Towards Process-Oriented, Modular, and Versatile Question Generation that Meets Educational Needs

Xu Wang **Simin Fan** **Jessica Houghton** **Lu Wang**
Computer Science and Engineering
University of Michigan
Ann Arbor, MI
{xwanghci, oliviaaa, houghj, wangluxy}@umich.edu

Abstract

NLP-powered automatic question generation (QG) techniques carry great pedagogical potential of saving educators' time and benefiting student learning. Yet, QG systems have not been widely adopted in classrooms to date. In this work, we aim to pinpoint key impediments and investigate how to improve the usability of automatic QG techniques for educational purposes by understanding how instructors construct questions and identifying touch points to enhance the underlying NLP models. We perform an in-depth need finding study with 11 instructors across 7 different universities, and summarize their thought processes and needs when creating questions. While instructors show great interests in using NLP systems to support question design, none of them has used such tools in practice. They resort to multiple sources of information, ranging from domain knowledge to students' misconceptions, all of which missing from today's QG systems. We argue that building effective human-NLP collaborative QG systems that emphasize instructor control and explainability is imperative for real-world adoption. We call for QG systems to provide process-oriented support, use modular design, and handle diverse sources of input.

1 Introduction

Decades of educational research has shown the benefit of *active learning* in improving students' learning outcomes where students actively engage with the materials, e.g., through question answering and problem solving (Crouch and Mazur, 2001; Deslauriers et al., 2019; Koedinger et al., 1997), compared with passively receiving lectures or reading texts (Chi and Wylie, 2014; Freeman et al., 2014). However, most instructors still use traditional teaching methods, e.g., lecturing, in large-enrolment college courses (Henderson and Dancy, 2007; Handelsman et al., 2004; Stains et al., 2018), due to their limited time and resources in creating opportunities for active learning (Dancy and Henderson, 2007; Silverthorn et al., 2006; Fagen et al., 2002). Meanwhile, there is a growing interest in leveraging NLP methods to serve educational needs. Specifically, automatic question generation (QG) is promising in enhancing students' active learning experience by creating problem-solving activities at scale (Alsubait et al., 2016).

Still, the adoption of automatic QG systems in classrooms is low (Kurdi et al., 2020; Alsubait et al., 2016), mainly because those models are only suitable for specific domains, e.g., language learning, and the generated questions are often of low quality and limited in types and difficulty levels (Kurdi et al., 2020; Alsubait et al., 2016). For QG systems to provide more meaningful support to instructors' work, we need to address at least two fundamental challenges. First, designing questions to assist learning is a highly *complex, creative, and knowledge intensive* process. Typical college instructors have received years of training on the subject domain, and their question construction process is hard to be fully automated by one single model. Second, classroom is a high-stake environment, and instructors often have *pre-defined goals*. As a result, they may not prefer imperfect and fallible AI models (Holstein and Aleven, 2021).

To fill the gap between the design of NLP-powered QG systems and the reality in classrooms, we conducted a novel **need finding study**, with two major goals: (i) To *understand how instructors create questions* of high educational value, including decision making processes and information sources used throughout the procedure. (ii) To identify key touch points to *improve, design, and re-create NLP systems* that instructors would find useful in their practice.

Concretely, we interviewed and observed 11 instructors from 7 different universities as they created questions based on readings to support student learning of the content. We piloted the interview study protocol for several rounds, before finalizing a **two-phase protocol**. In the first phase, we conducted an in-depth case study of one instructor's quiz design experience for a semester-long course. We extracted the instructor's output and manually labeled the context of each action. We then propose a novel method: **text replay enactment**, where we replayed the instructor's process of generating the output and annotated what NLP tasks could be used to accomplish the instructor's text transformation operations. This method successfully addresses the challenge that users are unable to directly articulate what they want from AI (Yang et al., 2020). Next, we summarized a list of NLP tasks that this instructor found useful, and used them as design probes in a subsequent interview study with 10 instructors. We then analyzed our data using affinity diagrams. On a subset of these tasks, we reported a qualitative analysis of current NLP models' performance and pointed out major issues that should be addressed to allow classroom adoption.

Our major findings include: 1) None of the instructors interviewed are currently using any type of automatic question generation tools, though they expressed a strong desire to find better ways to design quality questions to support active learning and effective reading. 2) Instructors' natural behaviors and thought processes reveal multiple reasons why existing automatic QG techniques fall short: They leverage multiple sources of information that cover domain knowledge, educational goals, and student misconceptions; They view question creation as an iterative process; They often apply a suite of techniques and strategies. All of these traits are largely ignored by existing end-to-end QG models. 3) Finally, instructors also reported challenges when constructing questions, and showed strong interests in receiving suggestions and support from NLP systems if the outputs are of high quality and interpretable, and that they have sufficient control when interacting with the system.

Based on these findings, we argue that developing *effective human-NLP collaborative QG systems* is a promising direction, and propose **three design implications**. First, it is critical to provide *process-oriented support* to instructors, rather than automating the process end to end. Second, we advocate for *modular system design with different NLP components* that give instructors strong control over which NLP components to use to meet their goals. Third, we argue that future QG systems should accept *diverse data input*, including expert domain knowledge, students' background, educational goals, and pedagogical context.

Finally, we stress the value of understanding stakeholders' needs to better design and deploy NLP systems. As anticipated, users have concerns about how well NLP can do, and are hesitant to use AI-based tools in high-stake environments. Therefore, an important first step in developing human-NLP collaborative systems is to understand user needs and increase user buy-in. As we learned from the interviews, using simple models with reliable performance in the beginning while requesting more user input may be a good approach to bootstrap. We also encourage our community to collaborate with HCI researchers to develop user-friendly interfaces and investigate users' adoption and preferences of NLP tools in context of use.[1]

2 Related Work

2.1 Automatic Question Generation (QG)

Here we focus on describing QG models for educational purposes, and refer readers to a detailed literature review in Kurdi et al. (2020). Existing QG systems primarily produce questions that target *low-level cognitive skills*, such as factual wh-questions and cloze questions that can be answered with short phrases (Heilman and Smith, 2010; Chali and Hasan, 2015; Olney et al., 2012). These simple questions only contain limited concepts (Song and Zhao, 2016) and, as a consequence, offer limited opportunities to control question difficulty level (Alsubait et al., 2016) and are limited for assessment only, leaving many learning opportunities unfulfilled. Additionally, prior work largely develops domain-specific techniques for learning language, math, and medicine knowledge, relying on existing domain ontology (Alsubait et al., 2016; Stasaski and Hearst, 2017).

Our study investigates multiple-choice question (MCQ) generation (Ch and Saha, 2018). MCQs can be graded automatically and offers immediate feedback to students, which has profound benefits to scale active learning. Prior research has shown compelling results on the educational value of MCQs in comparison to open-ended question items. For example, Smith and Karpicke found that students performed equally well on English reading tasks no matter whether they practiced with multiple-choice, short-answer, or hybrid questions

[1] Data, code, and models used in this paper are released at: https://github.com/Olivia-fsm/P2MCQ.

(Smith and Karpicke, 2014). Similarly, multiple-choice questions are shown to provide a win-win situation compared to open-ended cued-recall tests on English reading tasks (Little and Bjork, 2015; Little et al., 2012). The authors find that both open-ended and cued-recall tests foster retention of previously tested information, but multiple-choice tests also facilitated recall of information pertaining to incorrect alternatives, whereas cued-recall tests did not. Recent work (Wang et al., 2021) demonstrates that even for less well-defined domains, MCQs could exercise critical thinking elements and require students to evaluate the quality of MCQ options, especially when common student misconceptions were used as distractors (Wang et al., 2019).

This poses unique challenges in automatic generation of MCQs. First, creating rich-content and meaningful correct options that help students retrieve correct and relevant information. Second, generating plausible distractors that help students internalize incorrect information. Well-designed and meaningful distractors require students to evaluate and contrast options (Wang et al., 2019), that involves going beyond surface traits and considering deeper connecting principles (Schwartz et al., 2011). However, only phrase-level replacement has been commonly studied (Papasalouros et al., 2008) in automatic generation of MCQ options and distractors. Recent automatic QG systems are built on end-to-end trained neural generation models, where a single question is generated from a given context (Sun et al., 2018; Zhou et al., 2019; Zhang and Bansal, 2019). Although questions that require multi-hop reasoning (Pan et al., 2020; Su et al., 2020) or reading long text (Bi et al., 2021; Cheng et al., 2021; Cao and Wang, 2021) have been studied, these systems do not offer control over question difficulty nor consider different sources of knowledge. In this work, we propose to modularize the automatic QG process with different components, e.g., summarization, simplification, or contradiction generation, to offer flexible interface for instructors to control various aspects of the produced questions.

2.2 NLP Tasks for Education

NLP systems have been developed to support a variety of educational applications, including writing assistance (Bellino and Bascuñán, 2020; Frankenberg-Garcia et al., 2018), reading comprehension support (Ross et al., 1991; Vodolazova and Lloret, 2019; Vajjala and Lucic, 2019; Siddharthan and Katsos, 2010; Chatzipanagiotidis et al., 2021), language learning systems (Tweissi, 1998; Petersen and Ostendorf, 2007; Katinskaia and Yangarber, 2021; Üksik et al., 2021), and generating feedback for programming and design (Wang et al., 2018; Kang et al., 2019). Here we provide a brief overview of the state-of-the-art NLP models that are relevant to the question generation process.

Summarization, the ability of condensing a reading material into a concise passage, has been used for evaluating and improving students' reading comprehension ability (Edmonds et al., 2009; Vaughn et al., 2011; Stevens et al., 2019; Hwang et al., 2019). Similarly, **paraphrasing** also demonstrates students' content comprehension skills (Haynes and Fillmer, 1984) since it requires the ability of conveying the same semantic meaning with different languages. However, both tasks are rarely studied for question generation (Lyu et al., 2021). **Simplification**, on the other hand, has demonstrated its values in language learning and reading comprehension (Tweissi, 1998; Inui et al., 2003; Petersen and Ostendorf, 2007; Rets and Rogaten, 2021), and is often treated as a preprocessing step to convert complex sentences into simpler versions before creating questions (Majumder and Saha, 2015; Patra and Saha, 2019). State-of-the-art NLP models for these three tasks are all based on neural generation systems, which are known to suffer from errors and lack of controllability (Maynez et al., 2020). Therefore, their usefulness for QG will largely depend on output quality and new designs with enhanced explainability and easy-to-use interface.

2.3 Human-AI Systems for Education

The concept of human-AI systems is recently introduced into the education domain, where human inputs are continuously solicited and there exists a collaborative relationship between humans (often teachers) and AI algorithms to provide effective instruction to students. They differ from prior AI for education systems where human (teacher and student) input is often elicited before the development of the system (Koedinger et al., 1997; Kumar et al., 2007). Human-AI systems have been developed to help human teachers more easily identify the students that are struggling (Holstein et al., 2018), design higher quality assignment questions (Wang et al., 2019), and offer aggregated feedback to students' programs (Glassman et al., 2015). They all demonstrate the advantages of combining both humans (robust and flexible) and AI (low-cost and quick) in addressing real-world challenges in education, and it inspires us to develop human-AI systems for QG.

Also relevant to this work is the abundant literature on human-AI interaction (Horvitz, 1999; Amershi et al., 2019; Yang et al., 2020), which highlights the need for considering user's goals and behaviors and points to general design guidelines when prototyping and designing algorithmic experiences. We consider this work to be an instantiation of the idea in a specific context, examining human-NLP interaction when instructors design questions for educational purposes. We expect our findings to contribute to the future development of human-AI systems that address this specific educational problem, and to the greater body of work on human-AI interaction and design principles.

3 Need Finding Study Methodology

We wanted to understand how instructors construct questions that align with their educational needs. We also wanted to probe on when instructors think an intelligent system might offer support or their work. To address these needs, we chose to conduct a qualitative study consisting of observations and semi-structured interviews. It has become a standard HCI approach when designing new software systems that address human needs.

In this work, we conducted an IRB-approved two-phase need finding study. We went through rounds of piloting before reaching at the final protocol, which addressed two challenges that emerged during the pilots. 1) An interview-alone approach is insufficient since it is hard for instructors to recollect all details of question design. In addition, it requires concrete textual input for us to understand how humans use texts when designing questions. 2) Instructors are not able to directly articulate their NLP needs (Yang et al., 2020). To address the first challenge, we propose a specific scenario where instructors design multiple-choice style quiz questions to support active reading. On one hand, it targets an authentic problem since text reading is a passive learning experience that occurs everyday in college classrooms. Active reading opportunities are much needed to support students' comprehension and learning. On the other hand, it is a question design scenario where the text input is explicit. To address the second challenge, we propose a text replay enactment method where we, as NLP researchers, review users' operation on the text and annotate possible NLP tasks that can be applied to make the text transformation.

3.1 Phase 1: Case Study and Replay Enactment of Text

The case study contains an in-depth analysis of one instructor's quiz design experience for a semester-long course on human-computer interaction at the University of Michigan. We will refer to the course as HCI101 for the rest of this paper. The course has 1-2 required readings per lecture, including academic papers and book chapters. The instructor (Instructor A) designed quiz questions for each reading text to facilitate students' reading and understanding of the material.

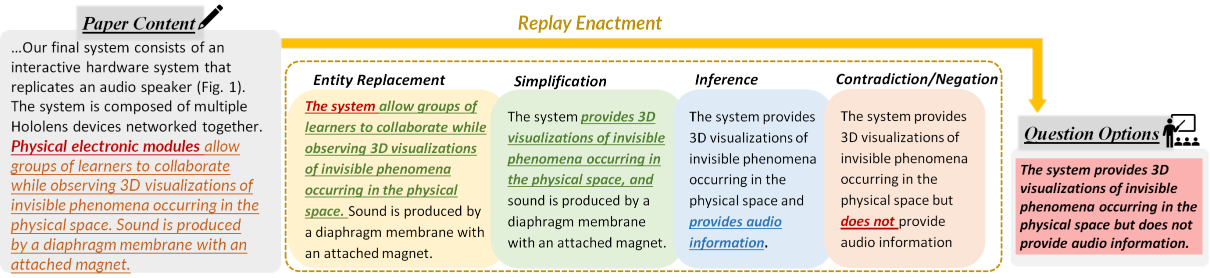
3.1.1 P2MCQ Dataset

We obtained all questions that Instructor A created during HCI101, including 160 multiple-choice questions with 629 question options in total (197 correct answers and 432 incorrect answers or distractors). These quizzes were manually written based on 30 reading materials (24 conference papers and 6 book chapters) and were assigned throughout HCI101. Example questions can be seen in Figure 2.

Context Annotation One annotator aligned each question option with its supporting context in the original reading material. Both sentence-level and paragraph-level contexts are extracted. For each single option, the sentence-level context include sentences with its supportive evidence, and the paragraph-level context refers to the whole paragraph containing those sentences. During context annotation, questions and options without supportive evidence in the given material (i.e., they were designed based on the instructor's prior knowledge, not from the text) were removed. Instructor A also checked our annotations for accuracy. This resulted in a dataset of 128 multiple-choice questions, including 439 question options (110 target options and 329 distractors). We release the P2MCQ dataset with this paper.

3.1.2 Text Replay Enactment

We are inspired by user-centered design methods, including user enactment (Odom et al., 2012) and replay enactment (Holstein et al., 2020), through which designers imagine possible futures and use



...Our final system consists of an interactive hardware system that replicates an audio speaker (Fig. 1). The system is composed of multiple Hololens devices networked together. Physical electronic modules allow groups of learners to collaborate while observing 3D visualizations of invisible phenomena occurring in the physical space. Sound is produced by a diaphragm membrane with an attached magnet.

The system allow groups of learners to collaborate while observing 3D visualizations of invisible phenomena occurring in the physical space. Sound is produced by a diaphragm membrane with an attached magnet.

The system provides 3D visualizations of invisible phenomena occurring in the physical space, and sound is produced by a diaphragm membrane with an attached magnet.

The system provides 3D visualizations of invisible phenomena occurring in the physical space and provides audio information.

The system provides 3D visualizations of invisible phenomena occurring in the physical space but does not provide audio information

The system provides 3D visualizations of invisible phenomena occurring in the physical space but does not provide audio information.

Figure 1: An example Text Replay Enactment process. For every context-option pair, we as NLP researchers annotate what NLP modules are needed to generate the final question option from the original context. In this case, 5 NLP modules: Extraction, Entity Replacement, Simplification, Inference, and Negation are needed.

them as conversation starters to elicit users' feedback on technologies that do not exist in users' previous experience. We propose a text replay enactment method where researchers analyze user's written text and the context, to annotate what possible NLP tasks could help the user achieve this outcome. A sample enactment process is shown in Fig 1. To ensure that the annotation process is consistent with the instructor's intention, we started with a reflective interview with Instructor A on a sample of the P2MCQ dataset (10 question stems and 40 options). Instructor A described how they arrived at the question stem and each of the options. We then applied the text replay enactment process on the entire P2MCQ dataset. All four researchers first constructed the annotation scheme collaboratively with 5% of the dataset. Two researchers then completed another 13% of the dataset and verified inter-rater reliability. One researcher further completed the annotation. The resulted annotation scheme with NLP tasks and examples can be found in Table 3. These NLP tasks are used in phase 2 to probe into users' needs for NLP support.

3.2 Phase 2: Interview and Observation Study

We recruited participants through social media (including mailing lists and social groups of professors) and offline correspondences. 10 instructors from 7 different universities participated in the study. 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. 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. We provided question templates sampled from the P2MCQ dataset to help them jump start. 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. For each of the NLP tasks in Table 3, we presented the task in the context of the user's own quiz design experience, and asked the participant to share to what extent did they find the support useful. We transcribed the interview recordings and analyzed our data using affinity diagrams (Moggridge and Atkinson, 2007), which is a commonly used method in HCI to identify emerging themes from qualitative data and discover opportunities for technologies to enhance future.

4 Findings

Through affinity diagramming, we observe emerging themes from the interview data. We present our findings and use participants' quotes to illustrate these themes. We refer to participants as P1-P10.

4.1 Supporting active reading is desirable yet expensive

All participants mentioned that there should be better ways to support students in reading. 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 is not scalable as grading can be challenging; collaborative annotation platforms such as Perusall (Perusall, 2021) encourage participation but do not necessarily make sure students get the key messages. Most participants expressed interest in using this approach if question design becomes less expensive and more accessible, as P2 said "*I wouldn't have the time to do it this way, but if you were to give me the questions, I think it would be amazing.*"

4.2 Why Automatic Question Generation Techniques Are Insufficient

Instructors often use the questions for specific educational goals, leverage diverse data sources when designing questions, and iteratively improve the content. We point out five categories of reasons on why existing QG techniques are insufficient, as they overlook these critical factors behind making questions of high educational value.

4.2.1 I want the questions to be aligned with my teaching goals

Instructors are more interested in targeting higher-level concepts instead of specific details, when supporting active reading. P2 mentioned "*students complain when asking specific questions about the reading, because they feel like we're not testing them for knowledge, we're testing them for obscure pieces of information in a paper.*" Other instructors also worry that using detail-oriented quiz questions may make the students perceive it as a reading comprehension task and mindlessly answer the questions without thinking about the broader implications of the reading. Instructors require the quiz questions to meet their objectives of assigning the reading and help students "get a gist" and "think deeply" of the material. For example, P10 mentioned "*I'm never trying to just see how much they can memorize. We are really trying to get at the conceptual stuff.*" P1 wanted the students to think critically about the assigned text: "*This paper itself has all of these pretty deep flaws, and I want students to see the flaws.*"

4.2.2 This is not based on the text. It's from my own experience.

An emerging theme from the interviews is that instructors rely on their prior knowledge when designing questions that they think are of educational value. This is a frequent quote we get from participants "*So this [question option] I'm not getting from the article. This is from my own experience.*" There are three categories of information that instructors leverage: 1) key messages in the text; 2) common student misconceptions; 3) course syllabus and prerequisite relationships.

First, instructors rely on their memory of what are **the key takeaways from the text**. They would often skim the reading text and say "*I want people to get right away [a concept].*"

Second, instructors use **students' common misconceptions** when designing questions and distractors. For example, P5 said "*Most students got it wrong, because they have a shallow understanding of the content.*" P10 also made similar comments "*Students have trouble distinguishing what's a milestone and what's a mechanism, so I could imagine a question. . .*"

Third, instructors consider **syllabus information** and think about how the questions are connected to previous or subsequent activities in the course. For example, P2 said "*[I'm interested in] quiz questions that are heavily integrating a broad set of concepts throughout the course. I want to test them on the whole understanding of the material. . .*"

4.2.3 The distribution and sequence of questions are also important

In addition to the questions themselves, instructors also keep thinking about the meta-level properties of the quiz, such as the distribution and sequence of questions. E.g., P4 said "*I want to make sure my questions cover all three sections in that chapter.*" P6 paid attention to the sequence of questions as they were in particular interested in using the questions as a reading guide.

4.2.4 It is an iterative process

Another frequent quote from the quiz design processes is that "*I'll write it down first and see whether I like it.*" Multiple participants mentioned that "*this may not be the final question*" or "*I'd like to make this better.*" Question design is an iterative process where instructors keep revising the text and may jump between questions.

4.2.5 Instructors combine a suite of techniques and strategies

One emerging theme from both phases of our study is that instructors combine a variety of strategies when designing questions and apply different strategies across contexts. From text replay enactment, we found that multiple text operations are often needed (in 57% of the dataset) in order to produce the target text used in the question (Figures 3 and 4).

4.3 Challenges in Human QG

The two mostly mentioned challenges are 1) **figuring out the key messages** that instructors want students to take away from the text; 2) **coming up with distractors** (incorrect options). Multiple participants described how identifying the key messages is critical and hard. P4 expressed "*I think the challenging thing is to find out the outcome I want students to have after reading the article*." P5 also mentioned "*I think one challenge is like, I realized there's a lot of information in the reading, and seeing what are the key things I want students to get across*." Additionally, all participants found coming up with plausible distractors to be difficult. P5 commented "*coming up with the distractors, it is challenging. Because I don't know what would distract the student, but also not be too confusing*."

For novice instructors, **producing the question stem** is difficult. We saw participants often referred to the question templates for ideas. Participants also requested **meta-level monitoring**, e.g., making sure the difficulty of the questions is reasonable, the questions have a good coverage of the content, etc. Some instructors mentioned it was challenging to **ensure that questions are "good"**. P7 expressed their frustration that "*What I always want, and never had in a multiple-choice quizzing tool is a way to say if they've chosen this answer, they have demonstrated these learning objectives and failed to demonstrate these learning objectives*." P5 said "*maybe I write a question, and then there could be support like 'it could have been framed better.' Bringing in learning sciences principles and suggest how to write good questions would be useful*." Finally, some encounter challenges in **phrasing and wording**. P6 mentioned, "*the hardest part for me is to decide how to rephrase it*."

4.4 Desires for NLP Support

In general, participants showed positivity towards receiving support in the process, and would want to spend the minimal time possible on question design. Participants had mixed opinions about the proposed NLP support concerning the level of control they could have and the performance of the NLP models. Instructors considered some models to be beneficial in some cases but not all. We also observed individual differences on which models they found useful. For example, for summarization, P7 did not think it would be useful since the reading text was already very condensed, but P6 and P9 felt that it would help shorten their own rereading time and extract useful information. Participants consider many tasks to be really hard to be automated, and questioned how the models would be able to produce the desirable outcomes. For example, P7 felt that in order to use simplification, the model performance would have to be perfect, because they are concerned with losing important nuances or misrepresenting ideas in the actual text. Ultimately, the participants were interested in trying the system, and thought that it could help them perform QG faster.

5 Evaluation of Existing NLP Tools

Given the observations from §4, here we select popular NLP models for the most frequent text transformation operations used by the instructors and investigate existing issues. We show both instructor-constructed transformations and system outputs in Table 1 in Appendix B. Model implementation and dataset collection details are given in Appendix A.

Here we use the context consisting of one or multiple sentences annotated on the quiz dataset as described in §3. We first conduct **sentence selection** and **abstractive summarization** based on **BERTSUMEXT** and **BERTSUMEXTABS** models (Liu and Lapata, 2019), which are extractive and abstractive summarization models fine-tuned on news articles.

To build an abstractive summarization model that is suitable for scientific domains, we further fine-tune the sequence-to-sequence BART model (Lewis et al., 2020) on our newly collected HCI article summarization dataset that consists of section-summary sentence pairs.

Given the same context, we then build models for text **simplification** using **ACCESS** (Martin et al., 2020) and **MUSS** (Martin et al., 2021), which are built on top of BERT and BART, respectively; a **paraphrasing** model that fine-tunes BART using *ParaSCI* (Dong et al., 2021) which contains paraphrase pairs from scientific papers; a **negation** generation model based on **CROSSAUG** (Lee et al., 2021) which fine-tunes BART to produce text that contradicts the given context.

Three major findings are made. For sentence selection and summarization models, it is often hard to interpret why the output is produced as is. For example, it is unclear why a sentence is treated as more important and thus selected from a given context. Second, for the generation-driven models, such as abstractive summarization, simplification, and paraphrasing, their output is sometimes only tangentially relevant to the content. The output frequently contains content that cannot be inferred from the input as well as errors, which can hurt instructor experience and even raise ethical concerns. Moreover, the diversity of the generations is rather limited. For example, simplification and negation models tend to focus on changing specific words. Overall, all these NLP models are trained without user needs being specified, thus do not offer control over which topic to summarize, what phrase or content to simplify or paraphrase, and which knowledge to be used as a pivot to create distractors. These issues point to the future directions for building summarization models with *explainability*, *guaranteed faithfulness and correctness*, as well as that *allow users to exert control* over where the transformation should be applied.

6 Implications for Research

Our study reveals instructors' natural processes of constructing questions, the challenges they have, and when they may benefit from NLP support. Specifically, we see a strong desire for user control, where humans provide input to NLP systems and can decide when to use NLP outcomes.

6.1 Implications for Developing Human-NLP Collaborative QG Systems

Recommendation 1: Instead of generating outcomes, providing process-oriented support is more desirable. First, instructors considered it to be highly critical for the questions to serve their teaching goals. As P10 put it "*Whatever the goals are, they are shaping those questions, because mostly, the texts, chapters or articles, they were not written with the course in mind. I have to take a reading that was meant for one purpose and pull it to my purpose.*" When QG systems do not align with instructors' goals, it's hard for them to meet the educational needs and support student learning. Second, all participants viewed question design as an iterative process and preferred to have the opportunity to keep revising and improving content. Third, we observed that instructors had challenges in doing QG themselves, e.g., when the text was dense, refreshing their memories of the key messages was time-consuming; when they wanted to include a distractor on one concept, figuring out the exact language for an option was hard; instructors may also jump around the whole article and make inferences based on text spread in the article.

We argue that it is more productive for NLP systems to provide process-oriented support to instructors instead of trying to generate complete outcomes immediately. As an example, most instructors liked the idea of highlighting key phrases in the text for the system to identify relevant content. Instructors also liked having the system summarize content for them to decide what questions to design. When asked about using Contradiction/Negation to create distractors, they applauded the idea of telling the system which parts to negate. We see a future direction of building human-NLP collaborative systems where expert input is continuously solicited to tell the system where to pay attention to.

Recommendation 2: Develop QG systems with NLP modules that provide the flexibility of applying modules depending on the context. Instructors had different preferences on which NLP modules to use depending on the goal of the questions, students' prior knowledge, etc. For example, instructors want to use easy questions at times as quick comprehension checks: "*This one [question] is actually quite easy. If they read the paper they'll know.*" Other times, they want to make challenging questions that provoke deep thinking, as we discussed in depth in §4.2.1. Even for a single question, instructors prefer using different strategies to design options. E.g., P5 suggested using Contradiction/Negation to generate one distractor and using Extracting Parallel Concepts to generate another. This ensures that a question contains rich information that benefits student learning. Additionally, we also observed substantive cases where instructors chained multiple NLP modules (as shown in Figure 1) to generate an option.

We argue that framing QG as a single NLP task does not align with educators' process of question creation. Instead, we encourage QG systems to provide modular NLP tools and give instructors the flexibility of choosing which NLP tools to use as they see fit. This approach will also greatly increase instructors' trust to the system and the interpretability of QG systems. We provide a list of NLP tasks that instructors find useful (Table 3), the frequency of these tasks in the P2MCQ dataset (Figure 3), and the preliminary results on the performance of existing NLP tools on these tasks (Table 2).

Recommendation 3: QG systems need diverse data sources as input. Traditional QG systems mostly use only one text source as input. We observed instructors relying on diverse data sources when designing questions. Some instructors wanted to use student-created examples in questions to reflect common misconceptions. P10 said "*Since I have lots of examples of what students have said, I could use student examples and say, which ones of these are good examples?*" P7 mentioned an interest in using the whole course content as a data source: "*I could imagine a system taking not only the paper in but also taking in the rest of my course content, just as secondary, like what is foreshadowing, what is he going to teach in about two weeks.*" Even within a single document, we observed that users may aggregate texts from non-adjacent locations as input. We encourage QG systems to take diverse data sources as input, including previous student answers, course syllabus, lecture notes, relevant reading materials, and give instructors the flexibility to select the input sources.

6.2 Implications for Research on Human-Centered NLP

6.2.1 Towards More Robust NLP Outcomes

As detailed in §5, we observed a considerable quality gap between human-generated options and machine-generated ones. First, end-to-end neural models often produce extraneous information. To address this, we propose to solicit user input and train models to focus on user-specified requirements. As an example, for sentence simplification, user may choose to keep several phrases and ask the system to simplify the rest of the content. We encourage researchers to work on models that support and comply with different forms of user control. Second, we argue for modular NLP systems. With the large language models becoming the de facto tool for NLP tasks of varying difficulty levels, dividing the tasks into steps and chaining the outputs could improve the quality of task outcomes (Wu et al., 2021). Our study revealed that instructors applied complex transformations on text when creating meaningful questions. We further proposed a list of pre-defined NLP tasks that instructors frequently employed, and evaluated fine-tuned language models for these tasks. In future work, we consider the inclusion of user input as part of the modular system design, e.g., allowing users to construct prompts or instructions (Floridi and Chiriatti, 2020; Wei et al., 2021) in-situ to interact with the large language models to better satisfy their needs.

6.2.2 Prototyping Human-Centered NLP Systems

In this study, we also surface challenges in designing and developing human-centered NLP systems. First, users are not able to articulate their needs around NLP. For example, users are not able to say they want the system to make an inference or complete an entity replacement at a certain point. We consider the text replay enactment approach to be effective in helping researchers understand user intentions and desires for NLP and encourage others to further explore and extend this method. Second, we found that having users directly evaluate NLP outcomes is hard, as it often requires them to read the context and outputs by several models. When we tried this in a pilot interview, a user found this to be too cognitively demanding. A more effective way is to visualize the NLP outcomes in context to reduce cognitive load, and investigate user adoption and preferences in real usage. We also encourage collaboration between NLP and HCI researchers in making progress on novel methods to prototype NLP experiences and build usable NLP systems.

7 Conclusion

QG has been an area of interest in the NLP community. However, the adoption of QG systems in classrooms is low. The goal of this work is to investigate this gap and explore directions for future QG systems that can meet stakeholders' needs. Our work reveals that existing QG systems do not take information that instructors deem critical when designing questions to support learning, including educational goals, and student misconceptions. We surface instructors' desires for receiving support during question design. We make recommendations on how to develop process-oriented, modular, human-NLP collaborative QG systems.

Acknowledgements

This work is supported in part by National Science Foundation through grant IIS-2046016. We thank all our participants for their time. We thank the anonymous reviewers for their valuable suggestions on various aspects of this work.

References

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

Saleema Amershi, Daniel S. Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi T. Iqbal, Paul N. Bennett, Kori Inkpen Quinn, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for human-ai interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*.

Alessio Bellino and Daniela Bascuñán. 2020. Design and evaluation of writebetter: A corpus-based writing assistant. *IEEE Access*, 8:70216–70233.

Sheng Bi, Xiya Cheng, Yuan-Fang Li, Lizhen Qu, Shirong Shen, Guilin Qi, Lu Pan, and Yinlin Jiang. 2021. Simple or complex? complexity-controllable question generation with soft templates and deep mixture of experts model. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4645–4654, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shuyang Cao and Lu Wang. 2021. Controllable open-ended question generation with a new question type ontology. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6424–6439, Online. Association for Computational Linguistics.

Dhawaleswar Rao Ch and Sujan Kumar Saha. 2018. Automatic multiple choice question generation from text: A survey. *IEEE Transactions on Learning Technologies*, 13(1):14–25.

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

Savvas Chatzipanagiotidis, Maria Giagkou, and Walt Detmar Meurers. 2021. Broad linguistic complexity analysis for greek readability classification. In *BEA*.

Yi Cheng, Siyao Li, Bang Liu, Ruihui Zhao, Sujian Li, Chenghua Lin, and Yefeng Zheng. 2021. Guiding the growth: Difficulty-controllable question generation through step-by-step rewriting. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5968–5978, Online. Association for Computational Linguistics.

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

Catherine H Crouch and Eric Mazur. 2001. Peer instruction: Ten years of experience and results. *American journal of physics*, 69(9):970–977.

Melissa Dancy and Charles Henderson. 2007. Framework for articulating instructional practices and conceptions. *Physical Review Special Topics-Physics Education Research*, 3(1):010103.

Louis Deslauriers, Logan S McCarty, Kelly Miller, Kristina Callaghan, and Greg Kestin. 2019. Measuring actual learning versus feeling of learning in response to being actively engaged in the classroom. *Proceedings of the National Academy of Sciences*, 116(39):19251–19257.

Qingxiu Dong, Xiaojun Wan, and Yue Cao. 2021. Parasci: A large scientific paraphrase dataset for longer paraphrase generation. In *EACL*.

Meaghan S Edmonds, Sharon Vaughn, Jade Wexler, Colleen Reutebuch, Amory Cable, Kathryn Klingler Tackett, and Jennifer Wick Schnakenberg. 2009. A synthesis of reading interventions and effects on reading comprehension outcomes for older struggling readers. *Review of educational research*, 79(1):262–300.

Adam P Fagen, Catherine H Crouch, and Eric Mazur. 2002. Peer instruction: Results from a range of classrooms. *The physics teacher*, 40(4):206–209.

Luciano Floridi and Massimo Chiriatti. 2020. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30(4):681–694.

A. Frankenberg-Garcia, Robert Lew, Jonathan C. Roberts, Geraint Paul Rees, and Nirwan Sharma. 2018. Developing a writing assistant to help eap writers with collocations in real time. *ReCALL*, 31:23 – 39.

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):8410–8415.

Elena L Glassman, Jeremy Scott, Rishabh Singh, Philip J Guo, and Robert C Miller. 2015. Overcode: Visualizing variation in student solutions to programming problems at scale. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 22(2):1–35.

Jo Handelsman, Diane Ebert-May, Robert Beichner, Peter Bruns, Amy Chang, Robert DeHaan, Jim Gentile, Sarah Lauffer, James Stewart, Shirley M Tilghman, et al. 2004. Scientific teaching.

John Earl Haynes and H. Thompson Fillmer. 1984. Paraphrasing and reading comprehension. *Literacy Research and Instruction*, 24:76–79.

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*, pages 609–617, Los Angeles, California. Association for Computational Linguistics.

Charles Henderson and Melissa H Dancy. 2007. Barriers to the use of research-based instructional strategies: The influence of both individual and situational characteristics. *Physical Review Special Topics-Physics Education Research*, 3(2):020102.

Kenneth Holstein and Vincent Aleven. 2021. Designing for human-ai complementarity in k-12 education. *arXiv preprint arXiv:2104.01266*.

Kenneth Holstein, Erik Harpstead, Rebecca Gulotta, and Jodi Forlizzi. 2020. Replay enactments: Exploring possible futures through historical data. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*, pages 1607–1618.

Kenneth Holstein, Bruce M McLaren, and Vincent Aleven. 2018. Student learning benefits of a mixed-reality teacher awareness tool in ai-enhanced classrooms. In *International conference on artificial intelligence in education*, pages 154–168. Springer.

Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 159–166.

Gwo-Jen Hwang, Mei-Rong Alice Chen, Han-Yu Sung, and Mengwei Lin. 2019. Effects of integrating a concept mapping-based summarization strategy into flipped learning on students' reading performances and perceptions in chinese courses. *Br. J. Educ. Technol.*, 50:2703–2719.

Kentaro Inui, Atsushi Fujita, Tetsuro Takahashi, Ryu Iida, and Tomoya Iwakura. 2003. Text simplification for reading assistance: a project note. In *Proceedings of the second international workshop on Paraphrasing*, pages 9–16.

Sungku Kang, Lalit Patil, Arvind Rangarajan, Abha Moitra, Tao Jia, Dean M. Robinson, and Debasish Dutta. 2019. Automated feedback generation for formal manufacturing rule extraction. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, 33:289 – 301.

Anisia Katinskaia and Roman Yangarber. 2021. Assessing grammatical correctness in language learning. In *BEA*.

Kenneth R Koedinger, John R Anderson, William H Hadley, Mary A Mark, et al. 1997. Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education*, 8(1):30–43.

Rohit Kumar, Carolyn Penstein Rosé, Yi-Chia Wang, Mahesh Joshi, and Allen Robinson. 2007. Tutorial dialogue as adaptive collaborative learning support. *Frontiers in artificial intelligence and applications*, 158:383.

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):121–204.

Minwoo Lee, Seungpil Won, Juae Kim, Hwanhee Lee, Cheoneum Park, and Kyomin Jung. 2021. Crossaug. *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*.

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 *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Jeri L Little and Elizabeth Ligon Bjork. 2015. Optimizing multiple-choice tests as tools for learning. *Memory & Cognition*, 43(1):14–26.

Jeri L Little, Elizabeth Ligon Bjork, Robert A Bjork, and Genna Angello. 2012. Multiple-choice tests exonerated, at least of some charges: Fostering test-induced learning and avoiding test-induced forgetting. *Psychological science*, 23(11):1337–1344.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. *ArXiv*, abs/1908.08345.

Chenyang Lyu, Lifeng Shang, Yvette Graham, Jennifer Foster, Xin Jiang, and Qun Liu. 2021. Improving unsupervised question answering via summarization-informed question generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4134–4148, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mukta Majumder and Sujan Kumar Saha. 2015. A system for generating multiple choice questions: With a novel approach for sentence selection. In *Proceedings of the 2nd workshop on natural language processing techniques for educational applications*, pages 64–72.

Louis Martin, Angela Fan, Éric de la Clergerie, Antoine Bordes, and Benoît Sagot. 2021. Muss: Multilingual unsupervised sentence simplification by mining paraphrases.

Louis Martin, Benoît Sagot, Eric Villemonte de la Clergerie, and Antoine Bordes. 2020. Controllable sentence simplification. In *LREC*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919.

Bill Moggridge and Bill Atkinson. 2007. *Designing interactions*, volume 17. MIT press Cambridge.

William Odom, John Zimmerman, Scott Davidoff, Jodi Forlizzi, Anind K Dey, and Min Kyung Lee. 2012. A fieldwork of the future with user enactments. In *Proceedings of the Designing Interactive Systems Conference*, pages 338–347.

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

Liangming Pan, Yuxi Xie, Yansong Feng, Tat-Seng Chua, and Min-Yen Kan. 2020. Semantic graphs for generating deep questions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1463–1475, Online. Association for Computational Linguistics.

Andreas Papasalouros, Konstantinos Kanaris, and Konstantinos Kotis. 2008. Automatic generation of multiple choice questions from domain ontologies. In *e-Learning*, pages 427–434. Citeseer.

Rakesh Patra and Sujan Kumar Saha. 2019. A hybrid approach for automatic generation of named entity distractors for multiple choice questions. *Education and Information Technologies*, 24(2):973–993.

Perusall. 2021. Perusall.

Sarah E. Petersen and Mari Ostendorf. 2007. Text simplification for language learners: a corpus analysis. In *SLaTE*.

Irina Rets and Jekaterina Rogaten. 2021. To simplify or not? facilitating english l2 users' comprehension and processing of open educational resources in english using text simplification. *J. Comput. Assist. Learn.*, 37:705–717.

Steven Ross, Michael H. Long, and Yasukata Yano. 1991. Simplification or elaboration? the effects of two types of text modifications on foreign language reading comprehension.

Daniel L Schwartz, Catherine C Chase, Marily A Oppezzo, and Doris B Chin. 2011. Practicing versus inventing with contrasting cases: The effects of telling first on learning and transfer. *Journal of educational psychology*, 103(4):759.

Advaith Siddharthan and Napoleon Katsos. 2010. Reformulating discourse connectives for non-expert readers. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1002–1010, Los Angeles, California. Association for Computational Linguistics.

Dee U Silverthorn, Patti M Thorn, and Marilla D Svinicki. 2006. It's difficult to change the way we teach: lessons from the integrative themes in physiology curriculum module project. *Advances in physiology Education*, 30(4):204–214.

Megan A Smith and Jeffrey D Karpicke. 2014. Retrieval practice with short-answer, multiple-choice, and hybrid tests. *Memory*, 22(7):784–802.

Linfeng Song and Lin Zhao. 2016. Question generation from a knowledge base with web exploration. *arXiv preprint arXiv:1610.03807*.

Marilyne Stains, Jordan Harshman, Megan K Barker, Stephanie V Chasteen, Renee Cole, Sue Ellen DeChenne-Peters, MK Eagan, Joan M Esson, Jennifer K Knight, Frank A Laski, et al. 2018. Anatomy of stem teaching in north american universities. *Science*, 359(6383):1468–1470.

Katherine Stasaski and Marti A Hearst. 2017. Multiple choice question generation utilizing an ontology. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 303–312.

Elizabeth A Stevens, Sunyoung Park, and Sharon Vaughn. 2019. A review of summarizing and main idea interventions for struggling readers in grades 3 through 12: 1978–2016. *Remedial and Special Education*, 40(3):131–149.

Dan Su, Yan Xu, Wenliang Dai, Ziwei Ji, Tiezheng Yu, and Pascale Fung. 2020. Multi-hop question generation with graph convolutional network. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4636–4647, Online. Association for Computational Linguistics.

Xingwu Sun, Jing Liu, Yajuan Lyu, Wei He, Yanjun Ma, and Shi Wang. 2018. Answer-focused and position-aware neural question generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3930–3939, Brussels, Belgium. Association for Computational Linguistics.

Adel I. Tweissi. 1998. The effects of the amount and type of simplification on foreign language reading comprehension. *Reading in a foreign language*, 11:191–204.

Tiiu Üksik, Jelena Kallas, Kristina Koppel, Katrin Tsepelina, and Raili Pool. 2021. Estonian as a second language teacher's tools. In *BEA*.

Sowmya Vajjala and Ivana Lucic. 2019. On understanding the relation between expert annotations of text readability and target reader comprehension. In *BEA@ACL*.

Sharon Vaughn, Janette K Klingner, Elizabeth A Swanson, Alison G Boardman, Greg Roberts, Sarojani S Mohammed, and Stephanie J Stillman-Spisak. 2011.

Efficacy of collaborative strategic reading with middle school students. *American educational research journal*, 48(4):938–964.

Tatiana Vodolazova and Elena Lloret. 2019. Towards adaptive text summarization: How does compression rate affect summary readability of l2 texts? In *RANLP*.

Ke Wang, Rishabh Singh, and Zhendong Su. 2018. Search, align, and repair: data-driven feedback generation for introductory programming exercises. *Proceedings of the 39th ACM SIGPLAN Conference on Programming Language Design and Implementation*.

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*, pages 1–14.

Xu Wang, Srinivasa Teja Talluri, Carolyn Rose, and Kenneth Koedinger. 2019. Upgrade: Sourcing student open-ended solutions to create scalable learning opportunities. In *Proceedings of the Sixth (2019) ACM Conference on Learning@ Scale*, pages 1–10.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Tongshuang Wu, Michael Terry, and Carrie J Cai. 2021. Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts. *arXiv preprint arXiv:2110.01691*.

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*, pages 1–13.

Shiyue Zhang and Mohit Bansal. 2019. Addressing semantic drift in question generation for semi-supervised question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2495–2509, Hong Kong, China. Association for Computational Linguistics.

Wenjie Zhou, Minghua Zhang, and Yunfang Wu. 2019. Question-type driven question generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6032–6037, Hong Kong, China. Association for Computational Linguistics.

A Details for NLP Models on Selected Tasks

Here we describe implementation details for the NLP models used in §5.

A.1 Sentence Extraction

We extract salient sentences from the given context using extractive summarization model. The automatic extractive summarization task is approached as sentence classification, where each sentence i is assigned a label $y_i \in \{0, 1\}$, with 1 indicating the sentence is predicted for inclusion in the summary, otherwise 0.

We use **BERTSUMEXT** (Liu and Lapata, 2019) that adds a sentence positional encoding schema to the large pre-trained BERT model. In our experiment, we take the released checkpoint trained on CNN/DailyMail dataset and only consider paragraph-level context as the input.

A.2 Abstractive Summarization

A.2.1 Pre-trained Summarization Model

We first use BERTSUMEXTABS (Liu and Lapata, 2019), which is first fine-tuned on the extractive summarization task and then on abstractive summarization. We again use a checkpoint of **BERTSUMEXTABS** fine-tuned on CNN-DailyMail released by the authors.

A.2.2 Summarization Model for Scientific Domain

To build an abstractive summarization model that is suitable for scientific papers, we build **BARTSUM-HCI** by fine-tuning BART (Lewis et al., 2020) on an automatically aligned section-summary pairs collected from HCI papers on *arXiv*.

Dataset Collection (*arXiv-HCI*). We first retrieve papers from `arXiv.org` with search query *cat:cs.HC* to identify all HCI relevant papers, resulting in 8658 academic papers. Since abstract is expected to condense the most salient information of a paper, we consider each sentence in abstract as a sentence-level summary of relevant sections in the paper. To align each abstract sentence with the corresponding section context, we apply BM25 to rank all sections by their lexicon similarity to the sentence and picked the top 2 sections to create two section-sentence pairs. This creates the *arXiv-HCI* dataset, which contains 72755 section-sentence pairs. The average numbers of tokens in context and summary are 468.62 and 23.01, respectively.

Continued Training We then fine-tune BART on *arXiv-HCI* with train/validation/test splits of 71301/727/727. We use the checkpoint *bart-base-finetuned-arxiv* released by HuggingFace[2], which has been trained on scientific papers.

The model is trained for 3 epochs, using Adam optimizer with default parameters (β_1, β_2)=(0.9, 0.999) and ϵ=1e-08. The learning rate is initialized as 3×10^{-5} with 2000 warm-up steps and the weight decay is set to 0.01. After training for 25000 steps with batch size 8, the model is evaluated with ROUGE-1 of 20.51, ROUGE-2 of 5.30, and ROUGE-L of 16.73 on the test set.

A.3 Simplification

We use the AudienCe-CEntric Sentence Simplification (**ACCESS**) model (Martin et al., 2020) for sentence simplification. In our experiments, we take the released checkpoint[3] of ACCESS, which is pre-trained and evaluated on WikiLarge dataset. During our inference process, the default control parameters by ACCESS are used.

We also experiment with the Multilingual Unsupervised Sentence Simplification (**MUSS**) model (Martin et al., 2021). The dataset used for MUSS training is a combination of WikiLarge dataset and mined paraphrases from CCNet.

A.4 Paraphrasing

We create a paraphrase generation model **BART-PARA-SCI** by fine-tuning the *bart-paraphrase*[4] checkpoint on the ParaSCI-ACL (Dong et al., 2021) dataset, which contains 33,981 paraphrase pairs from articles published in ACL conferences and workshops. The model is trained for 10 epochs, using Adam optimizer with default parameters (β_1, β_2)=(0.9, 0.999) and ϵ=1e-08. The learning rate is initialized as 3×10^{-5} with 2500 warm-up steps. The weight decay is set to 0.01.

A.5 Negation

To generate distractors, we use **CROSSAUG**, which is proposed as a data augmentation method by training BART to generate negative claims (Lee et al., 2021). It is fine-tuned on the WikiFactCheck-English dataset, with positive claims as the inputs and their corresponding negative claims as the outputs. We use the checkpoint[5] released by its author.

B MCQ Samples with Human-generated and Machine-generated MCQ Options

B.1 Context-option Generation Comparison Samples

Table 1 lists sample machine-generated options along with human-constructed transformed texts, based on the same context from our *P2MCQ* dataset (§3). For **Simplification**, **Paraphrasing**, **Negation** and **Abstractive Summarization** tasks, the sentence-level contexts are used as inputs; for the **Extraction** task, the paragraph-level contexts are used.

B.2 Selected Samples from *P2MCQ* Dataset

Figure 2 shows sample multiple choice questions in our *P2MCQ* dataset discussed in §3.

C NLP Text Transformation Operation Strategies

In Tables 2 and 3, we define our coding scheme for different text transformation operations, along with selected samples from our *P2MCQ* dataset.

D Distribution of Operations in *P2MCQ* Dataset

Figure 3 illustrates the distribution of different tasks used as single operation and along with other operations based on the *P2MCQ* dataset.

[2] https://huggingface.co/mse30/bart-base-finetuned-arxiv

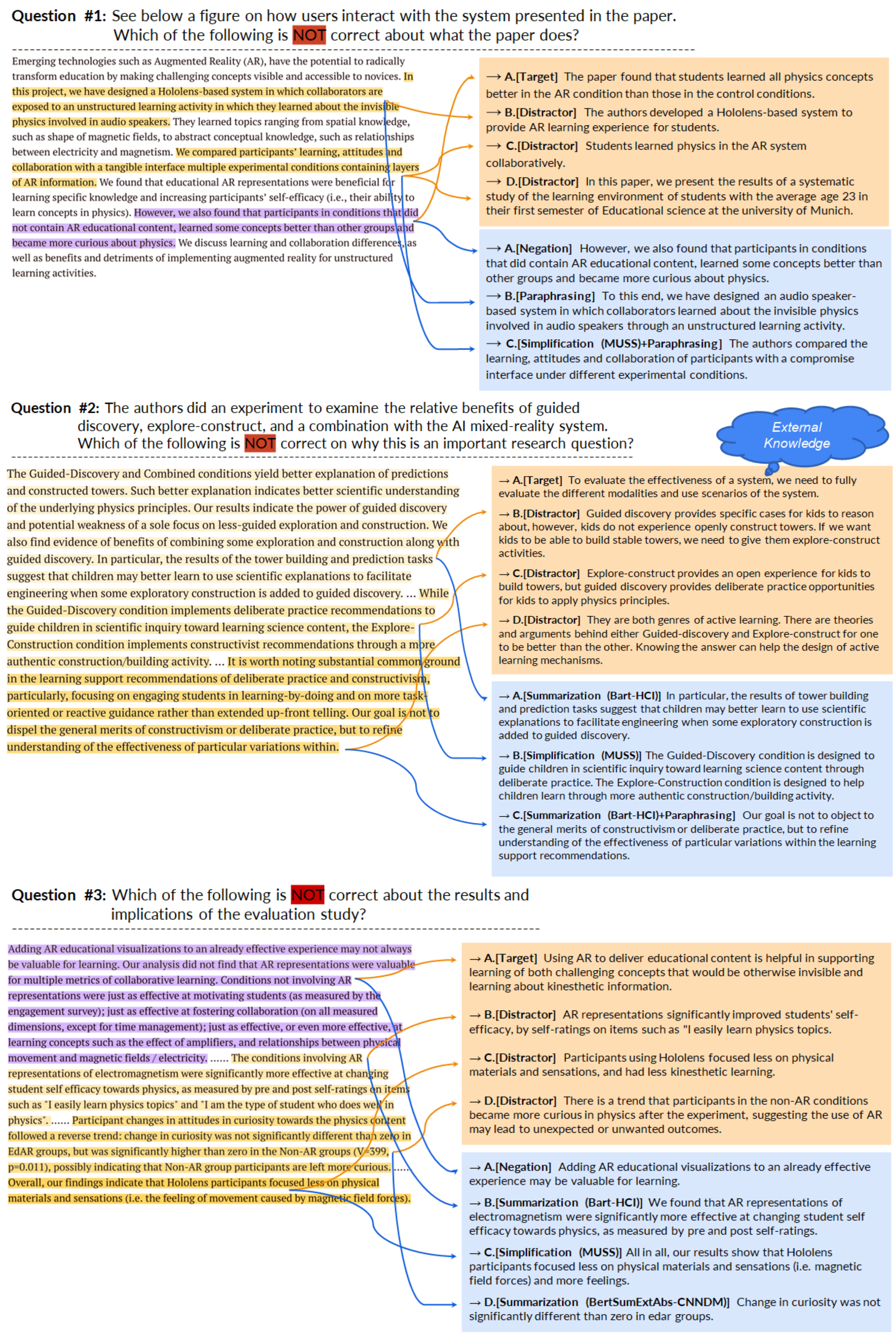
[3] http://dl.fbaipublicfiles.com/access/best_model.tar.gz

[4] https://huggingface.co/eugenesiow/bart-paraphrase

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

Context	Human Generated Option	Machine Modification
Other studies have found that helping people to think about themselves as having multiple identities, in particular, focusing on those facets of their identity that are in-group (e.g., college student) rather than outgroup (e.g., female), improves performance for those at risk of stereotype threat.	[**Simplification**] Studies have found that helping people to think about themselves as having multiple identities could improve performance for those at risk of stereotype threat	**ACCESS** Other studies have found that helping people to think about themselves as having many identities, in particular, using those facets of their identity that are in-group (e.g. college student), rather than outgroup (e. g , female) , improves performance for those at risk of stereotype threat . **MUSS** Other studies have found that helping people to think about themselves as having multiple identities, in particular, focusing on those facets of their identity that are in-group (e.g., female) helps them to remember who they are, even if they do not feel like they are.
In the neoclassical theories and Soar, once knowledge leaves working memory it is permanently lost, whereas in ACT the knowledge remains but is inactive.	[**Simplification**] Once declarative knowledge leaves working memory, it's permanently lost.	**ACCESS** In this case, theories and Soar, once knowledge leaves working memory it is lost, whereas in ACT the knowledge remains but is still inactive. **MUSS** In the neoclassical theory, once knowledge leaves working memory, it is lost forever. In the ACT theory, the knowledge remains but is not used.
Cognitive skills are realized by production rules.	[**Paraphrasing**] The skill itself should be modeled as a set of production rules.	**BART-PARA-SCI** the cognitive skills can be found by production rules.
People must evaluate and improve their practice by seeing the effects of their efforts in the world.	[**Paraphrasing**] Learners need to receive in-time feedback form them to evaluate and improve their practice.	**BART-PARA-SCI** they must evaluate and improve their practice by the effects of their efforts in the world.
Increasing the number of raters per submission helps accuracy but quickly yields diminishing returns'	[**Negation**] Increasing the number of raters per submission would always help increase accuracy of median grades.	**CROSSAUG** Increasing the number of raters per submission decreases accuracy and quickly yields diminishing returns
The second finding is that it is not just the quantity of practice but also the quality. Chess duffers get good by playing. Grand masters get great by further studying famous games. Deliberate practice involves focusing on what is beyond one's current skill set rather than just executing what one is already able to do.	[**Extraction**] Deliberate practice involves focusing on what is beyond one's current skill set rather than just executing what one is already able to do.	**BERTSUMEXT** The second finding is that it is not just the quantity of practice but also the quality.
These phenomena interact with each other in complex ways while being invisible to the naked eye, thus making the concept difficult to understand. Yet these phenomena are critically important for understanding the physics of electromagnetism.	[**Abstractive Summarization**] Electromagnetism is a complex mechanism, and many physical phenomena such as electric current are invisible.	**BERTSUMEXTABS(CNNDM)** These phenomena interact with each other in complex ways while being invisible to the naked eye. **BERTSUMEXTABS(XSUM)** one of the world 's most important phenomena of the visual effects of last year has been explained by the un 's visual assessment of the effects of electromagneting in the solar system, according to the unforgettable motion of motion. **BART-HCI** These phenomena interact with each other in complex ways while being invisible to the naked eye.
We had Non-Hololens and Hololens-Simple groups in order to test the effect of novelty and excitement that may come with experiencing even basic AR technology. We also had two types of AR educational groups, AR Scaffold vs AR Full, because learning theories suggest that presenting increasingly complex representations facilitates learning	[**Abstractive Summarization**] There are 4 conditions in total, with a "Non-Hololens" condition, and three AR conditions that display different amount of information.	**BERTSUMEXTABS(CNNDM)** we had non-hololens and hololens-simple groups to test the effect of novelty and excitement that may come with even basic ar technology **BERTSUMEXTABS(XSUM)** we had two types of ar-signing information in the last few months of this year , according to a new report by the national library of sciences -lrb- mod -rrb-. **BART-HCI** We also tested the effect of novelty and excitement that may come with experiencing even basic AR technology.

Table 1: Sample outputs with **Simplification**, **Paraphrasing**, **Negation**, **Extraction** and **Abstractive Summarization** operations.



Question #1: See below a figure on how users interact with the system presented in the paper. Which of the following is NOT correct about what the paper does?

Emerging technologies such as Augmented Reality (AR), have the potential to radically transform education by making challenging concepts visible and accessible to novices. In this project, we have designed a Hololens-based system in which collaborators are exposed to an unstructured learning activity in which they learned about the invisible physics involved in audio speakers. They learned topics ranging from spatial knowledge, such as shape of magnetic fields, to abstract conceptual knowledge, such as relationships between electricity and magnetism. We compared participants' learning, attitudes and collaboration with a tangible interface multiple experimental conditions containing layers of AR information. We found that educational AR representations were beneficial for learning specific knowledge and increasing participants' self-efficacy (i.e., their ability to learn concepts in physics). 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. We discuss learning and collaboration differences, as well as benefits and detriments of implementing augmented reality for unstructured learning activities.

→ **A.[Target]** The paper found that students learned all physics concepts better in the AR condition than those in the control conditions.

→ **B.[Distractor]** The authors developed a Hololens-based system to provide AR learning experience for students.

→ **C.[Distractor]** Students learned physics in the AR system collaboratively.

→ **D.[Distractor]** In this paper, we present the results of a systematic study of the learning environment of students with the average age 23 in their first semester of Educational science at the university of Munich.

→ **A.[Negation]** However, we also found that participants in conditions that did contain AR educational content, learned some concepts better than other groups and became more curious about physics.

→ **B.[Paraphrasing]** To this end, we have designed an audio speaker-based system in which collaborators learned about the invisible physics involved in audio speakers through an unstructured learning activity.

→ **C.[Simplification (MUSS)+Paraphrasing]** The authors compared the learning, attitudes and collaboration of participants with a compromise interface under different experimental conditions.

Question #2: The authors did an experiment to examine the relative benefits of guided discovery, explore-construct, and a combination with the AI mixed-reality system. Which of the following is NOT correct on why this is an important research question?

External Knowledge

The Guided-Discovery and Combined conditions yield better explanation of predictions and constructed towers. Such better explanation indicates better scientific understanding of the underlying physics principles. Our results indicate the power of guided discovery and potential weakness of a sole focus on less-guided exploration and construction. We also find evidence of benefits of combining some exploration and construction along with guided discovery. In particular, the results of the tower building and prediction tasks suggest that children may better learn to use scientific explanations to facilitate engineering when some exploratory construction is added to guided discovery. ... While the Guided-Discovery condition implements deliberate practice recommendations to guide children in scientific inquiry toward learning science content, the Explore-Construction condition implements constructivist recommendations through a more authentic construction/building activity. ... It is worth noting substantial common ground in the learning support recommendations of deliberate practice and constructivism, particularly, focusing on engaging students in learning-by-doing and on more task-oriented or reactive guidance rather than extended up-front telling. Our goal is not to dispel the general merits of constructivism or deliberate practice, but to refine understanding of the effectiveness of particular variations within.

→ **A.[Target]** To evaluate the effectiveness of a system, we need to fully evaluate the different modalities and use scenarios of the system.

→ **B.[Distractor]** Guided discovery provides specific cases for kids to reason about, however, kids do not experience openly construct towers. If we want kids to be able to build stable towers, we need to give them explore-construct activities.

→ **C.[Distractor]** Explore-construct provides an open experience for kids to build towers, but guided discovery provides deliberate practice opportunities for kids to apply physics principles.

→ **D.[Distractor]** They are both genres of active learning. There are theories and arguments behind either Guided-discovery and Explore-construct for one to be better than the other. Knowing the answer can help the design of active learning mechanisms.

→ **A.[Summarization (Bart-HCI)]** In particular, the results of tower building and prediction tasks suggest that children may better learn to use scientific explanations to facilitate engineering when some exploratory construction is added to guided discovery.

→ **B.[Simplification (MUSS)]** The Guided-Discovery condition is designed to guide children in scientific inquiry toward learning science content through deliberate practice. The Explore-Construction condition is designed to help children learn through more authentic construction/building activity.

→ **C.[Summarization (Bart-HCI)+Paraphrasing]** Our goal is not to object to the general merits of constructivism or deliberate practice, but to refine understanding of the effectiveness of particular variations within the learning support recommendations.

Question #3: Which of the following is NOT correct about the results and implications of the evaluation study?

Adding AR educational visualizations to an already effective experience may not always be valuable for learning. Our analysis did not find that AR representations were valuable for multiple metrics of collaborative learning. Conditions not involving AR representations were just as effective at motivating students (as measured by the engagement survey); just as effective at fostering collaboration (on all measured dimensions, except for time management); just as effective, or even more effective, at learning concepts such as the effect of amplifiers, and relationships between physical movement and magnetic fields / electricity. ...... The conditions involving AR representations of electromagnetism were significantly more effective at changing student self efficacy towards physics, as measured by pre and post self-ratings on items such as "I easily learn physics topics" and "I am the type of student who does well in physics". ...... Participant changes in attitudes in curiosity towards the physics content followed a reverse trend: change in curiosity was not significantly different than zero in EdAR groups, but was significantly higher than zero in the Non-AR groups (V=399, p=0.011), possibly indicating that Non-AR group participants are left more curious. ...... Overall, our findings indicate that Hololens participants focused less on physical materials and sensations (i.e. the feeling of movement caused by magnetic field forces).

→ **A.[Target]** Using AR to deliver educational content is helpful in supporting learning of both challenging concepts that would be otherwise invisible and learning about kinesthetic information.

→ **B.[Distractor]** AR representations significantly improved students' self-efficacy, by self-ratings on items such as "I easily learn physics topics.

→ **C.[Distractor]** Participants using Hololens focused less on physical materials and sensations, and had less kinesthetic learning.

→ **D.[Distractor]** There is a trend that participants in the non-AR conditions became more curious in physics after the experiment, suggesting the use of AR may lead to unexpected or unwanted outcomes.

→ **A.[Negation]** Adding AR educational visualizations to an already effective experience may be valuable for learning.

→ **B.[Summarization (Bart-HCI)]** We found that AR representations of electromagnetism were significantly more effective at changing student self efficacy towards physics, as measured by pre and post self-ratings.

→ **C.[Simplification (MUSS)]** All in all, our results show that Hololens participants focused less on physical materials and sensations (i.e. magnetic field forces) and more feelings.

→ **D.[Summarization (BertSumExtAbs-CNNDM)]** Change in curiosity was not significantly different than zero in edar groups.

Figure 2: Three sample MCQs in our *P2MCQ* dataset. Human constructed options are created by the instructors, Machine generated options are from NLP model outputs with the corresponding contexts as inputs.

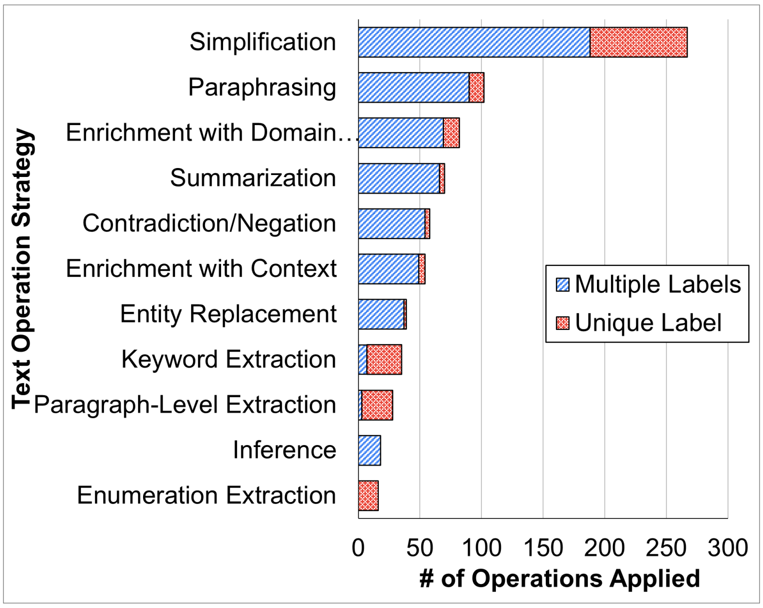


Figure 3: Annotation from Text Replay Enactment: for each row in the *P2MCQ* dataset, the context-option pairs were annotated according to the annotation scheme described in §3. The majority of rows had multiple operations applied (blue/striped in the diagram is non-exclusive), but there were a number of rows that only had one unique label (red/dotted in the diagram is exclusive).

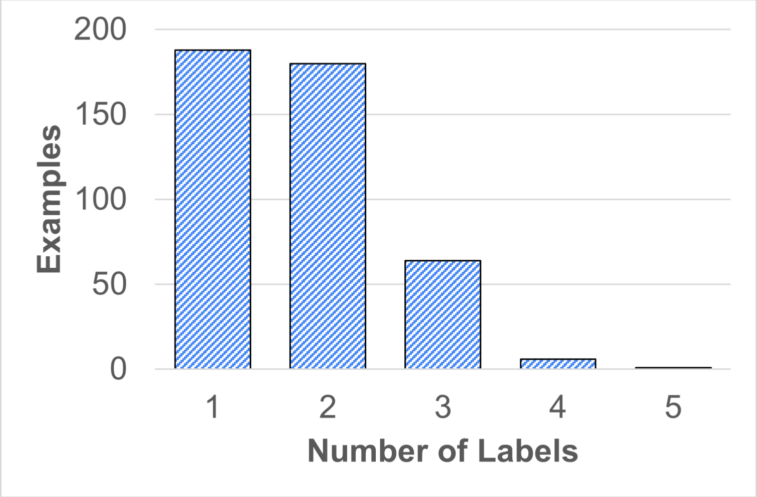


Figure 4: Annotation from Text Replay Enactment: In the *P2MCQ* dataset, the context-option pairs were annotated according to the annotation scheme described in §3. This figure shows the numbers of entries that have 1 to 5 labels. 5 means that in the enactment process, 5 NLP tasks was needed to produce the user-generated outcome.

Text Operation Strategy	**Before**	**After**
1. **Simplification**: reduce either the complexity of the language or the content density on a sentence level.	Other *studies have found that helping people to think about themselves as having multiple identities*, in particular, focusing on those facets of their identity that are in-group (e.g., college student) rather than outgroup (e.g., female), *improves performance for those at risk of stereotype threat.*	Studies have found that helping people to think about themselves as having multiple identities could improve performance for those at risk of stereotype threat.
	In the neoclassical theories and Soar, *once knowledge leaves working memory it is permanently lost*, whereas in ACT the knowledge remains but is inactive.	Once declarative knowledge leaves working memory, it's permanently lost.
2. **Paraphrasing**: restate the original context with near equivalent semantic meaning	Cognitive skills are realized by production rules.	The skill itself should be modeled as a set of production rules
	People must evaluate and improve their practice by seeing the effects of their efforts in the world.	Learners need to receive in-time feedback for them to evaluate and improve their practice
3. **Enrichment with Domain Knowledge**: include additional content to serve as an explanation, definition, or example from the instructor's own expert and domain knowledge.	The system is composed of multiple Hololens devices networked together.	The system is composed of multiple Hololens devices networked together, **displaying multiple signals.**
	Contrasting cases can help students learn to "see" where they should and should not use their knowledge.	Giving students contrasting cases **(e.g., when to use median and when to use mean)** would help them understand where they should and shouldn't use their knowledge.
4. **Summarization**: reduce the content density on a multiple sentence or paragraph level.	These phenomena interact with each other in complex ways while being invisible to the naked eye, thus making the concept difficult to understand. Yet these phenomena are critically important for understanding the physics of electromagnetism.	Electromagnetism is a complex mechanism, and many physical phenomena such as electric current are invisible.
	We had Non-Hololens and Hololens-Simple groups in order to test the effect of novelty and excitement that may come with experiencing even basic AR technology. We also had two types of AR educational groups, AR Scaffold vs AR Full, because learning theories suggest that presenting increasingly complex representations facilitates learning.	There are 4 conditions in total, with a "Non-Hololens" condition, and three AR conditions that display different amount of information.
5. **Contradiction/Negation**: add, remove, or change words in the original text to logically modify the original meaning to serve as distractors.	In all conditions we measure participant learning, collaboration and attitudes.	This paper focuses on understanding students' attitudes towards AR technology without measuring student learning.
	Increasing the number of raters per submission helps accuracy but quickly yields diminishing returns	Increasing the number of raters per submission would always help increase accuracy of median grades.
6. **Enrichment with Context**: include additional content to serve as an explanation, definition, or example from other sections of the current paper.	Learning partners did not know each other before the experimental session.	**The experiment manipulated the variable** of whether the learning partners know each other before the experimental session.
	Overall, our findings indicate that Hololens participants focused less on physical materials and sensations (i.e. the feeling of movement caused by magnetic field forces).	Participants using Hololens focused less on physical materials and sensations, and had less kinesthetic learning.

Table 2: Text transformation operation strategy definitions and examples (part 1).

7, **Entity Replacement**: replace some subject in the context with a different word that maintains the original meaning.	*We* compared participants' learning, attitudes and collaboration with a tangible interface through multiple experimental conditions containing varying layers of AR information.	**The paper** compared participants' learning, attitudes and collaboration through multiple experimental conditions containing varying layers of AR information.
	We also compared students' self-grade with their median peer grade to measure whether students rate themselves differently than their peers.	**The authors** also compared students' self-grade with their median peer grade to measure whether students rate themselves differently from their peers.
8. **Keyword Extraction**: extract a single important word or phrase from the context.	In addition, the theoretical/descriptive analysis focuses more on procedures required for good performance because it focuses on the expert's problem-solving processes.	Theoretical/Descriptive
	The main components of a production rule model are its working memory and production rules.	working memory elements
9. **Paragraph-Level Extraction**: extract an entire sentence from the original text.	... In this project, we have designed a Hololensbased system in which collaborators are exposed to an unstructured learning activity in which they learned about the invisible physics involved in audio speakers. ...	The authors developed a Hololens-based system to provide AR learning experience for students.
	The second finding is that it is not just the quantity of practice but also the quality. Chess duffers get good by playing. Grand masters get great by further studying famous games. *Deliberate practice involves focusing on what is beyond one's current skill set rather than just executing what one is already able to do.*	Deliberate practice involves focusing on what is beyond one's current skill set rather than just executing what one is already able to do.
10. **Inference**: apply logical reasoning in order to reach a new conclusion based on the original text.	The second explanation could be that students felt an increased rapport, or sameness, with the agent in our system who spoke in their own dialect, as students typically learn from those who are more similar to themselves [33].	Students felt an increased rapport, or sameness, with the agent in our system who spoke in their own dialect, making them feel more comfortable learning and sharing.
	Furthermore, we investigate how much of the learning effects are due to the novelty of AR technology, by comparing a condition involving just physical interaction with the system without AR visualizations and the same physical system with simple AR visualizations (with no educational content).	The paper also investigates whether the novelty of AR technology itself would lead to learning benefits.
11. **Enumeration Extraction**: extract list items from the text.	RQ1: *Are participant attitudes influenced by the presence of educational AR representations?* RQ2: *Is the understanding of learning content influenced by the presence of educational AR representations?* RQ3: *Is group collaboration influenced by the presence of educational AR representations?* RQ4: *Does the mere presence of AR technology (without any educational content) affect participant experience?*	Are participant attitudes influenced by the presence of educational AR representations? Is the understanding of learning content influenced by the presence of educational AR representations? Is group collaboration influenced by the presence of educational AR representations? Does the mere presence of AR technology (without any educational content) affect participant experience?
	Cooke (1994) conducted one of the more extensive reviews of CTA. She identified three broad families of techniques: (a) *observation and interviews*, (b) *process tracing*, and (c) *conceptual techniques*.	observation and interviews, process tracing, conceptual techniques

Table 3: Text transformation operation strategy definitions and examples (part 2).