Towards Human-Like Educational Question Generation with Large Language Models

Zichao Wang[1(✉)], Jakob Valdez[1], Debshila Basu Mallick[2], and Richard G. Baraniuk[1,2]

[1] Rice University, Houston, USA
{jzwang,jpv3,richb}@rice.edu
[2] OpenStax, Houston, USA
debshila@rice.edu

Abstract. We investigate the utility of large pretrained language models (PLMs) for automatic educational assessment question generation. While PLMs have shown increasing promise in a wide range of natural language applications, including question generation, they can generate unreliable and undesirable content. For high-stakes applications such as educational assessments, it is not only critical to ensure that the generated content is of high quality but also relates to the specific content being assessed. In this paper, we investigate the impact of various PLM prompting strategies on the quality of generated questions. We design a series of generation scenarios to evaluate various generation strategies and evaluate generated questions via automatic metrics and manual examination. With empirical evaluation, we identify the prompting strategy that is most likely to lead to high-quality generated questions. Finally, we demonstrate the promising educational utility of generated questions using our concluded best generation strategy by presenting generated questions together with human-authored questions to a subject matter expert, who despite their expertise, could not effectively distinguish between generated and human-authored questions.

1 Introduction

Practice questions and quizzes have been vital instruments for the assessment of learning [1,20,27]. Engaging in retrieval practice by answering expert-designed questions has shown to be more effective at improving learning outcomes [9,10], by providing opportunities for recall of knowledge, applying knowledge to novel scenarios, and critical thinking and writing skills. The learning benefits are greater than other means of pedagogy such as passively re-reading course materials or studying notes [4,8–10,12,13] or watching instructional videos [21]. However, these questions are also known to be challenging to create: they usually take subject matter experts (SMEs) a significant amount of time, which is both costly and

Z. Wang and J. Valdez—Contributed equally.

M. M. Rodrigo et al. (Eds.): AIED 2022, LNCS 13355, pp. 153–166, 2022.
https://doi.org/10.1007/978-3-031-11644-5_13

labor-intensive [20]. Therefore, this question generation process does not easily generalize and scale to the continually expanding repositories of educational content that need large banks of assessments to be effective sources of instruction.

To create a scalable question generation process, several recent works leveraged artificial intelligence (AI) methods for *automatically* generating questions. For example, some prior works [5,25,26] focused on generating factual questions using recurrent neural network (RNN) architectures. [28] designed a method to select highly interesting phrases which a generated question is supposed to ask about. The implications of these works are far-reaching. In addition to reducing the labor and cost for producing assessment questions, automatic question generation methods have the potential to create a more engaging learning experience by generating (i) personalized questions that adapt to each student's learning trajectory [7] and (ii) real-time pop-up quizzes while the student is reading a textbook or watching instructional videos. Once trained, these methods have been shown to perform well on question generation tasks. However, they require custom model design and (sometimes significant) computational resources for training, making them a less appealing option for practitioners who desire a "plug-and-play" AI-assisted question generation process that allows them to easily interact with an AI system without the need for model training.

Recently, a new paradigm in text generation using large pretrained language models (PLMs), such as GPT-3 [3], is now making such "plug-and-play" question generation a possibility. These PLMs have been pretrained on web-scale data which equip the model with abundant knowledge of the language compared to their earlier counterparts. Furthermore, they can be easily and effectively adapted to various generation tasks via the "prompting" technique, where the user simply specifies the generation task that they would like to perform as a prompt. A *prompt* usually contains, in addition to a "query" from which the PLM will generate the outcome, a series of examples in an input-output structure that "teach" the model how to generate the output given the input specific to a particular task. Figure 1 gives an example of using prompting to adapt a PLM for machine translation and arithmetic question answering. Prompting provides an easy interface and high controllability for users to interact with PLMs and customize it for different generation tasks. Because of its simplicity and practicality, prompting techniques to adapt PLMs for downstream generation tasks have attracted increasing attention in the past few years [11,16,18,19]. Figure 1 shows an example of prompting for machine translation, question answering.

Unfortunately, using prompts to adapt PLMs for question generation is challenging due to the open-ended nature of the process, i.e., it does not have a clearly defined input-output structure. This poses certain challenges such as, what content should the questions be generated from, how should we deal with the fact that multiple different questions can be asked about the same concept, etc. This open-ended nature makes question generation unique in contrast with other generation tasks commonly studied in existing literature (e.g., in machine translation, input and output are simply texts in the source and target languages, respectively). As a result, unlike other generation tasks where adapting PLMs via prompting is straightforward (e.g., see Fig. 1 for an illustration), it is

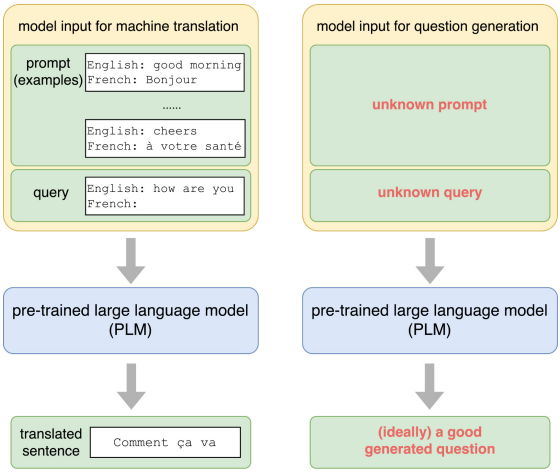

Fig. 1. Illustrations of adapting PLMs for machine translation and the challenges in designing prompts to adapt PLMs for educational question generation.

unclear how to design effective prompts for PLMs in order for question generation. To the best of our knowledge, to date no existing literature has investigated the modification of prompting strategy for question generation. To harness the power of AI for educational question generation, prompt design for question generation by PLMs is an exciting open problem.

1.1 Contributions

In this paper, we investigate the problem of effectively prompting a PLM to generate desirable, high-quality, educational practice questions. An effective prompt strategy will enable us to leverage the power of PLMs with minimal effort and without having to conduct model training with large volumes of domain-focused content. We start with the core question: how do we design prompts such that a PLM can generate the most desirable and effective practice questions? We answer this question by proposing 5 different generation settings with a specific prompting strategy for each. We conduct a series of manual examinations of the generated questions as well as automatic evaluations, which lead to the empirical conclusion of the best combinations of our prompting strategy. This strategy serves as an empirical guideline for practitioners to set up PLMs to generate the best practice questions for educational purposes. Furthermore, we evaluated the educational value of PLM-generated questions by presenting them alongside human-authored questions for SMEs to discern the human-authored from machine-authored questions. Evaluation by the respective SMEs (biology, psychology, and history) demonstrated that the generated questions achieved similar educational value relative to the human-authored ones, setting a strong case for their practical utility. In essence, we emulate how real practitioners and educators might be able to use these models to generate questions that meet their need in a practical setting.

1.2 Background: Large Pretrained Language Models and Prompting

We focus on large pretrained language models (PLMs) in this paper, specifically, auto-regressive PLMs, such as GPT that have become the dominant tools for text generation. These models learn a distribution over text, which can be decomposed auto-regressively as follows:

$$\boldsymbol{x} \sim p_\theta(\boldsymbol{x}) = p_\theta(x_1) \prod_{t=2}^{T} p_\theta(x_t|x_1, \ldots, x_{t-1}) \,. \tag{1}$$

where p_θ is the LM where θ represents all model parameters. In this paper, we focus on an LM that is already trained on massive data and thus assume p_θ is fixed throughout this paper.

In practice, we will give the model some initial texts called a "prompt" as input which instructs the model to generate specific texts. This is possible because of the decomposition in Eq. 1. To see this, let $\boldsymbol{c} := [c_1, \ldots, c_L]$ denotes the prompt which consists of L ordered tokens c_l. Then the LM models a conditional distribution as follows:

$$p_\theta(\boldsymbol{x}|\boldsymbol{c}) = p_\theta(x_1|\boldsymbol{c}) \prod_{t=2}^{T} p_\theta(x_t|x_1, \ldots, x_{t-1}, \boldsymbol{c}) \,. \tag{2}$$

Equation 2 makes it possible to adapt an LM for a wide range of generation tasks: depending on the interpretation of $\boldsymbol{c}$, we can adapt a pretrained LM for a wide range of tasks. [3] shows that, without further fine-tuning p_θ, simply changing $\boldsymbol{c}$ for different tasks perform on par with fine-tuning p_θ. This makes it very easy to use the LM because we only need to change the input to the model to adapt it for a variety of tasks. See Fig. 1 for an illustration. The question now is how to design such a prompt for question generation.

2 Exploring Prompting Strategies in Question Generation

Table 1. Summary of the four factors in our prompting strategy and the choices under consideration for each factor.

Example structure for question generation	**Data source in the examples**	**Number of examples**	**Lengths of context and question in each example**
CAQ: context (C) and an answer (A) and the output contains a question (Q)	Content agnostic (SQuAD)	One-Shot	Small (avg. 15 words)
CTQA: (C) and a target (T) and the output contains a question (Q) and an answer (A)	Content specific	Few-Shot	Medium (avg. 25 words)
		Five-Shot	Large (40 and above)
		Seven-Shot	

In the remainder of the paper, we set out to answer the question: how do we design effective prompts for educational question generation? Answers to this question will provide practitioners with clear guidance on how to better control off-the-shelf PLMs for high-quality question generation. We take an empirical approach and design a series of experiments to systematically investigate various factors that impact the effectiveness of prompting strategies for question generation with PLMs. We propose four factors that are crucial considerations to prompt design for question generation. Below, we detail these factors and the possible choices that we study for each factor (see Table 1 for a high-level summary). In contrast to automated prompting methods as in existing literature, our prompting design is interpretable and flexible, enabling practitioners to explicitly control and iteratively refine the generation process as needed.

2.1 Example Structure for Question Generation

The first factor we investigate is the question generation formulation, i.e., the input-output structure in each example that we will use to instruct and adapt the PLM for question generation. Different formulations will likely impact the generated questions' quality. In this work, we focus on contextualized question generation, in which a question is asked and the answer to it can be found within a given paragraph. We compare two different generation setups. In the first setup, labeled as CAQ, the input contains a context (C) and an answer (A) and the output contains a question (Q). The context can be a short excerpt from a textbook and the answer should correctly answer the generated question. This setup has been considered in a wide range of question generation tasks [5,26,28]. In the second setup, referred to as CTQA, the input contains a context (C) and a target (T) and the output contains a question (Q) and an answer (A). The target does not need to be the answer to the generated question but guides the model to generate a question to ask *about* the particular part in the context specified by the target. The model also generated an answer in addition to the question. The intuition behind this setup is that the model may generate more on-topic and relevant questions because it is forced to also generate the answer. This setup is reminiscent of prior work that leverages question answering modeling for question generation [6,17].

2.2 Data Source in the Examples

The second factor we investigate is the data source in each example, i.e., where do the context, question, answer (target) come from? This question arises when a user wants to generate questions for different subjects; depending on the subject, the examples in the prompt may need to change so that PLM is given the appropriate domain knowledge. We are most interested in whether we can use the same set of examples that come from a generic source for question generation across different subjects/content. We thus compare a *content-agnostic and a*

content-specific selection of examples. In the content-agnostic setup, we choose examples from SQuAD [24], a generic, widely used question answering dataset that can also be used for question generation. In the content-specific setup, we choose examples in the same subject as the one in which the PLM will generate questions.

2.3 Number of Examples

The third factor we investigate is the number of examples to include in the prompt. Usually, PLMs' performance improves with more examples. Nevertheless, because of the open-ended nature of question generation, it is unclear to what point increasing the number of examples will help. We thus consider four setups including One-shot, Few-shot, Five-shot, Seven-shot where "shot" refers to the number of examples.

2.4 Lengths of Context and Question in Each Example

The last factor we investigate is the length of context and question in each example. A context or question that is too short may limit the diversity and complexity of the generated questions. A context or question that is too long may contain irrelevant information which may confuse the PLMs, potentially leading to generated questions that are irrelevant or off-topic. We thus compare three different setups including small, medium, large contexts and questions depending on the length of texts they contain. Small corresponded to questions about 15 words in length, medium questions were around 25 words long, and large questions were about 40 words long on average. Small contexts consist of around 2 sentences, medium contexts around 4–5 sentences of information, and large contexts usually a full paragraph or multiple paragraphs.

3 Experiments

We recommend the best prompt setting for each generation strategy that yielded the best-generated questions. Code scripts, additional clarifications, and additional results such as examples of generated questions are publicly available.[1]

Experiment Setup. We choose biology as the subject to generate questions and use the Openstax Biology 2e (Bio 2e) Textbook as the source for most of our example content. In this paper, we focus on generating open-ended questions of Bloom's level below three because higher-order Bloom's questions typically involve making connections across larger content [2,14]. Generating diverse types of potentially more challenging questions is left for future work. We also limit

[1] https://github.com/openstax/research-question-generation-gpt3.

our investigation to textual content and remove images, tables, links, and references from the textbook. During generation, we first pre-select a fixed number of examples from the textbook (and SQuAD, for the data source experiment; see Sect. 2.2). During generation for all setups under each factor, we randomly pick a fixed number of examples to serve as the prompt and another two queries, i.e., with only the context (possibly also the target; see Sect. 2.1) from which the PLM is asked to generate questions. Unless otherwise noted, for each query in each setup under each factor, the PLM generates 75 questions for evaluation. When generating questions for a factor, all the other factors are set to the same value to ensure fairness in comparison. Throughout our experiments, we use the GPT-3 Davinci API from OpenAI with temperature = 0.9 and top_p = 1.

Evaluation Protocol. We primarily evaluate the quality and diversity of the generated questions. For quality, we report **perplexity** and **grammatical error**. Perplexity is inversely related to the coherence of the generated text; the lower the perplexity score, the higher the coherence. To make the process computationally efficient, we computed perplexity using a GPT-2 language model for all generations. We computed grammatical error using the Python Language Tool [22] which counts the number of grammatical errors averaged over all generated questions in each setup under each factor. For diversity, we report the **Distinct-3** score [15], which counts the average number of distinct 3-grams in the generated questions. Furthermore, we believe that ensuring the generated questions are safe, i.e., without profanity or inappropriate language is critical for high-stakes educational applications. Therefore, we report the **toxicity** of the generated questions, using the Perspective API [23], which is often missing from the evaluation in existing question generation literature. Last but not least, we perform a preliminary human evaluation to mark **percentage of acceptable questions** for each setup under each factor. A question is considered acceptable if it is coherent, on-topic, answerable, grammatically correct, and appropriate. We conduct a more comprehensive human evaluation in Sect. 3.3.

3.1 Empirical Observations

Table 2. Results for the example structure comparisons, which show that the CTQA structure is distinctly better than the CAQ structure.

Gen. format	Diversity ↑	Perplexity ↓	Toxicity ↓	Gramm. error ↓	% acceptable ↑
CAQ	0.895	64.683	0.153	**0.053**	26.7%
CTQA	**0.898**	**29.900**	0.153	0.080	**54.7%**

Structure of Examples in the Prompt. Recall that this experiment compared CAQ and CTQA structures of the examples in the prompt (Sect. 2.1). The results, presented in Table 2, show that, although the CTQA structure produces questions of comparable diversity, quality, and toxicity, it generates about twice as many acceptable questions as the CAQ structure. This comparison suggests that CTQA is a superior example structure and confirms our earlier hypothesis that asking PLMs to generate the answer in addition to only the question is beneficial for improving the quality of generated questions. Additionally, the generated answers can be potentially useful for evaluating a student's performance on the generated question. Ensuring that the generated answer correctly answers the generated question is important ongoing work.

Table 3. Results for the example data source comparisons. Using content specific examples gives superior generation performance compared to content agnostic example.

Gen. format	Diversity ↑	Perplexity ↓	Toxicity ↓	Gramm. error ↓	% acceptable ↑
SQuAD	0.884	102.840	0.201	0.093	18.0%
OpenStax	**0.895**	**64.683**	**0.153**	**0.053**	**26.7%**

Data Source in Examples. Recall that this experiment compared whether the examples come from the same subject (Bio 2e) as the query or a generic dataset (SQuAD) (Sect. 2.2). The results in Table 3 showed that when a prompt consists of examples from the same subject, the PLM can generate questions about twice as effective as when using SQuAD examples across all metrics. These results suggest that a generic set of examples may not adapt to question generation for various domains and that appropriately choosing examples from desired subjects is a better setup for question generation.

Table 4. Results for the number of examples comparisons. Five- and seven-example settings yield better questions compared to one- and three-example settings.

# Examples	Diversity ↑	Perplexity ↓	Toxicity ↓	Gramm. error ↓	% acceptable ↑
1 example	0.897	37.954	0.384	0.182	24.9%
3 examples	0.924	36.586	0.232	0.151	37.8%
5 examples	**0.938**	35.990	0.208	0.119	**51.6%**
7 examples	0.918	**30.731**	**0.176**	**0.076**	44.9%

Number of Examples. Table 4 shows the results comparing one-, three-, five-, and seven-shots, i.e., the number of examples in the prompt. The results show that one- and three-shots are ineffective; we observe that they produce a majority of unacceptable questions. The five-shot condition results were optimal followed

closely by the seven-shot, with the one-shot being most inefficient. We prefer using the five-shot condition because here, the PLM generated more varied questions that are also of high quality. For example, although the model was only given free-response questions, it could produce a small number of multiple-choice or true-or-false questions.

Table 5. Results for the context and question length comparisons. We see that, in general, short context and question lengths in the examples improve generation quality.

Context length	Diversity ↑	Perplexity ↓	Toxicity ↓	Gramm. error ↓	% acceptable ↑
Short	0.861	33.452	0.329	**0.380**	22.0%
Medium	**0.878**	30.692	**0.214**	0.410	**24.0%**
Long	0.877	**30.385**	0.331	0.420	**24.0%**
Question length	Diversity ↑	Perplexity ↓	Toxicity ↓	Gramm. error ↓	% acceptable ↑
Short	**0.906**	34.275	**0.246**	**0.377**	**30.0%**
Medium	0.893	33.704	0.318	0.487	23.7%
Long	0.885	**30.38**	0.295	0.610	14.7%

Lengths of Context and Question in Each Example. Table 5 shows the results comparing different lengths of the question and context in each example, respectively. In terms of question lengths, results suggested that a smaller question length generally yields the best performance. In terms of context lengths, results are mixed. This is likely because longer contexts contain information that is not directly useful for generating questions and because longer texts lead to longer prompts, which makes it more difficult to instruct the model to adapt to the question generation task.

3.2 Discussions

From the above quantitative results, we obtain a good understanding of how the different choices, while constructing the prompt for each generation strategy, will impact the quality of the generated questions. It is clear that when preparing examples to instruct and adapt PLMs for question generation, the PLM is likely to generate higher quality questions given the prompt design: if prompt contains five to seven examples that are in CTQA format, are chosen from the desired subject, rather than generic content, and contain relatively short contexts and questions. This recommendation has the potential to serve as a guideline for practitioners when adapting off-the-shelf PLMs for their unique question generation needs.

3.3 Human Expert Evaluation for Multiple Subjects

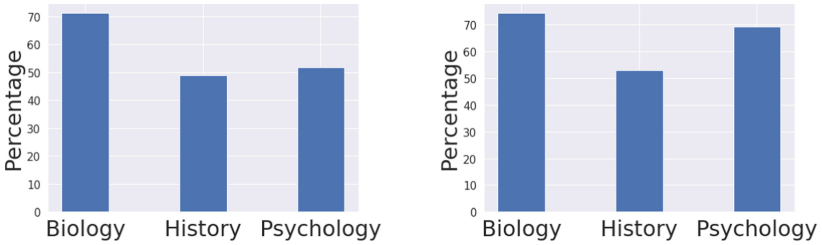

Fig. 2. Human evaluation results. **Left**: the percentage of PLM-generated questions that are recognized as human-authored by SMEs. **Right**: the percentage of PLM-generated questions that SMEs considered as ready-to-use in their classes.

To validate the utility of the generated questions as well as to investigate whether our best prompt strategy would result in good question generations across domains (e.g., natural sciences, social sciences, and humanities), we engaged biology, psychology, and history subject matter experts (SMEs) respectively to evaluate the quality of questions from these domains generated using the best prompting strategy.

For each domain, we chose 5 examples as the prompt and another 5 examples with only the question and target as the query. For each query, we generated and selected the 10 best questions. We then mixed these 50 questions together with 5 reference human-authored questions (one for each query) and presented them to SMEs. SMEs were asked to (1) distinguish whether each question is machine or human authored and (2) evaluate whether they would use the question in the assessment they create for their class.

The results, shown in Fig. 2, suggest that the majority of PLM-generated questions look "human-like" for SMEs and can be readily used in their respective classes. This result is encouraging because it suggests that our prompting strategy can effectively adapt a publicly available, general-purpose text generation model for high-quality educational question generation. Table 6 demonstrates a few successful generated questions.

Table 6. Examples of three generated questions using our recommended prompting strategy along with the context, target, and reference human-authored questions.

Context (Biology): At 0.1 to 5.0 μm in diameter, prokaryotic cells are significantly smaller than eukaryotic cells, which have diameters ranging from 10 to 100 μm (Fig. 4.6). The prokaryotes' small size allows ions and organic molecules that enter them to quickly diffuse to other parts of the cell. Similarly, any wastes produced within a prokaryotic cell can quickly diffuse. This is not the case in eukaryotic cells, which have developed different structural adaptations to enhance intracellular transport
Target: size
Reference human-authored question: Some unusually large cells do exist. How might such cells solve the problem of diffusion?
Machine-generated question: Compare the size of prokaryotic and eukaryotic cells to describe why diffusion is more rapid through prokaryotic cells
Context (Psychology): Sleep debt and sleep deprivation have significant negative psychological and physiological consequences. As mentioned earlier, lack of sleep can result in decreased mental alertness and cognitive function. In addition, sleep deprivation often results in depression-like symptoms. These effects can occur as a function of accumulated sleep debt or in response to more acute periods of sleep deprivation. It may surprise you to know that sleep deprivation is associated with obesity, increased blood pressure, increased levels of stress hormones, and reduced immune functioning
Target: sleep deprivation
Reference human-authored question: Which animal would be most likely to die from a lack of sleep?
Machine-generated question: Sleep debt and sleep deprivation are associated with significant negative psychological and physiological consequences. What are some of the potential consequences of a person having less than optimal sleep?
Context (U.S. History): Opponents made clear their resistance to Garrison and others of his ilk; Garrison nearly lost his life in 1835, when a Boston anti-abolitionist mob dragged him through the city streets. Anti-abolitionists tried to pass federal laws that made the distribution of abolitionist literature a criminal offense, fearing that such literature, with its engravings and simple language, could spark rebellious Black people to action. Their sympathizers in Congress passed a "gag rule" that forbade the consideration of the many hundreds of petitions sent to Washington by abolitionists. A mob in Illinois killed an abolitionist named Elijah Lovejoy in 1837, and the following year, ten thousand protestors destroyed the abolitionists' newly built Pennsylvania Hall in Philadelphia, burning it to the ground
Target: anti-abolitionists
Reference human-authored question: How did anti-abolitionists respond to the abolition movement?
Machine-generated question: What were some of the responses to the abolitionist movement?

3.4 Limitations and Future Work

Question generation is complex and is an active ongoing area of research. In this work, we proposed four factors as the prompt design framework and conducted a controllable scientific study on those factors. However, our investigation was by no means exhaustive and several other factors have the potential to impact question generation such as question difficulty, type (i.e., multiple-choice, fill in the blank, etc.), and modality (i.e., potentially including tables and equations). Investigations of more factors and complex interactions among them are left for future work. In addition, our human evaluation was a small-scale experiment because we were only able to engage the SMEs for a short time. The next step is to conduct a large-scale evaluation that involves both instructors and students

Table 7. Examples of failed cases and the failing reasons. Our prompting strategy can still generate questions that contain grammatical errors and other types of errors.

(Biology): What is the correct statement is about centrosomes? (Multiple-choice question with no options and bad grammar)
(Psychology): Sleep deprevation can lead to serious changes in the body. Which one of these changes characterized by sleep deprivation? (grammatical and spelling errors)
(History): During the Gold Rush, the Forty-Niners did not find wealth so easy to come by, most did not. (not a question)

in a safe environment to obtain a better understanding of the educational utility of machine-generated questions. Lastly, our prompting strategy generated questions with grammatical errors and other problems at times; we show some failed examples in Table 7. A promising future direction is to develop automated filters capable of removing undesirable generated questions and only select the highest quality ones, preferably also personalized to each student and instructor.

4 Conclusion

In this work, we investigate the best practices to prompt a PLM for educational question generation. We develop and empirically study a prompting strategy consisting of four different factors. Based on a series of quantitative experiments, we recommended the choices for each factor under our prompting strategy that led to high-quality generated questions. Human evaluations by subject experts in three different educational domains suggest that most of the questions generated by a PLM with our recommended prompting strategy are human-like and ready-to-use in real-world classroom settings. Our results indicate that properly prompting existing off-the-shelf PLMs is a promising direction for high-quality educational question generation with many exciting future research directions.

Acknowledgements. This work is supported by NSF grants 1842378, 1917713, 2118706, ONR grant N0014-20-1-2534, AFOSR grant FA9550-18-1-0478, and a Vannevar Bush Faculty Fellowship, ONR grant N00014-18-1-2047. We thank Prof. Sandra Adams (Excelsior College), Prof. Tyler Rust (California State University), Prof. Julie Dinh (Baruch College, CUNY) for contributing their subject matter and instructional expertise. Thanks to the anonymous reviewers for thoughtful feedback on the manuscript.

References

1. Adesope, O.O., et al.: Rethinking the use of tests: a meta-analysis of practice testing. Rev. Educ. Res. **87**(3), 659–701 (2017)
2. Bloom, B.S., Engelhart, M.D., Furst, E., Hill, W.H., Krathwohl, D.R.: Handbook I: Cognitive Domain. David McKay, New York (1956)

3. Brown, T., et al.: Language models are few-shot learners. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) Advances in Neural Information Processing Systems, vol. 33, pp. 1877–1901 (2020)
4. Connor-Greene, P.A.: Assessing and promoting student learning: blurring the line between teaching and testing. Teach. Psychol. **27**(2), 84–88 (2000)
5. Du, X., Shao, J., Cardie, C.: Learning to ask: neural question generation for reading comprehension. In: Proceedings of the ACL, pp. 1342–1352 (July 2017)
6. Duan, N., Tang, D., Chen, P., Zhou, M.: Question generation for question answering. In: Proceedings of the Conference on EMNLP, pp. 866–874 (September 2017)
7. Huang, Y.T., Chen, M.C., Sun, Y.S.: Bringing personalized learning into computer-aided question generation (2018)
8. Karpicke, J.D.: Retrieval-based learning: active retrieval promotes meaningful learning. Curr. Dir. Psychol. Sci. **21**(3), 157–163 (2012)
9. Karpicke, J.D., Blunt, J.R.: Retrieval practice produces more learning than elaborative studying with concept mapping. Science **331**(6018), 772–775 (2011)
10. Karpicke, J.D., Roediger, H.L., III.: The critical importance of retrieval for learning. Science **319**(5865), 966–968 (2008)
11. Keskar, N.S., McCann, B., Varshney, L.R., Xiong, C., Socher, R.: CTRL: a conditional transformer language model for controllable generation (2019)
12. Koedinger, K.R., Kim, J., Jia, J.Z., McLaughlin, E.A., Bier, N.L.: Learning is not a spectator sport: Doing is better than watching for learning from a MOOC. In: Proceedings of the Conference on Learning at Scale, pp. 111–120 (2015)
13. Kovacs, G.: Effects of in-video quizzes on MOOC lecture viewing. In: Proceedings of the Conference on Learning at Scale, pp. 31–40 (2016)
14. Krathwohl, D.R.: A revision of bloom's taxonomy: a overview. Theor. Pract. **41**(4), 212–218 (2002)
15. Li, J., Galley, M., Brockett, C., Gao, J., Dolan, B.: A diversity-promoting objective function for neural conversation models. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 110–119 (Jun 2016)
16. Li, X.L., Liang, P.: Prefix-tuning: optimizing continuous prompts for generation. In: Proceedings of the ACL. pp. 4582–4597 (August 2021)
17. Li, Y., Duan, N., Zhou, B., Chu, X., Ouyang, W., Wang, X.: Visual question generation as dual task of visual question answering. arXiv e-prints (2017)
18. Liu, P., et al.: Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing (2021)
19. Liu, X., Ji, K., Fu, Y., Du, Z., Yang, Z., Tang, J.: P-tuning v2: prompt tuning can be comparable to fine-tuning universally across scales and tasks (2021)
20. 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)
21. Martin, L., Mills, C., D'Mello, S.K., Risko, E.F.: Re-watching lectures as a study strategy and its effect on mind wandering. Exp. Psychol. **65**(5), 297–305 (2018)
22. Morris, J.: Python language tool (2021). https://github.com/jxmorris12/language_tool_python
23. Perspective: Using machine learning to reduce toxicity online (2021). https://www.perspectiveapi.com/
24. Rajpurkar, P., et al.: SQuAD: 100,000+ questions for machine comprehension of text. In: Proceedings of the Conference on EMNLP, pp. 2383–2392 (November 2016)

25. Serban, I.V., et al.: Generating factoid questions with recurrent neural networks: the 30M factoid question-answer corpus. In: Proceedings of the ACL, pp. 588–598 (August 2016)
26. Wang, Z., Lan, A.S., Nie, W., Waters, A.E., Grimaldi, P.J., Baraniuk, R.G.: QG-Net: a data-driven question generation model for educational content. In: Proceedings of the Conference on Learning at Scale (2018)
27. Wiklund-Hörnqvist, C., Jonsson, B., Nyberg, L.: Strengthening concept learning by repeated testing. Scand. J. Psychol. **55**(1), 10–16 (2014)
28. Willis, A., et al.: Key phrase extraction for generating educational question-answer pairs. In: Proceedings of the Conference on Learning at Scale (2019)