and high-stakes task that heavily relies on subject matter expertise. Additionally, Holstein et al. proposed that involving practitioners at all stages and iteratively improving the system is critical when designing human-AI systems [41]. Yang et. al [89] showed that users prefer high-recall systems to high-precision systems [47], which suggests that users need control over when and how to use the AI outcomes. The design of ReadingQuizMaker was informed by a thorough formative study with 11 university professors, and we went through rounds of pilot testing to make sure the human-AI interactions are intuitive and desirable to the users.

3 FORMATIVE INVESTIGATIONS

We performed a formative study with 11 college instructors to understand their natural workflows of creating questions, with the goal of understanding the unique challenges instructors have when they hand-write questions and summarizing the design requirements for developing a user-centered system to support the process.

3.1 Participants and Procedure

The formative study is IRB-approved. 11 instructors from 7 different universities participated in the study (6 male, 5 female). The instructors have teaching experience ranging from 2 to 40 years and are from disciplines including computer science, information science, data science, education, developmental psychology, and political science.

The interviews were done through Zoom and each lasted between 50 and 75 minutes. Participants were given a $50 Gift Card. We first asked participants to share how they approached reading assignments. The majority of the session was spent having the participant design questions based on a reading text of their choice. No support was provided during the session. Specifically, we asked them to design questions (MCQ preferred) that could help their students understand and learn from the content. We asked the participants to think aloud throughout the process.

Participants were able to design 3 to 10 multiple-choice questions (with one question stem and four options) during the session. We then asked the participant to reflect on how they arrived at each question stem and option, and shared the challenges they had encountered throughout the process. At last, the researcher asked the participant to imagine there being an intelligent system to provide support alongside the quiz design process. Specifically, the researcher asked the user's attitudes towards a list of NLP tasks contextualized in their question creation process, e.g., "What do you think if the system can paraphrase this sentence for you?" We transcribed the interview recordings and analyzed our data using affinity diagrams [61].

3.2 Results

3.2.1 *Designing High Quality Reading Questions is Desirable yet Inaccessible.*

All participants mentioned that there should be better ways to support students to read. For example, P5 said "*It is definitely a problem that instructors face*." Participants shared the techniques they have used to support reading and the limitations of such approaches. For example, reading summaries are not scalable as grading can be challenging; collaborative annotation platforms such as Perusall [68] encourage participation but do not necessarily make sure students get the key messages. P7 appreciated the many functionalities of Perusall while raising a concern that "*similar collaborative annotation tools can treat readings as mechanistic and measure whether they did the reading not whether they found the insight*". Several participants were already using quiz questions to enhance reading. Most participants expressed interest in using quiz questions if question design becomes less expensive and more accessible.

3.2.2 *I Want to Use Reading Questions to Guide the Students Think.*

While some instructors liked using quiz questions for readings, they also emphasized the questions should be designed in the right way. 1) Instructors wanted to ask about integrative knowledge instead of questions that require students to pattern match. P2 mentioned he was interested in asking "*why*" questions, e.g., "*why is it difficult*". P1 mentioned that critical thinking was important "*But this paper itself has all of these pretty deep flaws. And I'd want students to see the flaws*". P5 mentioned they would encourage students' thoughts outside of the material.

3.2.3 I Face Challenges in Quiz Authoring. Participants reported a variety of challenges in their quiz authoring process. Almost all participants said that this is a time consuming task for them, and that they could spend a considerable amount of time on it. The two most salient challenges are 1) Identifying question opportunities; 2) Coming up with distractors. Most participants first skim through the text to identify content that they think is important for students to know and then design questions for it. For example, P4 said "*I need to first think about what are the key concepts in the text"*. Almost all participants said that coming up with distractors was hard. Instructors wanted distractors to be thought provoking (P7), convincing (P6), and reveal student misconceptions and bring up opportunities to discuss concepts in class (P5). P5 said they often were not sure whether the distractors they came up with were good, and considered open-ended questions to be better at helping instructors elicit student misconceptions.

Other challenges mentioned by the participants include coming up with question stems, getting a comprehensive coverage, and being accurate with content details. Instructors found writing the question stem to be difficult because they should be tailored to the type of paper and the learning outcomes. Multiple participants mentioned that they wanted to get a relatively comprehensive coverage of the content, but it was hard for them to gauge the coverage at different points. Moreover, instructors wanted to be accurate with the question options, and make sure that correct answers were entirely true, and wrong answers were wrong.

3.2.4 NLP Support could be Useful if they are Good, Controllable and Transparent. Participants generally responded positively to the idea of using Natural Language Processing (NLP) tools to support their process. While at the same time, participants expressed desire for controls in the process to decide whether they will use the NLP outcomes. Instructors showed mixed opinions on when they would use different NLP models. For example, P6 said "*I think paraphrasing would be helpful, because I think that's the hardest part for me"*. He also thought summarization would be helpful for him to shorten the rereading time, "if I were able to have [a section of the text] summarized then, if I trusted the intelligence, it would be like, here are the main points of this. So it just saves time." Participants mentioned that their use of the support would depend on the performance of the models. They thought negation could be helpful in some contexts, especially if they [the user] could provide a keyword, noting "*that would make it more accurate and more towards what I need*". P7 emphasized that they want the intelligent system to provide suggestions while they will maintain control.

3.3 Design Requirements

Based on the need-finding study, we summarize the following design requirements for developing an instructor-centered quiz design system. The design requirements also correspond to prior literature on human-AI interaction design guidelines [10, 90].

- **Support instructors in creating convincing distractors.** Many participants expressed difficulty in generating distractors, P4 said "*I want [the distractor] to not to be very easy for them to guess, but I don't want it to be too tricky"*. Most participants said that they would like to receive support on creating meaningful distractors.
- **Provide process-oriented support and enable instructors to incorporate their expert knowledge.** Instructors did not like the traditional end-to-end AI approaches in question creation. Instructors wanted the flexibility to make decisions and incorporate their expertise when needed, e.g., the context of the course, the background knowledge of the students, etc.
- **Question creation needs to be quick and integrate with instructors' current workflows.** Most participants wanted to shorten their time in question creation. For example, P7 and P9 mentioned that they would like to spend as little time writing questions as possible, and P10 suggested that having a system to take care of the lower level problems would be helpful to their workflow.
- **Enable instructors to easily write feedback for the questions.** Feedback was extremely important to instructors, with P5 discussing how they "*would usually give the students some comments"* when they notice the student had a misunderstanding. Participants said that incorporating a mechanism to give feedback, especially for incorrect answers was vital.
- **Give instructors a sense of control when interacting with AI** When asked about their willingness to receive AI support, instructors shared that they would like to remain in control (P6). P7 commented that it would work if the AI provides suggestions, but they would prefer to maintain control and embed their own knowledge when creating questions. This aligns with multiple principles in the human-AI interaction guidelines [10] to support efficient invocation, dismissal and correction.
- **Give instructors flexibility in making decisions** We observed that different instructors had diverse strategies for creating questions. Instructors have preferences over different question types. For example, they may use multiple-choice questions for large classes, and open-ended questions for seminar-style graduate-level classes. Instructors also wanted the question pool to have a mixed level of difficulty.

4 READINGQUIZMAKER

Based on the user challenges and design requirements identified in the formative study, we develop ReadingQuizMaker, a system that provides process-oriented support for instructors to create high quality reading quiz questions that align with their educational goals. ReadingQuizMaker adapts to an instructor's natural quiz design process. ReadingQuizMaker places the reading text and question creation panel side by side, aiming to shorten the time it takes for users to peruse text, as shown in Figure 1. We first describe a user journey with the system. We will then describe each system component aligned with the design requirements above.

4.1 An Example User Journey

Alice teaches at a university and assigns a paper on the topic of Augmented Reality (AR) in her class as a required reading. She wants to create reading quiz questions that can guide the students to read and help students understand the main takeaways from the paper. Alice uses ReadingQuizMaker to create questions. She loads the HTML file of the paper into the system and starts creating questions.

As Alice reads through the abstract, she gets an idea for a question. She wants the students to know what this paper did and what the authors found. She picks a question stem from the menu "Which of the following is NOT correct about what the paper does?" Alice then starts to find options in the Paper Panel. Upon her selection of a sentence, ReadingQuizMaker makes suggestions for paraphrasing the sentence. Alice likes the paraphrased result. Here is an example, the original sentence "We discuss learning and collaboration differences, as well as benefits and detriments of implementing augmented reality for unstructured learning activities" is paraphrased as " The benefits and drawbacks of augmented reality for learning are discussed by the authors." After Alice successfully creates several correct options, she needs to come up with a distractor. Alice opens the Transform Menu in the Question Authoring Panel, and tries to negate an option. Alice thinks the negation result is OK, and makes minor changes. Alice now completes the first question. As Alice has more questions, she wants to check the content coverage. Alice reviews the Navigation Bar to see which parts of the reading need more attention. Alice then reviews all the questions she created and downloads them. Alice can now import the questions to Canvas and use the quiz as a pre-class reading assignment.

4.2 Detailed Design

4.2.1 Compatible with Any Article with an HTML Source. ReadingQuizMaker limits the reading sources to HTML files. Most academic publications since 2018 in ACM Digital Library (e.g., ACM CHI, UIST, ICER) have online HTML versions, following the ACM Publishing System (TAPS) to enhance accessibility [6, 33]. Other publishers are also supporting online HTML versions for academic publications, including Springer [3] and Taylor&Francis Online [4]. ReadingQuizMaker also works well with online documentation and tutorials which are frequently used in programming courses, online textbooks [11], and articles from news platforms such as Washington Post and Vox [5].

ReadingQuizMaker (RQM) uses an iframe to display the HTML file of the article and keeps the original formatting. It removes all Javascript code in the HTML source file to avoid conflicts with the system's functionalities. For academic publications in ACM DL, we parsed the structure of the paper through HTML tags and extracted the section titles (e.g., Abstract, Introduction, Related Work, etc.) and displayed the section titles in the Question Authoring Panel. For content from other sources, in the evaluation study, the first author manually extracted the section titles. However, users can choose to use the Plain Mode, in which the section titles are not displayed. This makes ReadingQuizMaker compatible with any articles that have an HTML version.

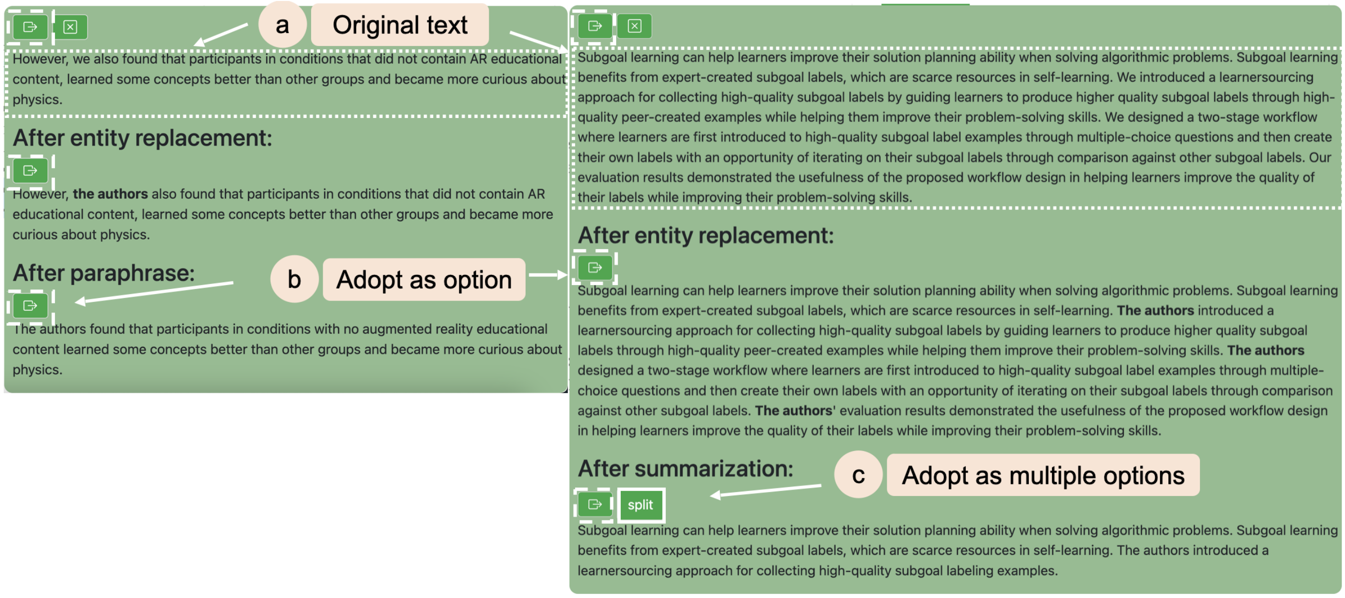
4.2.2 The Navigation Bar Supports Users' Read of Content Coverage and Navigation to Figures and Tables. Many users in the formative study specifically mentioned that they wanted students to read tables and figures. The Navigation Bar highlights the tables and figures in a reading text as green blocks, the size of the blocks is proportional to the size of the figure. This was implemented using pagemap, an npx package for minimap [45]. Tables and figures are automatically extracted based on HTML tags when the file is loaded into the interface. In addition, the Navigation Bar also displays user highlights to indicate what content in the reading has been used in the questions. Users can review the Navigation Bar to check which part of the reading needs more attention.

4.2.3 Provide Immediate Preview of NLP Suggestions. ReadingQuizMaker offers AI suggestions to users based on their text selection. Specifically, users read the content in the Paper Panel. When they see sentences they want to peruse in a question, they can select the text and transfer it to the Question Authoring Panel. The system provides two ways of selecting text: 1) drag the mouse over text for exact selection; 2) double-click within a sentence for a whole sentence selection. Immediately upon the user selects some text, the system previews AI suggestions based on the user selection, as shown in Figure 2. No explicit action is needed from the user to see the AI suggestions. Specifically, three NLP-based transforms are offered. When the sentence contains first-person pronouns such as "we" or "our design", it's replaced with "The authors" or "The design" using regular expression. For short sentences, the system provides a paraphrase suggestion. For longer texts or paragraphs, the system provides a concise summary. If the user likes to use the AI suggestions, they can transfer them to the Question Authoring Panel. Or users have the flexibility to send the original text. When a user sends text to be an option, more AI suggestions will appear automatically beneath the added option. The user can decide whether they want to use them.

4.2.4 Adapt to Users' Natural Workflows and Enable Users' to Write Feedback for Options. The Question Authoring Panel is designed to align with users' natural flow of creating questions. An overview of the panel is shown in Figure 3. The system supports three types of questions multiple-choice questions, multiple-response questions, and open-ended questions, as shown in Figure 3. For all question types, users can transfer text from the Paper Panel, transfer images from the Paper Panel, or freely add content by themselves(Figure 1). Based on the formative study, we also make sure it is easy for users to add feedback to question options, and enable users to do high-level planning by adding notes to the Notes field.

Many participants in the formative study mentioned that it was difficult for them to come up with question stems. To address this, ReadingQuizMaker offers a question stem bank that is crowdsourced from the formative study, as shown in Figure 4. The question stem bank contains 28 question templates. The user can choose from the stem bank, or they can type a new stem and the system offers suggestions based on keyword match. The system offers different stem suggestions based on the sections the users are in (e.g., introduction, related work), the question type (MCQ, MRQ, OEQ), and whether the user selects a figure. Table 1 shows example question stems. Through the interface, users can also negate a question stem by adding or removing the word "NOT".

4.2.5 NLP Toolbox. In the formative study, we found that participants chose to use different strategies when creating questions depending on the context. To support versatile question creation, in ReadingQuizMaker, we provide an NLP Toolbox that users can receive further AI assistance, as shown in Figure 5. The toolbox first loads the original text in an option. Users can choose to apply



However, we also found that participants in conditions that did not contain AR educational content, learned some concepts better than other groups and became more curious about physics.

After entity replacement:

However, **the authors** also found that participants in conditions that did not contain AR educational content, learned some concepts better than other groups and became more curious about physics.

After paraphrase:

The authors found that participants in conditions with no augmented reality educational content learned some concepts better than other groups and became more curious about physics.

Subgoal learning can help learners improve their solution planning ability when solving algorithmic problems. Subgoal learning benefits from expert-created subgoal labels, which are scarce resources in self-learning. We introduced a learnersourcing approach for collecting high-quality subgoal labels by guiding learners to produce higher quality subgoal labels through high-quality peer-created examples while helping them improve their problem-solving skills. We designed a two-stage workflow where learners are first introduced to high-quality subgoal label examples through multiple-choice questions and then create their own labels with an opportunity of iterating on their subgoal labels through comparison against other subgoal labels. Our evaluation results demonstrated the usefulness of the proposed workflow design in helping learners improve the quality of their labels while improving their problem-solving skills.

After entity replacement:

Subgoal learning can help learners improve their solution planning ability when solving algorithmic problems. Subgoal learning benefits from expert-created subgoal labels, which are scarce resources in self-learning. **The authors** introduced a learnersourcing approach for collecting high-quality subgoal labels by guiding learners to produce higher quality subgoal labels through high-quality peer-created examples while helping them improve their problem-solving skills. **The authors** designed a two-stage workflow where learners are first introduced to high-quality subgoal label examples through multiple-choice questions and then create their own labels with an opportunity of iterating on their subgoal labels through comparison against other subgoal labels. **The authors**' evaluation results demonstrated the usefulness of the proposed workflow design in helping learners improve the quality of their labels while improving their problem-solving skills.

After summarization:

Subgoal learning can help learners improve their solution planning ability when solving algorithmic problems. Subgoal learning benefits from expert-created subgoal labels, which are scarce resources in self-learning. The authors introduced a learnersourcing approach for collecting high-quality subgoal labeling examples.

Figure 2: ReadingQuizMaker gives users immediate AI suggestions based on their text selection, including 1) entity replacement that replaces first-person pronouns as third-person pronouns; 2) paraphrase for single sentences (left); 3) summarization for paragraphs or long texts (right). Users can choose to adopt any of the AI suggestions as a question option. Users can also choose to split the summarization result into multiple options.

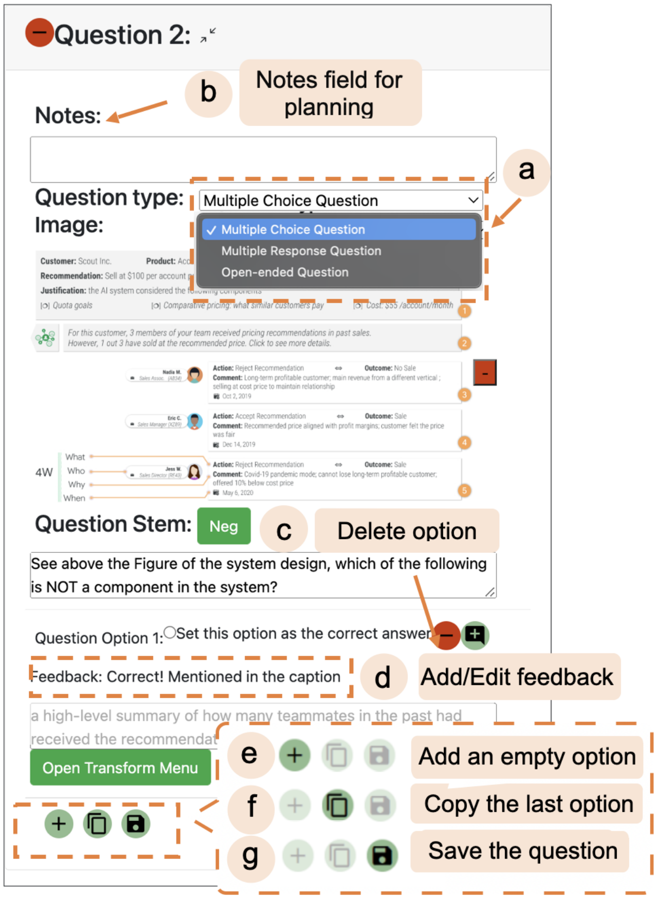
Table 1: Example question stems from the question stem bank of ReadingQuizMaker. Users see question stem suggestions when they are at the corresponding sections of the reading text.

Question Stem	Paper Section
Which of the following is NOT a correct description of the motivation of this paper?	Abstract, Introduction
Why is this a hard problem to solve?	Introduction, Open-Ended
How is the outcome measured in the evaluation study?	Methods, Multiple-Response
See the Figure above, which of the following is correct?	Figure
The authors claimed [], what is a justification for that?	Findings, Open-Ended
Which of the following are the findings and takeaways of this paper? (Select all that apply)	Findings, Multiple-Response
Which of the following is NOT correct about the limitations of the paper?	Discussion

Paraphrase, Summarize, Negation operations to transform the text. Since most participants found coming up with distractors to be challenging, we introduced the Negation model to give them suggestions. We also implemented a user-controllable version of Negation where the user can decide which word to negate in a sentence, as shown in Figure 6. The user selects the word "greater" to negate. When the user selects a word, the system extracts 7 words before and after the parameter, and sends these 15 words to the negation model. The system then replaces the negated word(s) in the original sentence.

In addition to single operations, users can also combine multiple operations through chaining. Figure 6 shows an example of chaining two operations. The paraphrased result is sent as the input of the negation model. The user can always load the original text if they change their mind.

4.2.6 *Review and Output.* During the user's question creation process, they can review all the questions created as shown in Figure 7. After the user has enough questions, they can download the questions into a .csv file, which is formatted to be ready to transfer into a .QTI package. Quiz in QTI packages is compatible with most learning management systems, including Canvas. Users can easily import questions they created in ReadingQuizMaker to Canvas.



See above the Figure of the system design, which of the following is NOT a component in the system?

Feedback: Correct! Mentioned in the caption

a high-level summary of how many teammates in the past had received the recommendat

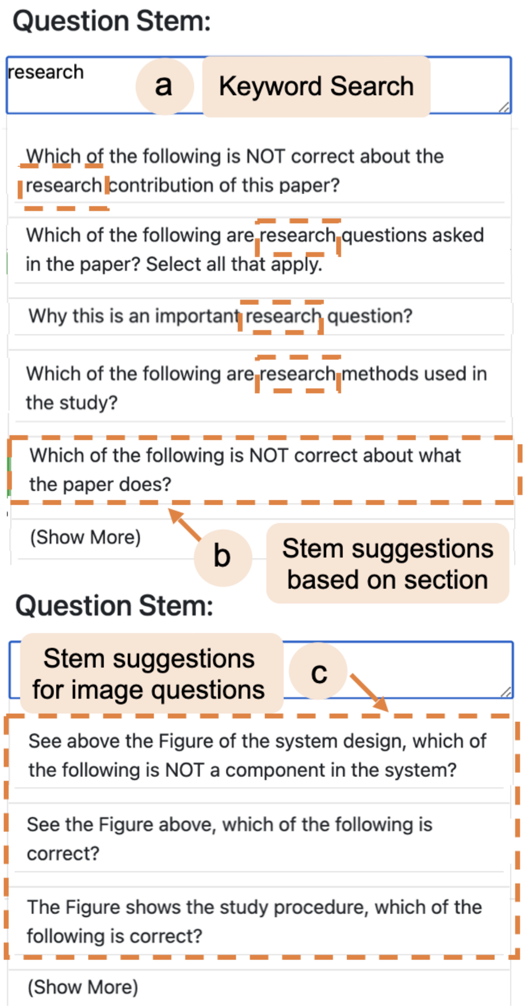
Figure 3: Question Authoring Panel in ReadingQuizMaker is collapsible. The system supports 3 question types (a). Users can add notes for high-level planning (b), add/delete options (e, c), add/delete feedback for each option (d), copy a previous option (f), and save(g)

This also addresses the design requirement revealed in the formative study that instructors want a seamless approach that saves their time.

4.3 NLP Models

A collection of transformer-based NLP models were applied to give users suggestions in the ReadingQuizMaker system, including (1) an Abstractive Summarization model to condense long paragraphs; (2) a Paraphrase model to paraphrase and simplify sentences; (3) a Negation model to generate incorrect options. The details of the models are shown below.

- **Abstractive Summarization** We used a fine-tuned **BART** [52] model on CNN-DailyMail to condense the content given a paragraph as input. We use the checkpoint *bart-large-cnn* from HuggingFace[1].
- **Paraphrase** We used a paraphrase model pretrained on **PEGASUS** [93] to rephrase a sentence while remaining its semantic information. We take the released checkpoint *pegasus_paraphrase* from HuggingFace[2].
- **Negation** We applied a **BART**-based negative claim generation model fine-tuned on WikiFactCheck-English [75]. We take the checkpoint released by the authors on HuggingFace[3].



Which of the following is NOT correct about the research contribution of this paper?

Which of the following are research questions asked in the paper? Select all that apply.

Why this is an important research question?

Which of the following are research methods used in the study?

Which of the following is NOT correct about what the paper does?

See above the Figure of the system design, which of the following is NOT a component in the system?

See the Figure above, which of the following is correct?

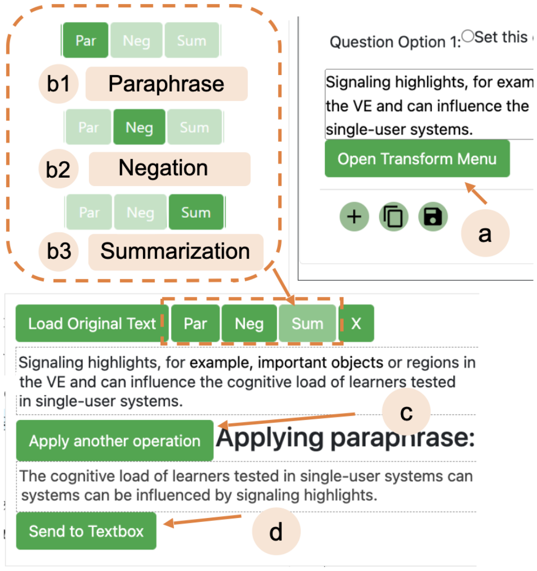
The Figure shows the study procedure, which of the following is correct?

Figure 4: ReadingQuizMaker helps users come up with question stems. The system offers suggestions of stems based on keywords (a). The system offers different suggestions based on the sections the users are in (e.g., introduction, related work), the question type (MCQ, MRQ, OEQ), and whether the user selects a figure (b, c).

[1] https://huggingface.co/facebook/bart-large-cnn

[2] https://huggingface.co/tuner007/pegasus_paraphrase

[3] https://huggingface.co/minwhoo/bart-base-negative-claim-generation



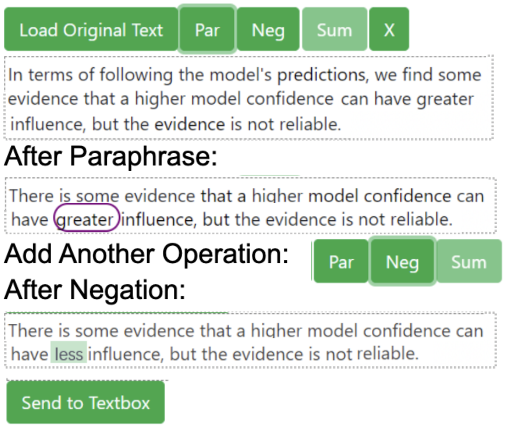
Signaling highlights, for exam
the VE and can influence the
single-user systems.

Signaling highlights, for example, important objects or regions in the VE and can influence the cognitive load of learners tested in single-user systems.

Applying paraphrase:

The cognitive load of learners tested in single-user systems can systems can be influenced by signaling highlights.

Figure 5: The Transform Menu can be opened by clicking the button (a) under each option. The menu takes the option as the original text where users can apply Paraphrase (b1), Negation (b2) or Summarize (b3) transformation. They can choose to apply another operation on the result (c) or use the AI-transformed result (d)



In terms of following the model's predictions, we find some evidence that a higher model confidence can have greater influence, but the evidence is not reliable.

After Paraphrase:

There is some evidence that a higher model confidence can have greater influence, but the evidence is not reliable.

Add Another Operation:

After Negation:

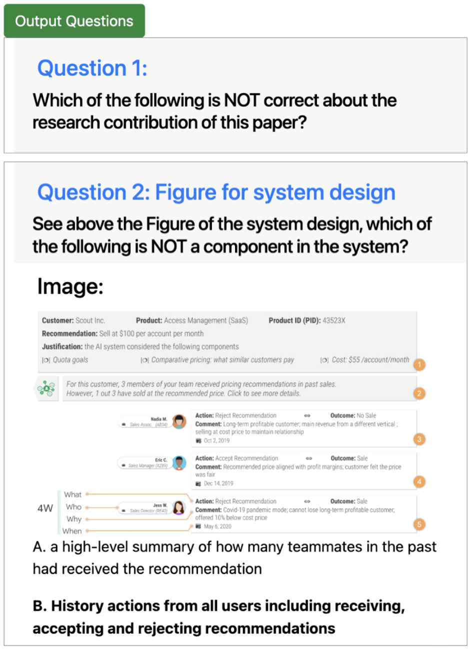
There is some evidence that a higher model confidence can have less influence, but the evidence is not reliable.

Figure 6: An example of chaining NLP operations to generate a distractor. The user chains a paraphrase operation with a negation operation. In the negation operation, the user selects the word "greater" to negate.

4.4 System Iteration

We did three rounds of pilot testing to interactively improve the system design. Here are the main things we iterated on.

4.4.1 Increase Discoverability of the AI features. During the pilot studies, we realized that users tended not to click buttons to apply NLP transformations. However, when we encouraged the users to evaluate the NLP suggestions, they found them to be satisfying.



Question 1:

Which of the following is NOT correct about the research contribution of this paper?

Question 2: Figure for system design

See above the Figure of the system design, which of the following is NOT a component in the system?

Image:

Customer: Scout Inc. **Product:** Access Management (SaaS) **Product ID (PID):** 43523X
Recommendation: Sell at $100 per account per month
Justification: the AI system considered the following components
Quota goals Comparative pricing: what similar customers pay Cost: $55 /account/month

For this customer, 3 members of your team received pricing recommendations in past sales. However, 1 out 3 have sold at the recommended price. Click to see more details.

Action: Reject Recommendation ⇔ **Outcome:** No Sale
Comment: Long-term profitable customer; main revenue from a different vertical ; selling at cost price to maintain relationship
Oct 2, 2019

Action: Accept Recommendation ⇔ **Outcome:** Sale
Comment: Recommended price aligned with profit margins; customer felt the price was fair
Dec 14, 2019

Action: Reject Recommendation ⇔ **Outcome:** Sale
Comment: Covid-19 pandemic mode; cannot lose long-term profitable customer; offered 10% below cost price
May 6, 2020

A. a high-level summary of how many teammates in the past had received the recommendation

B. History actions from all users including receiving, accepting and rejecting recommendations

Figure 7: The Question Review Panel gives an overview of all the questions created.

To increase the discoverability of the AI features, we introduced the immediate preview of NLP suggestions, as shown in Figure 2. This avoids extra clicks from users to receive support, which is in alignment with the human-AI interaction guidelines [10] to support efficient invocation, dismissal and correction.

4.4.2 Visualize NLP Transformations. We found that participants needed to spend time comparing the NLP suggestions with the original text. We then visualize the changes before and after an operation for the Entity Replacement and Negation transformations, as shown in Figure 2 and Figure 6. We did not implement a visualization for Abstractive Summarization and Paraphrase operations, since they are based on generative models, and the difference is often more dramatic than a couple of words. We will leave it to future work to explore better visualization techniques to help end-users read and use NLP outcomes.

4.4.3 Introducing Thresholds to Increase NLP Performance. In the pilot studies, we found that when users apply the summarization operation to a relatively short text, it does not work well. Based on several iterations, we introduced a threshold of 400 characters to decide which models we use to give users suggestions. If the user-selected text is longer, the system will give a summarization suggestion, otherwise, the system will give a paraphrase suggestion, as shown in Figure 2. In the NLP Toolbox, the summarization operation is disabled when the option length is below the threshold.

4.5 Implementation

ReadingQuizMaker is implemented as a full-stack web application with a back-end server for hosting the NLP models. The user interface is written in React.js [31] and Django [30] frameworks. The web app is connected to a back-end server implemented in python, which accepts API calls from the web app (e.g., paraphrase), applies the NLP operation, and returns the result. The web app is deployed through DigitalOcean [2], and the back-end server is deployed as an AWS EC2 instance [1].

5 EVALUATION STUDY

We performed an IRB-approved evaluation study to understand the usability and usefulness of ReadingQuizMaker. We are also interested in understanding how users respond to ReadingQuizMaker as a human-AI collaborative system compared to an automatic approach. We address the following research questions.

- RQ1: Is ReadingQuizMaker usable? Can instructors use ReadingQuizMaker to create questions that they are satisfied with?
- RQ2: How would instructors perceive the AI suggestions? Are they of satisfying quality? Are they distracting? Do instructors find the AI suggestions to be useful or unsatisfying?
- RQ3: How do instructors compare the human-AI teaming approach provided by ReadingQuizMaker with an automatic question generation approach?
- RQ4: What challenges do users experience and what are the design implications to develop human-AI collaborative systems for education?

5.1 Participant Recruitment

We recruited participants through social media (including mailing lists and social groups of professors) and offline correspondences. 13 college instructors (9 male, 4 female) from 10 universities participated in the study. All participants have taught or designed a college-level course that requires readings, and have designed discussion or quiz questions to support students to read. They have an average experience of 4-5 years teaching college courses, with the longest being 17 years. They are from disciplines including education, information science, computer science, technical communication, engineering education, and political science. The study sessions lasted for 90-100 minutes via Zoom. Participants were compensated with a $50 Gift Card. Before the study session, we asked the participants to select a reading text that they would use in the session. The only requirement is that the text needs to have an online HTML source. Among the 13 participants, 7 used academic publications from ACM Digital Library, 4 used online textbook chapters, 1 used online tutorials, and 1 used news articles from a news and opinion website.

5.2 A Baseline Condition with Auto-Generated Questions

To address RQ3, how instructors perceive the human-AI teaming approach compared to an automatic generation approach, we introduce a baseline condition. We developed a pipeline to automatically generate multiple-choice questions from the user's choice of reading. We observed user behaviors during the pilot tests of ReadingQuizMaker that users may extract sentences from one or adjacent paragraphs, paraphrase the sentences as correct options, and negate a sentence as the incorrect option. We followed this pattern in designing the pipeline.

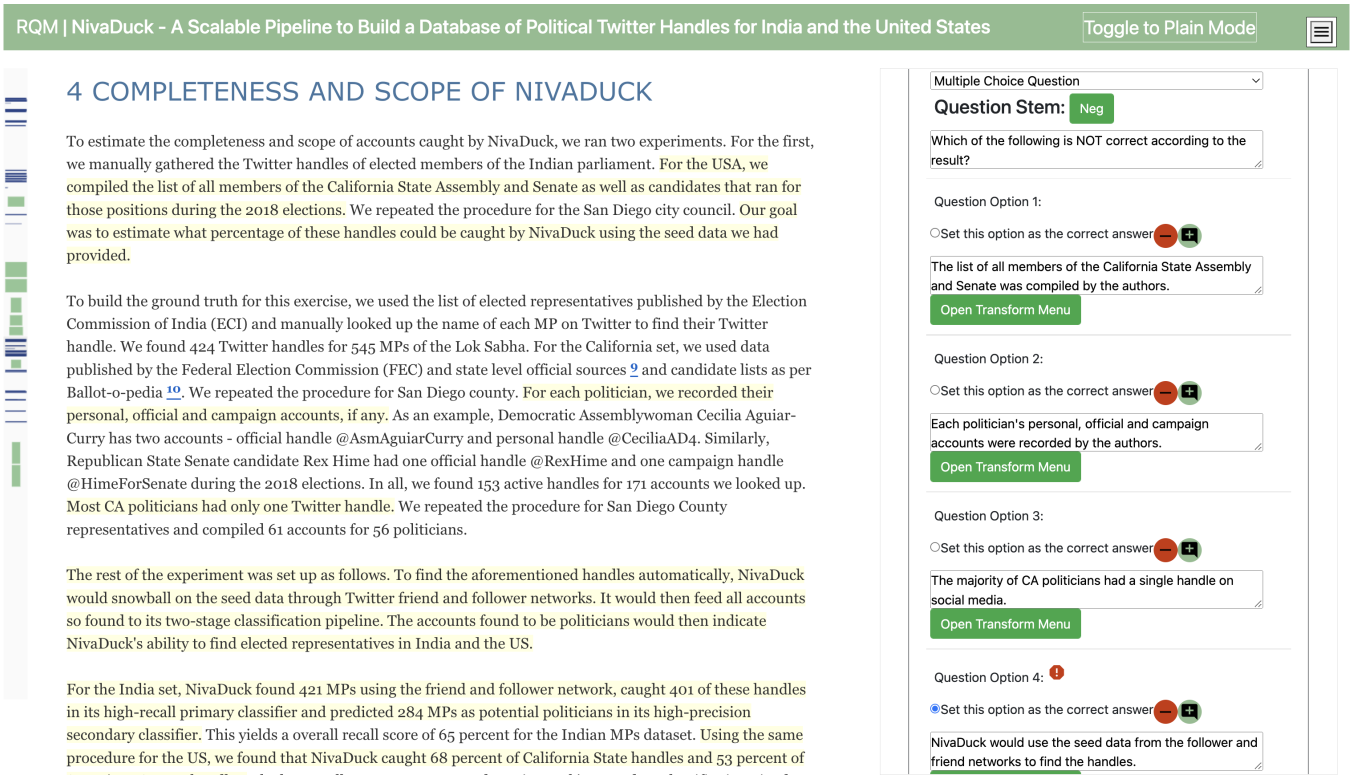
We first parsed the HTML files with BeatuifulSoup [69] to extract the paragraphs. With the paragraphs as input, we applied an Extractive Summarization model to extract salient sentences. We used **BertSumExt** [53], which employs a document-level encoder based on the pretrained BERT model. We take the released checkpoint trained on CNN-DailyMail dataset [40]. We used this model to extract two salient sentences for each paragraph. We then combined two adjacent paragraphs to generate options for one question, so that each question concerns two paragraphs. The question stem is automatically generated following a template as "Which of the following is NOT correct according to the [section heading]". For the four sentences extracted in the previous step, three were paraphrased and used as correct information. The last one was paraphrased, negated, and then paraphrased again, and was used as the incorrect information. The Paraphrase and Negation models are the same ones as those used in the ReadingQuizMaker system.

Following this pipeline, we generated 1-2 multiple choice questions for each section, resulting in 7-10 questions in total for the entire reading. The automatically generated questions are then displayed through the ReadingQuizMaker interface to enable users to make edits if they do not like these questions, as shown in Figure 8.

5.3 Procedure

Participants were asked to send us the reading that they wanted to use prior to the study session. The only requirement is that the reading has an HTML source. All participants experienced both the ReadingQuizMaker condition and the baseline condition. Participants always did the ReadingQuizMaker condition first, because we do not want the automatically generated questions to bias instructors' own design. During the session, we first get participants' consent, and then gave a demo on how to use the ReadingQuizMaker system. Participants then have 45 minutes to use ReadingQuizMaker (Task 1), and 20 minutes to review and potentially edit the automatically generated questions (Task 2), as shown in Figure 8. Participants were asked to share their screens the whole time and think out loud. At the end of each task, the researchers asked follow-up questions.

5.3.1 Task 1: Use the ReadingQuizMaker System to Create Reading Questions. Participants were asked to imagine that they were assigning this reading in a class and designing quiz questions to help students read. We emphasized that quality is more important than quantity so that they did not need to aim for a certain number of questions. We also encouraged the participants to design questions that target higher-order thinking, so that the questions can be a reading guide. We specifically suggested that they did not have to design easy questions to test whether the students had read the text or not. We also made it clear to the participants that they did not need to aim for a certain number of questions since we observed instructors had varying proficiency in question creation in our formative study. The goal here is to probe into whether ReadingQuizMaker can help instructors design thought-provoking questions that are of high educational value, instead of simple factual questions. At the end of the first task, participants were asked to share their experiences. We specifically asked them to comment on the quality of the questions, whether they are satisfied, the quality of the AI suggestions, and the challenges they had experienced in the process.



RQM | NivaDuck - A Scalable Pipeline to Build a Database of Political Twitter Handles for India and the United States

4 COMPLETENESS AND SCOPE OF NIVADUCK

To estimate the completeness and scope of accounts caught by NivaDuck, we ran two experiments. For the first, we manually gathered the Twitter handles of elected members of the Indian parliament. For the USA, we compiled the list of all members of the California State Assembly and Senate as well as candidates that ran for those positions during the 2018 elections. We repeated the procedure for the San Diego city council. Our goal was to estimate what percentage of these handles could be caught by NivaDuck using the seed data we had provided.

To build the ground truth for this exercise, we used the list of elected representatives published by the Election Commission of India (ECI) and manually looked up the name of each MP on Twitter to find their Twitter handle. We found 424 Twitter handles for 545 MPs of the Lok Sabha. For the California set, we used data published by the Federal Election Commission (FEC) and state level official sources [9] and candidate lists as per Ballot-o-pedia [10]. We repeated the procedure for San Diego county. For each politician, we recorded their personal, official and campaign accounts, if any. As an example, Democratic Assemblywoman Cecilia Aguiar-Curry has two accounts - official handle @AsmAguiarCurry and personal handle @CeciliaAD4. Similarly, Republican State Senate candidate Rex Hime had one official handle @RexHime and one campaign handle @HimeForSenate during the 2018 elections. In all, we found 153 active handles for 171 accounts we looked up. Most CA politicians had only one Twitter handle. We repeated the procedure for San Diego County representatives and compiled 61 accounts for 56 politicians.

The rest of the experiment was set up as follows. To find the aforementioned handles automatically, NivaDuck would snowball on the seed data through Twitter friend and follower networks. It would then feed all accounts so found to its two-stage classification pipeline. The accounts found to be politicians would then indicate NivaDuck's ability to find elected representatives in India and the US.

For the India set, NivaDuck found 421 MPs using the friend and follower network, caught 401 of these handles in its high-recall primary classifier and predicted 284 MPs as potential politicians in its high-precision secondary classifier. This yields a overall recall score of 65 percent for the Indian MPs dataset. Using the same procedure for the US, we found that NivaDuck caught 68 percent of California State handles and 53 percent of

Question Stem:

Which of the following is NOT correct according to the result?

Question Option 1:

○ Set this option as the correct answer

The list of all members of the California State Assembly and Senate was compiled by the authors.

Question Option 2:

○ Set this option as the correct answer

Each politician's personal, official and campaign accounts were recorded by the authors.

Question Option 3:

○ Set this option as the correct answer

The majority of CA politicians had a single handle on social media.

Question Option 4: !

● Set this option as the correct answer

NivaDuck would use the seed data from the follower and friend networks to find the handles.

Figure 8: We developed a pipeline to automatically generate multiple-choice questions for the readings our participants picked. The resulting questions are then displayed through the ReadingQuizMaker interface inviting user edits. In the baseline condition, participants review and edit these questions. The source sentences used in the options are highlighted.

5.3.2 *Task 2: Review and Edit Automatically-Generated Questions.* Participants were given 7-10 automatically generated questions on the same reading as they used in Task 1. Participants were asked to read through the questions, share their opinions, make edits when they felt like, or directly abandon questions if the quality was low. Participants used an interface similar to ReadingQuizMaker as shown in Figure 8. The source sentences for the options were highlighted in the Paper Panel. At the end of Task 2, we asked participants to share their thoughts on the automatically generated questions and compare their experiences of ReadingQuizMaker (the human-AI collaborative approach) versus the automatic approach.

5.4 Data Analysis Methods

5.4.1 *Analysis of System Logs.* The system keeps logs of the API calls sent to the back-end on each of the NLP models. Two researchers watched the user study recordings to label for each of the API calls sent, whether the user read or used the suggestion, and what modifications the user made based on the AI suggestion. We labeled an AI suggestion to be adopted if the user sends the AI suggestion as an option to the Question Authoring Panel instead of using the original sentence. We labeled an AI suggestion to be read if the user paused on the interface and read the result. Since there were times participants directly sent the original text, without waiting for the AI suggestions to display. The system also logged whether the adoption of the AI suggestions was from the Immediate Preview function, or from the NLP Toolbox, which requires explicit clicking.

Similarly, we logged users’ adoption of the question stem suggestions. We recorded whether the user checked, and whether they adopted the suggestions, and their modifications to it. A stem is labeled as checked if the user scrolled down the Question Authoring Panel to see more options. And it is labeled as adopted if the user picked a stem from the menu. Users’ handwritten question stems are logged as well.

5.4.2 *Affinity Diagram for Think-Aloud Transcripts.* The recordings were transcribed and analyzed using affinity diagrams [61]. Two authors interpreted the transcripts, iteratively grouped the interpretation notes, and identified emerging themes from the data.

6 FINDINGS

We present findings corresponding to each research question.

6.1 RQ1: All participants successfully created questions that they were satisfied with using ReadingQuizMaker

All users successfully created questions using ReadingQuizMaker. In total, 89 questions were created, including 51 Multiple-Choice Questions, 28 Multiple-Response Questions, and 10 Open-Ended Questions. The 79 Multiple-Choice/Response questions contain 288 options altogether. Each question contains an average of 3.6455 options. We want to emphasize that, before the study we explicitly told the participants to not focus on quantity. During the study sessions, multiple participants mentioned that they saw some easy question opportunities but those did not target higher order thinking, so that they did not write them down.

6.1.1 Instructors were satisfied with the question quality.

All the participants were satisfied with the quality of the questions they created and expressed excitement that they would use them in their classes. For example, P12 said *"I would definitely assign these in my class, because it's [efforts] you know, reread the article. And with the help of the program, I think these are pretty solid to get to start the conversation."* P4 said *"I actually think they are good to go. Except, like, I want to shuffle some of the option orders."* Instructors also said that the tool helped them create more meaningful questions (P6). P10 also said *"that's helpful to me to think about as a question writer of like, what is a good multiple choice question, what is a good prompting question?... Like if I sit there and I don't know what to do with it, then it lets me, okay, well, negate this or do something different."* P5 said *"I think the question I made on sub goals using this tool was probably better than what I would create offhand because the summarization gave me three pretty concise sentences on the sub goals."* Example questions created by participants during their sessions are shown in Figure 9. Participants created questions with figures, open-ended questions, multiple-choice and multiple-response questions that prompt students to think deeper behind the text.

6.1.2 Question creation is perceived as easier and quicker.

Most participants found ReadingQuizMaker to be time-saving compared to their usual instructional design processes. P13 said *"I can easily create questions after I complete reading the paper, I almost complete the questions right with answers. So this is definitely going to be like a time saver for me."* P12 said *"I do think that this was helpful in terms of the timeliness of it because it was faster to come up with and I thought the interface was fairly easy."* Participants also mentioned that reading or re-reading the content is time consuming. Multiple users said it would have been better if they read the text before the session, because they would usually have a higher level of familiarity with the content they will assign, thus creating questions faster.

Almost all participants mentioned that the interface was easy to use. P2 found the process to be smooth and easy to follow, and the Question Review Panel is helpful for them to see the big picture. P9 found the flow of text selection and adding to options to be "pretty clear". P3 said *"it's a new tool, it takes some time to just get acquaintance with, you know, all the basic things, but it was quick".* P8 said *"I think having the paper and also these other supports, really helped me, like ground me in creating these questions. And it also makes it easier for me to do so. Typically, I don't think I enjoyed creating questions before."* Participants found the system was designed to serve their natural flows and was flexible. P10 said *"you kind of saw my process that I find something and then make that an answer and then try to build the question backward from it. And I think the system's pretty flexible to be able to do that."*

6.2 RQ2: Instructors find the AI suggestions to be useful and desirable.

We first present a log data analysis on how instructors used the stem suggestions and AI suggestions provided by ReadingQuizMaker. Among the 89 questions created, users checked the question stem menu 52 times, and picked one from the menu 41 times. Users expressed appreciation for the question stems being suggested. P3 said *"I saw the recommendations, then I thought, Oh, well, why can't I use those recommended questions and build on those things?"* P8 thought the question bank gave him "inspiration of where to get started", and P10 found it "nice to inspire me without having to be created myself."

6.2.1 Users adopted 60% of AI suggestions.

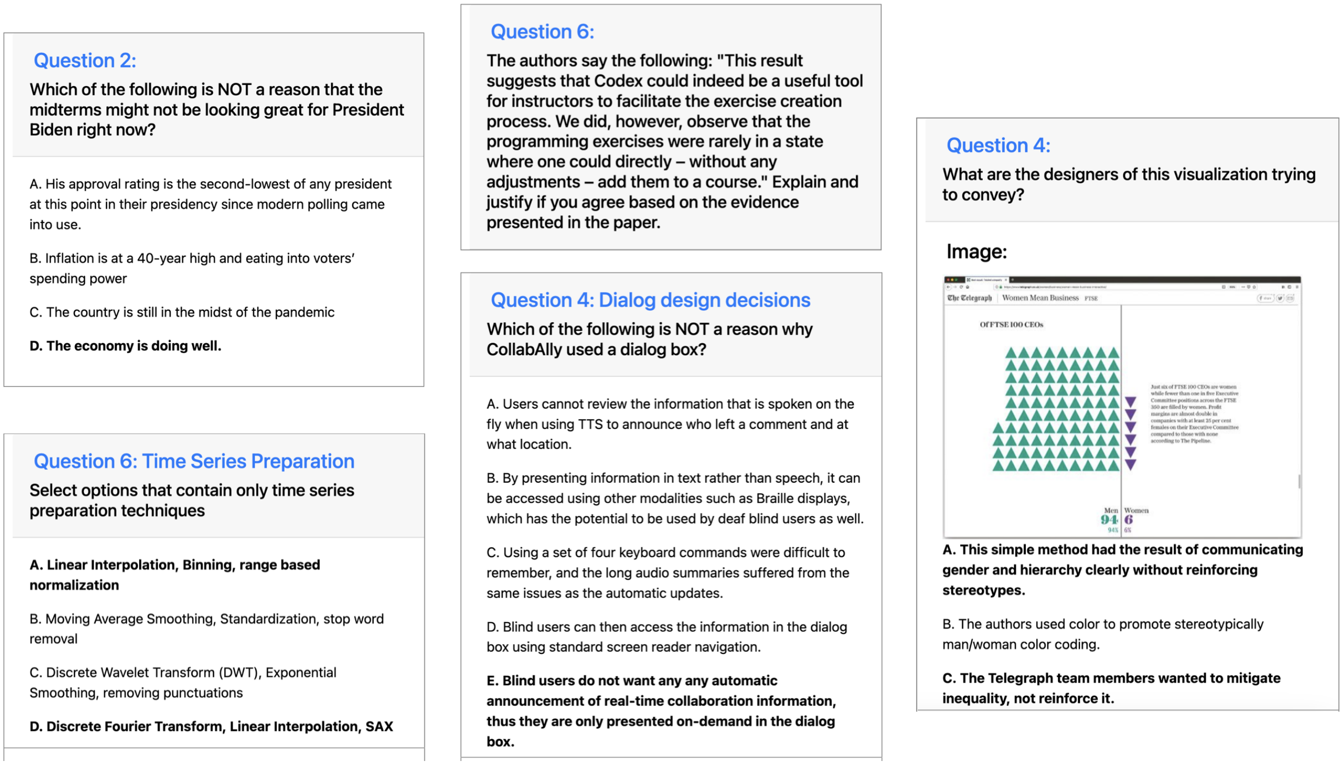
Summarization. In all 13 study sessions, the summarization model was triggered 37 times, and checked by the users 37 times. This means that every time the summarization model was triggered, the users checked to see the result. We found 19 out of the 37 suggestions were adopted as options, among which 10 were used directly, and 9 were split to multiple options. The reason that the summarization model was not triggered very often was because the user needed to select a whole paragraph or multiple paragraphs. In the study, we observed that users mostly selected sentences, for which the default suggestion was a paraphrase.

Paraphrase. The paraphrase model was triggered 197 times throughout all the user study sessions. Only 6 of them were triggered through the NLP Toolbox, and the rest were through the Paper Panel preview of the AI suggestions, as shown in Figure 2. We consider the preview of AI suggestions upon user selection to be successful in helping users discover AI support. 143 of the 191 paraphrase suggestions were checked by the user, which is about 75%, and 88 were accepted and adopted in question creation, which is 59% of the suggestions that were checked. For the paraphrase suggestions that were adopted, 23 were edited further by the user.

Negation. Across all the study sessions, 49 negation operations were applied and checked by the user, among which 29 were used, taking up 59% of the triggered ones. Most of the negations were triggered manually through the NLP Toolbox, and only 2 were automatically triggered when the user sent a new option. Instructors used the controllable version which required a keyword input 9 times, 5 of them were accepted. 3 of the 4 rejections were done on the same sentence as an input, where P8 wanted to negate a specific word in the sentence however the result wasn't as expected. P8 ended up negating it himself. After users applied the negation operation, some of them made further edits on the option. There were significant edits on 11 of them.

6.2.2 Instructors found the AI suggestions to be useful and inspiring.

Following Task 1, we asked participants to comment on the quality of the AI suggestions. Almost all participants commented that AI gave them useful suggestions and helped make question creation easier. Participants also mentioned that they would read the AI results first before using them. For example, P10 said *"I like the large summarization and the split. I think that's good, especially the way this paper is written that you can take one paragraph or two paragraphs and then get a lot of easy question stubs out of that so I*



Question 2:

Which of the following is NOT a reason that the midterms might not be looking great for President Biden right now?

A. His approval rating is the second-lowest of any president at this point in their presidency since modern polling came into use.

B. Inflation is at a 40-year high and eating into voters' spending power

C. The country is still in the midst of the pandemic

D. The economy is doing well.

Question 6: Time Series Preparation

Select options that contain only time series preparation techniques

A. Linear Interpolation, Binning, range based normalization

B. Moving Average Smoothing, Standardization, stop word removal

C. Discrete Wavelet Transform (DWT), Exponential Smoothing, removing punctuations

D. Discrete Fourier Transform, Linear Interpolation, SAX

Question 6:

The authors say the following: "This result suggests that Codex could indeed be a useful tool for instructors to facilitate the exercise creation process. We did, however, observe that the programming exercises were rarely in a state where one could directly – without any adjustments – add them to a course." Explain and justify if you agree based on the evidence presented in the paper.

Question 4: Dialog design decisions

Which of the following is NOT a reason why CollabAlly used a dialog box?

A. Users cannot review the information that is spoken on the fly when using TTS to announce who left a comment and at what location.

B. By presenting information in text rather than speech, it can be accessed using other modalities such as Braille displays, which has the potential to be used by deaf blind users as well.

C. Using a set of four keyboard commands were difficult to remember, and the long audio summaries suffered from the same issues as the automatic updates.

D. Blind users can then access the information in the dialog box using standard screen reader navigation.

E. Blind users do not want any any automatic announcement of real-time collaboration information, thus they are only presented on-demand in the dialog box.

Question 4:

What are the designers of this visualization trying to convey?

Image:

The Telegraph
Women Mean Business
FTSE
Of FTSE 100 CEOs
Just six of FTSE 100 CEOs are women while fewer than one in five Executive Committee positions across the FTSE 350 are filled by women. Profit margins are almost double in companies with at least 25 per cent females on their Executive Committee compared to those with none according to The Pipeline.
Men
Women
94
6
94%
6%

A. This simple method had the result of communicating gender and hierarchy clearly without reinforcing stereotypes.

B. The authors used color to promote stereotypically man/woman color coding.

C. The Telegraph team members wanted to mitigate inequality, not reinforce it.

Figure 9: Example questions the participants created using ReadingQuizMaker. This includes both multiple-choice and open-ended questions. Participants considered the questions to have satisfying educational value and aligned with their goals.

could see myself using that" P9 found the AI results to be reasonably satisfying, especially with longer inputs:"I think overall, I was reasonably satisfied with the AI." P12 in particular liked the paraphrase suggestions *"Sometimes I was actually very impressed with a lot of the paraphrasing that the AI did. Um, so I really would probably use that. I would read it first."*

We also asked participants to specifically comment on whether the AI suggestions are distracting or got in their natural way of thinking and creating questions. All participants shared that the AI suggestions were not distracting at all, and they had enough control to decide when and how to use the suggestions. P5 said *"I mean, it wasn't really distracting because like, I had to pull it up by clicking this transform menu, and then do the stuff"* Participants also shared that they were a bit worried about relying too much on the AI so that they needed to proof-read the AI suggestions well before adopting them. P6 said *"distracting? No, although I could see myself making an error by relying on it, because at least a couple of times I felt like it deleted a phrase in a way that changed the meaning."*

6.2.3 *Instructors further modify AI suggestions when the results are unsatisfying.*

Instructors also shared that sometimes the AI results were not satisfying. AI may generate half-baked drafts that they needed to further modify to make them usable. For example, P11 said *"Sometimes it worked well, sometimes it didn't like. sometimes it gave me a good half baked draft that I can take advantage of, so I don't think I would use it as is."* Participants shared that sometimes even if they found the results to be not satisfying, the AI suggestions gave them inspiration and helped them think alternatively. As an example, P4 created their own distractors based on the negation result. P6 said *"I'm sure that the AI also has a set number of things, but they are a different set of things than what I have. And so it's, really nice to provide those alternate versions"*

6.2.4 *Discoverability of AI is critical for adoption.*

During the study, we found that users were more likely to check and adopt AI suggestions when they were readily available and did not require extra actions from the users. For example, the paraphrase suggestion is automatically displayed once the user selects a sentence. For example, P2 selected a sentence, hoping to get the pronouns replaced. After seeing the paraphrase suggestion, they went with it instead. Similarly, P4 did not intend to paraphrase and did not wait for paraphrase suggestions to load at first, but gradually as they saw more paraphrase suggestions after making a selection, they found the paraphrases to be satisfying and went with them. Later in the study, P4 more intentionally wait for the paraphrase result to load or used the NLP Toolbox to apply the paraphrase operation. However, on the other hand, the summarization and negation operations take extra steps for users to apply and receive less attention. For summarization in particular, users need to select multiple paragraphs. P12 mentioned that in the middle of the study, she forgot how to apply the summarization operation.

Table 2: Example usage of the AI suggestions from the study. In some cases, users modified the AI suggestions after adopting them. The bolded texts in the "Result" column visualize the difference between the original text, and the AI-transformed text produced by the model. The bolded texts in the User "Modification" column indicates the user's modification based on the AI suggestion. One frequent type of user modification is to provide more context to make the text more specific and accurate (as shown in the last paraphrase example).

Operation	Original Text	Transformed Outcome	User Modification
Negation	How often did a subject "follow" a model's prediction?	How often did a **subject's prediction fail to match up to a model's**	How often did a subject's prediction fail to match up to a model's **prediction**
Negation	The above programming exercise, its solution, and the code explanation were all generated automatically by OpenAI codex	The above programming exercise, its solution, **but not** the code explanation were all generated automatically by OpenAI codex	**It can generate** programming exercise, its solution, but not the code **explanations**
Negation	Since World War II, the president's party has lost 17 House seats	Since World War II, the president's party has **won** 17 House seats	
Paraphrase	Higher income growth tends to be associated with better midterm outcomes for the president's party	The president's party tends to benefit from higher income growth	
Paraphrase	The author see a need for sports sensing interfaces to communicate the background knowledge necessary to understand seeing metrics and couple it closely to the visualisation. This was apparent in the dashboard interface. While textual descriptions of the metrics were available, none of the participants used them	In the dashboard interface, there was a need for background knowledge to be communicated to the participants	
Paraphrase	In the asymmetric system, the teachers' detachable spectator view made communication difficult for both participants as it was not obvious for them what exactly their peer was viewing at the moment	Communication was difficult for both participants as they didn't know what their peer was watching at the time	Communication was difficult for both participants **in the asymmetric system** as they didn't know what their **partner** was **attending to** at the time

6.3 RQ3: Instructors prefer the human-AI teaming approach provided by ReadingQuizMaker

At the end of the study, we asked the participants to compare their experiences with ReadingQuizMaker which uses a human-AI collaborative approach versus the automatic approach. 12 out of 13 instructors preferred the human-AI teaming approach provided by ReadingQuizMaker.

6.3.1 Having control is important. Participants shared that they liked the human-AI teaming approach because they felt they were more in control. For example, P11 said "*I would, again, prefer the first task rather than this one, because I would have control over what's being generated.*". P5 mentioned that "*it looks like I'm doing a little bit more on myself, but actually, that meets my expectation.*". Users wanted to be in control not only to keep the questions aligned with their goals, but also to avoid errors. As P1 said, *"if the questions generated were not accurate, then I will never use it. Because without myself to review it again. Because you know, like in education, any mistakes you made in the class will reduce your authority."*

6.3.2 Reviewing and editing automatically-generated questions limits my creativity. Although editing a list of pre-templated questions may seemingly reduce instructors' work, users think their imagination and creativity are limited after seeing the automatically generated questions.

P5 said "*I feel like it hinders my creativity, but it's like I see the easy route, and it's like the easy route is to just keep these questions at the remember, understand level.* ". Similarly, P8 mentioned *"there's a trade off, it's like, you know, if you present me with something, then I'm limited to thinking about these things... So you basically have a ceiling there for me. If I create from scratch, then I can try to create better ones."*

6.3.3 Automatically generated questions are of lower quality. Users find the automatically generated questions to be of lower quality compared to the ones they created through the ReadingQuizMaker system. One significant shortcoming is that the options in the automatically generated questions are not self-contained and are out of context. P4 said "*I feel like these sentences are being picked from different sections, but when you put them together and out of the context it's really hard to judge.*". Participants also shared concerns that even when the options were accurate, they did not make good question options since they did not target their teaching goals. P10 said *"I would, again, kind of look at this and say, for each of these it is representative of the text in there. But does it make sense standing alone as a question answer?"* Similarly p9 said *"I would not include such a question. There's no learning that's coming out of that. It's just a story.".* P8 commented *"These two are not relevant. They are the backgrounds, but I wouldn't include them in the question because I do not need the students to know the background.".* To summarize, the main drawback of the automated approach as revealed in the study is that, even though it is capable of generating logically sensible questions and options, the options could be out of context and do not align with the instructor's educational goals.

6.3.4 Some questions are good if I don't have time. Although most participants preferred to create questions themselves through the ReadingQuizMaker interface, some of them found the automatically generated questions to be acceptable if they had limited time. For example, P3 said "*I think the first question is great. And options are actually, you know, very analytical in sense*" Some participants found similarities between the automatically generated questions and the questions they created in Task 1. P8 mentioned that "*this is still pretty good. I think the question I created was also, like, I had this option, I had this option, but with more details.* " Users admitted that they might be willing to use these automatically generated questions if they did not have enough time.

6.4 RQ4: User challenge and experiences in ReadingQuizMaker and design implications

Even though participants in general found the ReadingQuizMaker system to be easy to use and help them create higher quality questions, they reported challenges in the process. Users generally found coming up with distractors to be challenging. P5 said " *I think distractors generation is like the hardest part of MCQ, both for like models and for humans.* ". Indeed, many instructors struggled when creating distractors. For example, P10 scrolled up and down, reading through sections for minutes to find one distractor. Another challenge participants experienced was finding a proper question stem. We saw in the study that the question stem suggestions were well utilized by participants. However, the question stem bank had a focus on academic papers. When participants design quiz questions for online tutorials, textbook chapters, and news articles, they did not find the question stem suggestions to be useful. P8 had an idea of what to ask about, but struggled to phrase the question stem. Similarly, P5 said "*knowing how I wanted the question to be worded*" is challenging. But he found the question stem suggestions to be helpful, saying "*it was cool that I got the auto complete to kind of give me ideas*".

7 DEMONSTRATIONS OF USE CASES

We demonstrate two scenarios where students can use the resulting questions from the ReadingQuizMaker system to learn.

7.1 Use as Formative or Summative Assessments through Existing Learning Management Platforms

The questions created in ReadingQuizMaker can be downloaded as a .csv file and converted to a .QTI package, which can be directly imported into existing learning management systems, such as Canvas. This enables instructors to assign the questions they created in the system as formative or summative assessments to students.

7.2 Use as a Reading Guide for Students

The questions created in ReadingQuizMaker can also be made into an interactive reading guide. Students can answer the generated questions through an extended ReadingQuizMaker interface that supports question answering. If they get a question wrong, the interface will give them feedback highlighting where the content locates in the reading, which was already logged in the question creation process.

8 DISCUSSION

In this section, we discuss potential future directions. In the evaluation study, we found ReadingQuizMaker to be well received by our participants. Instructors found the system easy and intuitive to use, the question stem suggestions and the AI suggestions useful, and that they were able to create high quality questions with the support from the system. Instructors also showed a strong preference for the human-AI teaming approach provided by ReadingQuizMaker compared to an automatic approach. Here we discuss some remaining challenges in instructors' question creation process and propose future directions to address them.

8.1 Increasing Discoverability, Visualization, and Explainability of AI Output

In the study, we found that users were more likely to check and adopt AI suggestions when they were readily available and did not require extra actions from the users. For example, the paraphrase suggestion, which is automatically displayed once the user selects a sentence, is more frequently checked and adopted compared to the summarization and negation operations. This aligns with the human-AI interaction guidelines to support efficient invocation, dismissal, and correction of AI outcomes. Additionally, we found that some users found it hard to parse the NLP outcomes, especially for summarization. The user needs to read the original paragraph and read the summary to make sure it is accurate. In ReadingQuizMaker, we implemented visualizations to highlight the changes for entity replacement and negation. However, we did not provide visualizations for summarization and paraphrase models since some involve dramatic changes. Future work needs to investigate better visualization techniques to help users perceive NLP outcomes more efficiently. Lastly, some users found it hard to understand how the model generated the result. This especially applies to the negation model. Some users were wondering why the model picked a certain word to negate, while others preferred the user-controllable version of the negation model we provided. Future work needs to explore techniques to explain generative models to end-users and provide methods for users to control an NLP model.

8.2 Human-AI Collaborative Approaches to Support Educational Content Creation

Our study joins prior work in demonstrating the power of human-AI collaborative approaches to support education [41, 42, 83]. We found strong user preference for Human-AI teaming to automatic approaches in question creation. In ReadingQuizMaker, the user has full control over the process. This not only gives the user a sense of security, but also makes the process more fluent for them. Some users said they would like to keep track of the process, e.g., what texts they have used, which part of the reading each option comes from, etc. On the other hand, users find the automatically generated questions to be of lower quality. One of the reasons is that it is hard for AI to identify what to focus on, i.e., what is an important piece of content to create questions for. When the input used for the AI models does not align with users' expectations, the paraphrased or negated sentences will lose context and are less satisfying as question options. The automated approach may generate logically correct multiple-choice questions, but the instructors find the options to be out of context and meaningless. We argue that for educational content authoring which requires expertise and creativity, human involvement to specify the input to AI systems is necessary and helps improve the adoption of AI suggestions. As large language models (LLM) such as ChatGPT become more prevalent and researchers have been exploring the use of LLMs to serve educational goals, our results offer important suggestions that for high-stakes tasks such as educational content creation, allowing users to provide input and giving users sufficient control over the process is more preferable to fully automated approaches.

Another direction is to use controllable AI models, where extra parameters are specified by the user. We implemented a simple controllable version for negation, which resulted in acceptable results. However, there were times where the negation result was not as satisfying. For example, the users may choose a specific word to be negated, whereas the model picked an adjacent word to negate. Future work needs to develop and incorporate controllable NLP models and help users specify parameters and obtain outcomes that align with their goals. For educational content creation, including a "focus" word for both negation and paraphrase models can reduce the risks of losing important information and generate results that match users' expectations.

8.3 Implications on Improving Question Creation Systems

One biggest challenge in creating questions is knowing what to ask. Some participants find the summarization provided by ReadingQuizMaker to be inspiring and give them ideas, but knowing what to summarize and how to convert the summarization result into a question is still challenging. Future work could explore ways to help people identify question opportunities. Many users find it hard to come up with good distractors. Although many participants used negation to generate distractors, coming up with a plausible incorrect option is still challenging. Based on the strategies people have displayed, we suggest future work to explore pulling out information from a specific location of the text, e.g., the related work section, to generate distractors. In the current version of ReadingQuizMaker, question stems are mainly designed for academic papers, and not tailored for other reading texts such as tutorials and textbooks. As users of ReadingQuizMaker increase and diversify, the crowdsourced question bank will grow and give users more versatile suggestions in the future. In ReadingQuizMaker, the user needs to wait 2-3 seconds for the AI suggestions to load. In the evaluation study, we did not observe users to be annoyed or distracted by the latency since they usually spent the time reading the original text. Future work that incorporates large language models in user interfaces needs to reduce latency and examine the potential effects on user experiences.

9 LIMITATION

1) The participant sample is small. In the future, we plan to run a larger study to understand how instructors use tools like ReadingQuizMaker to create questions. It also requires a longitudinal study to understand how instructors may develop trust with the system, and whether they become more proficient in using the system over time. 2) In the evaluation study, we asked users to self-report whether the approach saved their time on question creation. This is meant to serve as an investigation into how likely instructors are to adopt a system like ReadingQuizMaker in practice. A more comprehensive quantitative study is needed to fully investigate the time-saving factor of the system. 3) This work focused on a teacher-facing evaluation study where the goal was to investigate the usability and utility of ReadingQuizMaker from an instructor's perspective. A large-scale student-facing experiment is needed to understand whether the resulting questions are beneficial for students' reading comprehension and learning.

10 CONCLUSION

We propose ReadingQuizMaker that supports instructors to conveniently design high-quality questions to help students comprehend readings. ReadingQuizMaker adapts to instructors' natural workflows of creating questions, while providing NLP-based process-oriented support. ReadingQuizMaker enables instructors to decide when and which NLP models to use, select the input to the models, and edit the outcomes. In an evaluation study, instructors found the resulting questions comparable to their previously designed quizzes. Instructors praised ReadingQuizMaker for ease of use, and considered the NLP suggestions to be satisfying and helpful. We compared ReadingQuizMaker with a baseline condition where instructors were given automatically generated questions to edit. Instructors showed a strong preference for the human-AI teaming approach provided by ReadingQuizMaker. Compared to ReadingQuizMaker, instructors found the automatically generated questions to be of lower quality and the content in the questions to be out of context and meaningless. Our findings offer important suggestions for the use of large language models to support education. We argue that for high-stakes tasks such as educational content creation, allowing users to provide input and giving users sufficient control over the process is more preferable to fully automated approaches.

ACKNOWLEDGMENTS

We would like to thank our participants for sharing their thoughts and feedback on our system, and the reviewers for their valuable feedback.

REFERENCES

[1] [n.d.]. Amazon Web Services (AWS). https://aws.amazon.com/ Accessed: 2022-09-15.

[2] [n.d.]. Digital Ocean. https://www.digitalocean.com/ Accessed: 2022-09-15.

[3] [n.d.]. Springer. https://link.springer.com/ Accessed: 2022-09-15.

[4] [n.d.]. Taylor & Francis Online. https://www.tandfonline.com/ Accessed: 2022-09-15.

[5] [n.d.]. The Washington Post. https://www.washingtonpost.com/ Accessed: 2022-09-15.

[6] ACM. 2022. The ACM Publishing Systems (TAPS) Best Practices. https://www.acm.org/publications/taps/taps-best-practices

[7] Hamed S. Alavi and Pierre Dillenbourg. 2012. An Ambient Awareness Tool for Supporting Supervised Collaborative Problem Solving. *IEEE Transactions on Learning Technologies* 5, 3 (2012), 264–274. https://doi.org/10.1109/TLT.2012.7

[8] Tahani Alsubait, Bijan Parsia, and Ulrike Sattler. 2016. Ontology-based multiple choice question generation. *KI-Künstliche Intelligenz* 30, 2 (2016), 183–188.

[9] Susan A Ambrose, Michael W Bridges, Michele DiPietro, Marsha C Lovett, and Marie K Norman. 2010. *How learning works: Seven research-based principles for smart teaching*. John Wiley & Sons.

[10] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300233

[11] Eric Whitmire Amy J. Ko, Jacob O. Wobbrock. 2022. *User Interface Software and Technology*. Retrieved September 15, 2022 from https://faculty.washington.edu/ajko/books/user-interface-software-and-technology/

[12] Russell Baker, Jeffery Harrison, Barry Thornton, and Rhett Yates. 2011. An Analysis Of The Effectiveness Of Podcasting As A Supplemental Instructional Tool: A Pilot Study. *College Teaching Methods and Styles Journal* 4 (03 2011). https://doi.org/10.19030/ctms.v4i3.5535

[13] Thomas Berry, Lori Cook, Nancy Hill, and Kevin Stevens. 2010. An exploratory analysis of textbook usage and study habits: Misperceptions and barriers to success. *College Teaching* 59, 1 (2010), 31–39.

[14] TAXONOMY MADE EASY BLOOM'S. 1965. *Bloom's taxonomy of educational objectives*. Longman.

[15] Karen Wilken Braun and R Drew Sellers. 2012. Using a "daily motivational quiz" to increase student preparation, attendance, and participation. *Issues in Accounting Education* 27, 1 (2012), 267–279.

[16] Virginia Braun and Victoria Clarke. 2012. Thematic analysis. (2012).

[17] Brian Brost and Karen Bradley. 2006. Student compliance with assigned reading: A case study. *Journal of the Scholarship of Teaching and Learning* (2006), 101–111.

[18] Colin M Burchfield and John Sappington. 2000. Compliance with required reading assignments. *Teaching of Psychology* (2000).

[19] David M. Carkenord. 1994. Motivating Students to Read Journal Articles. *Teaching of Psychology* 21, 3 (1994), 162–164. https://doi.org/10.1177/009862839402100309 arXiv:https://doi.org/10.1177/009862839402100309

[20] Amy G Carney, Sara Winstead Fry, Rosaria V Gabriele, and Michelle Ballard. 2008. Reeling in the big fish: Changing pedagogy to encourage the completion of reading assignments. *College Teaching* 56, 4 (2008), 195–200.

[21] Yllias Chali and Sadid A Hasan. 2015. Towards topic-to-question generation. *Computational Linguistics* 41, 1 (2015), 1–20.

[22] Chih-Ming Chen and Sheng-Hui Huang. 2014. Web-based reading annotation system with an attention-based self-regulated learning mechanism for promoting reading performance. *British Journal of Educational Technology* 45, 5 (2014), 959–980.

[23] Michelene TH Chi and Ruth Wylie. 2014. The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist* 49, 4 (2014), 219–243.

[24] Michael A Clump, Heather Bauer, and Catherine Bradley. 2004. The extent to which psychology students read textbooks: A multiple class analysis of reading across the psychology curriculum. *Journal of Instructional Psychology* 31, 3 (2004), 227–233.

[25] Andy Coenen, Luke Davis, Daphne Ippolito, Emily Reif, and Ann Yuan. 2021. Wordcraft: a Human-AI Collaborative Editor for Story Writing. *ArXiv* abs/2107.07430 (2021).

[26] Patricia A Connor-Greene. 2000. Assessing and promoting student learning: Blurring the line between teaching and testing. *Teaching of Psychology* 27, 2 (2000), 84–88.

[27] Bidyut Das, Mukta Majumder, Santanu Phadikar, and Arif Ahmed Sekh. 2021. Automatic question generation and answer assessment: a survey. *Research and Practice in Technology Enhanced Learning* 16 (2021), 1–15.

[28] Richard R Day and Jeong-suk Park. 2005. Developing Reading Comprehension Questions. *Reading in a foreign language* 17, 1 (2005), 60–73.

[29] Cynthia S Deale and Seung Hyun Lee. 2022. To read or not to read? Exploring the reading habits of hospitality management students. *Journal of Hospitality & Tourism Education* 34, 1 (2022), 45–56.

[30] Django Software Foundation. [n.d.]. *Django*. https://djangoproject.com

[31] Facebook. [n.d.]. React - A JavaScript library for building user interfaces. 2018. React-AJavaScriptlibraryforbuildinguserinterfaces.(2018).https://reactjs.org/ Accessed: 2022-09-15.

[32] Nancy W Fordham. 2006. Crafting questions that address comprehension strategies in content reading. *Journal of Adolescent & Adult Literacy* 49, 5 (2006), 390–396.

[33] Pamela Fox. 2022. The Benefits of HTML Slides for Programming Lectures. In *Proceedings of the 53rd ACM Technical Symposium on Computer Science Education V. 2*. 1055–1055.

[34] Scott Freeman, Sarah L Eddy, Miles McDonough, Michelle K Smith, Nnadozie Okoroafor, Hannah Jordt, and Mary Pat Wenderoth. 2014. Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the national academy of sciences* 111, 23 (2014), 8410–8415.

[35] Sarah J. Hatteberg and Kody Steffy. 2013. Increasing Reading Compliance of Undergraduates: An Evaluation of Compliance Methods. *Teaching Sociology* 41, 4 (2013), 346–352. https://doi.org/10.1177/0092055X13490752 arXiv:https://doi.org/10.1177/0092055X13490752

[36] Andrew Head, Kyle Lo, Dongyeop Kang, Raymond Fok, Sam Skjonsberg, Daniel S Weld, and Marti A Hearst. 2021. Augmenting scientific papers with just-in-time, position-sensitive definitions of terms and symbols. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–18.

[37] Andrew Head, Amber Xie, and Marti A Hearst. 2022. Math Augmentation: How Authors Enhance the Readability of Formulas using Novel Visual Design Practices. In *CHI Conference on Human Factors in Computing Systems*. 1–18.

[38] Neil T Heffernan and Cristina Lindquist Heffernan. 2014. The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education* 24, 4 (2014), 470–497.

[39] 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*. Association for Computational Linguistics, Los Angeles, California, 609–617. https://www.aclweb.org/anthology/N10-1086

[40] Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching Machines to Read and Comprehend. In *Advances in Neural Information Processing Systems (NIPS)*. http://arxiv.org/abs/1506.03340

[41] Kenneth Holstein, Vincent Aleven, and Nikol Rummel. 2020. A conceptual framework for human–AI hybrid adaptivity in education. In *International conference on artificial intelligence in education*. Springer, 240–254.

[42] Kenneth Holstein, Bruce M. McLaren, and Vincent Aleven. 2018. Student Learning Benefits of a Mixed-Reality Teacher Awareness Tool in AI-Enhanced Classrooms. In *Artificial Intelligence in Education*, Carolyn Penstein Rosé, Roberto Martínez-Maldonado, H. Ulrich Hoppe, Rose Luckin, Manolis Mavrikis, Kaska Porayska-Pomsta, Bruce McLaren, and Benedict du Boulay (Eds.). Springer International Publishing, Cham, 154–168.

[43] Cathy HC Hsu and Minglong Li. 2017. Effectiveness and usage frequency of learning methods and tools: Perceptions of hospitality students in Hong Kong. *Journal of Hospitality & Tourism Education* 29, 3 (2017), 101–115.

[44] SuHua Huang, Matthew Capps, Jeff Blacklock, and Mary Garza. 2014. Reading Habits of College Students in the United States. *Reading Psychology* 35, 5 (2014), 437–467. https://doi.org/10.1080/02702711.2012.739593 arXiv:https://doi.org/10.1080/02702711.2012.739593

[45] Lars Jung. 2022. pagemap. https://larsjung.de

[46] Mary Margaret Kerr and Kristen M Frese. 2017. Reading to learn or learning to read? Engaging college students in course readings. *College teaching* 65, 1 (2017), 28–31.

[47] Rafal Kocielnik, Saleema Amershi, and Paul N. Bennett. 2019. Will You Accept an Imperfect AI? Exploring Designs for Adjusting End-User Expectations of AI Systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300641

[48] Kenneth R Koedinger, Albert T Corbett, and Charles Perfetti. 2012. The Knowledge-Learning-Instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive science* 36, 5 (2012), 757–798.

[49] 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, 1 (2020), 121–204.

[50] Robert N Leamnson. 1999. *Thinking about teaching and learning: Developing habits of learning with first year college and university students.* Stylus Publishing, LLC.

[51] Simon A Lei, Kerry A Bartlett, Suzanne E Gorney, and Tamra R Herschbach. 2010. Resistance to Reading Compliance Among College Students: Instructors' Perspectives. *College Student Journal* 44, 2 (2010).

[52] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *ACL*.

[53] Yang Liu and Mirella Lapata. 2019. Text Summarization with Pretrained Encoders. *ArXiv* abs/1908.08345 (2019).

[54] Mukta Majumder and Sujan Kumar Saha. 2014. Automatic selection of informative sentences: The sentences that can generate multiple choice questions. *Knowledge Management & E-Learning: An International Journal* 6 (2014), 377–391.

[55] Mukta Majumder and Sujan Kumar Saha. 2015. A System for Generating Multiple Choice Questions: With a Novel Approach for Sentence Selection. In *NLP-TEA@ACL/IJCNLP*.

[56] Roberto Martinez-Maldonado. 2019. A handheld classroom dashboard: Teachers' perspectives on the use of real-time collaborative learning analytics. *International Journal of Computer-Supported Collaborative Learning* 14, 3 (2019), 383–411.

[57] Roberto Martinez-Maldonado, Andrew Clayphan, Kalina Yacef, and Judy Kay. 2015. MTFeedback: Providing Notifications to Enhance Teacher Awareness of Small Group Work in the Classroom. *IEEE Transactions on Learning Technologies* 8, 2 (2015), 187–200. https://doi.org/10.1109/TLT.2014.2365027

[58] Emma Mercier. 2016. Teacher orchestration and student learning during mathematics activities in a smart classroom. *International Journal of Smart Technology and Learning* 1, 1 (2016), 33–52.

[59] Frances Miley and Andrew Read. 2011. Using word clouds to develop proactive learners. *J. Scholar. Teach. and Learn.* 11 (01 2011).

[60] Kelly Miller, Brian Lukoff, Gary King, and Eric Mazur. 2018. Use of a social annotation platform for pre-class reading assignments in a flipped introductory physics class. In *Frontiers in education*. Frontiers, 8.

[61] Bill Moggridge and Bill Atkinson. 2007. *Designing interactions*. Vol. 17. MIT press Cambridge.

[62] Teresa Murden and Cindy S Gillespie. 1997. The role of textbooks and reading in content area classrooms: What are teachers and students saying. *Exploring literacy* (1997), 87–96.

[63] Rebekah Nathan. 2006. *My freshman year: What a professor learned by becoming a student.* Penguin.

[64] Kenton O'Hara and Abigail Sellen. 1997. A Comparison of Reading Paper and On-Line Documents. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '97)*. Association for Computing Machinery, New York, NY, USA, 335–342. https://doi.org/10.1145/258549.258787

[65] Andrew M Olney, Arthur C Graesser, and Natalie K Person. 2012. Question generation from concept maps. *Dialogue & Discourse* 3, 2 (2012), 75–99.

[66] Jennifer Pearson, George Buchanan, and Harold Thimbleby. 2011. The Reading Desk: Applying Physical Interactions to Digital Documents. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 3199–3202. https://doi.org/10.1145/1978942.1979416

[67] Laura W Perna. 2010. Understanding the working college student. *Academe* 96, 4 (2010), 30–33.

[68] Perusall. 2021. Perusall. https://perusall.com/

[69] Leonard Richardson. 2007. Beautiful soup documentation. *April* (2007).

[70] Nikol Rummel and Kenneth R. Koedinger. 2014. Adaptive Intelligent Support to Improve Peer Tutoring in Algebra. *International Journal of Artificial Intelligence in Education* 24 (2014). https://doi.org/10.1007/s40593-013-0001-9

[71] André A Rupp, Tracy Ferne, and Hyeran Choi. 2006. How assessing reading comprehension with multiple-choice questions shapes the construct: A cognitive processing perspective. *Language testing* 23, 4 (2006), 441–474.

[72] Tracey E Ryan. 2006. Motivating novice students to read their textbooks. *Journal of Instructional psychology* 33, 2 (2006).

[73] John Sappington, Kimberly Kinsey, and Kirk Munsayac. 2002. Two studies of reading compliance among college students. *Teaching of psychology* 29, 4 (2002), 272–274.

[74] Sami Sarsa, Paul Denny, Arto Hellas, and Juho Leinonen. 2022. Automatic Generation of Programming Exercises and Code Explanations Using Large Language Models. In *Proceedings of the 2022 ACM Conference on International Computing Education Research-Volume 1*. 27–43.

[75] Aalok Sathe, Salar Ather, Tuan Manh Le, Nathan Perry, and Joonsuk Park. 2020. Automated Fact-Checking of Claims from Wikipedia. In *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 6874–6882. https://aclanthology.org/2020.lrec-1.849

[76] Emily Schnee. 2018. Reading across the curriculum at an urban community college: Student and faculty perspectives on reading. *Community College Journal of Research and Practice* 42, 12 (2018), 825–847.

[77] Daniel L. Schwartz, Jessica M. Tsang, and Kristen P. Blair. 2016. *The ABCs of How We Learn.* W. W. Norton & Company, New York, NY, USA.

[78] Jia Tracy Shen, Michiharu Yamashita, Ethan Prihar, Neil Heffernan, Xintao Wu, Ben Graff, and Dongwon Lee. 2021. Mathbert: A pre-trained language model for general nlp tasks in mathematics education. *arXiv preprint arXiv:2106.07340* (2021).

[79] Sheng Shen, Yaliang Li, Nan Du, X. Wu, Yusheng Xie, Shen Ge, Tao Yang, Kai Wang, Xingzheng Liang, and Wei Fan. 2020. On the Generation of Medical Question-Answer Pairs. In *AAAI*.

[80] Craig S. Tashman and W. Keith Edwards. 2011. LiquidText: A Flexible, Multitouch Environment to Support Active Reading. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 3285–3294. https://doi.org/10.1145/1978942.1979430

[81] Xu Wang, Simin Fan, Jessica Houghton, and Lu Wang. 2022. Towards Process-Oriented, Modular, and Versatile Question Generation that Meets Educational Needs. *arXiv preprint arXiv:2205.00355* (2022).

[82] Xu Wang, Carolyn Rose, and Ken Koedinger. 2021. Seeing Beyond Expert Blind Spots: Online Learning Design for Scale and Quality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.

[83] Xu Wang, Srinivasa Teja Talluri, Carolyn Penstein Rosé, and K. Koedinger. 2019. UpGrade: Sourcing Student Open-Ended Solutions to Create Scalable Learning Opportunities. *Proceedings of the Sixth (2019) ACM Conference on Learning @ Scale* (2019).

[84] Zichao Wang, Andrew S. Lan, Weili Nie, Andrew E. Waters, Phillip J. Grimaldi, and Richard G. Baraniuk. 2018. QG-Net: A Data-Driven Question Generation Model for Educational Content. In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale* (London, United Kingdom) *(L@S '18)*. Association for Computing Machinery, New York, NY, USA, Article 7, 10 pages. https://doi.org/10.1145/3231644.3231654

[85] Zichao Wang, Jakob Valdez, Debshila Basu Mallick, and Richard G. Baraniuk. 2022. Towards Human-Like Educational Question Generation with Large Language Models. In *Artificial Intelligence in Education*, Maria Mercedes Rodrigo, Noburu Matsuda, Alexandra I. Cristea, and Vania Dimitrova (Eds.). Springer International Publishing, Cham, 153–166.

[86] Daijin Yang, Yanpeng Zhou, Zhiyuan Zhang, Toby Jia-Jun Li, and LC Ray. 2022. AI as an Active Writer: Interaction Strategies with Generated Text in Human-AI Collaborative Fiction Writing 56-65. In *IUI Workshops*.

[87] Kexin Bella Yang, LuEttaMae Lawrence, Vanessa Echeverria, Boyuan Guo, Nikol Rummel, and Vincent Aleven. 2021. Surveying Teachers' Preferences and Boundaries Regarding Human-AI Control in Dynamic Pairing of Students for Collaborative Learning. In *European Conference on Technology Enhanced Learning*. Springer, 260–274.

[88] Kexin Bella Yang, Zijing Lu, Vanessa Echeverria, Jonathan Sewall, Luettamae Lawrence, Nikol Rummel, and Vincent Aleven. 2022. Technology Ecosystem for Orchestrating Dynamic Transitions Between Individual and Collaborative AI-Tutored Problem Solving. In *International Conference on Artificial Intelligence in Education*. Springer, 673–678.

[89] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-Examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376301

[90] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-examining whether, why, and how human-AI interaction is uniquely difficult to design. In *Proceedings of the 2020 chi conference on human factors in computing systems*. 1–13.

[91] Iman Yeckehzaare, Tirdad Barghi, and Paul Resnick. 2020. QMaps: Engaging Students in Voluntary Question Generation and Linking *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376882

[92] Ke Yuan, Dafang He, Zhuoren Jiang, Liangcai Gao, Zhi Tang, and C. Lee Giles. 2020. Automatic Generation of Headlines for Online Math Questions. In *AAAI*.

[93] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. arXiv:1912.08777 [cs.CL]