# Reading Comprehension Quiz Generation using Generative Pre-trained Transformers

Dijkstra, R.; Genç, Z.; Kayal, S.; Kamps, J.

**Publication date**
2022

**Document Version**
Final published version

**Published in**
Proceedings of the Fourth International Workshop on Intelligent Textbooks 2022

# Reading Comprehension Quiz Generation using Generative Pre-trained Transformers

Ramon Dijkstra[1,2], Zülküf Genç[2], Subhradeep Kayal[2], and Jaap Kamps[1]

[1] University of Amsterdam, Amsterdam, The Netherlands
[2] Prosus, Amsterdam, The Netherlands
ramon.dijkstra@hotmail.com, zulkuf.genc@prosus.com, deep.kayal@prosus.com, kamps@uva.nl

**Abstract.** Recent advances in AI have resulted in large pre-trained language models with superior performance on text generation tasks, prompting the question of whether we can use them to generate educationally useful text completions. This holds the potential to generate relevant quizzes for any educational text, greatly complementing current formative and summative tests from education professionals. We explore pre-trained language models for quiz generation on reading comprehension texts and propose **EduQuiz**, an end-to-end quiz generator based on a GPT-3 model fine-tuned on text-quiz pairs, able to generate a complete multiple-choice question, with the correct and distractor answers. We observed that the majority of generated quizzes is reasonable, and that generation of high-quality distractors is more challenging than question and answer generation. More generally, while it may be too early to replace manually generated tests for summative feedback and grading with automatic quiz generation, EduQuiz already has potential value for formative feedback and to increase engagement during the learning phase by enhancing textbooks with assessments.

**Keywords:** Quiz Generation · Question Generation · Question Answering · Distractor Generation

## 1 Introduction

This paper investigates the problem of quiz generation on reading comprehension texts: given any arbitrary educational text as input, can we automatically generate complete quizzes for potential use as formative or summative feedback in an educational setting?

Figure 1 illustrates this problem. Figure 1(a) shows a slightly abbreviated version of a text used in a reading comprehension test setting. Recent advances in AI have resulted in very large pre-trained language models based on a transformer architecture, which have demonstrated superior text generation capabilities [4, 23, 30]. Figure 1(b) shows an example of text completion based on the

Chinese young people love their smart phones. ... But now, a book has let them forget about their phones. The book is called Secret Garden. It is a colouring book for adults'. ... Many young people bought this book. They like it so much that they stop playing games or surfing the Internet on their phones. ... Drawing the book can help them become less nervous. In the past, people thought colouring books are only for children. With Secret Garden getting more and more popular, many young adults also begin to paint as a hobby.

**(a) Prompt EQG-Race #649 (shortened)**

... begin to paint as a hobby. *An author from UK who came up with the idea told Beijing Morning that he wanted to let people forget about the stressful world for a while. He also thought people would feel relaxed while doing the work. The artist was very busy. He had little time to meet and paint with his friends.*

**(b) Text completion (unconstrained)**

... begin to paint as a hobby. ***Question:*** *Why do many young people like to colour the book?* ***True answer:*** *Because it is a way to relax themselves.* ***False answer:*** *Because they think it is interesting.* ***False answer:*** *Because they like to show their talent.* ***False answer:*** *Because they can earn much money.*

**(c) EduQuiz quiz generation (template in bold)**

**Fig. 1.** Quiz generation use case: (a) prompt from EQG-RACE, (b) text completion by GPT-3, and (c) EduQuiz is trained on text-quiz pairs, and generates a complete quiz on never-seen before text.

example reading comprehension text, demonstrating high levels of text quality and relatedness to the text prompt, but also that unconstrained text completion is very "creative" and the relation to the text at hand may be lost very quickly. We propose an end-to-end quiz generator, EduQuiz, based on a GPT-3 model fine-tuned on text-quiz pairs to complete an educationally relevant quiz template [4]. Figure 1(c) shows the output for the example input, in this case successfully generating a relevant question, with a correct answer, and three different false answers. This paper addresses the quiz generation problem head-on, trying to answer the question: *Can generative pre-trained transformers learn to generate a reading comprehension quiz?*

EduQuiz is created for students and teachers and can be seen as an initial step towards fully automatic quiz generation. This could reduce the burden of manually creating quizzes for teachers or educational content creators. Moreover, students could use this tool to test their knowledge while learning from textbooks. This is beneficial for students as asking exam-like questions to learners is proven to be the way to test the real knowledge of learners [2, 10]. Besides, active learning and testing knowledge after reading an educational text have shown to be beneficial for learning [1, 22].

Quiz generation is a complex problem, with each quiz consisting of a question, a true answer, and several false answers that are closely related to the true answer. We will call these false answers distractors. We performed our research on the EQG-RACE dataset, which contains examination-like questions from the

**Table 1.** Quiz Generation Tasks.

| *Task* | *Input* | *Output* |
|---|---|---|
| SWQG | 1. context | question |
| | 2. context, question | answer |
| | 3. context, question, true answer | distractors |
| EEQG | context | question, true answer, distractors |

original RACE dataset [13, 15]. On this dataset, we compared a general-purpose model called Macaw-11b and task-specific fine-tuned GPT-3 models on the tasks in Table 1 [4, 27].

As shown in Table 1, we introduce the tasks of Step-Wise Quiz Generation (SWQG) and End-to-End Quiz Generation (EEQG). SWQG generates a question based on the context, uses both the context and generated question to generate the corresponding answer, and uses the context, generated question and answer to generate distractors. EEQG fulfills all of this in one go by generating a quiz directly from the context.

To evaluate the performances, we used the common metrics BLEU-4, ROUGE-L, and METEOR, against the human reference ground-truth [3, 18, 21].

Our main contributions can be summarized as follows:

- We propose the end-to-end quiz generation as a key research problem with a large potential impact on the education domain, helping both learners and educators to increase learning effectiveness in every situation where professional quizzes are not readily available, and thereby also contributing to the enhancement of assessments in textbooks.
- We propose an end-to-end quiz generator based on GPT-3, EduQuiz, where we observed that the majority of generated quizzes is reasonable, and that generation of high-quality distractors is more challenging than question and answer generation [4].
- We encourage other researchers to reproduce and expand our results and share all the used data and code on Github.[1]

## 2 Related work

Quiz Generation combines Question Generation (QG), Question Answering (QA), and Distractor Generation (DG). QG, QA, and DG all have a comparable research history. Early models within these domains were rule-based [11, 20, 24]. Later, the paradigm switched to neural methods [8, 17, 32]. With the rise of transformers, the focus within QG, QA, and DG switched again [30]. As of

[1] https://github.com/RamonDijkstra/EduQuiz

current, large pre-trained language models such as BERT, T5, and GPT-3 have shown superior performances on these text generation tasks [4, 6, 23]. A general-purpose model called Macaw can perform QG, QA, and DG as it is trained on these different angles [27].

Generating full quizzes for educational purposes has been performed on fill-in-the-blank questions, knowledge bases, and listening comprehension [12, 19, 25, 26]. Previous research also aimed to generate assessments from textbooks [7, 28]. Within the educational domain, strictly generating a quiz based on an educational text using large pre-trained language models has not been done before. Closely related, the research by Lelkes et al. generates quiz questions, answers, and distractors on news articles [16]. In their approach, they first apply Question-Answer Generation (QAG) and then DG. Related to quiz generation is a system proposed by Khan et al. where the user can generate assessment content by interactively rating generated text [14]. This approach requires the user to have the domain knowledge to specify exactly what it wants to generate whereas we focus on full automation of this process. To the best of our knowledge, SWQG has not been done before by concatenating QG, QA, and DG and no previous research has been performed on EEQG within the educational domain.

## 3 Approach

GPT-3 is a generative pre-trained transformer that can be fine-tuned to downstream tasks using the API[2] of OpenAI [4]. The API allows users to submit training and validation data files as JSONL documents where each training/validation instance contains a prompt-completion pair. OpenAI takes care of the training itself. In our case, the prompt is an educational text, and the completion a quiz. Previous research has shown that fine-tuning with instructions or with a template is needed to perform unseen tasks [9, 31]. For end-to-end quiz generation, we propose to use the template specified below as our completion. We propose an end-to-end quiz generator based on GPT-3, EduQuiz, which is fine-tuned on text-quiz pairs to generate quizzes on never-seen-before texts.

GPT-3 comes in four different versions called Ada, Babbage, Curie, and Davinci which contain 350M, 1.3B, 6.7B, and 175B parameters respectively[3]. We used Curie as manual tests have shown that similar results can be achieved with a smaller model when fine-tuning. Lastly, we kept the default fine-tuning hyperparameters.

| **End-to-End Quiz Generation Template** |
| --- |
| Question: ... |
| True answer: ... |
| False answer: ... |
| False answer: ... |
| False answer: ... |

[2] https://beta.openai.com/docs/guides/fine-tuning
[3] https://blog.eleuther.ai/gpt3-model-sizes/

## 4 Experimental Setup

The main question that is addressed in this paper is: *Can generative pre-trained transformers learn to generate a reading comprehension quiz?* We aim to answer this question by evaluating SWQG and EEQG. In this section, we will elaborate on the dataset used, our evaluation method, and the tested models.

### 4.1 Dataset

We perform our experiments on the EQG-RACE dataset[3] [13]. It is a processed RACE dataset where only examination questions are kept. As we are focusing on the education domain, this dataset is a good fit for our purposes. During processing, Jia et al. removed the distractors from the data and only kept the questions and answers. We combine the questions from EQG-RACE with the original data in RACE[4] to extract the distractors again.

The EQG-RACE dataset contains 18,501 train, 1,035 validation, and 950 test instances. After connecting it to the original RACE dataset, each data instance consists of a reading comprehension text and a quiz, containing a question, one true answer, and three distractors, as specified in the template above.

### 4.2 Automatic Evaluation

For automatic evaluation, we follow previous work on QG, QA, and DG and use the existing evaluation methods BLEU-4, ROUGE-L, and METEOR as our metrics [3, 18, 21]. BLEU-4 measures the 4-gram similarity between a prediction and ground truth instances [21]. ROUGE-L measures the longest common sub-sequence between the prediction and ground truth instances [18]. METEOR is similar in comparison to BLEU-4 but also takes synonyms, stemming, and paraphrasing into account [3]. For BLEU-4 and ROUGE-L we use the Huggingface implementation[5]. As the newest version of the METEOR metric is not in Huggingface, we used the implementation from the original website[6].

### 4.3 Tested Models

To test our approach from Section 3, we compared both a general-purpose model called Macaw-11b, and task-specific fine-tuned GPT-3 models [4, 27] on SWQG and EEQG. For SWQG, we concatenate the QG, QA, and DG configurations of Macaw-11b to generate a quiz in a step-wise manner. Macaw-11b could not be used for EEQG[7]. To perform SWQG with GPT-3, we have fine-tuned the QG, QA, and DG models according to the different prompt-completion pairs from the earlier Table 1. We again concatenate these models to perform SWQG.

[3] https://github.com/jemmryx/EQG-RACE
[4] https://www.cs.cmu.edu/~glai1/data/race/
[5] https://huggingface.co/metrics
[6] https://www.cs.cmu.edu/~alavie/METEOR

**Table 2.** Automatic Evaluation on Quiz Generation.

| *Models* | BLEU-4 | ROUGE-L | METEOR |
|---|---|---|---|
| Macaw-11b SWQG | 1.93 | 27.00 | 20.78 |
| GPT-3 SWQG | 9.67 | 34.92 | 24.88 |
| EduQuiz | **11.61** | **36.11** | **25.42** |

We compare the SWQG models with EduQuiz, an end-to-end quiz generator we proposed in Section 3. For clarification, the name of each experiment will be attached to the models. For example, Macaw-11b SWQG is the general-purpose Macaw-11b model configured for SWQG and GPT-3 SWQG is a GPT-3 model fine-tuned on SWQG.

## 5 Experimental Results

In this section, we will discuss the results of the experiments. We have performed SWQG and EEQG with Macaw-11b and fine-tuned GPT-3 models. Table 2 shows the automatic evaluation of the generated quizzes by the different models. From this table, we would expect that EduQuiz outperforms Macaw-11b SWQG and GPT-3 SWQG. We do however see that the BLEU-4 score is remarkably low for all tested models. The ROUGE-L and METEOR score are higher as the template specified in Section 3 boosts the scores. The next example showcases the BLEU-4 score for generated quizzes.

**Test instance ID: #46**
**Original quiz:**
Question: What is Edell's strongest character?
True answer: Liking to try new things.
False answer: Liking to remember things.
False answer: Liking to help others.
False answer: Liking to think of problems differently.

**Macaw-11b SWQG generated quiz (BLEU-4 = 0.0):**
Question: The writer's mother told her that if she wanted to be successful in life, she should be prepared for any kind of change. What did Edell do to prepare for the Asian MTV Awards?
True answer: she read the biographies of the guests
False answer: she practiced her English with her friends
False answer: she made a list of the names of the guests
False answer: she made a plan to talk with the guests

[7] Macaw-11b did have the configurations to perform EEQG but our experiments showed that only 14% of the generated quizzes contained all the elements of the quiz. Therefore, we will not use Macaw-11b for EEQG.

**GPT-3 SWQG generated quiz (BLEU-4 = 0.0):**
Question: What advice does Edell give us?
True answer: Get ready for any opportunity in life.
False answer: Try to live a colorful life.
False answer: Take any chance that comes up.
False answer: Explore your potential talents.

**EduQuiz-generated quiz (BLEU-4 = 0.0):**
Question: What advice does Edell give to young people?
True answer: Try to get yourself well-prepared in life.
False answer: Have a rich collection of CDs.
False answer: Never miss an opportunity to learn ballet.
False answer: Be a hostess of the Asian MTV Awards.

The example shows generated quizzes and the original quiz. We see that the generated quizzes from the different models are different from the original quiz. Therefore, when automatically comparing n-gram similarity with BLEU-4, there is little overlap with the original quiz and the BLEU-4 scores are 0.0. Thus, the automatic evaluation is problematic. The generated quizzes seem reasonable but the BLEU-4 scores are uninterpretable as they only represent n-gram similarity. Therefore, it is hard to evaluate the quality of generated quizzes with these automatic scores. From this example, we can see that Macaw-11b SWQG first generates a sentence regarding the prompt before asking the question. Also, the punctuation from the GPT-3 models seems better in comparison to the Macaw-11b SWQG model. Detailed analysis is needed to conclude which model generates the highest quality quizzes.

## 6 Analysis of generated quizzes

The experimental results showed that automatic evaluation on quiz generation is limited in expressing the quality of generated quizzes. We have seen that generated quizzes seem reasonable even when they have little overlap with the original quiz. In this section, we will analyse the experimental results with human evaluation, as this is the golden standard to evaluate natural language generation tasks [29].

In our analysis, we rely on previously used human evaluation approaches in QG and DG to evaluate generated quizzes [13, 33]. As shown in Table 3, the question, answer, and distractors are all evaluated on whether they are grammatical and fluent. Furthermore, the question is also rated on whether it is relevant to the passage and if the passage contains the answer. The answer is also rated on whether the generated text contains correct information, and whether the answer contains all information to answer the question properly. The distractors are rated on whether they are coherent with the text and whether they can mislead the learner to choose a wrong answer. Whenever there is true information in the distractors, the distractors fail in their distracting ability. Each of the metrics in Table 3 is binary rated. Binary rating the metrics will result in a hard cut-off.

**Table 3.** Human Evaluation metrics on the quality of generated quizzes.

| *Sub-part* | *Metric* | *Description* |
|---|---|---|
| Question | Fluency | whether the question is grammatical and fluent. |
| | Relevancy | whether the question is semantic relevant to the passage. |
| | Answerability | whether the question can be answered by the right answer. |
| Answer | Fluency | whether the answer is grammatical and fluent. |
| | Correctness | whether the answer contains correct information. |
| | Valid | whether the answer is a correct answer to the question. |
| Distractors | Fluency | whether the distractors are grammatical and fluent. |
| | Coherence | whether the distractors are relevant to the article and the question. |
| | Distracting Ability | whether the distractors can mislead the learner and if there is no true information in the distractors. |

**Table 4.** Human Evaluation on Quiz Generation. Metrics (%)

| | Question | | | Answer | | | Distractor | | | Average | | | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Models* | Flu | Rel | Ans | Flu | Cor | Val | Flu | Coh | Dis | Q | A | D | Avg |
| Macaw-11b SWQG | 98.0 | 64.0 | 39.0 | 93.7 | 47.7 | 31.7 | 98.0 | 55.3 | 37.0 | 67.0 | 57.7 | 63.4 | 62.7 |
| GPT-3 SWQG | 99.0 | **99.0** | 90.3 | **99.7** | **93.7** | 73.3 | 99.0 | 95.7 | 53.7 | 96.1 | 88.9 | 82.8 | 89.3 |
| EduQuiz | **99.3** | 97.3 | **92.3** | 99.0 | 91.7 | **77.7** | **99.7** | **97.3** | **60.3** | **96.3** | **89.4** | **85.8** | **90.5** |

Also, as we have built upon previous work, there is a slight overlap between the metrics. Lastly, we experienced that the output of Macaw-11b is often different from GPT-3 which made double-blind evaluation impossible.

The human evaluation scores are an average score of three human annotators on 100 test instances for that specific task. The annotators rated 82.0% of all instances similarly. The Cohen kappa ($\kappa$) scores between editorial judges 1-2, 2-3, and 1-3 on all rated instances were 0.63, 0.53, and 0.49 respectively [5].

### 6.1 Step-Wise Quiz Generation (SWQG)

Table 4 shows that the quizzes generated by Macaw-11b SWQG mostly contain grammatical and fluent language. On all other metrics, Macaw-11b SWQG shows low performance. In contrast, GPT-3 SWQG generates quizzes that only lack the validness of the answer (73.7%) and the distracting ability of the distractors (53.7%). Overall, GPT-3 SWQG seems to perform better (89.3%) in comparison to Macaw-11b SWQG (62.7%) on the total average score.

Macaw-11b SWQG often first summarizes the text before it asks the question. GPT-3 SWQG generates stand-alone questions. The following example showcases where Macaw-11b SWQG and GPT-3 SWQG both generated a high-quality quiz. We do see that Macaw-11b SWQG lacks interpunction as it misses a capital letter and does not have a period at the end of the sentence. In contrast, GPT-3 SWQG does this correctly. When the models generate a relevant and answerable question, it directly becomes easier to generate a high-quality quiz. This is intuitive as the step-wise quiz generation starts with the task of creating a high-quality question. When the question is of low quality, the whole quiz will be of low quality.

**Test instance ID: #424**
**Macaw-11b SWQG:**
Question: Vampires have been a part of fiction for centuries. What is the most recent vampire story that has been published?
True answer: the twilight saga
False answer: Dracula
False answer: Frankenstein
False answer: The Sun Also Rises
**GPT-3 SWQG:**
Question: Why are vampires attractive to women?
True answer: Because they are noble and handsome.
False answer: Because they are powerful and rich.
False answer: Because they are the same as Count Dracula.
False answer: Because they are the containers into which we can pour our cultural desires.

The next example showcases that both models also generate quizzes of low quality. Here, Macaw-11b SWQG generated an irrelevant question which leads to an unusable quiz. GPT-3 SWQG asks a really easy question as the answer is in the question. This tricked GPT-3 SWQG to generate irrelevant answers and distractors. Macaw-11b SWQG and GPT-3 SWQG both suffer from repetition. Especially Mascaw-11b SWQG repeats the same answer option multiple times.

**Test instance ID: #847**
**Macaw-11b SWQG:**
Question: If a student is given a list of words that begin with the letter A, what will be the most likely sentence?
True answer: a sentence
False answer: a paragraph
False answer: a list
False answer: a chart
**GPT-3 SWQG:**
Question: What's Men Xue's family name?
True answer: He
False answer: She
False answer: They
False answer: We don't know

**Table 5.** EduQuiz-generated Quizzes. High quality generations: #82 and #379. Low question quality: #685. Low answer quality: #388. Low distractors quality: #55

| *ID* | *EduQuiz-generated Quizzes* |
|---|---|
| #82 | Question: Why did Everett feel optimistic?<br>True answer: Because Mr. Hanson seemed to be responsive to the idea.<br>False answer: Because he had already persuaded Mr. Hanson.<br>False answer: Because he knew Mr. Hanson well.<br>False answer: Because he was sure that the skateboard park could be built. |
| #379 | Question: Why did Mira run into the house when she saw the turtles?<br>True answer: Because she wanted to turn off the lights in the porch.<br>False answer: Because she wanted to get some food for the turtles.<br>False answer: Because she wanted to take off her shoes and socks.<br>False answer: Because she wanted to turn on the radio. |
| #685 | Question: If you want to have a nice Saturday, which is not ?<br>True answer: busy<br>False answer: expensive<br>False answer: boring<br>False answer: sad |
| #388 | Question: Which is NOT mentioned in the text?<br>True answer: Having a positive attitude in life.<br>False answer: Not being afraid of taking risks.<br>False answer: Making sure to do all the things on the list.<br>False answer: Writing down a list of things you want to do. |
| #55 | Question: How did Jocelyn disappear?<br>True answer: She disappeared from the spot where she was playing.<br>False answer: She disappeared when she was playing with her friends.<br>False answer: She disappeared when she was getting her bike.<br>False answer: She disappeared from her grandmother's apartment. |

Macaw-11b SWQG generated low-quality quizzes whereas GPT-3 SWQG generated quizzes of high quality. For GPT-3 SWQG, there is room for improvement on the validity of the answer and the distracting ability of the distractors.

### 6.2 End-to-End Quiz Generation (EEQG)

The evaluation results in Table 4 showcase that EduQuiz generates quizzes of comparable quality to GPT-3 SWQG on total average scores. The highest scores in the columns are interleaved between GPT-3 SWQG and EduQuiz. For SWQG, we needed to fine-tune three different models. To achieve the same results, we could just fine-tune one GPT-3 model which reduces the cost by a factor of three.

Table 5 shows generated quizzes by EduQuiz. The test instances #82 and #379 showcase that EduQuiz generates quizzes of high quality. The generated quizzes

are different from the ground truth as multiple quizzes can be valid for the same piece of text. The test instances #685, #388, and #55 showcase examples where EduQuiz generated low-quality quizzes. The first example of low-quality quizzes contains a question that is irrelevant and unanswerable. Therefore, the full quiz is of low quality. In the second row, EduQuiz generated a false answer in the place of a true answer as it has difficulties with the negation in the question. The third example shows that EduQuiz tends to switch true and false answers. The question and true answer are of high quality. However, the distractors also contain true information so they fail in distracting ability as the distractors could have been the true answer. Referring back to our human evaluation in Table 4, EduQuiz has the most difficulties with generating valid answers (77.7%) and generating distractors with a distracting ability (60.3%). EduQuiz sometimes lists facts about the question instead of generating a quiz by clearly separating the answer and distractors.

EduQuiz generates quizzes of comparable quality (90.5%) to GPT-3 SWQG (89.3%) and can generate a complete quiz in one pass rather than the three inference steps required by GPT-3 SWQG. It still has difficulties generating valid answers and distractors with a distracting ability.

## 7 Discussion & Conclusions

In this paper, we explored pre-trained language models for quiz generation on reading comprehension texts and propose an end-to-end quiz generator, EduQuiz, where we observed that the majority of generated quizzes is reasonable, and that generation of high-quality distractors is more challenging than question and answer generation. We have performed a comparative study on Step-Wise Quiz Generation (SWQG) and End-to-End Quiz Generation (EEQG). We performed automatic evaluation and analysed generated quizzes using human evaluation. We proposed two new tasks for text generation in the educational domain, SWQG, and EEQG. The first task is a concatenation of QG, QA, and DG. The latter is the generation of a full quiz only based on the context. Here, GPT-3 SWQG and EduQuiz outperformed Macaw-11b SWQG. This can be explained by the fact that this task is far more difficult and fine-tuning is needed otherwise it will not work. Besides that, GPT-3 is a strong larger pre-trained language model. GPT-3 SWQG and EduQuiz generated quizzes on the same quality level. Over all the domains, it is remarkable that almost all our generations contain fluent language. Traditional NLP pipelines often resulted in mixed quality text generation, but the large pre-trained language models seem to handle this very well with a lot of variation and expressiveness over rigid filled-out templates. EduQuiz generated questions that are relevant and answerable. The generated answer contains correct information but is not always a valid answer to the question. The generated distractors are coherent to the text but sometimes lack distracting ability.

There are some limitations to our research. The used models are mostly trained on the English language. Therefore, they will not fully generalize to other languages. Some manual tests have shown that the models can fulfill the trick in another language but the used language is not that expressive in comparison to English. Another limitation is that the fine-tuned GPT-3 models are costly. However, once the model is trained, it can be used for text generation on the fly with a far smaller completion cost. Moreover, the models are domain-specific and only create comparable questions and answers to the training dataset. There are two solutions to this problem. The first is to train the model on a really broad domain so that it can generate quizzes on all educational domains. Another solution is to create domain-specific quiz generators. The latter would probably generate better results for each domain specifically but it comes with a cost. Lastly, one could also argue that we have not made a fair comparison between the models. GPT-3 is fine-tuned on the task whereas Macaw-11b is used off the shelf.

While our experimental results are very encouraging and the model generates many useful quiz questions, our evaluation also reveals that not every quiz is perfect yet and the quality is sometimes lower than human-generated quizzes. This is issuing a call to caution to replace education professionals in particular for summative feedback and grading. This is also suggesting clear directions to further improve quiz generation for education, both directly by further improving the model and training regime, and indirectly in terms of the exact use case (current models may be helpful for formative rather than summative feedback), introducing ways of filtering out "bad" quizzes (as we can generate multiple candidates), or using it in a human-in-the-loop setting in an educator support system.

We hope to encourage other researchers to work on quiz generation as a research field with large potential impact on students and teachers, and with many applied research opportunities. We aimed to set the first step towards replacements of labor-intensive quiz generation by automatic quiz generation, thereby also contributing to the enhancement of textbooks with assessments.

### Acknowledgments

*We thank the reviewers for their insightful comments. Kamps is funded in part by the Netherlands Organization for Scientific Research (NWO CI # CISC.CC.016), and the Innovation Exchange Amsterdam (POC grant). Views expressed in this paper are not necessarily shared or endorsed by those funding the research.*

## References

[1] Anderson, R.C., Biddle, W.B.: On asking people questions about what they are reading. In: Psychology of learning and motivation, vol. 9, pp. 89–132, Elsevier (1975)

[2] Andre, T.: Does answering higher-level questions while reading facilitate productive learning? Review of Educational Research **49**(2), 280–318 (1979)

[3] Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, pp. 65–72 (2005)
[4] Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. arXiv preprint arXiv:2005.14165 (2020)
[5] Cohen, J.: A coefficient of agreement for nominal scales. Educational and psychological measurement **20**(1), 37–46 (1960)
[6] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
[7] Dresscher, L., Alpizar-Chacon, I., Sosnovsky, S., et al.: Generation of assessment questions from textbooks enriched with knowledge models. In: CEUR Workshop Proceedings, vol. 2895, pp. 45–59, CEUR WS (2021)
[8] Du, X., Shao, J., Cardie, C.: Learning to ask: Neural question generation for reading comprehension. arXiv preprint arXiv:1705.00106 (2017)
[9] Gao, T., Fisch, A., Chen, D.: Making pre-trained language models better few-shot learners. arXiv preprint arXiv:2012.15723 (2020)
[10] Hamaker, C.: The effects of adjunct questions on prose learning. Review of educational research **56**(2), 212–242 (1986)
[11] Heilman, M., Smith, N.A.: Good question! statistical ranking for question generation. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 609–617 (2010)
[12] Huang, Y.T., Tseng, Y.M., Sun, Y.S., Chen, M.C.: Tedquiz: automatic quiz generation for ted talks video clips to assess listening comprehension. In: 2014 IEEE 14Th international conference on advanced learning technologies, pp. 350–354, IEEE (2014)
[13] Jia, X., Zhou, W., Sun, X., Wu, Y.: Eqg-race: Examination-type question generation. arXiv preprint arXiv:2012.06106 (2020)
[14] Khan, S., Hamer, J., Almeida, T.: Generate: A nlg system for educational content creation. In: EDM (2021)
[15] Lai, G., Xie, Q., Liu, H., Yang, Y., Hovy, E.: Race: Large-scale reading comprehension dataset from examinations. arXiv preprint arXiv:1704.04683 (2017)
[16] Lelkes, A.D., Tran, V.Q., Yu, C.: Quiz-style question generation for news stories. In: Proceedings of the Web Conference 2021, pp. 2501–2511 (2021)
[17] Liang, C., Yang, X., Dave, N., Wham, D., Pursel, B., Giles, C.L.: Distractor generation for multiple choice questions using learning to rank. In: Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications, pp. 284–290 (2018)
[18] Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out, pp. 74–81 (2004)
[19] Liu, D., Lin, C.: Sherlock: a semi-automatic quiz generation system using linked data. In: International Semantic Web Conference (Posters & Demos), pp. 9–12, Citeseer (2014)
[20] Mitkov, R., Le An, H., Karamanis, N.: A computer-aided environment for generating multiple-choice test items. Natural language engineering **12**(2), 177–194 (2006)

[21] Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pp. 311–318 (2002)
[22] Prince, M.: Does active learning work? a review of the research. Journal of engineering education **93**(3), 223–231 (2004)
[23] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv preprint arXiv:1910.10683 (2019)
[24] Riloff, E., Thelen, M.: A rule-based question answering system for reading comprehension tests. In: ANLP-NAACL 2000 Workshop: Reading Comprehension Tests as Evaluation for Computer-Based Language Understanding Systems (2000)
[25] Rodríguez Rocha, O., Faron Zucker, C., Giboin, A.: Extraction of relevant resources and questions from dbpedia to automatically generate quizzes on specific domains. In: International Conference on Intelligent Tutoring Systems, pp. 380–385, Springer (2018)
[26] Sakaguchi, K., Arase, Y., Komachi, M.: Discriminative approach to fill-in-the-blank quiz generation for language learners. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 238–242 (2013)
[27] Tafjord, O., Clark, P.: General-purpose question-answering with macaw. arXiv preprint arXiv:2109.02593 (2021)
[28] Van Campenhout, R., Dittel, J.S., Jerome, B., Johnson, B.G.: Transforming textbooks into learning by doing environments: an evaluation of textbook-based automatic question generation. In: Third Workshop on Intelligent Textbooks at the 22nd International Conference on Artificial Intelligence in Education (2021)
[29] Van Der Lee, C., Gatt, A., Van Miltenburg, E., Wubben, S., Krahmer, E.: Best practices for the human evaluation of automatically generated text. In: Proceedings of the 12th International Conference on Natural Language Generation, pp. 355–368 (2019)
[30] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems, pp. 5998–6008 (2017)
[31] Wei, J., Bosma, M., Zhao, V.Y., Guu, K., Yu, A.W., Lester, B., Du, N., Dai, A.M., Le, Q.V.: Finetuned language models are zero-shot learners. arXiv preprint arXiv:2109.01652 (2021)
[32] Yin, J., Jiang, X., Lu, Z., Shang, L., Li, H., Li, X.: Neural generative question answering. arXiv preprint arXiv:1512.01337 (2015)
[33] Zhou, X., Luo, S., Wu, Y.: Co-attention hierarchical network: Generating coherent long distractors for reading comprehension. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 9725–9732 (2020)