International Journal of Assessment Tools in Education
2022, Vol. 9, Special Issue, 72–87
https://doi.org/10.21449/ijate.1124382
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Automatic story and item generation for reading comprehension assessments with transformers

Okan Bulut[1,*], **Seyma Nur Yildirim-Erbasli**[2]

[1]University of Alberta, Centre for Research in Applied Measurement and Evaluation, Edmonton, AB Canada
[2]Concordia University of Edmonton, Faculty of Arts, Department of Psychology, Edmonton, AB Canada

ARTICLE HISTORY

Received: June 1, 2022
Revised: Sep. 15, 2022
Accepted: Sep. 21, 2022

Keywords:
Reading comprehension,
Natural language processing,
Automatic item generation,
Language modeling,
Text generation.

Abstract: Reading comprehension is one of the essential skills for students as they make a transition from learning to read to reading to learn. Over the last decade, the increased use of digital learning materials for promoting literacy skills (e.g., oral fluency and reading comprehension) in K-12 classrooms has been a boon for teachers. However, instant access to reading materials, as well as relevant assessment tools for evaluating students' comprehension skills, remains to be a problem. Teachers must spend many hours looking for suitable materials for their students because high-quality reading materials and assessments are primarily available through commercial literacy programs and websites. This study proposes a promising solution to this problem by employing an artificial intelligence (AI) approach. We demonstrate how to use advanced language models (e.g., OpenAI's GPT-2 and Google's T5) to automatically generate reading passages and items. Our preliminary findings suggest that with additional training and fine-tuning, open-source language models could be used to support the instruction and assessment of reading comprehension skills in the classroom. For both automatic story and item generation, the language models performed reasonably; however, the outcomes of these language models still require a human evaluation and further adjustments before sharing them with students. Practical implications of the findings and future research directions are discussed.

1. INTRODUCTION

Reading comprehension is one of the essential skills that all students need to foster in K-12 education because their learning and success in other subjects (e.g., math, social studies, and history) are strongly associated with their proficiency in reading comprehension (Bigozzi et al., 2017). Reading comprehension is also the key ability that students need to master to make the transition from "learning to read" to "reading to learn" by understanding, analyzing, and applying information gathered through reading different materials (e.g., books, articles, and newspapers). Students without adequate reading comprehension skills may not be able to understand what they read and fail to make this transition.

*Corresponding Author: Okan Bulut ✉ bulut@ualberta.ca University of Alberta, Centre for Research in Applied Measurement and Evaluation, Edmonton, AB, Canada

e-ISSN: 2148-7456 /© **IJATE** 2022

Developing reading proficiency requires students to read more texts with varying volumes, genres, and difficulties (Allington et al., 2010; Duke et al., 2011; Kim & White, 2008). To help students develop reading comprehension skills, teachers give students various texts (e.g., fables, fairy tales, and stories) and ask them to read these texts repeatedly. Students who struggle with reading comprehension might have to practice their skills by reading more texts until they become fluent readers. Once students can read the text fluently, teachers also provide a set of items related to the text to measure students' understanding of the text. This suggests that teachers may need new reading materials and items to continuously monitor students' growth in reading. Teachers attempt to find a suitable text from the literature to meet this need efficiently. If they try to find a text from the literature, they need to go through many pieces of literature to find a suitable text, but this is a very time-consuming process. Also, it is not easy to find free reading materials because most of the materials on the Internet are commercially available.

Alternatively, the teachers may attempt to develop their own text and items associated with each text. However, writing original texts with different volumes, genres, or complexities is a highly complex task, even for a professional writer. In addition to finding a suitable text or creating an authentic text, developing high-quality items related to the text is another tedious task. Teachers must formulate high-quality items related to the text by targeting different difficulty levels and ensuring that each item is strongly associated with the text. Therefore, a more practical and sustainable solution is necessary to help teachers find suitable reading materials for their students.

1.1. Story and Item Generation

Writing and telling stories have been central to the human experience in every culture. As humans attempt to make sense of the world surrounding them, they make discoveries and learn new information. Storytelling is one of the most popular communication tools for gathering and sharing the knowledge gained through such valuable experiences. However, writing stories or narratives is not necessarily an easy task for humans. Even good writers struggle with creating a story that is not only syntactically and semantically sound but also describes the chain of events in a meaningful way. Also, finding the correct language elements leading to the generation of a good story is challenging. For example, the type of text (e.g., narrative vs. expository text) and readability (e.g., sentence and passage length) may affect how accurately individuals with differential reading abilities can comprehend a story (Begeny & Greene, 2014; Sáenz & Fuchs, 2002).

In schools, storytelling has always been a part of children's language and literacy development, especially in terms of oral fluency and reading comprehension (Agosto, 2016; Miller & Pennycuff, 2008; Peck, 1989). Both fluency and comprehension are highly essential skills for learning other subjects because students' ability to understand what they read in these subject areas is strongly associated with their reading fluency and comprehension (Bigozzi et al., 2017). Teachers typically use a variety of literature selections to improve children's oral fluency and comprehension skills and help them make the transition from learning to read to reading to learn. With the emergence of online or digital reading materials, teachers have also begun to use learning and assessment tools focusing on online reading comprehension (Bulut et al., 2022). Therefore, teachers always need new learning resources (i.e., online reading materials) and assessment tools to gauge children's academic growth in online reading comprehension.

Researchers found that the development of reading comprehension skills depends highly on the quality of reading materials teachers select for their students (Taylor et al., 2003; Tivnan & Hemphill, 2005). Teachers must look for digital reading materials suitable for their students to support children's literacy development. However, this is costly because most digital literacy

materials are commercial and thus require a paid subscription. Also, teachers need to develop items based on each reading material that could help them evaluate students' reading comprehension skills. Traditional procedures for creating items for reading comprehension assessments (e.g., manually developing items starting with where, when, when, who, and so on) are laborious, challenging, and costly. Emerging technologies can facilitate the search for appropriate reading materials and items for teachers, such as text generation using language models and automatic item generation (see Das et al. [2021] for a detailed summary of the state-of-the-art techniques used to generate items automatically).

1.2. Current Study

Previous studies indicated that students could improve their reading comprehension skills when they practiced reading frequently (Allington et al., 2010; Duke & Pearson, 2009; Duke et al., 2011; Guthrie, 2004; Kim & White, 2008; Rasinski, 2012; Taylor et al., 2000). In K-12 education, teachers use different kinds of grade-appropriate texts (e.g., fables, fairy tales, and short stories) to help students develop reading comprehension skills. This approach is essential for students who struggle with reading comprehension because they need to practice their reading skills more often by reading more texts. Because intensive reading is necessary for students with or without adequate reading comprehension skills, teachers need new reading materials constantly. Finding a relevant text from the literature is time-consuming because teachers must go through many pieces of printed or digital literature, and most materials are commercially available. In addition, the digital learning environment in the 21st century requires digital tools, including the availability of digital reading materials that can support teaching and learning activities. Therefore, there is a need in K-12 education to leverage the potential of digital instructional materials to foster students' reading comprehension skills. To address this need and provide a practical and sustainable solution, we aimed to build a story generation system to help teachers find suitable reading materials for their students. The primary objective of our study was to create an artificial intelligence (AI) system that can analyze existing reading materials to develop new stories and related items to improve students' reading comprehension skills.

2. METHOD

Emerging technologies, such as digital learning platforms and intelligent tutoring systems, have reshaped education during the past decade. These tools are frequently used in the classroom by K-12 teachers, and it is vital to design more digital tools to suit the learning needs of students. One of these learning needs is to provide reading resources and items to help students improve their reading comprehension skills. However, there is only a limited number of open-access digital reading resources available, and thus, teachers would have to spend a significant amount of time searching for appropriate materials for their students. This study aims to create an AI-based system that can analyze existing reading materials to create new, authentic texts and related items that can be used to improve and assess elementary students' reading comprehension skills. To achieve our goals, we fine-tune a pre-trained transformer model to generate new texts (i.e., reading passages) based on existing reading materials and create related items for the texts generated by the transformer model. The following sections will describe the story and item generation sections in detail.

2.1. Story Generation

We fine-tuned a pre-trained transformer model using classic children's books to perform story generation through a decoding approach. We searched reading materials (i.e., fairy tales and fables) that were freely available on the Internet and saved the grade-appropriate examples. In total, the dataset consisted of 3,700 human-written stories. During the training process, the

Adamax optimizer was applied with a learning rate of 5e-5, the batch size was 32, and the total number of training epochs was 3.

2.1.1. *Transformer Model: GPT-2*

Large-scale neural language models such as Bidirectional Encoder Representations from Transformers (BERT; Devlin et al., 2019), Generative Pre-trained Transformer (GPT; Radford et al., 2018) and GPT-2 (Radford et al., 2019) have been extensively trained on massive amounts of text to be used for complex language tasks. Pre-trained transformer language models demonstrate state-of-the-art performance across different natural language tasks such as text generation, summarization, and translation. These models can be expected to generate fluent and diverse texts due to the large amounts of data they were trained on (See et al., 2019). The reason behind the success of the transformer-based models for different natural language tasks is the diversity of the training dataset. They generate texts representative of the corpora on which they were trained. A common approach is to fine-tune these language models to a specific domain of interest by providing different corpora of exemplars. These transformer-based models can effectively learn from training data and generate high-quality texts by fine-tuning pre-trained models. This study uses GPT-2 model—a neural language model that achieves state-of-the-art performance across different tasks. The GPT-2 language model was trained with 1.5 billion parameters on a dataset of 8 million web pages to predict the next word for the previous words within a text (Radford et al., 2019).

2.1.2. *Decoding Algorithms*

Neural text decoding algorithms highly influence the quality of text generated (Holtzman et al., 2019; Kulikov et al., 2018). During decoding, a vector is applied to the softmax function to convert it into a probability for each word:

$$P(x|x_{1:i-1}) = \frac{\exp\,(u_i)}{\sum_j \exp\,(u_i)}, \tag{1}$$

where x is a token (e.g., words, characters, or subwords) at timestep i and u is a vector that contains the numerical value of every token in the vocabulary V. Considering the critical role of decoding algorithms in improving the performance of language models, we experimented with different decoding algorithms (beam search, random sampling with and without temperature, top-k sampling, and top-p sampling) with different parameters for each method (e.g., $p = 0.90$, $p = 0.92$, or $p = 0.95$ for top-p sampling) because the correct decoding algorithm is needed to generate high-quality and meaningful texts.

2.1.2.1. Beam Search. Beam search generates all possible tokens in a vocabulary list and then chooses the top B number of candidates with the highest probability at each timestep (Holtzman et al., 2019). However, the search may fail to choose between the two words or phrases and yield a text that repeats the same word or phrase. Therefore, it tends to produce low-quality texts with short sentences and excessive repetitions (Fan et al., 2018; Basu et al., 2020).

2.1.2.2. Random Sampling. This method uses the probability of each token from the softmax function to generate the next token (Holtzman et al., 2019). Thus, it samples directly from probabilities estimated by the model and can generate incoherent texts (Holtzman et al., 2019).

2.1.2.3. Sampling with Temperature. A probability distribution can be shaped through temperature (Holtzman et al., 2019). Temperature increases the probability of probable tokens while decreasing the likelihood of less probable tokens. It has been widely applied to text generation (Fan et al., 2018). Higher temperature values result in higher

randomness in the generated text. Temperature is used to scale the value of each token before going into a softmax function. Thus, given the temperature t, the softmax is re-estimated as follows:

$$P(x|x_{1:i-1}) = \frac{\exp(u_i/t)}{\sum_j \exp(u_i/t)}. \quad (2)$$

2.1.2.4. Top-*k* Sampling. Top-k sampling samples the next word from the k most likely words (Fan et al., 2018; Holtzman et al., 2018). Thus, top-k sampling involves a fixed number of most likely words and ensures that less probable words are not sampled. Because the top-k sampling restricts selection to the k-most likely words, the k subset of vocabulary, V, maximizes the probability of selected words:

$$\sum_{x \in V^{(k)}} P(x|x_{1:i-1}). \quad (3)$$

2.1.2.5. Top-*p* Sampling. Top-p or nucleus sampling restricts the sampling process to the smallest possible set of words whose cumulative probability exceeds the probability threshold (Holtzman et al., 2019). Top-p sampling distributes the probability among this set of words, and thus, the number of words in that set can dynamically increase or decrease based on the subsequent probability distribution, indicating that it involves a dynamic number of words based on a fixed p value:

$$\sum_{x \in V^{(p)}} P(x|x_{1:i-1}) \geq p, \quad (4)$$

where $V^{(p)}$ is the smallest possible set of words, $P(x|x_{1:(i-1)})$ is the probability of generating word x given the previously generated words x from 1 to $(i-1)$. This shows that the model selects the highest probability set of words whose cumulative probability exceeds the pre-chosen threshold p. Similar to the beam search, top-k and top-p sampling methods sometimes repeat words in a generated text for small values of k and p, while similar to random sampling, they generate incoherent text for large values of k and p (Basu et al., 2020).

2.1.2.6. Hybrid Sampling. We also tested a hybrid sampling approach (i.e., the combinations of top-k and top-p sampling).

2.1.3. *Model Evaluation*

To evaluate each story generation model, we performed human evaluation by rating the quality of generated stories based on five criteria: fluency, coherence, grammar, logical ordering of events, and human-sounding. We used a 5-point scale with the following score categories: 1 = Fundamental errors and no meaning; 2 = Fundamental errors and difficult to understand the meaning; 3 = Moderate errors but reasonably easy to understand the meaning; 4 = Minor errors and reasonably easy to understand the meaning; and 5 = Minor errors and easy to understand the meaning. To facilitate human evaluation, we generated stories with 100 words and selected a subsample of 15 texts for each prompt (prompt 1: "It was a beautiful day." and prompt 2: "Once upon a time"), resulting in 30 texts from each model (i.e., beam search, random sampling with and without temperature, top-k sampling, top-p sampling, and hybrid sampling) and a total of 180 texts. We selected the parameters of the fine-tuned model and decoding algorithms based on human evaluations.

In addition to human evaluation, we used the perplexity (PPX) index as a data-driven metric for evaluating automatic story generation models. The PPX index is widely used in natural language processing (NLP) for evaluating language models. It measures how well a language model predicts text (i.e., probabilities of selecting the right words for an unseen test set). The

PPX index is typically calculated as the inverse probability of a test set (i.e., a sequence of tokens produced by the language model), normalized by the number of words in the test set:

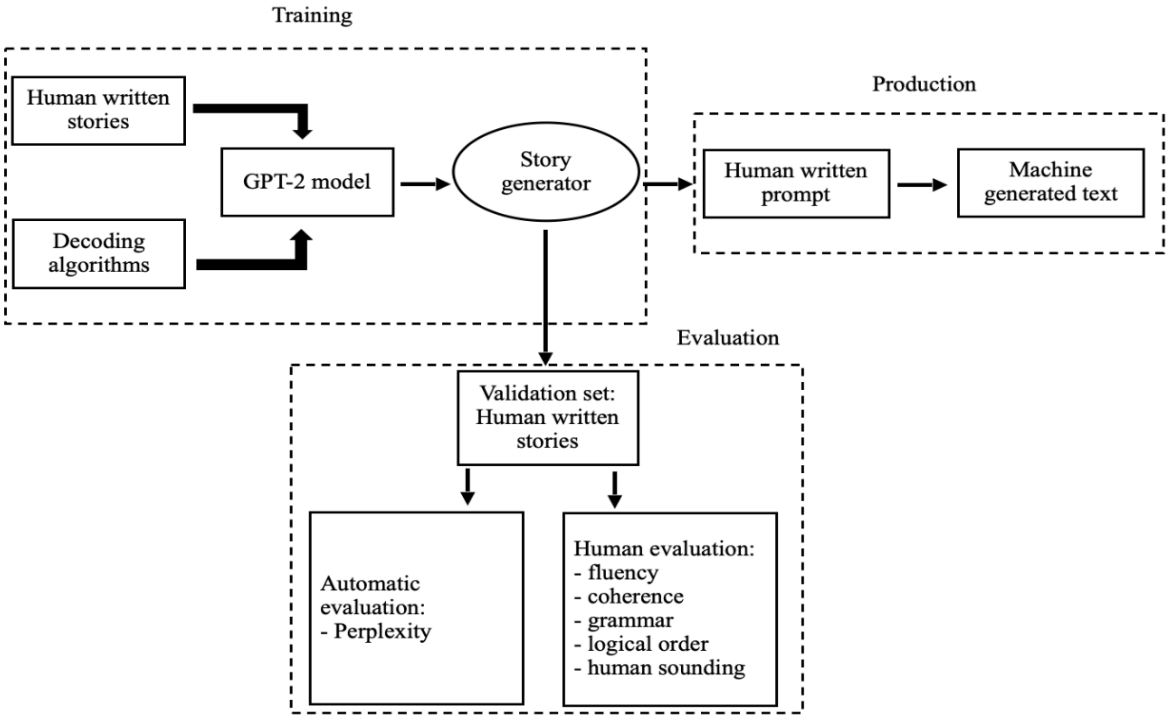
$$PPX(W) = P(w_1, w_2, \dots, w_N)^{-\frac{1}{N}} = \sqrt[N]{\frac{1}{P(w_1, w_2, \dots, w_N)}}, \quad (5)$$

where W is a tokenized sequence with N tokens, $W = (w_1, w_2, \dots, w_N)$ and $P(w_1, w_2, \dots, w_N)$ is the probability of observing a particular sequence of tokens. The lower the average value of PPX, the more accurate a language model. The PPX index can also be expressed as the exponential of the cross-entropy:

$$PPX(p, q) = -\sum_{X} p(x) log q(x), \quad (6)$$

where X refers to the language model's vocabulary of possible tokens (e.g., words or phrases), $p(x)$ is the target distribution for tokens, and $q(x)$ is the estimated distribution for tokens. PPX gets smaller as the predicted distribution becomes closer to the target distribution. In this study, the lower the perplexity of a story generation model, the better the model's accuracy when creating a new story. Figure 1 depicts the proposed framework for automatic story generation and model evaluation.

Figure 1. *The proposed framework for automatic story generation.*



2.2. Story Generation

We studied answer-aware item generation by jointly training item generation and answering and answer-agnostic item generation and compared their performance in terms of the quality of items generated. With answer-aware item generation, we aimed to design an algorithm that generates items and answers simultaneously and improves the performance of each other. We used a pre-trained transformer architecture to develop answer-aware and answer-agnostic item generation models. In terms of input, we used the texts generated from the story generation model and did not perform pre-processing (e.g., convert complex sentences into more straightforward sentences).

2.2.1. *Model Evaluation*

Item generation models aim to automatically generate a set of items that can be answered based on a particular content (Rus et al., 2012). This content can be a single sentence, paragraph, document, or database. Some researchers studied item generation and answer generation as dual tasks (e.g., Tang et al., 2017), while others generated items from texts without answers (Du & Cardie, 2017).

2.2.1.1. Answer-Aware Item Generation. Answer-aware item generation systems function with the content and generate items for target answers. However, generated items can be limited to certain types of items and focused on name entities (Dong et al., 2018) or arbitrary entities (Duan et al., 2017; Wang et al., 2020). Thus, answer-aware item generation approaches have the drawback of generating answers focusing on entities, and most items are easy to answer.

2.2.1.2. Answer- Agnostic Item Generation. Answer-agnostic item generation eliminates the requirement of the target answer before the items are generated. Answer agnostic item generation approaches reduce the bias toward entities while expanding the model flexibility (Wang et al., 2020). Although this approach is likely to generate more diverse items, it may also generate unanswered items (Sun et al., 2018; Wang et al., 2020).

2.2.2. *T5: Text-To-Text Transfer Transformer*

There are three approaches to item generation: rule-based, neural-based, and transformer-based. Item generation with a rule-based approach involves manually written rules for item generation based on heuristic rules and linguistic knowledge. The rule-based item generation systems can transform declarative sentences into interrogative sentences (e.g., Heilman & Smith, 2010; overgenerate and rank approach). However, these models are brittle and heavily depend on human effort. Therefore, rule-based models cannot be easily adapted to other domains (Zhou *et al*., 2018). Although rule-based models were more prevalent in generating items until the mid-2010s, there has been an increase in using neural networks since then (Pan *et al*., 2019).

Item generation with a neural-based approach trains a neural network based on a sequence-to-sequence framework from scratch. For example, Du et al. (2017) used a neural language model with an encoder-decoder architecture of the sequence-to-sequence model to generate items without relying on hand-crafted rules. An input sentence and its containing paragraph are encoded, and an item is generated by the decoder. Their proposed model outperformed the rule-based models (e.g., Heilman & Smith, 2010). However, the inherent sequential nature of these models makes it difficult to process long sequences. The sequence-to-sequence models cannot capture paragraph-level content, which is necessary to generate high-quality items. A generated item does not explicitly connect with the context of the target answer, and thus, includes a substantial portion of the target answer (Liu, 2020). Existing item generation models (e.g., Du et al., 2017) mostly use sentence-level content to generate items because models show significant performance degradation when applied to paragraph-level or long content. The transformer-based models address these problems.

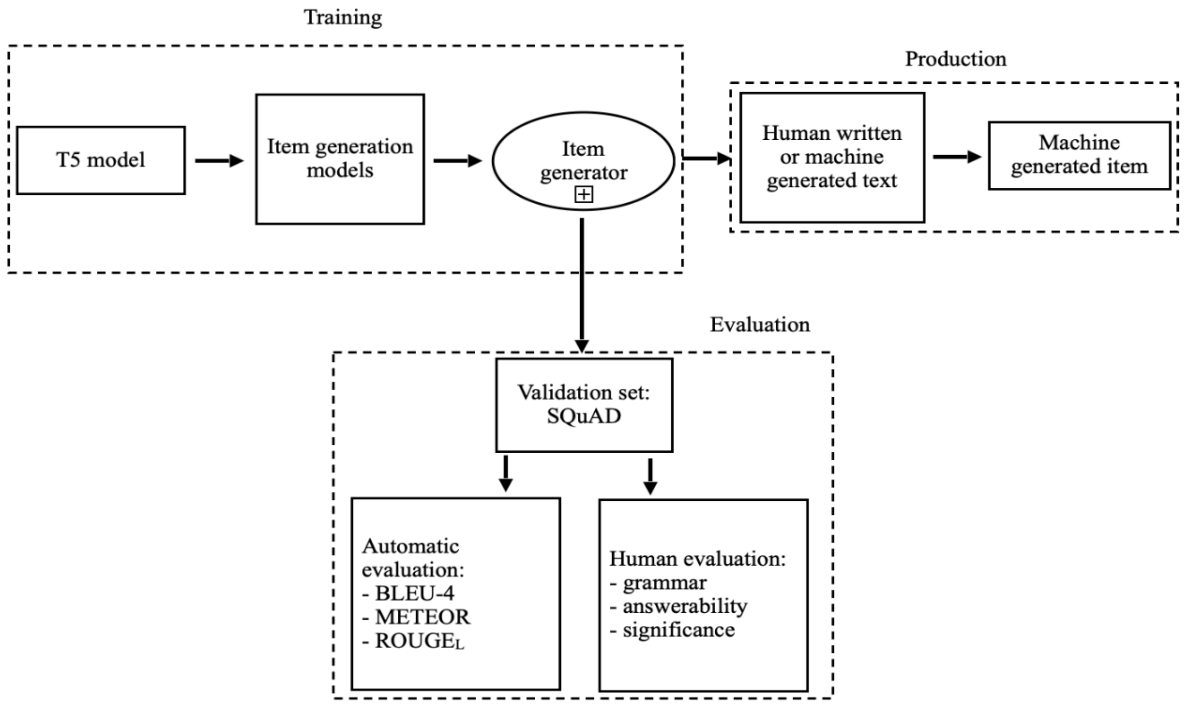
Transformers train and provide pre-trained models that show significant performance improvements in the NLP tasks (Radford et al., 2018). With transformer-based models, it is possible to improve the importance of item generation and to process paragraph-level content for item generation. We used T5: Text-to-Text Transfer Transformer that uses a text-to-text framework (i.e., takes text as input and generates new next as output) (Raffel et al., 2019). The T5 model is pre-trained on Colossal Clean Crawled Corpus (C4) and can be fine-tuned to achieve state-of-the-art results on different NLP tasks (Raffel et al., 2019). We trained the T5-small model for answer-aware and answer-agnostic item generation models and compared their performances.

2.2.3. *Model Evaluation*

We performed data-driven and human evaluations to analyze the performance of the item generation models. In terms of data-driven evaluation, we computed and reported BLEU, METEOR, and ROUGE scores using the SQuAD dataset (Rajpurkar et al., 2016). These metrics assign a score by measuring n-grams (i.e., sequence of words) and their frequency by comparing generated text with reference text. BLEU score is a more precision-based metric that provides an overall assessment of model quality by measuring the similarity of the generated text to the reference texts without considering semantic similarity (Papineni et al., 2002). BLEU-n (e.g., BLEU-4) counts co-occurrences by using up to n-grams. METEOR is a more recall-based metric that provides the similarity between generated texts and reference texts by considering synonyms, stemming, and paraphrases (Denkowski & Lavie, 2014). ROUGE is a more recall-oriented metric that compares generated text against reference text (Lin, 2004). $ROUGE_L$ measures the longest co-occurrence in n-grams by considering sentence-level structure similarities. For all three indices, larger values indicate better results.

In addition to data-driven evaluation based on the BLEU, METEOR, and $ROUGE_L$ scores, the generated items from the answer-aware and answer-agnostic models were also subject to human evaluation. We randomly selected 20 sets of items from each model using the inputs generated by the story generation model with the hybrid sampling approach. Two human evaluators rated the quality of the items based on the following criteria: grammar, answerability (i.e., the item can be answered based on the paragraph), and significance (i.e., the item relies on an essential piece of information from the paragraph). We used a 5-point scale ranging from 1 (very poor) to 5 (very strong) in the human evaluation of generated items. Figure 2 depicts the proposed framework for automatically generating items based on reading passages.

Figure 2. *The proposed framework for automatic item generation.*



3. RESULTS

3.1. Results for Automatic Story Generation

Table 1 shows two example stories generated using each decoding algorithm for two prompts: "It was a beautiful day." and "Once upon a time." Texts generated by the model with beam search showed extreme repetitions. The first example shows that the same sentence was repeated throughout the text, and the second example includes the repetition of words and sentences. Other researchers also reported a similar finding regarding beam search (e.g., Fan et al., 2018). In terms of the texts generated by the model with random sampling, although the texts may initially seem acceptable, when taking a closer look, they are not coherent and human-sounding. The reason might be that random sampling generates the next token by randomly sampling word sequences. When we tried random sampling with temperature to increase the likelihood of high probability words and decrease the likelihood of low probability words, the generated texts became coherent. However, the generated text still seemed problematic based on other criteria, particularly the logical ordering of the events.

Table 1. *Samples of generated texts for different decoding methods*.

Decoding Methods	Generated Texts
Beam search	*It was a beautiful day.* It was a beautiful day. It was a beautiful day. It was a beautiful day. It was a beautiful day. It was a beautiful day. It was a beautiful day. It was a beautiful day. It was a beautiful day. It was a beautiful day. It was a beautiful day. It was a beautiful day. It was a beautiful day. It was a beautiful day. It was a beautiful day. It was a beautiful day … *Once upon a time*, there was a man in the middle of the street. He was dressed in a black hooded sweatshirt and a white T-shirt. He had a black hooded sweatshirt and a white T-shirt. He had a black hooded sweatshirt and a white T-shirt. He had a black hooded sweatshirt and a white T-shirt. He had a black hooded sweatshirt and a white T-shirt … *It was a beautiful day.* The only real question is, will it be an issue when the Jets play at their pre-season home for the first time since 2011? The Jets have one more chance Sunday, when the San Francisco 49ers play host to the Indianapolis Colts. If they can't land a first-round draft pick from Chicago in Philadelphia as they did with the New York pick, this is just an opportunity to build something interesting with an actual head coach … *Once upon a time*, when one felt certain of his or her own powerlessness and self-hatred for others, he or she acted from a position of obedience to the law, but in another way acted from a position of obedience to his or her own and for the common good. Thus it is in the course of the history of the United States that it is possible to find and articulate an attitude of gratitude and admiration for others for acts of kindness …
Random sampling	*It was a beautiful day.* We had fun and we enjoyed it as much as anyone else. We were just very fortunate that we had a great time. The team's goal was to win the game and then make it 5-0. But it didn't work out that way. The Red Bulls, who won their first three games in Seattle, were 2-0 down at the break for the last time, and the Red Bulls were 0-2 down on their way to … *Once upon a time*, when we went to a place like this, it felt like a paradise. It was so cool to go to. It felt like heaven. It felt like a place where you could live. The only thing that I did that was drive past and watch the sunset. I can't remember anything about it. I just remember being in a place where the sun is shining. And I never saw the sunset before …
Sampling with temperature	

Table 1. *Continued*.

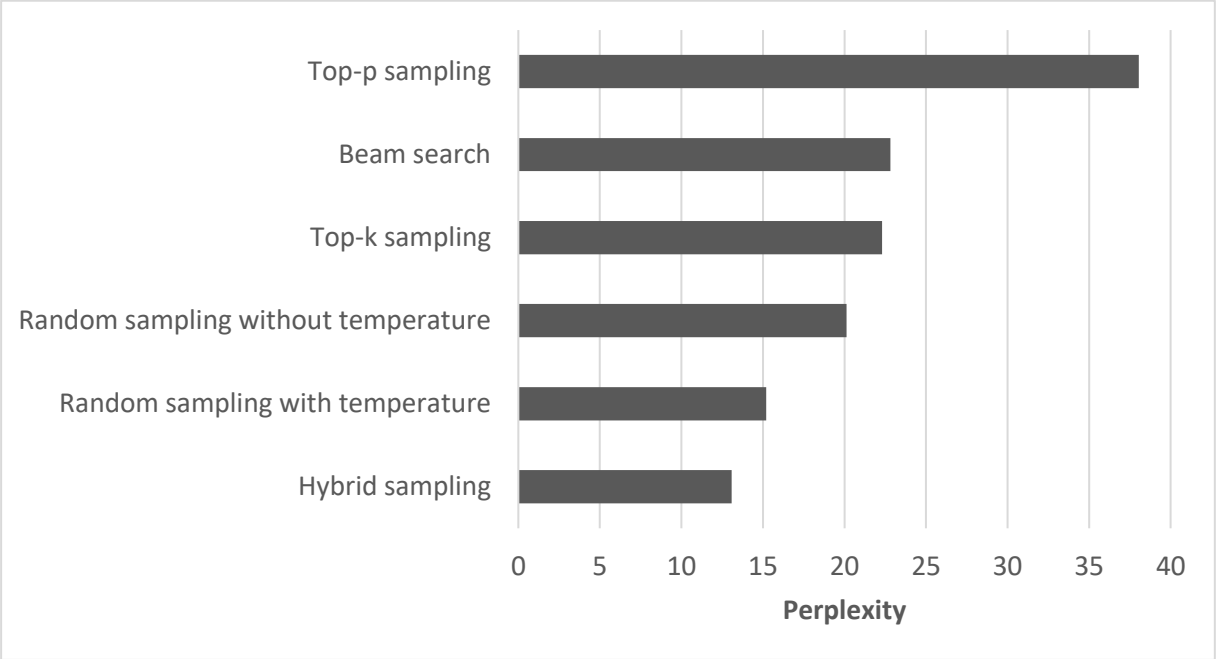
Decoding Methods	Generated Texts
Top-*k* sampling	*It was a beautiful day*. It reminded me that even though it was beautiful to me it was also a little bit hard to understand. I wasn't really ready to learn how to speak or speak Spanish so naturally I was getting into the habit of saying this to friends and family and sometimes I would say something like, "This isn't my first time speaking Spanish." But not anymore. The realization that my accent is more than just a language barrier made me realize I probably could do better … *Once upon a time* the same man had the power to force the enemy to surrender. He had no power to restrain them when he was in charge, so as to not give an enemy the chance of destroying him. The enemy, therefore, was allowed to escape without being put to any use. On this account the men of the village were ordered also to put to death their prisoners, and to eat of the corpses …
Top-*p* sampling	*It was a beautiful day*. I remember once in college the walls of the MIT Building were covered with graffiti and we all went down the hallway. So we all had fun around the MIT Building—people going to get their snowman blankets and they were breaking windows. In the morning I would go back to my office to read the fan pages that I would roll out in the middle of the night and go to class … *Once upon a time*, when thinking was a bold idea and no longer the tool of choice, power processes had a wider arc than experience and knowledge. Now we are beginning to recognize that power processes (as opposed to wishful thinking) are just a subset of working memory—each of them operates on the same information field, but with different energies. The architect of higher-order concepts has his or her own practices, decisions, and combinations of interests that could …
Hybrid sampling	*It was a beautiful day*. I was just amazed and happy to see this amazing team in action. I'm so thankful for the support and the training that they have given me over the years. I'm really happy that I have received so many support from the whole team and I'm so thankful for the support that they have given me … *Once upon a time*, the people of this country are working to create a world where the public is comfortable and the private sector is able to manage the economy. And that includes giving our youth a voice. And that includes educating them about the importance of social responsibility and the role of government in managing our economy. And that includes making sure that they understand the important role of private sector employees in our economy …

Compared to the other decoding algorithms, the model with top-*k* sampling generated higher-quality texts. This approach is more powerful in text generation as it filters only *k* most likely words and distributes probability among those *k* following words. The model generated texts with higher fluency, coherence, grammar, logical ordering of events, and human-sounding. However, it still shows some problems in terms of human-sounding (e.g., "how to speak or speak Spanish so naturally"). The reason can be that top-*k* sampling does not involve a dynamic number of words as it uses a fixed *k* number of words, limiting creativity in the model.

Using top-*p* sampling to sample from the smallest possible set of words instead of sampling only from the most likely k words produced texts with a wide range of words. Although both top-*k* and top-*p* produced high-quality texts, top-*p* seems to be a better decoding algorithm than top-*k* in theory (i.e., dynamic number of words). Finally, we had better results when we tried a hybrid of top-k and top-p sampling. After human evaluation of models by two raters, we selected the hybrid sampling—a combination of top-*p* and top-*k* sampling. The hybrid sampling was substantially more effective than other approaches because it generated texts with better fluency, coherence, grammar, logical ordering of events, and human-sounding.

In addition to human evaluation, we also used the PPX index to make a data-driven comparison among the story generation approaches. Figure 3 shows the perplexity results for each decoding method. The hybrid sampling approach yielded the smallest PPX value, suggesting that the text generated by this approach had the least amount of randomness based on the underlying language model. Surprisingly, top-*p* sampling yielded the largest perplexity value, followed by beam search. This finding indicates that the text generated by the top-*p* decoding method did not necessarily follow the underlying language model accurately. In other words, the text generated by top-*p* sampling included a high amount of randomness.

Figure 3. *Perplexity values for generated texts by different decoding algorithms.*



3.2. Results for Automatic Item Generation

Table 2 shows a sample of answer-agnostic and answer-aware items produced by the T5 model. The results showed that although the answer-agnostic item generation method allows for more flexibility in the language model (Wang et al., 2020), the answer-aware item generation method produced a higher number of quality items. Both methods (i.e., answer-agnostic and answer-aware item generation) yielded reading comprehension items with "what" and "how" questions. We identified two problems with the answer-agnostic items. First, some items already included the answer within the item stem. Second, some items involved a question statement that cannot be answered based on the available information from the reading passage. The items generated by the answer-aware item generation method were also problematic in other ways. For example, some items (e.g., "What kind of day was it?") required very simple or trivial answers. Also, some items were unnecessarily long, such as "What do the people of this country do to create a world where the public is comfortable and the private sector is able to manage the economy?". Further revisions are necessary to make these items helpful in evaluating students' reading comprehension.

Table 2. *Sample items generated by the answer-agnostic and answer-aware models.*

Model Input	Answer-Agnostic Items	Answer-Aware Items
It was a beautiful day. I was just amazed and happy to see this amazing team in action. I'm so thankful for the support and the training that they have given me over the years. I'm really happy that I have received so many supports from the whole team and I'm so thankful for the support that they have given me.	What was a beautiful day for me? What team has given me the support they have given me over the years? How many support did I receive from the whole team?	Q: What kind of day was it? A: Beautiful Q: What team was in action? A: Amazing Q: What is the reason I'm so thankful for? A: The support and the training that they have given me Q: Who has given me so much support? A: The whole team
Once upon a time, the people of this country are working to create a world where the public is comfortable, and the private sector is able to manage the economy. And that includes giving our youth a voice. And that includes educating them about the importance of social responsibility and the role of government in managing our economy. And that includes making sure that they understand the important role of private sector employees in our economy.	What are the people of this country working to create a world where the public is comfortable and the private sector is able to manage the economy? What does that include giving our youth a voice? How do the youth learn about the importance of social responsibility and the role of government?	Q: What are the people of this country working to create a world? A: The public is comfortable and the private sector is able to manage the economy Q: What do the people of this country do to create a world where the public is comfortable and the private sector is able to manage the economy? A: Giving our youth a voice Q: What is the key to educating youth about the importance of social responsibility and the role of government in managing our economy? A: Making sure that they understand the important role of private sector employees in our economy

Q: Question; A: Answer.

Table 3 shows the model evaluation indices for the items generated by the answer-agnostic and answer-aware methods. The results show that the answer-aware item generation performed slightly better than the answer-agnostic item generation; however, the difference between the two methods was negligible. Overall, the findings of our study appear to broadly support the work of other studies in automatic item generation. In our study, the answer-agnostic method yielded unanswerable items and failed to generate diverse items (Sun et al., 2018; Wang et al., 2020). Also, the answer-aware method yielded simple items that do not necessarily require higher levels of reading comprehension to find the correct answer.

Table 3. *Evaluation indices for the items generated by the answer-agnostic and answer-aware methods.*

Item Generation Model	BLEU-4	METEOR	$ROGUE_L$
Answer-Agnostic	18.3	24.7	39.9
Answer-Aware	18.6	24.9	40.2

4. DISCUSSION and CONCLUSION

Pre-trained transformer models can generate high-quality texts and items due to the large amounts of corpus they are trained on (See et al., 2019). In this study, we fine-tuned pre-trained transformer models to generate new stories and related items to enhance and assess students' reading comprehension skills. The proposed story and item generation models attain a fine-tuned understanding to produce human-like stories and items. However, it should be noted that the models might generate stories with repetitive words or unnatural changes in the topic. These weaknesses of language models remain a common challenge for the NLP community (Radford et al., 2019).

Our story generation model with hybrid sampling showed promising results in producing fluent, coherent, grammatically correct, logical, and human-sounding stories that students could use to practice and enhance their reading comprehension skills. Also, our answer-aware item generation model showed promising results in producing grammatically correct, answerable, and significant items. These language models for automatic story and item generation could enable teachers to generate authentic stories and items on the fly and share them with their students easily, without having to look for freely available printed or digital materials for hours. However, it should be noted that the generated items may still require a human evaluation and further adjustments before sharing them with students as they are likely to involve semantic errors (i.e., grammatically correct but nonsensical text). Also, the generated items may not be suitable for measuring complex reading skills such as inferencing, analyzing, and critiquing. Overall, the proposed models provide a feasible solution to the problem of finding new texts from the limited printed or digital materials and related items to the texts.

There are several limitations of this study. First, we used GPT-2 small and T5-small (i.e., the smallest versions of GPT-2 and T5) to generate stories and items due to their relatively less demand for computing power. It is possible that more advanced versions of the GPT-2 (e.g., GPT-2 large) and T5 (e.g., T5-base and T5-large) could generate higher-quality stories and items. Second, this study used a training dataset that involved freely available reading materials (i.e., fairy tales and fables) available on the Internet. A larger-size training dataset including more diverse reading materials (e.g., short stories, articles, or novels) could help fine-tune a transformer model more effectively and yield more consistent results in story and item generation stages. Finally, the sample stories and items generated in this study were not shared with students. Future studies on automatic story and item generation could involve students who can provide feedback on the readability and clarity of the generated stories and items. The feedback from students could facilitate the fine-tuning of pre-trained language models.

Acknowledgments

The research leading to these results received funding from the University of Alberta's Endowment Fund for the Support for the Advancement of Scholarship (SAS) Program.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors.

Authorship Contribution Statement

Okan Bulut: Investigation, Resources, Methodology, Software, Formal Analysis, and Writing-original draft. **Seyma Nur Yildirim-Erbasli**: Methodology, Software, Formal Analysis, and Writing-original draft.

Orcid

Okan Bulut https://orcid.org/0000-0001-5853-1267
Seyma Nur Yildirim-Erbasli https://orcid.org/0000-0002-8010-9414

REFERENCES

Agosto, D.E. (2016). Why storytelling matters: Unveiling the literacy benefits of storytelling. *Children and Libraries*, *14*(2), 21-26. https://doi.org/10.5860/cal.14n2.21

Allington, R.L., McGill-Franzen, A., Camilli, G., Williams, L., Graff, J., Zeig, J., Zmach, C., & Nowak, R. (2010). Addressing summer reading setback among economically disadva ntaged elementary students. *Reading Psychology, 31*(5), 411-427. https://doi.org/10.108 0/02702711.2010.505165

Basu, S., Ramachandran, G.S., Keskar, N.S., & Varshney, L.R. (2020). Mirostat: A neural text decoding algorithm that directly controls perplexity. *arXiv preprint*. https://doi.org/ 10.48550/arXiv.2007.14966

Begeny, J.C., & Greene, D.J. (2014). Can readability formulas be used to successfully gauge difficulty of reading materials? *Psychology in the Schools*, *51*(2), 198-215. https://doi.or g/10.1002/pits.21740

Bigozzi, L., Tarchi, C., Vagnoli, L., Valente, E., & Pinto, G. (2017). Reading fluency as a predictor of school outcomes across grades 4-9. *Frontiers in Psychology, 8*(200), 1-9. https://doi.org/10.3389/fpsyg.2017.00200

Bulut, H.C., Bulut, O., & Arikan, S. (2022). Evaluating group differences in online reading comprehension: The impact of item properties. *International Journal of Testing*. Advance online publication. https://doi.org/10.1080/15305058.2022.2044821

Das, B., Majumder, M., Phadikar, S., & Sekh, A.A. (2021). Automatic question generation and answer assessment: A survey. *Research and Practice in Technology Enhanced Learning* , *16*(1), 1-15. https://doi.org/10.1186/s41039-021-00151-1

Denkowski, M., & Lavie, A. (2014, June). Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation* (pp. 376-380).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirec tional transformers for language understanding. *arXiv preprint*. https://doi.org/10.48550 /arXiv.1810.04805

Dong, X., Hong, Y., Chen, X., Li, W., Zhang, M., & Zhu, Q. (2018, August). Neural question generation with semantics of question type. In *CCF International Conference on Natural Language Processing and Chinese Computing* (pp. 213-223). Springer, Cham.

Du, X., & Cardie, C. (2017, September). Identifying where to focus in reading comprehension for neural question generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 2067-2073). https://doi.org/10.18653/v1/ D17-1219

Du, X., Shao, J., & Cardie, C. (2017). Learning to ask: Neural question generation for reading comprehension. *arXiv preprint*. https://doi.org/10.48550/arXiv.1705.00106

Duan, N., Tang, D., Chen, P., & Zhou, M. (2017, September). Question generation for question answering. In *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 866-874). https://doi.org/10.18653/v1/D17-1090

Duke, N.K., & Pearson, P.D. (2009). Effective practices for developing reading comprehension. *Journal of Education, 189*(1/2), 107-122. https://doi.org/10.1177/002205740918900 1-208

Duke, N.K., Pearson, P.D., Strachan, S.L., & Billman, A.K. (2011). Essential elements of fostering and teaching reading comprehension. *What research has to say about reading instruction*, *4*, 286-314.

Fan, A., Lewis, M., & Dauphin, Y. (2018). Hierarchical neural story generation. *arXiv preprint*. https://doi.org/10.48550/arXiv.1805.04833

Guthrie, J.T. (2004). Teaching for literacy engagement. *Journal of Literacy Research, 36*(1), 1-30. https://doi.org/10.1207/s15548430jlr3601_2

Heilman, M., & Smith, N.A. (2010, June). 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).

Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2019). The curious case of neural text degeneration. *arXiv preprint*. https://doi.org/10.48550/arXiv.1904.09751

Holtzman, A., Buys, J., Forbes, M., Bosselut, A., Golub, D., & Choi, Y. (2018) Learning to write with cooperative discriminators. *arXiv preprint*. https://doi.org/10.48550/arXiv.1805.06087

Kim, J.S., & White, T.G. (2008). Scaffolding voluntary summer reading for children in grades 3 to 5: An experimental study. *Scientific Studies of Reading, 12*(1), 1-23. https://doi.org/10.1080/10888430701746849

Kulikov, I., Miller, A.H., Cho, K., & Weston, J. (2018). Importance of search and evaluation strategies in neural dialogue modelling. *arXiv preprint*. https://doi.org/10.48550/arXiv.1811.00907

Liu, B. (2020, April). Neural question generation based on Seq2Seq. In Proceedings of *the 2020 5th International Conference on Mathematics and Artificial Intelligence* (pp. 119-123).

Lin, C.Y. (2004, July). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out* (pp. 74-81).

Miller, S., & Pennycuff, L. (2008). The power of story: Using storytelling to improve literacy learning. *Journal of Cross-Disciplinary Perspectives in Education*, *1*(1), 36-43.

Pan, L., Lei, W., Chua, T.S., & Kan, M.Y. (2019). Recent advances in neural question generation. *arXiv preprint arXiv:* https://doi.org/10.48550/arXiv.1905.08949

Papineni, K., Roukos, S., Ward, T., & Zhu, W.J. (2002, July). 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).

Peck, J. (1989). Using storytelling to promote language and literacy development. *The Reading Teacher*, *43*(2), 138-141. https://www.jstor.org/stable/20200308

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. *OpenAI tech report*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI tech report*.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P.J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint*. https://doi.org/10.48550/arXiv.1910.10683

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 2383–2392).

Rasinski, T.V. (2012). Why reading fluency should be hot! *The Reading Teacher, 65*(8), 516-522. https://doi.org/10.1002/TRTR.01077

Rus, V., Wyse, B., Piwek, P., Lintean, M., Stoyanchev, S., & Moldovan. C. (2012). A detailed account of the first question generation shared task evaluation challenge. *Dialogue and Discourse, 3*(2), 177–204. https://doi.org/10.5087/dad

Sáenz, L.M., & Fuchs, L.S. (2002). Examining the reading difficulty of secondary students with learning disabilities: Expository versus narrative text. *Remedial and Special Education*, *23*(1), 31-41.

See, A., Pappu, A., Saxena, R., Yerukola, A., & Manning, C.D. (2019). Do massively pretrain ed language models make better storytellers? *arXiv preprint*. https://doi.org/10.48550/ar Xiv.1909.10705

Sun, X., Liu, J., Lyu, Y., He, W., Ma, Y., & Wang, S. (2018). Answer-focused and position-aware neural question generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 3930-3939).

Tang, D., Duan, N., Qin, T., Yan, Z., & Zhou, M. (2017). Question answering and question generation as dual tasks. *arXiv preprint*. https://doi.org/10.48550/arXiv.1706.02027

Taylor, B.M., Pearson, P.D., Clark, K., & Walpole, S. (2000). Effective schools and accomplished teachers: Lessons about primary-grade reading instruction in low-income schools. *The Elementary School Journal, 101*(2), 121-165. https://doi.org/10.1086/4996 62

Taylor, B.M., Pearson, P.D., Peterson, D.S., & Rodriguez, M.C. (2003). Reading growth in high-poverty classrooms: The influence of teacher practices that encourage cognitive engagement in literacy learning. *The Elementary School Journal, 104*(1), 3-28. https://d oi.org/10.1086/499740

Tivnan, T., & Hemphill, L. (2005). Comparing four literacy reform models in high-poverty schools: Patterns of first-grade achievement. *The Elementary School Journal*, *105(5)*, 419-441. https://doi.org/10.1086/431885

Wang, B., Wang, X., Tao, T., Zhang, Q., & Xu, J. (2020, April). Neural question generation with answer pivot. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, No. 05, pp. 9138-9145).

Zhou, Q., Yang, N., Wei, F., Tan, C., Bao, H., & Zhou, M. (2017, November). Neural question generation from text: A preliminary study. In *National CCF Conference on Natural Language Processing and Chinese Computing* (pp. 662-671). Springer, Cham.