Comparing Neural Question Generation Architectures for Reading Comprehension

E. Margaret Perkoff and **Abhidip Bhattacharyya** and **Jon Z. Cai** and **Jie Cao**
University of Colorado Boulder
Margaret.Perkoff@colorado.edu

Abstract

In recent decades, there has been a significant push to leverage technology to aid both teachers and students in the classroom. Language processing advancements have been harnessed to provide better tutoring services, automated feedback to teachers, improved peer-to-peer feedback mechanisms, and measures of student comprehension for reading. Automated question generation systems have the potential to significantly reduce teachers' workload in the latter. In this paper, we compare three different neural architectures for question generation across two types of reading material: narratives and textbooks. For each architecture, we explore the benefits of including question attributes in the input representation. Our models show that a T5 architecture has the best overall performance, with a RougeL score of 0.536 on a narrative corpus and 0.316 on a textbook corpus. We break down the results by attribute and discover that the attribute can improve the quality of some types of generated questions, including *Action* and *Character*, but this is not true for all models.

1 Introduction

The task of Automated Question Generation (AQG) has been proven to have significant potential for reducing teacher workload while effectively assessing reading comprehension for students (Kurdi et al., 2020). Reading comprehension is indicative of a student's understanding of a subject, making it a critical metric for ensuring their future academic success. Originally, advancements in Question Generation were isolated to broad question answering datasets including SQuAD (Rajpurkar et al., 2016) and NarrativeQA (Kočiský et al., 2018). In recent years, AQG models pre-trained on these datasets have been applied to education-specific corpora to help generate questions that are more useful in the classroom setting.

The domain shift from generic corpora to education-specific datasets is critical to model the unique characteristics of classroom discourse, but there is still much room for improvement. Classroom texts vary greatly in terms of the age of the students, the subject material, and their discourse structure. Prior work in question generation for education has focused on how language models perform on a single corpus with a single subject (Xu et al., 2022a), but not on how these models perform across different subjects. This research also considers how different discourse representations for a particular neural architecture can improve the quality of generated questions as opposed to evaluating multiple systems. Here, we analyze how different neural models perform on two corpora: the FairytaleQA Corpus (Xu et al., 2022b) and the Textbook Question Answering (Kembhavi et al., 2017) dataset. The FairytaleQA Corpus is representative of narrative comprehension, whereas the Textbook Question Answering dataset focuses on scientific topics including Physical, Earth, and Life Sciences. In addition to covering different subjects these datasets are quite different in terms of the passage structure. Earlier research on AQG has also considered how different forms of discourse representation, such as question type and event summarization, can improve the quality of questions generated (Zhao et al., 2022; Zhou et al., 2019). The FairytaleQA corpus distinguishes questions by seven attribute types. These attributes indicate the semantic nature of the question as well as the type of information that the reader is searching for either implicitly or explicitly from the text. We incorporate the question attribute into each of our model architectures to see whether the attribute has more significant impact when combined with a particular neural structure.

In this paper, we compare the performance of three different neural AQG architectures across two different datasets. We train baseline models for AQG on the FairytaleQA Corpus, (Xu et al., 2022b) a narrative dataset for K-12 reading comprehension, and the Textbook Question Answering (Kembhavi et al., 2017) dataset focused on middle school science. These models include a T5 (Raffel et al., 2019), BART (Lewis et al., 2019), and GPT-2 (Radford et al., 2019). We also investigate the impact of incorporating question attributes into these different model types for the FairytaleQA corpus. The T5 models achieve the highest metric rankings across both datasets, with BART outperforming GPT-2 on both as well. Including the question attribute as part of the input for training and inference improves the overall results for all model types, but leads to greater performance improvements for the GPT-2 and T5 models than for the BART model. Additionally, our by attribute breakdown found that including question attribute does not increase ROUGE scores for setting attribute questions. To our knowledge, this is the first comparison of a broader set of neural architectures for AQG in the education domain. These baselines are intended to inform future work on AQG in the classroom while taking into account the nuances of different subjects.

2 Related Works

2.1 Question Generation for the Education Domain

Significant amount of prior work addressed automated question generation (AQG) methods in the classroom. A review by Kurdi et al. (2020) concluded that AQG had the potential to provide significant benefit to teachers and students. Teachers can leverage question generation methods to automate assessment creation and reduce their workload. Question generation can also benefit students when used in tutoring or student-led learning contexts. Wang et al. (2018) introduced QG-Net, the earliest application of a model pretrained on a more general dataset (in this case SQuAD (Rajpurkar et al., 2016)) to the classroom material. They fine-tuned their model on the OpenStax textbooks [1]. The work of Zou et al. proposed an unsupervised method to generate true / false questions for reading comprehension. They compared a template-based framework and a pretrained BART model for text infilling. In human evaluations, the framework models outperformed the generative model in all categories except Fluency.

In 2022, Xu et al. (2022a) introduced the FairytaleQA dataset that we use for fine-tuning and establish a baseline for generating questions with a fine-tuned BART (Lewis et al., 2019) question answering model. They discovered that fine-tuning on the FairytaleQA dataset outperforms the BARTQA model fine-tuned on the NarrativeQA and FairytaleQA (Kočiský et al., 2018) corpora. Additionally, the distribution of attributes of the generated questions more closely resembled the distribution of the questions generated by expert human annotators. Their work implied the importance of fine-tuning models on domain specific datasets with high quality questions for reading comprehension. Rathod et al. introduced the concept of Multi Question Generation in the educational domain to create more lexically diverse questions that have the same answer.

This prior work in the education space has focused experiments largely on a single model architecture - BART, but has not considered more recent improvements in neural generative architectures. Grover et al. (2021) explored the use of a pre-trained T5 transformer model for the task of question generation without answer supervision. Their model was designed to take a passage as input and output multiple question-answer pairs related to the passage. It was trained and evaluated on the SQuAD dataset (Rajpurkar et al., 2016) for general question answering, but was not applied to education specific datasets. Based on their results, we evaluate the effectiveness of T5 in the education domain in our experiments.

Laban et al. (2022) looked beyond just generating quiz questions and conducted an experiment to evaluate generation errors. The result questions are categorized define a hierarchy of errors with three top-level justifications: *disfluent*, *off target*, and *wrong context*. Included among their models are three GPT-2 based models as well as two BART models - all of which are fine-tuned on the SQuAD dataset. The BART-large model has the second lowest rate of errors under their system with the GPT-2 based models all performing at the lower end of the range. Their experiment setup for both BART and GPT-2 does not fine-tune on pedagogical texts, so we will be able to explore if this boosts performance in our experiments.

2.2 Question Generation With Question Type or Attribute

Researchers have explored the use of question types or attributes to enhance question generation both

[1] https://openstax.org/k12

Dataset	Train	Valid	Test
FairytaleQA	6000	504	485
Textbook Question Answering	3346	1029	1074

Table 1: Breakdown of the datasets by training, validation, and test splits. Each sample includes an answer, a gold question, and section text from the relevant reading.

within and outside of a learning context. Zhou et al. proposed a model that would jointly predict the question type and generate a question. They distinguish between 8 types - seven types for different question words (*what*, *who*, *when*, *why*, *how*, *which*, *where*) and one *others* category. Their unified model outperformed earlier AQG methods on both the SQuAD and MARCO (Nguyen et al., 2016) datasets. Wang et al. sought to improve the diversity of generated questions by leveraging a conditional variational auto-encoder (CVAE) that incorporates the question types proposed in (Zhou et al., 2019). The CVAE approach demonstrated that incorporating question type did improve diversity of responses both on SQuAD and NewsQA (Trischler et al., 2017). Zhao et al. applied the idea of question type informed AQG to the FairytaleQA corpus. However, their approach involves taking a story passage as input and predicting the distribution of question types (noted as attributes in the context of the FairytaleQA data) to inform question generation. This distribution is then fed to an event-centric summary generation model and ultimately that output is passed on to a BART-based question generation model. Most of the aforementioned models were built with a pre-trained BART backbone, and none of these approaches considered using a T5 or GPT-model for the generation step. In our experiments, we incorporate the attribute value from the FairytaleQA corpus into all three of these model variants to compare the impact across different architectures.

3 Experimental Set Up

3.1 Datasets

We used two datasets for our experiments: Fairytale QA Corpus and Textbook Question Answering (TQA). A brief summary of the datasets is presented in Table 1. For the AQG task, we required having our data in the format of a story passage, or context C, an anticipated answer a, and a gold standard question g. During training, the goal is to generate a question that is as close (syntactically and semantically) to the gold question.

FairytaleQA
story: It so happened that Finn and his gigantic relatives were all working at the Giant's Causeway in order to make a bridge, ... **question:** Why were Finn and his gigantic relatives at the Giant's Causeway? **answer:** to make a bridge **attribute:** causal relation
TQA
context: A cold front occurs when a cold air mass runs into a warm air mass. This is shown in Figure 16.7. The cold air mass moves faster than the warm air mass and lifts the warm air mass out of its way. As the warm air rises, its water vapor condenses ... **question:** A warm front occurs when **answer:** a warm air mass slides over a cold air mass

Table 2: Question-Answer pair examples from FairytaleQA and TQA dataset

3.1.1 Fairytale QA Corpus

We use the FairytaleQA Corpus (Xu et al., 2022b) to assess the ability of our models to create meaningful questions based on narratives. This corpus contains 10,580 question-answer pairs based on 278 children-friendly stories. These pairs were created by annotators with expertise in education, cognitive science, and/or psychology. Each pair is labelled with the relevant story section. The corpus is further broken down into seven types of attributes: *character*, *setting*, *action*, *feeling*, *causal relationship*, *outcome resolution*, and *prediction*. All questions are also annotated as explicit or implicit - based on whether or not the answer to the question is explicitly stated in the corresponding text passage. Table 2 depicts an example from the fairytale dataset.

3.1.2 Textbook Question Answering

The Textbook Question Answering (TQA) dataset (Kembhavi et al., 2017) is based on questions from middle school textbooks in life science, earth science, and physical science. The original version contains 26,260 questions that can be used to train models for text-based and visual question answering. It is structured such that questions are associated with a particular lesson, but not the text passage from which the answer is drawn. Each lesson contains a set of topics along with a description of topic content. The questions are in a multiple-choice format and includes questions that refer to figures that are present in the text. We go

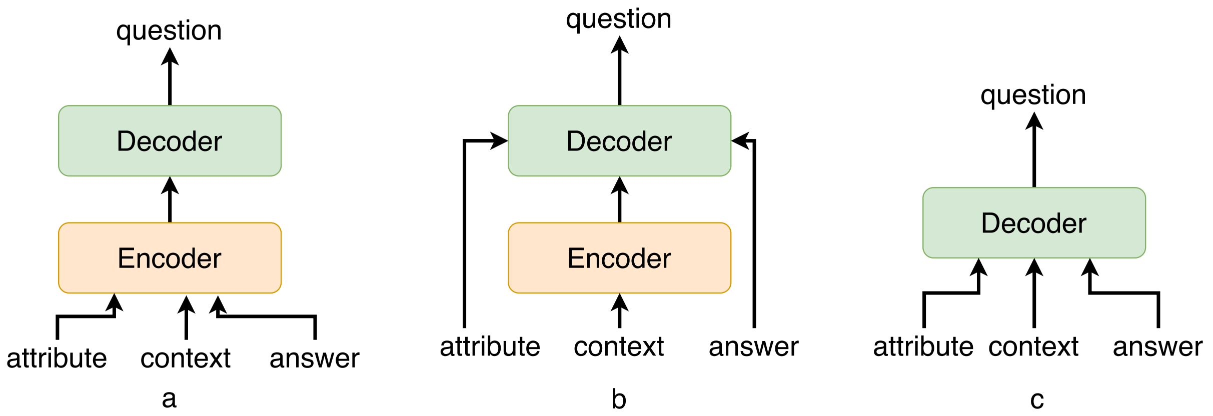

Figure 1: Architecture comparison: (a) represents T5 architecture, where the encoder takes attribute, context and answer text all together as input. (b) represents BART architecture, where encoder takes context as input and decoder takes attribute and answer as generation prefix. (c) represents GPT-2 architecture, where the decoder takes attribute, context and answer as generation prefix.

through the following preprocessing steps to make the dataset work for our text-based AQG task.

1. Remove any question that refers to a diagram.
2. Remove any question-answer pairs that require knowledge of more than one of the answer options, such as *Which of the following is false*, *None of the above* or *Answers B and C*.
3. For the remaining questions:
   (a) Extract the text from the correct answer label to use as the answer
   (b) Select the text passage or passages (in the case of a tie) with the highest word overlap between the passage and the question and answer to use as the context

The resulting dataset contains 3,346 question-answer pairs with context for training. Table 2 depicts an example from TQA dataset.

3.2 Models

We use three different pre-trained language models as our base model to further fine-tune on FairyTaleQA and TQA datasets to test the impact of different architectures and pre-training objectives to question generation task.

3.2.1 T5 Models

We use the T5 base model (Raffel et al., 2019) available from the huggingface library as the first example of a sequence-to-sequence architecture. T5 models treat all tasks as a text-to-text format, where the encoder takes source sequence as input and the decoder learns the generate output sequence. For question generation, the input text includes at minimum the question task indicator, a context passage, and then outputs a question. The encoder-decoder architecture can be seen in 1 a. The model is fine-tuned separately on each dataset for a total of 10 epochs with a learning rate of $1e^{-4}$. The attribute-based model for the FairytaleQA dataset includes the attribute along with a special token `attribute:`.

3.2.2 BART Models

Our second model is another encoder-decoder model. We use the BART base model (Lewis et al., 2019). To be specific, we deployed BartForConditionalGeneration from the huggingface library. Unlike the T5 model, we provide only the context text as input to the encoder. The attribute and answer was given to the decoder. The motivation was to enable the encoder to create a holistic representation of the context which can further be queried by the decoder with specific information. We trained the model for 50 epochs to learn both question generation and answer generation. During this training period for each data, with 50% probability the mode will be switched to either question generation or answer generation. During question generation the decoder will have the ground truth attribute and the answer. For answer generation, the decoder will have the ground truth attribute and the question. We further fine-tuned the model for 10 more epochs for question generation.

3.2.3 GPT-2 Models

The third model is a pre-trained GPT-2 model that leverages a pure decoder architecture (Radford et al., 2019). GPT-2(GPT-2 base model, 117M parameters) was trained on large amount of text with left-to-right Language Modeling objective, namely modeling the joint probability of a sequence of tokens in a left-to-right fashion of decomposition. The simplistic pre-training paradigm has been adopted by bigger and more powerful model successors such as GPT3, GPT4 and Llama(Brown et al., 2020; Touvron et al., 2023). We choose to test the viability of encoding Question generation task with GPT-2 given the amount of resources available and cost. We fine-tuned the GPT-2 model for the question generation task encoded with a prompt. See Sec.4.2 for more details about how we encode the question generation task for GPT-2.

4 Experiments

4.1 Evaluation

We evaluate our results based on both automated metrics and qualitative analysis. To compare our results with those of previous work, we use two standard evaluation metrics: BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004). BLEU measures the similarity between the generated sentence and one or more reference sentences based on n-gram overlaps. ROUGE also considers n-gram overlaps but is a recall-focused measure, while BLEU is precision-focused. ROUGE gives more weight to n-gram matches that occur in multiple references. In the context of response generation, this means that if multiple candidate responses include a particular phrase, it will have a greater impact on the scoring of a specific response. While these metrics are useful for comparison purposes, they have been shown to have limited correlation with human judgments (Liu et al., 2016). In future work, we intend to evaluate responses with a group of human annotators with teaching and reading comprehension expertise. As part of our initial human evaluation, we have included a qualitative analysis to break down some of the responses generated in each domain.

4.2 Prompting

Each of the models uses different input prompts as visualized in Table 3. Prompting for a specific model was inspired by the model's pre-training task. The T5 models are initially trained for multiple text-to-text tasks so they require a special token for the task, an input text value, and an output for each training example. To fine-tune on our datasets, we use the special token `ask_question` for the task, the input text includes a special token for `answer:` followed by the answer and a special token for the `context:` including the relevant section text. The output text value is the anticipated question output followed by a special token to signal the end of the output `</s>`. In the BART model, the encoder was fed with a special task token $\tau \in both, ask\_question$ and the story as the context. The beginning of the context is marked with a special token `context`. The target prompt consisted of mainly two elements- i) question-delineated by special tokens `<q> and </q>` and ii)answer-delineated by special tokens `<a> and </a>`. During the training for τ as `both` the order of question and answer was changed with a probability of 50% to help the encoder to capture task agnostic information. During training under the task $ask\_question$, question was the last element in the target prompt. Note as per the design of BART, the decoder during training will have the target string right shifted by one position. During inference the decoder will have the question and it is expected to generate the correct target string with the question. In GPT-2 we encode the story context, answer, question and optionally attributes in natural language format as "`story section: {story_context} Now given an answer: {answer_text}, a good question would be {question_text}`" (without attribute). in which the placeholder variables within the curly parenthesis are filled with each story QA triplet. Table 3 depicts an example of input and output for each model. The vocabulary of GPT-2 differs from BART and T5 in terms of the special tokens that it contains only a end-of-sentence token in the existing vocabulary. We therefore follow the default vocabulary configuration and not include extra untrained tokens such as start-of-sentence and segmentation tokens.

4.3 Models with Attribute Input

For the experimental condition where we include the question attribute as part of the input, we modify the prompt for each model accordingly. We add an additional special token `attribute:`

	Training
T5	input_text: ask_question: answer: on Knockmany Hill context: Finn lived at this time on Knockmany Hill,... output_text: Where did Finn live </s>
BART	encoder: ask_question:context: Finn lived at this time on Knockmany Hill,... target: <s>attribute:<a>on Knockmany Hill</a><q>Where did Finn live</q></s> decoder: </s><s>attribute:<a>on Knockmany Hill</a><q>Where did Finn live</q>
GPT	story section: Finn lived at this time on Knockmany Hill,... Now given an answer: on Knockmany Hill and it is related to {attributes_text}, a good question would be Where did Finn live
	Inference
T5	input_text: ask_question: answer: on Knockmany Hill context: Finn lived at this time on Knockmany Hill... output_text: Where did Finn live</s>
BART	encoder: ask_question:context: Finn lived at this time on Knockmany Hill,... target: Where did Finn live</q></s> decoder: attribute:<a>on Knockmany Hill</a><q>
GPT	story section: Finn lived at this time on Knockmany Hill,... Now given an answer: on Knockmany Hill and it is related to {attributes_text}, a good question would be

Table 3: Comparison of training and inference prompt styles for the T5, BART, and GPT models. The gold standard question from the dataset is: "Where did Finn live?" and the gold answer is "on Knockmany Hill". The full context of the story includes mention to the main character, a giant named Finn, his wife Oonagh, and his gigantic relations who reside on Knockmany Hill in Ireland. For brevity in the examples, we do not include the entire passage in the prompt table above.

to the input text for the T5 model, which is then followed by the corresponding question attribute for each training sample. At inference time, the attribute is also included as part of the input, and there is no change to the output text values for training or inference. For BART, the output prompt is modified by prepending the specific attribute token. For GPT-2, the prompt is modified as ‶`story section: {story_context} Now given an answer: {answer_text} and it is related to {attributes_text}, a good question would be {question_text}`″, with all attributes concatenated with comma within the attributes_text variable.

Model	Dataset	RougeL	BLEU
T5	FairytaleQA	**0.536**	**0.307**
T5-attr	FairytaleQA	0.500	0.279
BART	FairytaleQA	0.372	0.175
BART-attr	FairytaleQA	0.372	0.191
GPT	FairytaleQA	0.281	0.086
GPT-attr	FairytaleQA	0.295	0.087
T5	TQA	**0.316**	**0.107**
BART	TQA	0.166	0.042
GPT	TQA	0.089	0.008

Table 4: Quantitative evaluation scores for each of the models on the different datasets. We use the average BLEU score and RougeL for comparison with previous baselines.

5 Results

The result BLEU and RougeL scores across both datasets can be seen in Table 4. We found that the T5 models outperform all of the BART and GPT variations on both datasets. Our BART architecture achieves higher performance than the GPT models in all cases. On FairytaleQA, we found that incorporating attribute into the model's input did not significantly impact the RougeL or BLEU scores in comparison to the original variation. The T5 model achieves a higher RougeL score than the BART-QG model (0.527) fine-tuned in (Yao et al.,

Attribute	T5	T5-att	BART	BART-att	GPT	GPT-att
Action	0.562	**0.567**	0.371	0.335	0.291	0.269
Causal Relationship	**0.525**	0.408	0.364	0.400	0.257	0.307
Character	0.481	**0.516**	0.331	0.340	0.226	0.283
Feeling	**0.529**	0.507	0.424	0.403	0.347	0.354
Outcome Resolution	**0.478**	0.420	0.357	0.414	0.271	0.298
Prediction	**0.457**	0.410	0.324	0.303	0.119	0.203
Setting	**0.676**	0.667	0.486	0.438	0.428	0.346

Table 5: Breakdown of the RougeL scores by question attribute for the models fine-tuned on the FairytaleQA corpus.

<table>
<tr><th colspan="4">FairytaleQA</th></tr>
<tr><th>GT question</th><th>attribute</th><th>answer</th><th>Generated question</th></tr>
<tr><td>What happened after Finn lopped off the roots and branches of a fir-tree?</td><td>outcome resolution</td><td>He made a walking-stick of it and set out on his way to Oonagh.</td><td><i>T5:</i>What did Finn do after lopping off the roots and branches?
<i>T5-attr:</i>How did Finn get a fir-tree back?
<i>BART:</i>What happened after Finn pulled up a fir-tree?
<i>BART-attr:</i>What happened after Finn pulled up a fir-tree?
<i>GPT:</i>What did Oonagh do while Finn lived at Tallinn?
<i>GPT-attr:</i>What happened after Finn pulled up a fir-tree?</td></tr>
<tr><td>Where did Finn live?</td><td>setting</td><td>on Knockmany Hill</td><td><i>T5:</i>Where did Finn live?
<i>T5-attr:</i>Where did Finn live?
<i>BART:</i>Where did Finn and his gigantic relatives live?
<i>BART-attr:</i>Where did Finn live at this time?
<i>GPT</i>:Where did Finn's large relatives live?
<i>GPT-attr:</i>Where did Finn's large relatives live?</td></tr>
<tr><td>What will a farmer tell you about why the buckwheat looks black and singed?</td><td>action</td><td>The lightning did that</td><td><i>T5:</i> Why did the farmer say the buckwheat looked black and singed?
<i>T5-attr:</i> What made the buckwheat look black and singed?
<i>BART:</i>Fire?
<i>BART-attr:</i> Did it?
<i>GPT:</i> What did the farmer write?
<i>GPT-attr:</i> How did the farmer know about the black bear and its singed appearance?</td></tr>
<tr><th colspan="4">TQA</th></tr>
<tr><td>Air masses over the United States usually move toward the</td><td>-</td><td>northeast</td><td><i>T5:</i>Air masses tend to move toward the
<i>BART:</i>Where are air masses located?
<i>GPT:</i>What direction does the air masses move?</td></tr>
<tr><td>Our own species evolved during the</td><td>-</td><td>Cenozoic Era</td><td><i>T5:</i>mammals evolved during the
<i>BART:</i>the age of mammals was
<i>GPT:</i>Where did the men of the Cen section dwell?</td></tr>
<tr><td>Winds always blow from an area of higher to lower</td><td>-</td><td>pressure</td><td><i>T5:</i>Air always flows from an area of higher pressure to an area of lower
<i>BART:</i>Winds are generated by
<i>GPT:</i>A higher pressure would cause a weaker wind to blow., a good question would be Why does the upper part of the North America sink? Hamburger</td></tr>
</table>

Table 6: Examples of questions generated by our models. The top three rows present examples of questions generated from the FairytaleQA dataset, while the bottom three rows depict examples of questions generated from the TQA dataset. We noted consistency in the performance of the T5 model across both datasets.

2022) on the test split. However, our fine-tuned BART model performs significantly worse than the one from (Yao et al., 2022).

5.1 Results on the FairytaleQA Corpus

As a whole, the T5 models produce more sensical and relevant questions than the other model variations on the FairytaleQA Corpus. When we take a look at some of the individual questions produced by the T5 model, we find that in some cases they are identical to the gold question or within one or two words. However, the automated metrics do not capture some critical semantic errors in the generated questions. In some cases, the T5 model hallucinates additional information in the questions. For example, for the anticipated question *Where did Granua live?*, both of T5 and T5-attr generate *Where did Oonagh and Granua live?*. Additionally, the models sometimes switch the proper nouns between the subject and agent positions, changing the meaning of the gold question such as *What did Granua want from Oonagh?* to *What did Oonagh ask for from her sister?*. For these cases, we anticipate encoding more detailed discourse representations in the input, such as the use of named entity recognizers or abstract meaning representations could be highly benefical.

5.1.1 By-Attribute Comparisons

Table 5 shows the by-attribute breakdown of RougeL scores for each of the model architectures. Similarly to the overall scores, the T5 variants outperform both BART and GPT, and BART variants outperform the GPT ones across all question attributes. All model variations have the highest scores for the *Setting* attribute questions. The generated samples for gold label questions such as

Where did Finn live? can be seen in Table 6. All of the generated questions start with ‘where’ or ‘when’, include the correct character, Finn, and the correct verb: ‘live’. The T5 model also achieves high scores on the *Action*, *Causal Relationship*, and *Feeling* questions. However, the BART baseline scores well on the attributes of *Outcome Resolution*, *Feeling*, and *Causal Relationship*, relative to its performance on the *Action* attribute. The BART model that encodes the attribute as part of the input outperforms the standard BART model for the *Outcome Resolution* and *Character* questions, but not for the other ones. The GPT model with attribute also achieves higher performance than the one without attribute for *Outcome Resolution* and *Character* questions suggesting that generating these questions may be more influenced by the type of question. The T5 model with attribute also outperforms the baseline variation for Character questions and *Action* as well. Unlike the BART and T5 variants, the GPT model with attribute exceeds the RougeL score of the majority of the questions. This suggests that GPT style models may benefit the most from including attribute information in the input step. One thing to consider when evaluating the attribute models is the fact that all of these models original pre-training procedures rely on input that does not include the attribute, so we are limited to exposing the model to this type of input in the fine tuning stage. We could hope to see performance improvements with attribute models with more attribute encoded data available for the fine-tuning stage.

5.2 Error Analysis of the TQA Dataset

As with the FairytaleQA dataset, we found that the T5 model outperformed both the BART and GPT models in terms of automated metrics. When we analyzed the generated questions, we observed that the T5 model incorporates more context into the questions than the other two models. Specifically, on this dataset, BART tended to produce shorter output questions or, in some cases, no output at all. In contrast, the GPT models frequently included unnecessary additions, such as one that randomly had the word ’Hamburger’ appended to it. Refer to the last example of Table 6. The context passages included in this dataset require more specific concepts to be referenced, since generalizations may not be able to be made across passages in the text. For example, if a book is talking about how animals in the great plains adapt to their environment, this information is not going to transfer to a passage about how animals in the tundra survive. Although these are both adaptations, we need the context specific values. This indicates the need to consider more complex models or additional ways of representing passage context. The use of a knowledge graph to represent facts introduced in the textbook could have significant benefit in this domain.

5.3 Cross-Corpus Comparison

All of the models tested performed significantly better on the FairytaleQA dataset than they did for the Textbook Question Answering dataset. There are a number of factors that could have contributed to this gap in performance. The Textbook Question Answering corpus was originally designed to help improve the visual question answering task, specifically for multiple choice questions. We have modified the dataset using automated methods to fit the open question generation task instead. Our preprocessing methods are automated and could use a human review to ensure that we are not trying to generate questions that require knowledge of other answers from the multiple choice setting. Furthermore, the corpus is a third of the size of the FairytaleQA Corpus. Both domains suffered from factual correctness errors with the model replacing key nouns or names in the generated question with incorrect ones. This is something that could potentially be addressed with the use of discourse relations that are embedding in input.

6 Limitations and Future Work

Future work on automated question generation for learning contexts could benefit from a number of potential research paths. In this paper, we tested three different architectures - but there are many more to be considered including those that incorporate knowledge graphs which have been shown to improve the richness and semantic correctness of generated questions (Bi et al., 2020). There is also room to explore different prompt strategies including a fill-in-the-blank approach which may be more appropriate fo the TQA data. For the attribute models, we used the single task objective of question generation, but it would be worthwhile to explore jointly generating the question attribute and the question itself. Additionally, document level Abstract Meaning Representations with resolved coreferences has been shown to improve the quality of knowledge based question generation (Kapanipathi et al., 2021). We also recognize that we focused on different context for the input, but not on the wide variety of generation strategies available for this task. On top of the variety of model architectures, we would like to evaluate a greater set of corpora that include additional topics such as history and economics. Reading comprehension is critical to these fields as well and there is limited, if any, research on question generation for these topics.

Additionally, in future work we will conduct evaluation with expert annotators to incorporate into more complex models. Ideally, we will have educators and students assess the output of our models for factual correctness, relevance, and fluency of the questions generated. This output can then be used to train an instruction fine-tuned model. In order to make a solution that is viable for the classroom, it is critical to think beyond the automated metrics and get real teacher feedback. This preliminary research demonstrates the potential for expanding automated question generation to multiple classroom subjects and the value of incorporating discourse information into different model architectures to produce high quality questions.

7 Conclusion

In this paper, we conduct an initial comparison of automated question generation architectures for narrative stories (fairytales) and science textbooks. For each corpus, we trained BART, GPT-2, and T5 models to see which would perform best in which context. Our results indicate that the T5 models achieve the highest scores in terms of automated metrics for both domains. The highest performing T5 model also outperforms the BART baseline for question generation on the FairytaleQA dataset put forth in (Xu et al., 2022b). We also evaluated the effectiveness of encoding question attribute information in different model architectures. We saw improvements in performance for both *Character* and *Outcome Resolution* questions when the attribute was included for multiple architectures suggesting that this information is beneficial for generating certain types of questions, but not all. Additionally, the inclusion of attribute information led to a more significant improvement across question types for the GPT architecture.

8 Acknowledgements

This research was supported by the NSF National AI Institute for Student-AI Teaming (iSAT) under grant DRL 2019805. The opinions expressed are those of the authors and do not represent views of the NSF. The authors would like to thank Dr. Alexis Palmer, Dr. Peter Foltz, Dr. James Martin, and Dr. Clayton Lewis for their insightful comments and suggestions throughout the experiments and paper writing process.

References

Sheng Bi, Xiya Cheng, Yuan-Fang Li, Yongzhen Wang, and Guilin Qi. 2020. Knowledge-enriched, type-constrained and grammar-guided question generation over knowledge bases. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2776–2786, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Khushnuma Grover, Katinder Kaur, Kartikey Tiwari, Rupali, and Parteek Kumar. 2021. Deep learning based question generation using T5 transformer. In *Advanced Computing*, pages 243–255. Springer Singapore.

Pavan Kapanipathi, Ibrahim Abdelaziz, Srinivas Ravishankar, Salim Roukos, Alexander Gray, Ramón Fernandez Astudillo, Maria Chang, Cristina Cornelio, Saswati Dana, Achille Fokoue, Dinesh Garg, Alfio Gliozzo, Sairam Gurajada, Hima Karanam, Naweed Khan, Dinesh Khandelwal, Young-Suk Lee, Yunyao Li, Francois Luus, Ndivhuwo Makondo, Nandana Mihindukulasooriya, Tahira Naseem, Sumit Neelam, Lucian Popa, Revanth Gangi Reddy, Ryan Riegel, Gaetano Rossiello, Udit Sharma, G P Shrivatsa Bhargav, and Mo Yu. 2021. Leveraging Abstract Meaning Representation for knowledge base question answering. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3884–3894, Online. Association for Computational Linguistics.

Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler, and Salam Al-Emari. 2020. A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30(1):121–204.

Philippe Laban, Chien-Sheng Wu, Lidiya Murakhovs'ka, Wenhao Liu, and Caiming Xiong. 2022. Quiz design task: Helping teachers create quizzes with automated question generation. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 102–111, Seattle, United States. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. BART: Denoising Sequence-to-Sequence pre-training for natural language generation, translation, and comprehension.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset. *CoRR*, abs/1611.09268.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified Text-to-Text transformer.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Manav Rathod, Tony Tu, and Katherine Stasaski. 2022. Educational multi-question generation for reading comprehension. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 216–223, Seattle, Washington. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada. Association for Computational Linguistics.

Zhen Wang, Siwei Rao, Jie Zhang, Zhen Qin, Guangjian Tian, and Jun Wang. 2020. Diversify question generation with continuous content selectors and question type modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2134–2143, Online. Association for Computational Linguistics.

Zichao Wang, Andrew S Lan, Weili Nie, Andrew E Waters, Phillip J Grimaldi, and Richard G Baraniuk. 2018. QG-net: a data-driven question generation model for educational content. In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*, number Article 7 in L@S '18, pages 1–10, New York, NY, USA. Association for Computing Machinery.

Ying Xu, Dakuo Wang, Mo Yu, Daniel Ritchie, Bingsheng Yao, Tongshuang Wu, Zheng Zhang, Toby Li, Nora Bradford, Branda Sun, Tran Hoang, Yisi Sang, Yufang Hou, Xiaojuan Ma, Diyi Yang, Nanyun Peng, Zhou Yu, and Mark Warschauer. 2022a. Fantastic questions and where to find them: FairytaleQA – an authentic dataset for narrative comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 447–460, Dublin, Ireland. Association for Computational Linguistics.

Ying Xu, Dakuo Wang, Mo Yu, Daniel Ritchie, Bingsheng Yao, Tongshuang Wu, Zheng Zhang, Toby Li, Nora Bradford, Branda Sun, Tran Hoang, Yisi Sang, Yufang Hou, Xiaojuan Ma, Diyi Yang, Nanyun Peng, Zhou Yu, and Mark Warschauer. 2022b. Fantastic questions and where to find them: FairytaleQA – an authentic dataset for narrative comprehension. In

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 447–460, Dublin, Ireland. Association for Computational Linguistics.

Bingsheng Yao, Dakuo Wang, Tongshuang Wu, Zheng Zhang, Toby Li, Mo Yu, and Ying Xu. 2022. It is AI's turn to ask humans a question: Question-Answer pair generation for children's story books. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 731–744, Dublin, Ireland. Association for Computational Linguistics.

Zhenjie Zhao, Yufang Hou, Dakuo Wang, Mo Yu, Chengzhong Liu, and Xiaojuan Ma. 2022. Educational question generation of children storybooks via question type distribution learning and event-centric summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5073–5085, Dublin, Ireland. Association for Computational Linguistics.

Wenjie Zhou, Minghua Zhang, and Yunfang Wu. 2019. Question-type driven question generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6032–6037, Hong Kong, China. Association for Computational Linguistics.

Bowei Zou, Pengfei Li, Liangming Pan, and Ai Ti Aw. 2022. Automatic true/false question generation for educational purpose. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 61–70, Seattle, Washington. Association for Computational Linguistics.