Assessing the Quality of Student-Generated Short Answer Questions Using GPT-3

Steven Moore(✉), Huy A. Nguyen, Norman Bier, Tanvi Domadia, and John Stamper

Carnegie Mellon University, Pittsburgh, PA 15213, USA
StevenJamesMoore@gmail.com

Abstract. Generating short answer questions is a popular form of learnersourcing with benefits for both the students' higher-order thinking and the instructors' collection of assessment items. However, assessing the quality of the student-generated questions can involve significant efforts from instructors and domain experts. In this work, we investigate the feasibility of leveraging students to generate short answer questions with minimal scaffolding and machine learning models to evaluate the student-generated questions. We had 143 students across 7 online college-level chemistry courses participate in an activity where they were prompted to generate a short answer question regarding the content they were presently learning. Using both human and automatic evaluation methods, we investigated the linguistic and pedagogical quality of these student-generated questions. Our results showed that 32% of the student-generated questions were evaluated by experts as high quality, indicating that they could be added and used in the course in their present condition. Additional expert evaluation identified that 23% of the student-generated questions assessed higher cognitive processes according to Bloom's Taxonomy. We also identified the strengths and weaknesses of using a state-of-the-art language model, GPT-3, to automatically evaluate the student-generated questions. Our findings suggest that students are relatively capable of generating short answer questions that can be leveraged in their online courses. Based on the evaluation methods, recommendations for leveraging experts and automatic methods are discussed.

Keywords: Question generation · Question quality · Question evaluation

1 Introduction

Students generating short answer questions has been proven to support their learning of new instructional content [4, 9]. As students generate questions, they deeply engage with the subject matter and utilize critical thinking skills [13]. This process leverages student engagement in ways that provide meaningful data around student interaction integrated with new student-generated learning assets that can support future learners [15]. This is known as a form of learnersourcing, where students complete activities that produce content which can then be leveraged by future learners [20]. Several systems to support students in the generation and sharing of questions have been leveraged by

I. Hilliger et al. (Eds.): EC-TEL 2022, LNCS 13450, pp. 243–257, 2022.
https://doi.org/10.1007/978-3-031-16290-9_18

thousands of students [14, 19]. This usage has led to the student-authoring of nearly a million questions, while also supporting research demonstrating that student question generation can lead to positive learning outcomes [18].

On the other hand, the quality of student-generated questions can widely vary [26]. While existing learnersourcing tools can scaffold this process and guide students towards generating better questions, they often require external systems [14, 19]. Additionally, evaluating the multitude of student-generated questions presents another challenge, with past research relying on experts, other students, or automated methods [24]. Automated methods often rely on the surface-level features of the question, such as the readability of text length, without including the pedagogical value it adds to a course. Recent research has developed and utilized a rubric for human evaluation of automatically generated questions that includes both linguistic and pedagogical criteria [16, 31]. However, these criteria have not seen wide adoption in automated evaluation methods, largely due to the difficulties associated with encoding them in a machine-interpretable way.

In this work, we explored how students could contribute short answer questions with minimal scaffolding and how we could assess their quality using machine learning models that match expert evaluations. We deployed a short answer question generation activity into seven instances of an online college-level chemistry course. From the student responses, we evaluated the quality of the short answer questions, determining if they were of sufficient quality, with respect to their pedagogical value, to be used in the course. The student-generated questions were also assessed for their cognitive level, in terms of Bloom's taxonomy [21]. Following this, we explored automatically evaluating the questions for their quality and cognitive level using a state-of-the-art language model. This study investigates the following research questions: (RQ1) *Can students generate high quality and educationally meaningful short answer questions?* (RQ2) *Can students generate short answer questions that target higher order cognitive processes with minimal prompting and scaffolding*? (RQ3) *Can we automatically assess the quality and cognitive level of a student-generated question with sufficient accuracy?*

Our work makes the following contributions towards learnersourcing and question evaluation. First, we demonstrate that students can create high-quality questions with a simple prompt that can be added to virtually any learning platform. Second, we present an expert evaluation process investigating the quality and cognitive level of student-generated questions. Third, we evaluate the usefulness of using a state-of-the-art language model in classifying educational questions, in an effort to make this process scalable and potentially saving instructor time. Ultimately, our work demonstrates how students can generate high quality questions with minimal scaffolding and how language models might be leveraged to assist in the quality and pedagogical evaluation of short answer questions.

2 Related Work

2.1 Students Generating Short Answer Questions

Previous work has explored leveraging learnersourcing for the creation of short answer questions and found that this process is beneficial to student learning, as it increases their engagement with the material and invokes critical thinking [9]. The quality of the student-generated questions can range depending on the study, influenced by factors such as the education level of the students and the domain of the course [4]. It is desirable to have students generate questions that assess the content in the course that they are making the questions for, but that has not already been assessed by an existing question, as it creates more practice opportunities [27]. Additionally, it is more beneficial for student learning if they generate questions that use higher order cognitive processes according to Bloom's revised Taxonomy. The revised Bloom's Taxonomy consists of six hierarchical categories, where each category corresponds to the cognitive processes that answering the question requires, from remembering a piece of information to combining information in a new way to create a new pattern or structure [21]. Research has shown that short answer questions typically assess at the lower levels of Bloom's Taxonomy, although it is possible for them to assess at all levels [12].

2.2 Evaluating the Quality of Student-Generated Questions

To evaluate student-generated questions, previous work typically leverages student performance data on the questions, using item response theory (IRT) techniques, or uses human experts to evaluate the questions according to a set of criteria [22]. Relying on IRT techniques that require student performance data on the questions can be detrimental to the learning process, because if the questions being used have not been first vetted for their quality, then they may be poorly constructed which can negatively impact students' performance and achievement [10]. In addition to IRT, previous research has leveraged experts or other students to review student-generated questions using a rubric consisting of different criteria such as language coherence, correctness, or the perceived difficulty [4, 23]. The criteria used in these past studies often focus on the surface-level aspects of the question, rather than including the pedagogical value of them, such as how well they might fit back into a given course or assess relevant content that has previously not been assessed. There has been a 9-item rubric used in two previous studies [16, 31] that assess both the linguistic and pedagogical qualities of questions in their expert evaluation. Unlike previous studies that utilize human evaluation, this rubric requires the evaluators to have domain knowledge of the questions and keep in mind how the question might be used in a course teaching the given domain. In the present study, we adopt this rubric to evaluate the student-generated short answer questions, as it is comprehensive, easy to interpret, and includes the pedagogical aspects of a question.

2.3 Automatically Evaluating Student-Generated Questions

A challenge in evaluating questions, whether automatically generated or created by students, is that human evaluation can be subjective, influenced by their prior knowledge and linguistic preferences [3]. To overcome this subjectivity, researchers commonly use automatic methods of evaluating questions [11]. These methods often utilize metrics related to the readability and explainability of the question, such as the popular natural language processing (NLP) ones of BLEU and METEOR [29]. These metrics are not appropriate for the present study, as we take a pedagogical perspective in evaluating the questions and previous research has indicated these metrics do not correlate with human evaluation [23]. Other automatic evaluation work has utilized deep learning methods and language models to evaluate the quality of questions, comparing it to the same evaluation done by a set of human experts [8, 28]. While these studies have achieved a model to expert human matching of 81%, surpassing the previous baseline of 42%, they focus their evaluation on the surface-level features of the questions, such as the length, word choice, or grammar, without considering the pedagogical value it might bring to a course [28].

In addition to automatically evaluating the quality of questions, previous work has looked to automatically classify questions according to which level of Bloom's Taxonomy they fit into [17, 30]. These studies have achieved classification accuracies ranging from 70% to 87%, however they note that the performance is limited by the training data used and that categorization was more accurate for the lower levels of Bloom's Taxonomy [37]. The automatic evaluation methods used in these, and many other prior studies are on questions that typically assess reading comprehension, at the lower cognitive levels of Bloom's Taxonomy, and do not require domain knowledge [2]. This is different from the questions used in the present study, which are at an advanced education level and contain domain knowledge, rather than the more basic recall and comprehension type questions traditionally used.

3 Learning Platform and Data Collection

The present study takes place in a digital courseware platform known as the Open Learning Initiative (OLI). OLI is an open-ended learning environment that offers courses from a variety of domains and consists of interactive activities and diverse multimedia content [5]. OLI consists of instructional content and low-stakes, also known as formative, activities. These activities consist of a variety of question types such as multiple-choice questions, short answer, and dropdown style questions. Students work through different modules in the system, akin to chapters in a textbook, where they are presented with instructional text and videos. Low-stakes activities are embedded throughout these instructional materials, providing the students with feedback and practice opportunities to assess the concepts they are learning.

The data used in this study was collected from a week-long module in seven instances of an introductory chemistry course taught at a community college in the western U.S. The course consists of first- and second-year undergraduates from varying degree backgrounds, with most of the students pursuing a chemistry-related degree. The data comes from the fall semester of 2021, when the introductory chemistry course was offered in the OLI system. In total, the data consists of 143 students and their contribution to the short answer generation activity. The OLI content the students used during the week when our data was collected covers the topics of pH, buffers, and amino acids. There are a total of 38 low-stakes activities embedded throughout the pages of this module. Every activity provides the students with detailed instructional feedback, for both incorrect and correct responses.

We focus on an activity that was added to the course that involves each student generating a short answer question. In the chemistry course, this activity is found on a page contains four paragraphs of instructional text, three worked examples, and eight multiple-choice questions. This activity is presented in the same low-stake format as the other activities found throughout the course, as students do not receive a grade for their participation or the quality of their response in the activity. It prompts students to generate a short answer question, by asking them to "*Create a short answer question that can be correctly answered based on the content covered in this module*". In the activity, students are first prompted to write the question text in the provided text box on the top part of the activity and then write the answer to the question in the bottom text box. The instructions for the self-explanation are intentionally brief and similar prompts have been used in related studies by [1, 36].

4 Data Analysis

4.1 Human Evaluation

The 143 student-generated short answer questions were evaluated by two experts to assess their quality and Bloom's taxonomy level. The two experts had content knowledge in chemistry, multiple years of teaching experience, familiarity with the OLI course, and ample previous experience coding qualitative student data. To first evaluate the quality of the questions, the two experts used a 9-item rubric that has been used in previous studies for assessing the linguistic and pedagogical quality of questions [16, 31]. This rubric contains 9 hierarchical criteria, shown in Table 1. These criteria are asked to the two experts in the order, from top to bottom, that they are presented in the table. Eight of the rubric criteria involve binary (yes/no) responses. The only non-binary item is *information needed*, which consists of three unique options, where each corresponds to the location of the information the students need to know in order to successfully answer the question.

The rubric items are hierarchical by nature, meaning that if certain criteria are answered as "no", then the remaining items will be marked as "not applicable". These criteria are bolded in Table 1. For example, if the experts answer "no" to the *answerable* rubric item, then the three items that follow will be marked as "not applicable". This contributes to avoiding distortion of the rubric criteria distributions for questions that are not ratable across certain items and helps to save the expert evaluators' time. The inter-rater reliability (IRR) values between the two evaluators for each rubric item are also reported in Table 1. It includes the percentage agreement and Cohen's Kappa κ statistic [25] as a measure of IRR for all rubric items. These items are at either a near perfect or substantial level of agreement between the two raters. Two of them, *domain related* and *central*, had perfect agreement, as all of the student-generated questions pertained to chemistry content covered in the current OLI module.

If the expert evaluators answer "yes" to all the binary rubric items and answer any of the three options for *information needed* then we consider that to be a high quality question. In line with previous research, meeting all the rubric criteria suggests that the question is both linguistically and pedagogically sound [16, 31]. Additionally, the last rubric criteria *would you use it* asks the evaluators if they would use the student-generated question if they were teaching the course and using the OLI materials. As the evaluators are familiar with the OLI content and have prior teaching experience, they can judge the pedagogical quality of the student-generated questions. However, we acknowledge that despite the two expert evaluators' backgrounds and high IRR they can still interpret

Table 1. The hierarchical 9-item rubric used to evaluate the questions; the bolded criteria stop the review process if answered as "no". The bracketed numbers indicate agreement percentage between raters and Cohen's κ value for each item.

Rubric item	Definition
Understandable (97.20%, $\kappa = 0.83$)	Could you understand what the question is asking?
DomainRelated (100%, $\kappa = 1.0$)	Is the question related to the Chemistry domain?
Grammatical (96.15%, $\kappa = 0.82$)	Is the question grammatically well formed, i.e. is it free of language errors?
Clear (98.46%, $\kappa = 0.83$)	Is it clear what the question asks for?
NotRephrasing (89.52%, $\kappa = 0.66$)	Does the question assess course content that has not been assessed by an existing question in the course?
Answerable (99.19%, $\kappa = 0.88$)	Are students probably able to answer the question?
InformationNeeded (88.14%, $\kappa = 0.73$)	(op) Information presented directly and in one place only in the text (dp) Information presented in different parts of the text (te) A combination of information from the text with external knowledge
Central (100%, $\kappa = 1.00$)	Do you think being able to answer the question is important to work on the topics covered by the current module?
WouldYouUseIt (82.35%, $\kappa = 0.62$)	If you were a teacher working with the OLI module in your class, would you include this question in the course?

the student-generated questions in different ways as influenced by their prior knowledge and linguistic preferences [3].

In order to assess the cognitive level of the student-generated questions, the two expert evaluators utilized Bloom's revised Taxonomy [21]. This taxonomy, shown in Table 2, has been applied to educational questions in prior research [17, 37]. It consists of six different levels, where each one corresponds to the cognitive processes involved in answering the question. Using these six taxonomy levels, the two expert evaluators classified each student-generated question to a level, depending on what cognitive process is required to answer it. Note, only student-generated questions that had no "non applicable" answers to the nine rubric criteria were evaluated in this way, resulting in a total of 120 of the 143 (84%) questions being assigned one of the six levels as agreed upon by the two expert evaluators. While there are six levels to the taxonomy, the student-generated questions in this study were all assigned to the first four levels, as none of the questions targeted the cognitive processes of *evaluate* or *create*. The omission of these two levels was not by design, however they are less common for short answer questions typically found in courses, which are more likely to assess the first four levels of Bloom's taxonomy [30]. Additionally, while assessing the questions using the 9-item rubric and for Bloom's taxonomy, the two expert evaluators had disagreements, as indicated by the Kappa values in Table 1. The discordant criteria for such questions were discussed between the two raters, resulting in them reaching a consensus on the categorization of the question.

Table 2. Six levels of Bloom's revised Taxonomy [21] in ascending cognitive order from lowest to highest, along with their operational definitions.

Bloom's level	Definition
Remember	Retrieve relevant knowledge from long-term memory
Understand	Construct meaning from instructional messages, including written communication
Apply	Carry out or using a procedure in a given situation
Analyze	Break down the learning material into constituent parts and determine how parts relate to one another and to an overall structure
Evaluate	Make judgments based on criteria and standards
Create	Put elements together to form a coherent whole or to reorganize into a new structure

The IRR between the two expert evaluators for applying Bloom's revised Taxonomy to the student-generated questions was assessed via percentage of agreement (81.67%) and Cohen's Kappa ($\kappa = .74$), suggesting a substantial level of agreement. This agreement level is akin to previous studies that applied Bloom's revised Taxonomy to student-generated questions [35]. In accordance with previous research [21, 34], we define a student-generated question as assessing a low cognitive level if it was evaluated to be at the *remember* or *understand* levels. Conversely the question is said to assess at a high cognitive level if it was evaluated to be at the *apply*, *analyze*, *evaluate*, or *create* levels. Typically, multiple-choice and short answer questions rely on the cognitive processes associated with lower cognitive levels, although both question types can assess higher levels [33]. It is desirable to have questions assessed at a higher level, as it is more beneficial for student learning [21].

4.2 Automatic Evaluation Using GPT-3

Our second evaluation method utilizes GPT-3, a language model with up to 175 billion parameters trained on a large dataset of text scraped from the internet [6]. We selected this language model for our evaluation due to it being state-of-the-art for multiple natural language processing tasks and being the largest publicly available transformer language model. It is a high-performing and popular language model choice for text classification, with recent applications in classifying emails [32] and determining if news articles were real or fake [7]. In this work, we used GPT-3 to perform classification on the student-generated questions in two different ways. We avoided using typical automated question generation evaluation criteria such as BLEU or METEOR, as they have been proven to not correlate with human evaluation and do not have pedagogical implications [29].

First, we used it for binary classification to see if it could classify the student-generated questions as being low or high quality, matching the evaluation of the two experts. To make this classification, we first fine-tuned a GPT-3 Ada model on the LearningQ dataset [8], which is publicly available and contains 5,600 student-generated short answer questions from Khan Academy. Each question in this dataset was evaluated by two expert instructors and assigned a label corresponding to if it was *useful for learning* or *not*. The researchers for the LearningQ dataset defined a question as being *useful for learning* akin to several of the rubric criteria we utilized in this study. They based their evaluation on the following three criteria: (i) concept-relevant, seeking information on the concepts taught in the course; (ii) context-complete, providing enough information to be answerable by other students; and (iii) not-generic, meaning the question asks about a course concept not on another topic or of another style, such as asking for learning advice. Additionally, the questions in the LearningQ dataset came from a variety of domains, which included STEM courses and a single humanity one. No preprocessing was performed on the questions used to fine–tune the model; they were used as-is from the publicly available dataset along with their corresponding binary labels. Fine-tuning the model with default hyperparameters suggested by the documentation[1] took approximately 10 min and incurred a cost of $0.21. Upon completion, we passed in the student-generated questions as the GPT-3 model's input, obtaining the output as a binary label indicating if it rated each question as useful for learning (*high quality*) or not (*low quality*).

Secondly, we used another instance of the GPT-3 Ada model to perform multiclass classification using Bloom's revised Taxonomy levels. We once again use GPT-3 Ada, which was selected due to its low cost and effectiveness at classification tasks that are less nuanced, with comparable performance to the Davinci model. We wanted to determine if GPT-3, fine-tuned on example questions from each level, could perform similarly to the two expert evaluators. To fine-tune the model, we utilized a dataset consisting of 100 questions mapped to each of six Bloom's revised Taxonomy levels, for a total of 600 questions [34]. These 600 questions were assigned to a level of Bloom's revised Taxonomy by a pedagogical expert and this dataset has been used in ample previous studies involving fine-tuning and classification tasks. In the present student, the expert evaluation of the student-generated questions only identified four of the six Bloom's levels that were applicable to the questions. However, we included questions from the two unused Bloom's levels in the fine-tuning process, because if the model was accurate, we could utilize it for future datasets that may contain questions at that cognitive level. For this dataset, we performed no preprocessing on the questions used to fine–tune the model; they were used as-is from the publicly available dataset along with their corresponding Bloom's revised Taxonomy labels. We once again fine-tuned the model with default hyperparameters which took approximately 5 min and incurred a cost of $0.08. Upon completion, the student-generated questions were passed as the GPT-3 model's input, outputting Bloom's labels for each question.

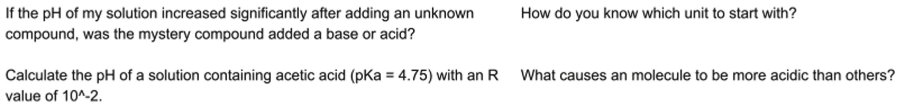
[1] We used the default hyperparameters as suggested in https://beta.openai.com/docs/guides/fine-tuning.

5 Results

We first begin with our human evaluation by experts, using the 9-item rubric, across all 143 student-generated short answer questions. As indicated in the Data Analysis section, the rubric criteria are hierarchical and they can be marked as "not applicable", causing the following rubric items to be ignored. For example, if a question was marked "not applicable" for the first rubric criteria of *understandable*, that would reduce the question pool for the other eight criteria. We report the percentage relative to the remaining questions, followed by the absolute percentage, i.e. (relative %/absolute %).

RQ1: *Can students generate high quality and educationally meaningful short answer questions?* We found that 91% of the student-generated short answer questions were rated *understandable*. All the questions rated as *understandable*, were also rated *domain related* (100%/91% total). Most questions were also free of *grammatical* errors (90%/82% total), which includes typos and punctuation mistakes. As a question's clarity is related to the understandability of the question, there were also many questions (95%/87% total) that were evaluated as being *clear*. If a question assessed course content that has not been assessed by an existing question found somewhere in the module, then it was marked as *not rephrasing* (84%/73% total). This is one of the lowest rubric criteria percentages and also presented a challenge for the evaluators to find agreement on, as they achieved a Cohen's Kappa of $\kappa = .66$.

The evaluation shows that most of the questions are rated as *answerable* by future students in the course (97%/84% total). Similar to the criteria about being domain related, the *central* criteria (100%/84% total) was perfect for the remaining pool of questions. This not only means the question relates to the chemistry, but it specifically targets a concept that is addressed in the current module. According to the evaluators, knowledge required for answering the questions is obtained in *one place* (68%/57% total) or in *different places* (30%/25% total) throughout the module. However, there were two questions that were evaluated as needing both the instructional *text and external knowledge* (2%/1% total).



If the pH of my solution increased significantly after adding an unknown compound, was the mystery compound added a base or acid?

How do you know which unit to start with?

Calculate the pH of a solution containing acetic acid (pKa = 4.75) with an R value of 10^-2.

What causes an molecule to be more acidic than others?

Fig. 1. The two questions on the left are evaluated as being high quality and the two questions on the right are low-quality, due to being vague (top) and grammatically incorrect (bottom).

As described in the Data Analysis section, a question was categorized as high quality if it passed all nine rubric criteria, including being evaluated as *would you use it* (38%/32% total). In total, 46/143 (32%) student-generated short answer questions met this criterion by passing all nine rubric items and were deemed to be of high quality. Figure 1 shows two questions evaluated as high quality and two questions evaluated as low-quality. The question in the upper-right was evaluated as not being *understandable* and the question in the bottom-right was not *grammatical*.

RQ2: Can students generate short answer questions that target higher order cognitive processes with minimal prompting and scaffolding? In order to assess the cognitive-level of the student-generated questions, the evaluators applied Bloom's Taxonomy to them. Due to some of the questions having certain rubric criteria marked as "not applicable" and thus ending the review, 120/143 (84%) student-generated questions were assigned a Bloom's Taxonomy level by the evaluators. The majority categorization was remember (52%), with understand (25%) and apply (20%) being tagged to a similar number of questions, followed by analyze (3%). An example of the student-generated questions corresponding to each of these four Bloom's Taxonomy levels is shown in Table 3.

Table 3. An example of a student-generated question assessed at each of the four levels of Bloom's Taxonomy present in this study.

Student-generated question	Bloom's level
What is the point in a titration curve that indicates the pKa value of a weak acid?	Remember
Imagine an acidic solution with a low pH. If a strong base is added to the solution, what happens to the pH in relation to the pKa?	Understand
If 10 mL of a diprotic weak acid is fully deprotonated with 20 mL of 0.5M NaOH, how many moles of the acid and NaOH are there?	Apply
When stomach acid enters the esophagus, typically with a pH of 1.5 to 3.5, calcium carbonate is often used to combat this. Why would calcium carbonate be a good substance for this problem?	Analyze

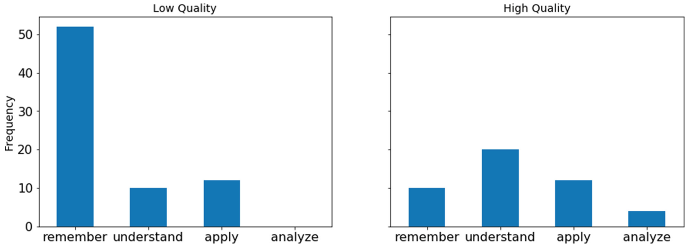


Fig. 2. The distribution of the four Bloom's Taxonomy levels between questions evaluated as low and high quality.

Prior research [21, 30] has indicated that questions at the *apply* level and above are categorized as targeting higher order cognitive processes. As a result, 28/120 (23%) questions tagged with Bloom's Taxonomy were evaluated as assessing at this higher level. Since Bloom's Taxonomy level was not included in the criteria for a high-quality question, we investigated if there was a correlation between the two measures. Fisher's exact test revealed that there was a strong statistically significant association between the quality of the question and the cognitive level ($p = .003$). Figure 2 shows the distribution of Bloom's Taxonomy levels between questions evaluated as being low and high quality.

RQ3: *Can we automatically assess the quality and cognitive level of a student-generated question with sufficient accuracy?* We utilized the first fine-tuned GPT-3 model to classify the quality of the student-generated questions as either low or high quality. The model agreed with the human evaluation for 57/143 questions (40%). In the cases they disagreed, 85/86 mismatches were interpreted as having high quality by GPT-3 but low quality by expert raters. There were only 13/143 questions (9%) the model classified as low quality, suggesting it was overestimating the quality of the questions, as 97/143 (68%) were evaluated by the experts as being low quality. Figure 3 provides a confusion matrix for the quality classifications made by the model.

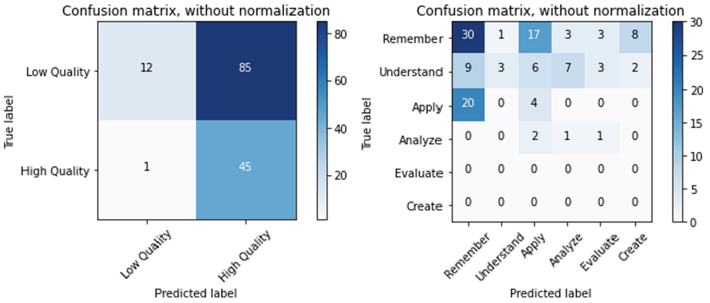


Fig. 3. Confusion matrices for the classification of a question's quality (left) and Bloom's revised Taxonomy (right).

We used the second fine-tuned GPT-3 model to classify the 120 student-generated questions to which the expert evaluators had assigned a Bloom's Taxonomy level. The results of the model compared to the expert evaluation, including the percentage of matches for each Bloom's Taxonomy level between the two, are shown in Table 4. In total, the model matched the expert evaluation for 38/120 (32%) student-generated questions. The GPT-3 model has a similar distribution of *remember* and *apply* questions, although they are often not correctly applied to the questions according to the expert evaluation. Additionally, GPT-3 classified 17 of the questions into the two highest cognitive levels that were not observed in our student-generated questions. Additionally, Fig. 3 also provides a confusion matrix for the classification of Bloom's revised Taxonomy between the expert human evaluators and the model.

Table 4. A breakdown of the six Bloom's revised Taxonomy and the number of questions the experts and GPT-3 tagged to each level.

Bloom's level	Remember	Understand	Apply	Analyze	Evaluate	Create
Expert Evaluation	62	30	24	4	0	0
GPT-3	59	4	29	11	10	7
Matching %	48%	10%	4%	25%	0%	0%

6 Discussion and Conclusion

In this research, we utilized human experts and automatic methods to evaluate the quality and cognitive level of student-generated short answer questions. We found that students were able to contribute high quality questions, as evaluated by a 9-item rubric that contained criteria assessing the linguistic and pedagogical features of the questions. In total, 32% of the student-generated short answer questions were evaluated as being high quality, indicating that the evaluators could use them in the course in their present condition. Students generated these questions through a simplistic prompt consisting of a single sentence instruction and two textboxes embedded into a digital learning platform. Previous research often has an overall lower percentage of high-quality questions and utilizes external systems or scaffolding methods that require the students to spend more time on the question generation activity [1, 4]. We believe that the implementation we used in this study keeps students more engaged in the learning process, by allowing them to create the question in a more natural context as they work through the instructional text and assessments in the platform.

The cognitive processes that the student-generated questions target were evaluated by the two expert evaluators, which identified 23% of the questions as assessing at a high cognitive level and the remaining 77% assessing the lower two cognitive levels. This majority distribution of the short answer questions assessing at the *remembering* and *understanding* cognitive levels is in line with findings from previous work [2, 37]. These questions that assess the first two cognitive levels can still be effective, particularly when students are first learning new concepts, where they might need to first learn essential terminology, methods, and formulas [21].

Automatic evaluation of the student-generated questions for both their quality and cognitive level was suboptimal compared to previous work leveraging different language models [8, 28], however, such prior research often evaluates questions that are mostly at the *remembering* cognitive level and often involve basic reading comprehension with no domain-related knowledge being assessed, which are more appropriate for students at lower education levels [22]. The student-generated questions in this study were at the post-secondary education level, assessed chemistry knowledge, and often included domain terminology. These differences between questions used in prior research in this study likely contributed to the difficulty the two GPT-3 models had, even when they were fine-tuned on relevant data for the classification tasks. The percentage of expert-matching classifications the models achieved for the quality (40%) and cognitive level (32%) could provide an initial estimation of the questions' value.

The main limitation of this study comes from the dataset, as the 143 student-generated short answer questions that were analyzed were all in the domain of chemistry. Including student-generated questions from other domains could lead to more generalizable findings. Question evaluation often entails human annotations as the ideal criterion to compare automatic methods against; however, there is a subjective nature to human ratings. While we tried to reduce subjectivity by using a detailed rubric for the human evaluation and achieving a high IRR for each criterion, there still lies the potential for different evaluation depending on who is doing the evaluation. Finally, the results of the GPT-3 model were suboptimal, often overestimating the quality of the student-generated questions or misclassifying the Bloom's revised Taxonomy level. The results of these classifications were influenced by the datasets used to fine-tune them, which was limited by public datasets that classify the educational quality of the question and the cognitive level.

This work demonstrates that students can generate short answer questions that are both linguistically and pedagogically sound without requiring an external tool or scaffolding. In total, we found that 32% of all the student-generated questions were evaluated as being high quality by the expert evaluators. Across all the questions that were classified according to Bloom's revised Taxonomy, 23% were evaluated as assessing high cognitive levels. Our results highlight how students in the context of an online course can create short answer questions that can readily be implemented into the course, providing new assessment opportunities for essential concepts. While the automatic evaluation may be improved with more robust datasets for fine-tuning, it offers a sufficient first pass classification that may assist experts in their evaluation of the questions. This research helps demonstrate one way to help scale online learning and improve educational resources, by leveraging the students in a course. It opens further opportunities for engaging students in the process of question generation and leveraging both humans and language models to assist in the evaluation process.

This research was supported by the National Science Foundation (Award #2135159).

References

1. Aflalo, E.: Students generating questions as a way of learning. Act. Learn. High. Educ. 1469787418769120 (2018)
2. Amidei, J., Piwek, P., Willis, A.: Evaluation methodologies in automatic question generation 2013–2018. In: Proceedings of the 11th International Conference on Natural Language Generation, pp. 307–317 (2018)
3. Amidei, J., Piwek, P., Willis, A.: Rethinking the agreement in human evaluation tasks. In: Proceedings of the 27th International Conference on Computational Linguistics, pp. 3318–3329 (2018)
4. Bates, S.P., Galloway, R.K., Riise, J., Homer, D.: Assessing the quality of a student-generated question repository. Phys. Rev. Spec. Top.-Phys. Educ. Res. **10**(2), 020105 (2014)
5. Bier, N., Moore, S., Van Velsen, M.: Instrumenting courseware and leveraging data with the open learning initiative. In: Companion Proceedings 9th International Conference on Learning Analytics & Knowledge, pp. 990–1001 (2019)
6. Brown, T., et al.: Language models are few-shot learners. In: Advances in Neural Information Processing Systems, vol. 33, pp. 1877–1901 (2020)
7. Chan, A.: GPT-3 and InstructGPT: technological dystopianism, utopianism, and "Contextual" perspectives in AI ethics and industry. AI Ethics 1–12 (2022)
8. Chen, G., Yang, J., Hauff, C., Houben, G.-J.: LearningQ: a large-scale dataset for educational question generation. In: Twelfth International AAAI Conference on Web and Social Media (2018)
9. Chin, C., Brown, D.E.: Student-generated questions: a meaningful aspect of learning in science. Int. J. Sci. Educ. **24**(5), 521–549 (2002)
10. Clifton, S.L., Schriner, C.L.: Assessing the quality of multiple-choice test items. Nurse Educ. **35**(1), 12–16 (2010)
11. Clinciu, M.-A., Eshghi, A., Hastie, H.: A study of automatic metrics for the evaluation of natural language explanations. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main, pp. 2376–2387 (2021)

12. Das, S., Mandal, S.K.D., Basu, A.: Identification of cognitive learning complexity of assessment questions using multi-class text classification. Contemp. Educ. Technol. **12**(2), ep275 (2020)
13. Denny, P.: Generating practice questions as a preparation strategy for introductory programming exams. In: Proceedings of the 46th ACM Technical Symposium on Computer Science Education, pp. 278–283 (2015)
14. Denny, P., Hamer, J., Luxton-Reilly, A., Purchase, H.: PeerWise: students sharing their multiple choice questions. In: Proceedings of the Fourth international Workshop on Computing Education Research, New York, NY, USA, pp. 51–58 (2008)
15. Denny, P., Tempero, E., Garbett, D., Petersen, A.: Examining a student-generated question activity using random topic assignment. In: Proceedings of the 2017 ACM Conference on Innovation and Technology in Computer Science Education, pp. 146–151 (2017)
16. Horbach, A., Aldabe, I., Bexte, M., de Lacalle, O.L., Maritxalar, M.: Linguistic appropriateness and pedagogic usefulness of reading comprehension questions. In: Proceedings of the 12th Language Resources and Evaluation Conference, pp. 1753–1762 (2020)
17. Huang, J., et al.: Automatic classroom question classification based on bloom's taxonomy. In: 2021 13th International Conference on Education Technology and Computers, pp. 33–39 (2021)
18. Khosravi, H., Demartini, G., Sadiq, S., Gasevic, D.: Charting the design and analytics agenda of learnersourcing systems. In: LAK21: 11th International Learning Analytics and Knowledge Conference, pp. 32–42 (2021)
19. Khosravi, H., Kitto, K., Williams, J.J.: RiPPLE: a crowdsourced adaptive platform for recommendation of learning activities. J. Learn. Anal. **6**(3), 91–105 (2019)
20. Kim, J.: Learnersourcing: improving learning with collective learner activity. Massachusetts Institute of Technology (2015)
21. Krathwohl, D.R.: A revision of Bloom's taxonomy: an overview. Theory Pract. **41**(4), 212–218 (2002)
22. Kurdi, G., Leo, J., Parsia, B., Sattler, U., Al-Emari, S.: A systematic review of automatic question generation for educational purposes. Int. J. Artif. Intell. Educ. **30**(1), 121–204 (2020)
23. van der Lee, C., Gatt, A., van Miltenburg, E., Krahmer, E.: Human evaluation of automatically generated text: Current trends and best practice guidelines. Comput. Speech Lang. **67**(2021), 101151 (2021)
24. Lu, O.H., Huang, A.Y., Tsai, D.C., Yang, S.J.: Expert-authored and machine-generated short-answer questions for assessing students learning performance. Educ. Technol. Soc. **24**(3), 159–173 (2021)
25. McHugh, M.L.: Interrater reliability: the kappa statistic. Biochemia Medica **22**(3), 276–282 (2012)
26. Moore, S., Nguyen, H.A., Stamper, J.: Examining the effects of student participation and performance on the quality of learnersourcing multiple-choice questions. In: Proceedings of the Eighth ACM Conference on Learning@ Scale, pp. 209–220 (2021)
27. Papinczak, T., Peterson, R., Babri, A.S., Ward, K., Kippers, V., Wilkinson, D.: Using student-generated questions for student-centred assessment. Assess. Eval. High. Educ. **37**(4), 439–452 (2012)
28. Ruseti, S., et al.: Predicting question quality using recurrent neural networks. In: Penstein Rosé, C., et al. (eds.) AIED 2018. LNCS (LNAI), vol. 10947, pp. 491–502. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-93843-1_36
29. Scialom, T., Staiano, J.: Ask to learn: a study on curiosity-driven question generation. In: Proceedings of the 28th International Conference on Computational Linguistics, pp. 2224–2235 (2020)

30. Shaikh, S., Daudpotta, S.M., Imran, A.S.: Bloom's learning outcomes' automatic classification using LSTM and pretrained word embeddings. IEEE Access **9**, 117887–117909 (2021)
31. Steuer, T., Bongard, L., Uhlig, J., Zimmer, G.: On the linguistic and pedagogical quality of automatic question generation via neural machine translation. In: De Laet, T., Klemke, R., Alario-Hoyos, C., Hilliger, I., Ortega-Arranz, A. (eds.) EC-TEL 2021. LNCS, vol. 12884, pp. 289–294. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-86436-1_22
32. Thiergart, J., Huber, S., Übellacker, T.: Understanding emails and drafting responses–an approach using GPT-3. arXiv e-prints (2021)
33. Wang, Z., Manning, K., Mallick, D.B., Baraniuk, R.G.: Towards blooms taxonomy classification without labels. In: Roll, I., McNamara, D., Sosnovsky, S., Luckin, R., Dimitrova, V. (eds.) AIED 2021. LNCS (LNAI), vol. 12748, pp. 433–445. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-78292-4_35
34. Yahya, A.A., Toukal, Z., Osman, A.: Bloom's Taxonomy–based classification for item bank questions using support vector machines. In: Ding, W., Jiang, H., Ali, M., Li, M. (eds.) Modern Advances in Intelligent Systems and Tools, pp. 135–140. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-30732-4_17
35. Yu, F.Y., Cheng, W.W.: Effects of academic achievement and group composition on quality of student-generated questions and use patterns of online procedural prompts. In: 28th International Conference on Computers in Education, ICCE 2020, pp. 573–581 (2020)
36. Yu, F.-Y., Liu, Y.-H.: Creating a psychologically safe online space for a student-generated questions learning activity via different identity revelation modes. Br. J. Educ. Technol. **40**(6), 1109–1123 (2009)
37. Zhang, J., Wong, C., Giacaman, N., Luxton-Reilly, A.: Automated classification of computing education questions using Bloom's taxonomy. In: Australasian Computing Education Conference, pp. 58–65 (2021)