KUIZ: Encouraging Modular Learnersourcing of Multiple Choice Questions through LLM Interventions

Hyoungwook Jin*
jinhw@kaist.ac.kr
School of Computing, KAIST
Daejeon, Republic of Korea

Haesoo Kim*†
hk778@cornell.edu
Cornell University
Ithaca, New York, USA
School of Computing, KAIST
Daejeon, Republic of Korea

Nathan Mekuria Haile
nathanmekuria@kaist.ac.kr
School of Computing, KAIST
Daejeon, Republic of Korea

Soyeong Min
ueaw@kaist.ac.kr
Industrial Design, KAIST
Daejeon, Republic of Korea

Juho Kim
juhokim@kaist.ac.kr
School of Computing, KAIST
Daejeon, Republic of Korea

ABSTRACT

The generation of multiple-choice questions (MCQs) is a well-documented pedagogical activity that fosters high-level thinking and understanding among learners. However, learners often find it challenging and less appealing compared to answering questions. We introduce KUIZ, a learnersourcing platform designed to encourage learner participation in MCQ generation through modular contributions and large language model (LLM) interventions. KUIZ breaks down the task into creating question stems and options, providing scaffolding and suggestions via LLMs to ease the process. A two-wave deployment in a university human-computer interaction class assessed the preliminary effect of the modular design and LLM features in KUIZ. Results indicated increased engagement and perceived learning benefits from the modular approach and LLM support.

CCS CONCEPTS

• **Human-centered computing → Collaborative and social computing systems and tools**.

KEYWORDS

Learnersourcing, Multiple-choice questions, Large language models

ACM Reference Format:
Hyoungwook Jin, Haesoo Kim, Nathan Mekuria Haile, Soyeong Min, and Juho Kim. 2024. KUIZ: Encouraging Modular Learnersourcing of Multiple Choice Questions through LLM Interventions. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale (L@S '24), July 18–20, 2024, Atlanta, GA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

*Both authors contributed equally to this research.
†This work was conducted while the author was at KAIST.

L@S '24, July 18–20, 2024, Atlanta, GA, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Despite the positive effects of question-generation activities in learning, many students still see them as challenging. Question generation requires high-level thinking and understanding of the subject, which can discourage participation from low-confidence learners [6]. When the multiple-choice question (MCQ) generation task is voluntary, only a small percentage of students opt to participate in the task [4]. Students often prefer answering questions to creating them, due to the larger perceived efficacy, although creating questions has a bigger learning effect [7, 14]. Highly motivated students will also create high-quality questions, but lesser motivated students will often create lower-quality ones or questions that require less cognitive involvement [14]. As a result, there is a divide in the benefits gained from the activity.

In this work, we explore methods of improving learners' motivation in MCQ activities by scaffolding the process through human-AI collaboration and co-creation. Previous research has looked into the impact of scaffolding activities on the quality of participation and the generated artifacts in learnersourcing [1]. We focus on two different methods of co-creation as instructional scaffolding: collaborative question-making through modular contributions and using large language models (LLMs) to provide guidance and suggestions. We developed KUIZ, a learnersourcing platform that introduces a modular, LLM-supported method of creating MCQs. In KUIZ, each question is divided into question stems and options, allowing learners to contribute in smaller or more flexible units, increasing engagement. KUIZ also facilitates the modular question-making process with LLM features, providing guidance for creative tasks and reducing the burden of collaboration.

We conducted our evaluation through two deployment sessions in a university-level human-computer interaction class. In the first session, we deployed the system without LLM features. In the second session, the system included the LLM features. 43 Students generated question stems and options based on assigned reading materials and submitted surveys about task preference, the utility of LLM features, and perceived learning effects. The evaluation results showed that the LLM features introduced increased students' preference for generating question stems, which is assumed to be more

difficult than generating options. We discuss the positive impact of LLM features and the challenges in our current design of KUIZ.

2 BACKGROUND

Learnersourcing is a method of crowdsourcing where the participants are learners who engage in pedagogical activities while also collectively creating practical artifacts for future learners [9, 13]. While learnersourcing provides a scalable and efficient method of crowdsourcing learning artifacts, learnersourcing is unique in that the individuals can benefit directly from the activity.

Much previous work has explored the use of learnersourcing to create MCQs. The most prominent example is PeerWise [3], a learning tool where students can create and evaluate multiple-choice questions. UpGrade [15] explored how to use MCQs as a structural scaffold to help students understand the material, while other systems such as RiPPLE [8] employed them among many diverse types of personalized learning resources. Such question-generating activities have been found to encourage higher-level thinking in students [14] as well as engagement with the class material [3, 5].

However, students often do not prefer to participate in generative tasks due to lack of confidence, motivation, or knowledge [7, 14]. For example, distractor writing is particularly difficult because it requires students to make subtle distinctions in the different concepts that they have learned [11]. Considering that the learning effect of MCQ generation also depends on their perceived difficulty [12, 18], there needs to be further emphasis on how to lower the barriers to participation for a scalable learning effect. Thus, exploring how to motivate learners is crucial to maintaining the sustainability and scalability of learnersourcing systems [4].

One way to mitigate this is to introduce microtasks. Many learnersourcing systems encourage each learner to complete complex, multi-step tasks, which are significantly harder to complete [13]. In comparison, we suggest collaborative learnersourcing could be more beneficial [17]. We suggest dividing multiple-choice questions into 2 modular components: (i) the question stem, where the learning objective and learners' task are defined; and (ii) the options, which are the 'multiple-choice' element. Maintaining the creative element while reducing the total workload for a single unit of contribution can lower the barriers to participating in the MCQ generation task. Furthermore, by actively engaging with others' questions or options, learners could interact with more diverse topics and learn to adapt through the collaborative learning process.

3 SYSTEM

We introduce *KUIZ*, a learning platform where students can learn by collaboratively generating multiple-choice questions with their peers under guidance and feedback from large-language models.

3.1 Collaborative and Modular MCQ Generation

KUIZ has two main parts, where learners can 1) create original question stems and 2) add distractors to the questions others have made. When creating question stems (Fig. 1, left), learners first fill out the learning objective of their questions. The learning objective is pre-templated; learners choose one of six levels in Bloom's taxonomy [10] and a learning topic from dropdown menus. Learners then write their questions, answers, and explanations of their intention behind those questions to guide later learners who will contribute to the questions. We designed KUIZ not to require distractors in question creation to reduce their initial load. Completed questions are displayed on a dashboard page where all the questions from peer learners are listed. Learners can check and solve each other's questions and improve them by adding distractors (Fig. 1, right). Learners read learning objectives, explanations, and questions from previous learners and add distractors or alternative answers that can add complexity to the questions.

3.2 LLM-Generated Scaffolds and Feedback

Learners receive support from KUIZ when creating questions and distractors. We had two base rules when designing the support. First, KUIZ should not provide finished content, encouraging learners to finish up based on the provided building blocks (e.g., keywords). This will prevent learners from copying and pasting and stimulate them to connect their prior knowledge to the given suggestions. Second, KUIZ should provide support only when learners request it. Request-based scaffolding can help learners control the difficulty of the question-generation task by selecting the level of assistance. Based on these rules, we devised two types of support—suggestions and rephrasing. Suggestions target learners who struggle to start and scaffold them to explore different ideas. Learners are given question templates (Fig. 1, A) or keywords to begin their questions (Fig. 1, B & E). Rephrasing encourages learners to refine their existing questions. Rephrasing includes low-level feedback, such as checking grammar (Fig. 1, C), and high-level feedback, such as ideas to improve created content (Fig. 1, D & F).

KUIZ uses OpenAI's *text-davinci-003* model to generate all the suggestions and rephrasing feedback. We wrote prompts for each scaffold with few-shot examples and provided LLMs with metadata of questions, such as learning objectives, explanations, and answers. We set the model temperature to 1.0, giving different outputs every time learners request feedback.

4 EVALUATION

We conducted deployment studies To explore the efficacy of the modular design and LLM features in KUIZ.

4.1 Methods

We conducted the deployment study through a university-level introductory human-computer interaction class. As one of the authors was an instructor of this course, in-depth identifiable data from the study, such as participation quality, was not shared with the instructor to ensure the integrity of the data, and the students were made aware of this before participation. Students were also provided with an opt-in modal upon accessing the system, asking for their consent to use the data for research purposes. 49 students participated in the activity, and 43 students agreed to use the data for research purposes. Two versions of the system were deployed, the first in April 2023 (Wave 1) and the second in May 2023 (Wave 2). The system at Wave 1 only included the modular MCQ generation pipeline, with no LLM features. The LLM features were added in Wave 2.

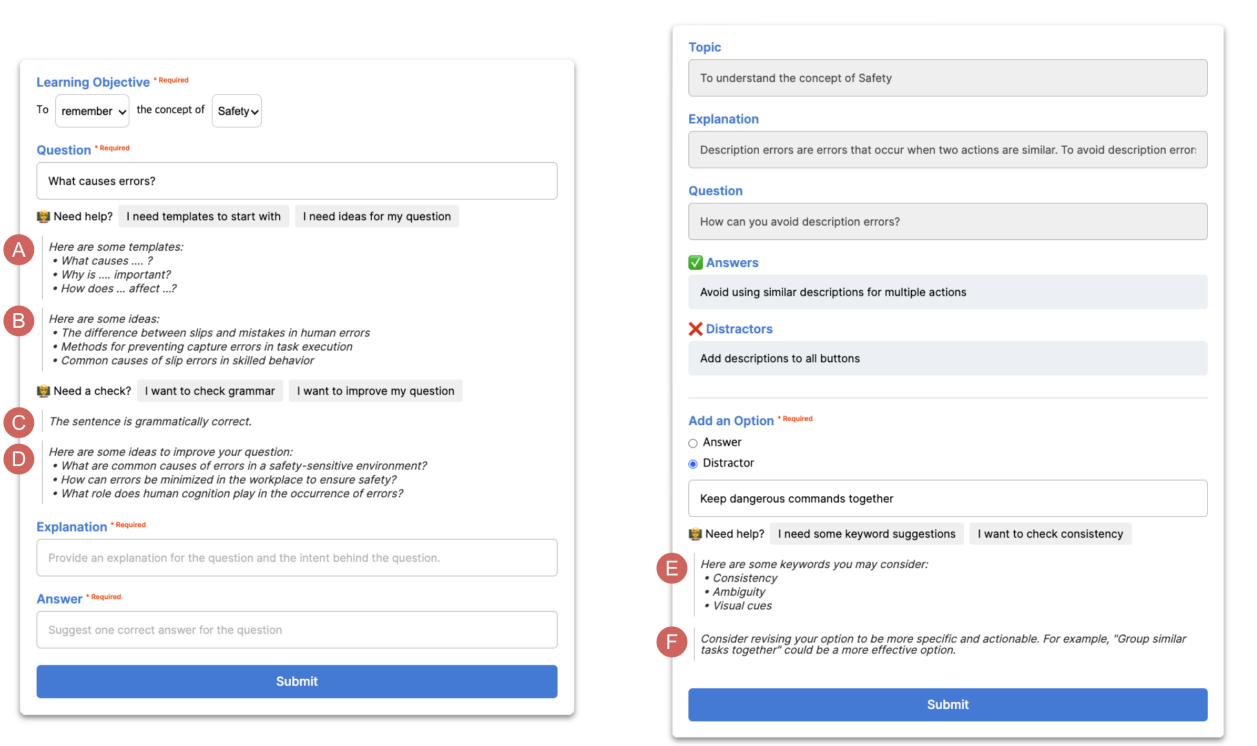

Figure 1: The main interface of KUIZ. On the left, learners can create new questions, receive template suggestions (A), get topic ideas for question improvements (B), check grammatical correctness (C), and improve their questions with detailed suggestions (D). On the right, learners can add options to existing questions, receive keyword suggestions (E), and get grammar and consistency checks (F).

As part of the course, students were asked to study an assigned reading material prior to attending the lecture. During the deployment sessions, students were asked to use KUIZ to generate questions on the subject of the reading material. To qualify for attendance, students were required to accumulate 6 participation points. 3 points were awarded for each generated question stem and 1 point for each option. Students were encouraged to create any combination of stems and options to qualify for the requirement. After the activity, students were provided with an optional post-assessment survey that asked their perception of the relative perceived task difficulty, learning effect, and task preference (Wave 1: $n = 24$, Wave 2: $n = 19$). In the second wave, the survey also included 5-point Likert scale questions on the usability and effects of each LLM feature. The survey also included free response questions for participants to elaborate on their responses and provide additional usability feedback on the system.

4.2 Results

We note that while the two waves of deployment were conducted on the same set of participants, we did not collect identifiers during the post-assessment survey to ensure the students' responses would not be affected by grading concerns. As the post-assessment survey was also optional, we cannot guarantee that the survey respondents were the same participants. Thus, we do not match the data from the two waves to conduct a formal between-subjects analysis, nor claim statistical significance of the results. Instead, we compared the average values in user responses to observe the impact of each version.

4.2.1 Generating MCQs. In the first deployment session, participants created 67 question stems and 225 question options. In the second session, participants created 62 stems and 191 options. In both versions of the system, a majority of the participants perceived the question stem generation task to be more beneficial to learning, with an average of 85% of participants considering the stem generation task to be equally or more beneficial to learning than the option generation task (Figure 2). Participants also generally perceived the stem generation task to be more difficult when compared to the option generation task. Noticeably, with the inclusion of LLM features in Wave 2, participants' preference for question stem generation (Wave 1: 20.8%, Wave 2: 31.6%) increased in comparison to their preference for question option generation (Wave 1: 41.7%, Wave 2: 21.1%) (Figure 3).

4.2.2 Effectiveness of LLM features. Out of the 19 respondents at Wave 2, 17 participants used at least one LLM helper feature to create question stems or options. Participants who used the LLM features were generally favorable towards their perceived usefulness in both the stem generation (Figure 4) and option generation tasks

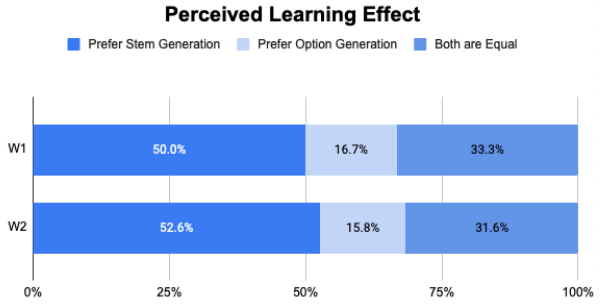

Figure 2: Participants' perception of relative learning effect at Wave 1 (n = 24) and Wave 2 (n = 19)

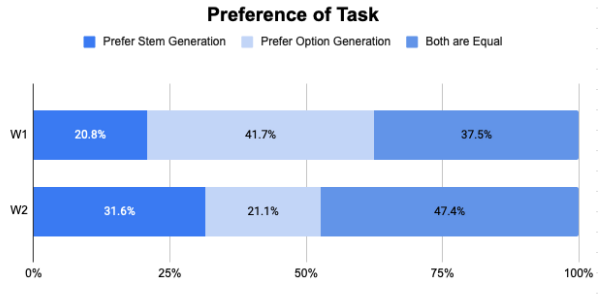

Figure 3: Participants' task preference at Wave 1 (n = 24) and Wave 2 (n = 19)

(Figure 5). In the question stem generation task, *Question Improvements* was the feature that generally received the highest evaluation of utility across all three categories, and *Suggesting Question Topics* received the lowest scores across all categories. Notably, different features aided in different elements of the question-generation process. For example, the *Grammar Checking* feature was perceived to help improve the quality of questions and make the question generation process easier, but less so in making more high-level questions. In the option generation task, the *Grammar and Consistency Check* function was considered marginally more helpful in making higher-quality options and making the option generation process easier, compared to the *Suggesting Keywords* feature. In making more high-level options, *Suggesting Keywords* was considered more helpful than *Grammar and Consistency Check*, although the effect was small.

Despite the perceived usefulness of the LLM features, some factors limited their usability. Lack of transparency in the LLM functions also limited their utility, as participants reported that they could not anticipate in what form the LLM features would present their outputs, or how they might be able to use them before trying it out. For supplementary functions such as *Grammar Checking*, some participants noted that such functions could be better utilized when they were always active, as in the case with many word processor programs and other grammar help tools such as Grammarly[1].

[1]https://www.grammarly.com/

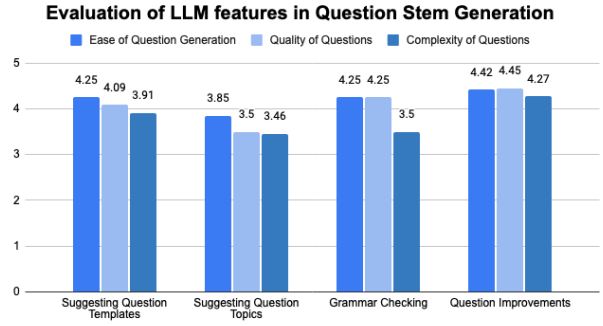

Figure 4: Participants' perception of the LLM features' utility in Question Stem generation. (Responses from participants who did not use a feature were excluded.)

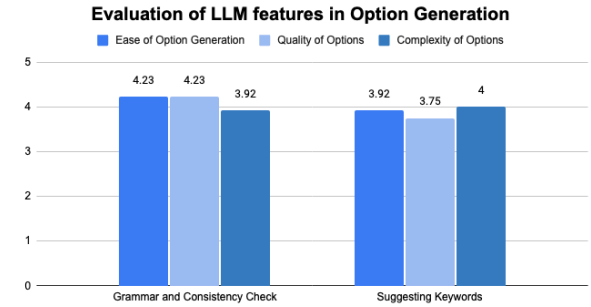

Figure 5: Participants' perception of the LLM features' utility in Option generation. (Responses from participants who did not use a feature were excluded.)

In some cases, participants noted that "AI did all the work for me", saying that they were not motivated to do additional work of editing or adapting the AI-generated output before their final contributions, while other participants noted that the suggestions made by the LLM features were too low-level. This implies variability in the utility of LLM features, potentially depending on the complexity of the questions and answers being generated.

5 DISCUSSION

Through our deployment, we could find challenges in building inter-student trust and error recovery in our collaborative MCQ generation design. Some participants noted that the crowdsourced nature of the questions and options made it harder for them to verify existing information and increased the burden of creating 'correct' questions. Potential concerns included the possibility of other learners (option generators) misunderstanding the intent of the question and faulty options (such as 'correct' distractors or 'wrong' answers) being included in the system. One possible solution is to add more social interactions in the collaborative process, such as discussing questions and clarifying the intentions of questions through a QA board [3]. We may also add a verification task where learners check each other's creation [2, 16].

We observed the positive signals of using LLMs in learner-driven MCQ generation. Participants often struggle with generating question stems, although they consider the stem generation task to have a larger effect on learning than the option generation task. However, comparing the task preference trends at Waves 1 and 2, LLM features increased the relative preference of the stem generation task over the option generation task. Although the LLM scaffolds concentrated in stem generation may affect the preference and hurt learning effects, it will be worth exploring different designs to motivate learners to participate in MCQ generation.

ACKNOWLEDGMENTS

This work was supported by the Center for Excellence in Learning and Teaching (CELT) at KAIST.

REFERENCES

[1] Simon P Bates, Ross K Galloway, Jonathan Riise, and Danny Homer. 2014. Assessing the quality of a student-generated question repository. *Physical Review Special Topics-Physics Education Research* 10, 2 (2014), 020105.

[2] Michael S Bernstein, Greg Little, Robert C Miller, Björn Hartmann, Mark S Ackerman, David R Karger, David Crowell, and Katrina Panovich. 2010. Soylent: a word processor with a crowd inside. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*. 313–322.

[3] Paul Denny, John Hamer, Andrew Luxton-Reilly, and Helen Purchase. 2008. PeerWise: students sharing their multiple choice questions. In *Proceedings of the Fourth international Workshop on Computing Education Research (ICER '08)*. Association for Computing Machinery, New York, NY, USA, 51–58. https://doi.org/10.1145/1404520.1404526

[4] Paul Denny, Fiona McDonald, Ruth Empson, Philip Kelly, and Andrew Petersen. 2018. Empirical Support for a Causal Relationship Between Gamification and Learning Outcomes. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3173885

[5] Judy Hardy, Simon P. Bates, Morag M. Casey, Kyle W. Galloway, Ross K. Galloway, Alison E. Kay, Peter Kirsop, and Heather A. McQueen. 2014. Student-Generated Content: Enhancing learning through sharing multiple-choice questions. *International Journal of Science Education* 36, 13 (Sept. 2014), 2180–2194. https://doi.org/10.1080/09500693.2014.916831 Publisher: Routledge _eprint: https://doi.org/10.1080/09500693.2014.916831.

[6] Vincent Hoogerheide, Justine Staal, Lydia Schaap, and Tamara van Gog. 2019. Effects of study intention and generating multiple choice questions on expository text retention. *Learning and Instruction* 60 (2019), 191–198.

[7] Ahmed Sayed Khashaba. 2020. Evaluation of the Effectiveness of Online Peer-Based Formative Assessments (PeerWise) to Enhance Student Learning in Physiology: A Systematic Review Using PRISMA Guidelines. *International Journal of Research in Education and Science* 6, 4 (2020), 613–628. https://eric.ed.gov/?id=EJ1271255 Publisher: International Journal of Research in Education and Science.

[8] Hassan Khosravi, Kirsty Kitto, and Joseph Jay Williams. 2019. RiPPLE: A Crowdsourced Adaptive Platform for Recommendation of Learning Activities. *Journal of Learning Analytics* 6, 3 (2019), 91–105. https://eric.ed.gov/?id=EJ1237639 Publisher: Society for Learning Analytics Research.

[9] Juho Kim. 2015. Learnersourcing: Improving Learning with Collective Learner Activity. (2015), 213.

[10] David R Krathwohl. 2002. A revision of Bloom's taxonomy: An overview. *Theory into practice* 41, 4 (2002), 212–218.

[11] Josh B Kurtz, Michael A Lourie, Elizabeth E Holman, Karri L Grob, and Seetha U Monrad. 2019. Creating assessments as an active learning strategy: what are students' perceptions? A mixed methods study. *Medical education online* 24, 1 (2019), 1630239.

[12] Steven Moore, Huy Anh Nguyen, and John Stamper. 2021. 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 (L@S '21)*. Association for Computing Machinery, New York, NY, USA, 209–220. https://doi.org/10.1145/3430895.3460140

[13] Anjali Singh, Christopher Brooks, and Shayan Doroudi. 2022. Learnersourcing in Theory and Practice: Synthesizing the Literature and Charting the Future. In *Proceedings of the Ninth ACM Conference on Learning @ Scale (L@S '22)*. Association for Computing Machinery, New York, NY, USA, 234–245. https://doi.org/10.1145/3491140.3528277

[14] Anjali Singh, Christopher Brooks, Yiwen Lin, and Warren Li. 2021. What's In It for the Learners? Evidence from a Randomized Field Experiment on Learnersourcing Questions in a MOOC. In *Proceedings of the Eighth ACM Conference on Learning @ Scale (L@S '21)*. Association for Computing Machinery, New York, NY, USA, 221–233. https://doi.org/10.1145/3430895.3460142

[15] 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 (L@S '19)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/3330430.3333614

[16] Sarah Weir, Juho Kim, Krzysztof Z Gajos, and Robert C Miller. 2015. Learnersourcing subgoal labels for how-to videos. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*. 405–416.

[17] Iman Yeckehzaare, Tirdad Barghi, and Paul Resnick. 2020. QMaps: Engaging Students in Voluntary Question Generation and Linking. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[18] Fu-Yun Yu and Chun-Ping Wu. 2008. Perceived Task Value and Task Difficulty of Online Student-Generated Questions: Its Effects and Implications for Instruction. Association for the Advancement of Computing in Education (AACE), 4516–4521. https://www.learntechlib.org/primary/p/29014/

A PROMPTS

We provide the prompts used to implement our LLM-powered scaffolding. The original prompts contain two few-shot examples. We present only one of them for brevity. The blue text represents the arguments that are programmatically filled in. The orange text represents the generated output.

A.1 Topic idea suggestion

```
Give three IDEAS (not questions) for making multiple choice
questions from MATERIAL.

MATERIAL:
Laws
Work and energy
Examples
- Uniformly accelerated motion
- Uniform circular motion
- Harmonic motion
- Objects with variable mass
Rigid-body motion and rotation
- Center of mass
- Rotational analogs of Newton's laws
- Multi-body gravitational system
Relation to other physical theories
- Thermodynamics and statistical physics
- Electromagnetism
- Special relativity
- General relativity
- Quantum mechanics

IDEAS:
Newton's laws of motion
Limitations to Newton's laws
Publication date of Newton's laws
```

A.2 Grammar check

```
Evaluate the grammar and punctuation of SENTENCE.

SENTENCE:
How we prevent mode errors?

EVALUATION:
It seems that you are missing a verb. Consider adding it. For
example, "How do we prevent mode errors?"

---

SENTENCE:
How can you avoid description errors?
```

```
EVALUATION:
The sentence is grammatically correct.
```

A.3 Feedback

```
Give three suggestions to improve QUESTION regarding
LEARING_OBJECTIVE and EXPLANATION.

QUESTION:
Which of the following is true?

LEARING_OBJECTIVE:
To understand the concept of Human cognition

EXPLANATION:
To understand different examples of human cognition

IMPROVED_QUESTION:
Which of the following is true about human cognition?
What is the definition of human cognition?
How does human cognition differ from other types of cognition?
```

A.4 Keyword suggestion

```
Give three possible keywords for a multiple choice question
QUESTION regarding LEARNING_OBJECTIVE, TYPE, and EXPLANATION.

QUESTION:
What is the correct way of writing error messages?

LEARNING_OBJECTIVE:
To remember the concept of Safety

EXPLANATION:
This question checks if a solver remembers the guidelines for
good error messages.

TYPE:
answer

KEYWORDS:
Recovery
Learnability
Actionable
```

A.5 Consistency check

```
Give feedback on OPTION's consistency with OTHER_OPTIONS
regarding QUESTION and EXPLANATION.

QUESTION:
Which of the following best describes slips and lapses in
human error?

EXPLANATION:
Better understanding for slips and lapses

OPTION:
problem solving

OTHER_OPTIONS:
Errors in planning or rule applicationForgetfulness or memory
lapses in skilled behavior

CONSISTENT_OPTION:
Be more specific. "Errors in problem solving or logical reasoning"
is a better option.
```