Fantastic Questions and Where to Find Them: FairytaleQA – An Authentic Dataset for Narrative Comprehension

Ying Xu[†*] University of California Irvine

Dakuo Wang[†] IBM Research

Mo Yu[†] WeChat AI, Tencent

Daniel Ritchie[†] University of California Irvine

Bingsheng Yao[†] Rensselaer Polytechnic Institute

Tongshuang Wu University of Washington

Zheng Zhang University of Notre Dame

Toby Jia-Jun Li University of Notre Dame

Nora Bradford University of California Irvine

Branda Sun University of California Irvine

Tran Bao Hoang University of California Irvine

Yisi Sang Syracuse University

Yufang Hou IBM Research Ireland

Xiaojuan Ma Hong Kong Univ. of Sci and Tech

Diyi Yang Georgia Institute of Technology

Nanyun Peng University of California Los Angeles

Zhou Yu Columbia University

Mark Warschauer University of California Irvine

Abstract

Question answering (QA) is a fundamental means to facilitate assessment and training of narrative comprehension skills for both machines and young children, yet there is scarcity of high-quality QA datasets carefully designed to serve this purpose. In particular, existing datasets rarely distinguish fine-grained reading skills, such as the understanding of varying narrative elements. Drawing on the reading education research, we introduce FairytaleQA[1], a dataset focusing on narrative comprehension of kindergarten to eighth-grade students. Generated by educational experts based on an evidence-based theoretical framework, FairytaleQA consists of 10,580 explicit and implicit questions derived from 278 children-friendly stories, covering seven types of narrative elements or relations. Our dataset is valuable in two folds: First, we ran existing QA models on our dataset and confirmed that this annotation helps assess models' fine-grained learning skills. Second, the dataset supports question generation (QG) task in the education domain. Through benchmarking with QG models, we show that the QG model trained on FairytaleQA is capable of asking high-quality and more diverse questions.

Story Title: ***Magic Apples***
Story Text:

[Sect 1] Once upon a time there was a lad who was better off than all the others. He was never short of money, for he had a purse which was never empty. ...
...

[Sect 6] When the king's daughter had eaten of the apples, she had a pair of horns. And then there was such a wailing in the castle that it was pitiful to hear. ...
But one day a foreign doctor from afar came to court. He was not from their country, he said, and had made the journey purposely just to try his luck here. But he must see the king's daughter alone, said he, and permission was granted him.
...

[Sect 8] ...

- Q1:Who will the foreign doctor turn out to be?

[explicit][prediction][sect 5, sect 6]
- A: The Lad.

- Q2:How did the princess feel when she had a pair of horns?

[implicit][feeling][sect 6]
- A: Upset.
- A: Angry.
- A: Horrified.

Table 1: Story and Question-Answer examples in FairytaleQA. Each question has meta info (implicitness, question type, and section origin), and may have multiple answers and span across multiple sections.

1 Introduction

Reading comprehension is a complex, multidimensional cognitive process (Kim, 2017). Question answering (QA) is fundamental for supporting humans' development of reading comprehen-

[*]Equal contributions ying.xu@uci.edu, dakuo.wang@ibm.com, moyumyu@tencent.com, drritchi@uci.edu, yaob@rpi.edu. Work done while Mo was at IBM. [*] Corresponding Author.

[1]Our dataset is available at https://github.com/uci-soe/FairytaleQAData.

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 447 - 460

sion skills, as questions serve as both instruments for evaluation and tools to facilitate learning. To achieve this goal, comprehension questions should be valid and reliable, meaning that all items are designed to cohesively assess comprehension rather than some other skills (e.g., text matching, paraphrasing, or memorization) (Roberts and Priest, 2006). Moreover, from the educational perspective, given that reading comprehension is a multi-component skill, it is ideal for comprehension questions to be able to identify students' performance in specific sub-skills, thus allowing teachers to provide tailored guidance (Francis et al., 2005).

However, creating a large and suitable set of questions for supporting narrative comprehension is both time-consuming and cognitively demanding. Some researchers have proposed developing models to automatically generate questions or QA-pairs that satisfy the need for a continuous supply of new questions (Kurdi et al., 2020; Yao et al., 2022), which can potentially enable large-scale development of AI-supported interactive platforms for the learning and assessment of reading comprehension skills (e.g., (Zhang et al., 2022)). However, existing datasets are not particularly suitable for training question generation (QG) models for educational purposes (Das et al., 2021). This is primarily because the datasets are not typically structured around the specific dimensions of reading comprehension sub-skills, nor do they provide sufficient information on what sub-skills are tested. Consequently, QG models built on these datasets only yield one single "comprehension" score without a more detailed breakdown of performance on comprehension sub-skills. This issue is compounded by the fact that many benchmarks rely on crowdsourced workers who may not have sufficient training or education domain knowledge needed to create valid questions in a consistent way.

To bridge the gap, we constructed FairytaleQA, an open-source dataset focusing on comprehension of narratives, targeting students from kindergarten to eighth grade. We focus on narrative comprehension for two reasons. First, narrative comprehension is a high-level comprehension skill strongly predictive of reading achievement (Lynch et al., 2008) and plays a central role in daily life as people frequently encounter narratives in different forms (Goldie, 2003). Second, narrative stories have a clear structure of specific elements and relations among these elements, and there are existing validated narrative comprehension frameworks around this structure, which provides a basis for developing the annotation schema for our dataset.

We employed education experts who generated 10,580 question-answer pairs based on a collection of 278 fairytale stories for young readers, following evidence-based narrative comprehension frameworks (Paris and Paris, 2003; Alonzo et al., 2009). Thereby, FairytaleQA contains questions that focus on several narrative elements and relations, increasing the validity and reliability of the assessment. In addition, FairytaleQA also contains both explicit questions that involve answers found explicitly in the text and implicit questions that require high-level summarization (Table 1), thus representing a relatively balanced assessment with questions of varying difficulty (Zucker et al., 2010; Raphael, 1986). Most importantly, our selection of annotators with education domain knowledge as well as the training and quality control process ensured that the aforementioned annotation protocol was consistently implemented. A subset of questions in our dataset has been validated with 120 pre-kindergarten and kindergarten students (IRB approved from the first author's institution), proving the questions' reliability and validity.

We show the utility of FairytaleQA through two benchmarking experiments. First, we used our data to train and evaluate state-of-the-art (SOTA) QA models and demonstrated that (1) FairytaleQA contains challenging phenomena for existing models, and (2) it can support finer-grained analysis on the different types of comprehension sub-skills, even for models trained on general QA datasets (NarrativeQA (Kočiskỳ et al., 2018)). We further calibrated model performances with human baseline, highlighting the most visible gap in models' reasoning capabilities on recognizing casual relationships and predicting event outcomes. Second, we used FairytaleQA to power question generation and showed that the QG model trained on ours was more capable of asking diverse questions and generating questions with higher quality.

2 Related Work

2.1 QA Datasets Focusing on Narratives

Despite a large number of datasets on reading comprehension, few focus on comprehension of narrative text. Table 2 reviews different narrative-related properties of existing popular QA datasets comparing with our proposed FairytaleQA dataset. Narra-

Dataset	Educ.	Narr.	Q. Type	A. Type	A. Source	Generation	Document Source
NarrativeQA	No	Yes	Open-ended	Natural	Free-form	Crowd-sourced	Movie Scripts, Literature (Full story or summary)
BookTest	No	Yes	Cloze	Mult. Choice	Entity/Span	Automated	Literature (Excerpt)
TellMeWhy	No	Yes	Open-ended	Natural	Free-form	Crowd-sourced	Short Fiction (ROCStories)
RACE	Yes	No	Open-ended	Mult. Choice	Free-form	Expert	(Partially) Literature (Short story or excerpt)
CLOTH	Yes	No	Cloze	Mult. Choice	Span	Expert	(Partially) Literature (Short story or excerpt)
FairytaleQA	Yes	Yes	Open-ended	Natural	Free-form & Span	Expert	Literature (Full story)

Table 2: Properties of existing datasets compared to FairytaleQA.

tiveQA (Kočiskỳ et al., 2018) is one of the representative datasets. It was generated by crowd workers who wrote QA pairs according to summaries of books or movie scripts, while the task takers are supposed to answer these questions based on their reading of original books or movie scripts. As such, this dataset is posited to evaluate a person's understanding of the underlying narrative, with a significant amount of event-related questions (Mou et al., 2021). However, NarrativeQA simply instructed crowd-sourced workers to generate questions as if they were to "test students" without using a detailed annotation protocol. It is questionable whether these workers actually had experiences in testing students, and the lack of protocol may have imposed too little control over the coverage of reading sub-skills.

BookTest (Bajgar et al., 2016) is an automatically constructed cloze-style QA dataset based on a collection of narrative texts retrieved from Project Gutenberg. The questions were generated by automatically removing a noun or entity in a sentence that has appeared in the preceding context. While cloze-style tests can be a valid instrument for assessing reading comprehension, their validity depends on the careful selection of words to be removed so that filling them in requires proper comprehension (Gellert and Elbro, 2013). It is unlikely that automatically constructed cloze tests would meet such standards.

Another dataset, TellMeWhy (Lal et al., 2021), aims to facilitate and assess understanding of *causal relationships*. This dataset contains "why" questions that are relatively challenging, given that they require additional information not directly provided in the text. However, TellMeWhy only addresses one narrative component type (i.e., causal relationship), whereas FairytaleQA provides seven evaluation components. Moreover, TellMeWhy was built upon ROCStories (Mostafazadeh et al., 2016) and thus only examine comprehension on incomplete story exerpts, which may have limited the dataset's ability to assess macro-level summarization and inference making.

2.2 QA Datasets for Reading Education

There are several benchmarks derived from sources for education purposes (e.g., exams or curricula). RACE (Lai et al., 2017) is a large-scale dataset consisting of comprehension questions from English exams for Chinese middle and high school students. RACE uses a mixture of narrative and informational paragraphs. These two genres require slightly different comprehension skills (Liebfreund, 2021), and students perform differently based on what genres of text they read (Denton et al., 2015). Mixing these two together in one dataset without annotating the specific genre of each story/question obscures the ability to offer a precise assessment. Moreover, RACE is in multiple-choice format, and paragraphs are usually shorter. These two characteristics may make the RACE dataset less challenging, and recent models have demonstrated close-to-human performance[2].

CLOTH (Xie et al., 2017) is a cloze-style dataset also collected from English exams with multiple choice fill-in-the-blank questions. CLOTH can be advantageous for educational QG as each question is labeled with the level of reasoning it involves. However, this dataset shares certain limitations inherent to multiple-choice formats (Klufa, 2015).

2.3 Non-QA Datasets for Narrative Comprehension

There are some datasets that are designed for assessing narrative comprehension skills but do not

[2] http://www.qizhexie.com/data/RACE_leaderboard.html

use QA as a form of evaluation. Several datasets, such as NovelChapters (Ladhak et al., 2020) and BookSum (Kryściński et al., 2021), evaluate models' comprehension through summarization tasks. However, there have been debates of whether comprehension can be assessed solely through summarization (Head et al., 1989), as summarization poses a high demand on writing that confounds the reading skills intended to be assessed. Two other recent datasets focus on singular specific elements in narratives. The LiSCU dataset (Brahman et al., 2021) targets readers' understanding of *characters*, and Sims et al. (2019) propose a dataset for detecting *events* in narratives. Given their focus on single narrative elements, these two datasets may not provide a comprehensive evaluation of narrative comprehension.

3 FairytaleQA

We developed the FairytaleQA dataset to address some of the limitations in existing benchmarks. Our dataset contains 10,580 QA pairs from 278 classic fairytale stories. In the remainder of this section, we report the dataset construction process and its key statistics.

3.1 Source Texts

The narrative texts utilized in the dataset are classic fairytales with clear narrative structures. We gathered the text from the Project Gutenberg website[3], using "*fairytale*" as the search term. Due to a large number of fairytales found, we used the most popular stories based on the number of downloads since these stories are presumably of higher quality.

To ensure the readability of the text, we made a small number of minor revisions to some obviously outdated vocabulary (e.g., changing "ere" to "before") and the unconventional use of punctuation (e.g., changing consecutive semi-colons to periods). For each story, we evaluated the reading difficulty level using the textstat[4] Python package, primarily based on sentence length, word length, and commonness of words. We excluded stories that are at 10th grade level or above.

These texts were broken down into small sections based on their semantic content by our annotators. The annotators were instructed to split the story into sections of 100-300 words that also contain meaningful content and are separated at natural story breaks. An initial annotator would split the story, and this would be reviewed by a cross-checking annotator. Most of the resulting sections were one natural paragraph of the original text. However, sometimes several paragraphs were combined (usually dialogue); and some exceptionally long paragraphs that contained more than one focal event were divided into multiple sections. On average, there are 15 sections per story, and each section has an average of 150 words (Table 4).

3.2 Schema for Question Annotation

Categorization via Narrative Elements or Relations FairytaleQA is intended to include QA pairs that capture the seven narrative elements/relations that are verified in prior educational research (Paris and Paris, 2003). Definitions of question types are shown below. Example questions for each type are in Appendix D.

- **Character** questions ask test takers to identify the character of the story or describe characteristics of characters.
- **Setting** questions ask about a place or time where/when story events take place and typically start with "*Where*" or "*When*".
- **Action** questions ask about characters' behaviors or information about that behavior.
- **Feeling** questions ask about the character's emotional status or reaction to certain events and are typically worded as "*How did/does/do . . . feel*".
- **Causal relationship** questions focus on two events that are causally related where the prior events causally lead to the latter event in the question. This type of questions usually begins with "*Why*" or "*What made/makes*".
- **Outcome resolution** questions ask for identifying outcome events that are causally led to by the prior event in the question. This type of questions are usually worded as "*What happened/happens/has happened...after...*".
- **Prediction** questions ask for the unknown outcome of a focal event, which is predictable based on the existing information in the text.

These labels are to ensure the presence of the variety of questions' sub-skills so that the models trained on this dataset can also generate the variety. The labels are not intended to aid the training of a model to classify questions. Some – but not all – of the labels may be determined by surface features. For example, *feeling* questions typically contain the words "feel" or "feels", while *action* questions

[3] https://www.gutenberg.org/
[4] https://pypi.org/project/textstat/

FairytaleQA Dataset	Train				Validation				Test			
	232 Books with 8548 QA-pairs				23 Books with 1025 QA-pairs				23 Books with 1007 QA-pairs			
	Mean	**S.D.**	**Min**	**Max**	**Mean**	**S.D.**	**Min**	**Max**	**Mean**	**S.D.**	**Min**	**Max**
# section per story	14.4	8.8	2	60	16.5	10.0	4	43	15.8	10.8	2	55
# tokens per story	2160.9	1375.9	228	7577	2441.8	1696.9	425	5865	2313.4	1369.6	332	6330
# tokens per section	149.6	64.8	12	447	147.8	56.7	33	298	145.8	58.6	24	290
# questions per story	36.8	28.9	5	161	44.5	29.5	13	100	43.7	28.8	12	107
# questions per section	2.8	2.440	0	18	2.9	2.3	0	16	3.0	2.4	0	15
# tokens per question	10.2	3.2	3	27	10.9	3.2	4	24	10.5	3.1	3	25
# tokens per answer	7.1	6.0	1	69	7.7	6.3	1	70	6.8	5.2	1	44

Table 3: Core statistics of the FairytaleQA dataset, which has 278 books and 10580 QA-pairs.

are more broad in their format.

Categorization via Source of Answers Orthogonal to the aforementioned question categories, questions in FairytaleQA are also categorized based on whether or not the answer source can be directly found in the text, namely explicit versus implicit questions. In general, explicit questions revolve around a specific story fact, and implicit questions require summarizing and making an inference based on information that is only implicit in the text. Using a combination of explicit and implicit questions yields an assessment with more balanced difficulty (Raphael, 1986; Zucker et al., 2010). In our data, explicit and implicit questions are defined as below (Examples in Appendix C):

- **Explicit** questions ask for answers that can be directly found in the stories. In other words, the source of answer are spans of text.
- **Implicit** questions ask for answers that cannot be directly found in the text. Answering the questions require either reformulating language or making inference. In other words, the answer source is "free-form", meaning that the answers can be any free-text, and there is no limit to where the answer comes from.

3.3 Annotation Process

Five annotators were involved in the annotation of QA pairs. All of these annotators have a B.A. degree in education, psychology, or cognitive science and have substantial experience in teaching and reading assessment. These annotators were supervised by three experts in literacy education.

Annotation Guidelines The annotators were instructed to imagine that they were creating questions to test elementary or middle school students in the process of reading a complete story. We required the annotators to generate only natural, open-ended questions , avoiding "yes-" or "no-" questions. We also instructed them to provide a diverse set of questions about 7 different narrative elements, and with both implicit and explicit questions. Each question in the dataset has a label on the narrative element/relation to be assessed and whether it is implicit or explicit.

We asked the annotators to also generate answers for each of their questions. We asked them to provide the shortest possible answers but did not restrict them to complete sentences or short phrases. For explicit questions, annotators extracted the shortest phrase from the text as the answer (i.e., span). For implicit questions, annotators provided at least two possible answers for each question (i.e., free-form). We also asked the annotators to label which section(s) the question and answer was from. We did not specify the number of questions per story to account for story length variability and to allow annotators to create meaningful questions rather than be forced to add unnecessary questions. However, we did ensure that the annotators broadly averaged 2-3 questions per *section* in order to guarantee dataset size.

Annotator Training and Cross-Checking All annotators received a two-week training in which each of them was familiarized with the coding template (described in the section below) and conducted practice coding on the same five stories. The practice QA pairs were then reviewed by the other annotators and the three experts, and discrepancies among annotators were discussed. At the end of the training session, the five annotators had a little disagreement with the questions generated by other coders. During the annotation process, the team met once every week to review and discuss each member's work. All QA pairs were cross-checked by two annotators, and 10% of the QA pairs were additionally checked by the expert supervisor. This process was to ensure that the questions focused on key information to the narrative and the answers to

	Mean	Min	Max	SD
Story Characteristics				
Sections / story	14.7	2	60	9.2
Tokens / story	2196.7	228	7577	1401.3
Tokens / section	149.1	12	447	63.6
Question Characteristics				
Tokens / question	10.3	3	27	3.3
Tokens / answer	7.2	1	69	6.1
Questions / story	38.1	5	161	29
Questions / section	2.9	0	18	2.4

Table 4: Various descriptive statistics for the length of stories and number of questions in the dataset.

Category	Count	Percentage (%)
Attributes		
character	1172	11.08
causal relationship	2940	27.79
action	3342	31.59
setting	630	5.95
feeling	1024	9.68
prediction	486	4.59
outcome resolution	986	9.32
Explicit vs Implicit		
explicit	7880	74.48
implicit	2700	25.52

Table 5: Breakdown of questions per category based on the schema in Section 3.2.

the questions were correct.

Agreement among Annotators The questions generated by the five coders showed a consistent pattern. All coders' questions have similar lengths (average length ranging from 8 to 10 words among the coders) and have similar readability levels (average readability between fourth to fifth grade among the coders). The distributions in narrative elements focused as well as implicit/explicit questions were also consistent. A detailed description of the distributions by coders is displayed in Appendix E. We chose not to use traditional inter-annotator agreement (IAA) metrics like Kappa coefficients because we explicitly asked the coders to generate questions and answers with variable language to aid QA and QG models based on this dataset. This language variability leads to inaccurate IAA metrics by traditional means (Amidei et al., 2018), leading to our decision.

Second Answer Annotation For the 46 stories used as the evaluation set, we annotate a second reference answer by asking an annotator to independently read the story and answer the questions generated by others. All questions were judged as answerable and thus answered by the second annotator. The second answers are used for both human QA performance estimation and for providing multiple references in automatic QA evaluation.

3.4 Statistics of FairytaleQA

We random split the FairytaleQA dataset into train/val/test splits with a QA ratio of roughly 8:1:1. Table 3 shows the detailed statistics of the FairytaleQA Dataset in train/val/test splits.

Overall, the resulting FairytaleQA dataset contained 10,580 questions from 278 fairytale stories. The description of story and question characteristics is presented in Table 4. In FairytaleQA, *action* and *causal relationship* questions are the two most common types, constituting 31.6% and 27.8%, respectively, of all questions. *Outcome resolution, character,* and *feeling* types each constitute about 10% of all questions. *Setting* and *prediction* questions are about 5% each. Our dataset contains about 75% explicit questions and 25% implicit questions (Table 5 for details).

Validation of FairytaleQA for Comprehension Assessment We validated the questions in FairytaleQA using established procedures in educational assessment development (Özdemir and Akyol, 2019) and have proven that our questions have high reliability and validity. Specifically, we sampled a small subset of the questions in our dataset (11 questions generated for one story) and tested them among 120 students in prekindergartens and kindergartens. This study was preapproved by the IRB in first author's institution. The Cronbach's coefficient alpha was 0.83 for the items in this story comprehension assessment; suggesting was high internal reliability. We also linked children's performance answering our questions to another validated language assessment (Martin and Brownell, 2011), and the correlation was strong 0.76 ($p<.001$), suggesting an excellent external validity.

4 Baseline Benchmark: Question Answering

In the following sections, we present a couple of baseline benchmarks on both the Question Answering (QA) task and the Question Generation (QG) task with FairytaleQA. We leveraged both pretrained neural models and models fine-tuned on different QA datasets, including NarrativeQA and our dataset, FairytaleQA. The baseline results show

Model	Validation / Test ROUGE-L F1
Pre-trained Models	
BERT	0.104 / 0.097
DistilBERT	0.097 / 0.082
BART	0.108 / 0.088
Fine-tuned Models	
BART fine-tuned on NarrativeQA	0.475 / 0.492
BART fine-tuned on FairytaleQA	0.533 / 0.536
Human‡	**0.651 / 0.644**

Table 6: Question Answering benchmarks on FairytaleQA validation and test splits. ‡Human results are obtained via cross-estimation between the two annotated answers, thus are underestimated. Still, they outperform models. We leave a full large-scale human study to future work.

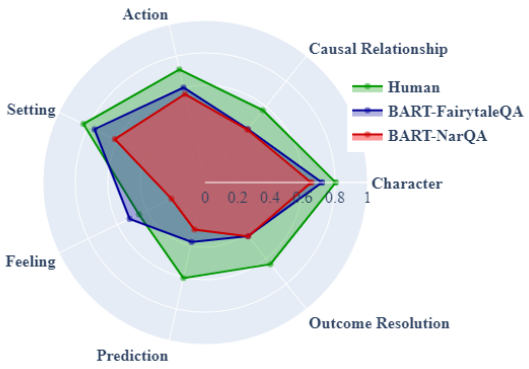

Figure 1: Decomposed QA results (Rouge-L) on 7 narrative elements on the validation split.

that our FairytaleQA demonstrates challenging problems to existing approaches, and those models fine-tuned on FairytaleQA can benefit from the annotations a lot to achieve significant performance improvement. We also report human performance by scoring one reference answer to the other.

4.1 Question Answering Task and Model

Question Answering (QA) is a straightforward task that our FairytaleQA dataset can contribute to. We leveraged the commonly-used Rouge-L F1 score for the evaluation of QA performances. For each QA instance, we compared the generated answer with each of the two ground-truth answers and took the higher Rouge-L F1 score.

4.2 Main Results

Here in Table 6, we show the QA performance of a few pretrained SOTA neural-model architectures: BERT (Devlin et al., 2018), BART (Lewis et al., 2019), and DistilBERT(Sanh et al., 2019). The quality of answers generated by these pre-trained models is on par with each other. Since BART outperformed other model architectures in the QA task of NarrativeQA (Mou et al., 2021), we decided to use BART as the backbone for our fine-tuned models.

We report the performance of fine-tuned BART models with the following settings: BART fine-tuned on NarrativeQA, which is the SOTA model reported in (Mou et al., 2021), and another BART model fine-tuned on FairytaleQA. We note that for the QA task, the model that was fine-tuned on FairytaleQA dataset performs much better than the model fine-tuned on NarrativeQA by at least 5%. Even the human performance is underestimated here because it is obtained via cross-estimation between two annotated answers, this result still leaves around 12% on both splits between human performance and the model fine-tuned with FairytaleQA, which demonstrates that the QA task is still a challenging problem for existing works on our FairytaleQA dataset. We leave a full large-scale human study for evaluating the accurate human performance to future work.

4.3 Analysis

Performance Decomposition Given that FairytaleQA has question type annotations on all the question-answer pairs, it supports the decomposition of performance on different types, thus resulting in a comprehensive picture of which reading skills the models lack the most.

Figure 1 presents the QA performance decomposition as a radar visualization. (The full results on both validation and test sets can be found in Table 10 in Appendix A). Compared to the model trained on NarrativeQA, our FairytaleQA led to the biggest improvement on dimensions of `Setting` and `Feeling` with more than 10% increase. The `Character` and `Prediction` dimensions were also improved by a large margin (7-8%). The large improvements in these dimensions suggested that despite the NarrativeQA dataset's overall focus on narrative comprehension, it might not include questions that sufficiently cover some of the fundamental elements, probably due to the lack of detailed annotating protocol and typical crowd workers' limited knowledge in reading assessment.

By comparison, on dimensions of `Action`, `Causal Relationship` and `Outcome Resolution`, our model fine-tuned on FairytaleQA resulted in smaller improvement compared

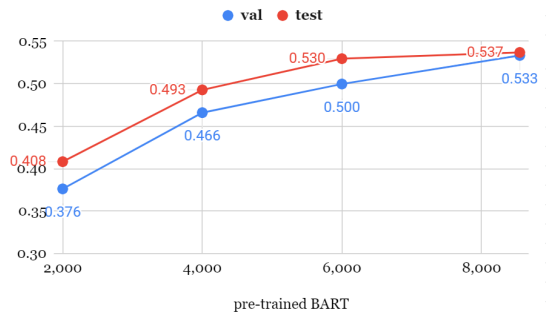

Figure 2: Learning curve of the QA model on FairytaleQA with varying size of training data.

to the model fine-tuned on NarrativeQA. This is likely due to the fact that most of the NarrativeQA questions are about event arguments and causal or temporal relations between events, as suggested by a human study (Mou et al., 2021).

Our performance decomposition also revealed substantial gaps between existing SOTA models and humans. Specifically, humans were 15-20% better on `Causal Relationship`, `Outcome Resolution` and `Prediction`. The model-human performance gaps on `Causal Relationship` and `Outcome Resolution` likely reflected the deficiency of current NLP models in understanding story plots, and the gap on `Prediction` might be due to the fact that this dimension asked the models to envision what would come next in the text, which required connecting commonsense knowledge with the content of the text. The model-human performance gaps on `Character` and `Setting` were also considerable, suggesting that the models' ability to understand these basic reading elements still has much room for improvement.

Finally, it was interesting that the model trained on our dataset outperformed humans on the `Feeling` dimension. This was likely because the answers to these `Feeling` questions were most explicitly described in the story. Therefore, it did not actually require reasoning of the character's mental states, but rather understanding which parts of the texts express the feelings. Another QA performance decomposition result based on explicit/implicit question types is provided in Appendix B.

Learning Curve Finally, we present the learning curve of the BART QA model on our FairytaleQA. Figure 2 plots the model performance on the validation set with different sizes of training data. The curve became flatter after training with 6,000 QA pairs in our dataset. This suggested that our dataset has a reasonably good size for fine-tuning a SOTA pre-trained model, and the performance gap between models and humans requires a more sophisticated reading model design rather than solely augmenting the training examples.

Model	Validation / Test ROUGE-L F1
BART fine-tuned on NarrativeQA	0.424 / 0.442
BART fine-tuned on FairytaleQA	**0.527 / 0.527**
BART fine-tuned on NarrativeQA and FairytaleQA	0.508 / 0.519

Table 7: Question Generation benchmarks on FairytaleQA-validation and test splits.

	Groundtruth	BART-NarQA	BART-FairytaleQA
Who	84	62	97
What	426	716	447
Why	287	144	304
How	178	59	129
Where	44	35	47
Other	6	9	1

Table 8: Distribution of question word in QG task for validation split by benchmark models.

5 Baseline Benchmark: Question Generation

5.1 Question Generation Task and Model

In terms of the QG performance on FairytaleQA, the task was to generate questions that correspond to the given answers and the context. This task has important empirical applications that in the future, models may help teachers to create questions in the educational settings.

Similar to the QA task, we fine-tuned a BART model to generate a question conditioned on each human-labeled answer and corresponding story section. The generated question is then evaluated with the corresponding ground-truth question. We used ROUGE-L F1 score as the evaluation metric. For this QG task, we compare the models fine-tuned on NarrativeQA, on FairytaleQA, and on both datasets.

5.2 Results and Analysis

Table 7 displays the QG results. The model fine-tuned on FairytaleQA demonstrated a clear advantage on Rouge-L over the model fine-tuned on NarrativeQA. It is worth noting that the model fine-tuned on both NarrativeQA and FairytaleQA performs worse than the model fine-tuned on FairytaleQA only; we would assume that NarrativeQA

Input story section: the wild people who dwell in the south-west are masters of many black arts. they often lure men of the middle kingdom to their country by promising them their daughters in marriage, but their promises are not to be trusted. once there was the son of a poor family, who agreed to labor for three years for one of the wild men in order to become his son-in-law.
Input Answer 1: The son of a poor family.
Ground-truth Question Who agreed to labor for three years for one of the wild men in order to become his son-in-law?
Outputs ***BART-NarQA:*** What was the son of a poor family? ***BART-FairytaleQA:*** Who agreed to labor for one of the wild men in order become his son-in law?
Input Answer 2: The wild people.
Ground-truth Question Who dwelled in the south-west and were masters of many black arts?
Outputs ***BART-NarQA:*** What dwells in the south-west? ***BART-FairytaleQA:*** Who dwell in the south-west are masters of many black arts?

Table 9: Qualitative analysis of QG models fine-tuned on NarQA or FairytaleQA dataset.

dataset introduces noises in question type distribution or semantics during the training process.

Further analysis (Table 8) examined the distribution of generated question types according to the beginning word of a question (wh- words). The questions generated by FairytaleQA more closely resembled the pattern of the ground-truth questions, suggesting that our dataset was able to improve the model's ability to mimic the education experts' strategy of asking questions that assess the seven elements of reading comprehension.

This result is further supported by qualitative analysis (as seen in examples in Table 9). Compared to the QG model trained with FairytaleQA, the baseline model trained with NarrativeQA dataset tended to generate vague questions that did not build upon specific contextual evidence within the narratives. These kinds of vague questions may not be suitable in educational settings, as improving students' skills to find text evidence to support their comprehension is a crucial aspect of reading education. The disparity between the two models might be attributed to how the QA-pairs were constructed in these two datasets: while NarrativeQA was constructed by crowd workers who only read the abstract of the stories, FairytaleQA required annotators to read the complete story before developing QA-pairs. As such, it is not surprising that models trained on FairytaleQA dataset could generate questions that are more closely related to the contextual evidence within the original text. In addition, we also observed that the model trained on NarrativeQA tended to generate questions with seemingly more correct grammar but were factually inaccurate (Table 12 Appendix C).

6 Conclusion and Future work

In summary, we constructed a large-scale dataset, FairytaleQA, for the context of children's narrative comprehension. The dataset was generated through a rigorous labeling process with educational domain experts. This dataset has been helpful to support preliminary works on QG tasks (Yao et al., 2022; Zhao et al., 2022) and already enabled possibilities for new downstream AI-for-Education applications (Zhang et al., 2022; Xu et al., 2021).

However, we acknowledge our work also has limitations that require future works to continue the exploration. As aforementioned, the human performance results for QA task are underestimated because they are obtained via cross-estimation between the two annotated answers. One possibility for future work is to conduct a large-scale human annotation to collect more answers per each question and then leverage the massive annotated answers to better establish a human performance evaluation. Another avenue of future work is to leverage our dataset to detect and remediate social stereotypes and biases represented in story narratives – the bias analysis in the children storybook corpus has been an underexplored research topic for the ML community, but it has profound societal impacts on the soceity. Through such analysis on our dataset, we may be able to answer "how do social stereotype and bias come into a child's mind?"

In sum, there are many new research and application opportunities enabled by our FairytaleQA dataset, and we welcome researchers from both NLP and education communities to join our effort to continue this endeavor.

Acknowledgements

We thank Schmidt Futures for providing funding for the development of the `FairytaleQA` dataset. This work is also supported by the National Science Foundation (Grant No. 1906321 and 2115382).

References

Julie Alonzo, Deni Basaraba, Gerald Tindal, and Ronald S Carriveau. 2009. They read, but how well do they understand? an empirical look at the nuances of measuring reading comprehension. *Assessment for Effective Intervention*, 35(1):34–44.

Jacopo Amidei, Paul Piwek, and Alistair Willis. 2018. Rethinking the agreement in human evaluation tasks. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3318–3329, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ondrej Bajgar, Rudolf Kadlec, and Jan Kleindienst. 2016. Embracing data abundance: Booktest dataset for reading comprehension. *arXiv preprint arXiv:1610.00956*.

Faeze Brahman, Meng Huang, Oyvind Tafjord, Chao Zhao, Mrinmaya Sachan, and Snigdha Chaturvedi. 2021. " let your characters tell their story": A dataset for character-centric narrative understanding. *arXiv preprint arXiv:2109.05438*.

Bidyut Das, Mukta Majumder, Santanu Phadikar, and Arif Ahmed Sekh. 2021. Automatic question generation and answer assessment: a survey. *Research and Practice in Technology Enhanced Learning*, 16(1):1–15.

Carolyn A Denton, Mischa Enos, Mary J York, David J Francis, Marcia A Barnes, Paulina A Kulesz, Jack M Fletcher, and Suzanne Carter. 2015. Text-processing differences in adolescent adequate and poor comprehenders reading accessible and challenging narrative and informational text. *Reading Research Quarterly*, 50(4):393–416.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

David J Francis, Jack M Fletcher, Hugh W Catts, and J Bruce Tomblin. 2005. Dimensions affecting the assessment of reading comprehension. In *Children's reading comprehension and assessment*, pages 387–412. Routledge.

Anna S Gellert and Carsten Elbro. 2013. Cloze tests may be quick, but are they dirty? development and preliminary validation of a cloze test of reading comprehension. *Journal of Psychoeducational Assessment*, 31(1):16–28.

Peter Goldie. 2003. One's remembered past: Narrative thinking, emotion, and the external perspective. *Philosophical Papers*, 32(3):301–319.

Martha H Head, John E Readence, and Ray R Buss. 1989. An examination of summary writing as a measure of reading comprehension. *Literacy Research and Instruction*, 28(4):1–11.

Young-Suk Grace Kim. 2017. Why the simple view of reading is not simplistic: Unpacking component skills of reading using a direct and indirect effect model of reading (dier). *Scientific Studies of Reading*, 21(4):310–333.

Jindrich Klufa. 2015. Multiple choice question tests–advantages and disadvantages. In *3rd International Conference on Education and Modern Educational Technologies (EMET)*, pages 39–42.

Tomáš Kočiskỳ, 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.

Wojciech Kryściński, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. 2021. Booksum: A collection of datasets for long-form narrative summarization. *arXiv preprint arXiv:2105.08209*.

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.

Faisal Ladhak, Bryan Li, Yaser Al-Onaizan, and Kathleen McKeown. 2020. Exploring content selection in summarization of novel chapters. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5043–5054.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*.

Yash Kumar Lal, Nathanael Chambers, Raymond Mooney, and Niranjan Balasubramanian. 2021. Tellmewhy: A dataset for answering why-questions in narratives. *arXiv preprint arXiv:2106.06132*.

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. *arXiv preprint arXiv:1910.13461*.

Meghan D Liebfreund. 2021. Cognitive and motivational predictors of narrative and informational text comprehension. *Reading Psychology*, 42(2):177–196.

Julie S Lynch, Paul Van Den Broek, Kathleen E Kremer, Panayiota Kendeou, Mary Jane White, and Elizabeth P Lorch. 2008. The development of narrative comprehension and its relation to other early reading skills. *Reading Psychology*, 29(4):327–365.

Nancy A Martin and Rick Brownell. 2011. *Expressive one-word picture vocabulary test-4 (EOWPVT-4)*. Academic Therapy Publications.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and evaluation framework for deeper understanding of commonsense stories. *arXiv preprint arXiv:1604.01696*.

Xiangyang Mou, Chenghao Yang, Mo Yu, Bingsheng Yao, Xiaoxiao Guo, Saloni Potdar, and Hui Su. 2021. Narrative question answering with cutting-edge open-domain qa techniques: A comprehensive study. *arXiv preprint arXiv:2106.03826*.

Ezgi Çetinkaya Özdemir and Hayati Akyol. 2019. The development of a reading comprehension test. *Universal Journal of Educational Research*, 7(2):563–570.

Alison H Paris and Scott G Paris. 2003. Assessing narrative comprehension in young children. *Reading Research Quarterly*, 38(1):36–76.

Taffy E Raphael. 1986. Teaching question answer relationships, revisited. *The reading teacher*, 39(6):516–522.

Paula Roberts and Helena Priest. 2006. Reliability and validity in research. *Nursing standard*, 20(44):41–46.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Matthew Sims, Jong Ho Park, and David Bamman. 2019. Literary event detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3623–3634.

Qizhe Xie, Guokun Lai, Zihang Dai, and Eduard Hovy. 2017. Large-scale cloze test dataset created by teachers. *arXiv preprint arXiv:1711.03225*.

Ying Xu, Dakuo Wang, Penelope Collins, Hyelim Lee, and Mark Warschauer. 2021. Same benefits, different communication patterns: Comparing children's reading with a conversational agent vs. a human partner. *Computers & Education*, 161:104059.

Bingsheng Yao, Dakuo Wang, Tongshuang Wu, Zheng Zhang, Toby Jia-Jun 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. Association for Computational Linguistics.

Zheng Zhang, Ying Xu, Yanhao Wang, Bingsheng Yao, Daniel Ritchie, Tongshuang Wu, Mo Yu, Dakuo Wang, and Toby Jia-Jun Li. 2022. Storybuddy: A human-ai collaborative agent for parent-child interactive storytelling with flexible parent involvement. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22. ACM.

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. Association for Computational Linguistics.

Tricia A Zucker, Laura M Justice, Shayne B Piasta, and Joan N Kaderavek. 2010. Preschool teachers' literal and inferential questions and children's responses during whole-class shared reading. *Early Childhood Research Quarterly*, 25(1):65–83.

A Decomposed QA results on 7 narrative elements for val/test splits

	BART-NarQA	BART-FairytaleQA	Human
Validation			
Character	0.65	0.720	0.804
Causal Relationship	0.417	0.422	0.570
Action	0.560	0.601	0.716
Setting	0.618	0.757	0.833
Feeling	0.231	0.517	0.453
Prediction	0.298	0.377	0.605
Outcome Resolution	0.425	0.423	0.645
Test			
Character	0.691	0.757	0.864
Causal Relationship	0.447	0.432	0.589
Action	0.559	0.608	0.710
Setting	0.683	0.696	0.755
Feeling	0.301	0.508	0.533
Prediction	0.275	0.300	0.366
Outcome Resolution	0.409	0.486	0.574

Table 10: Decomposed QA results on 7 narrative elements.

Table 10 shows the full decomposed QA results on 7 narrative elements for both validation and test splits, in terms of BART fine-tuned on NarrativeQA, BART fine-tuned on FairytaleQA, and human performance for the experts created ground-truth QA-pairs.

B Decomposed QA results on explicit/implicit question types for val/test splits

Model	Validation / Test ROUGE-L F1	
	Implicit	Explicit
BART-NarQA	0.280/0.278	0.548/0.563
BART-FairytaleQA	0.304/0.286	0.619/0.620
Human	0.363/0.330	0.760/0.750

Table 11: Decomposed QA results on implicit/explicit types.

We provided another QA performance decomposition based on explicit/implicit question types in Table 11. We noticed that the implicit questions are much more difficult for both humans and models to answer, which only achieves roughly half the performance compared with explicit questions. This result is consistent with our expectation, where answers to explicit questions can be directly found in the content while implicit questions require high-level summarization and inference.

Input story section: you see from this that the sparrow was a truthful bird, and the old woman ought to have been willing to forgive her at once when she asked her pardon so nicely. but not so.the old woman had never loved the sparrow, and had often quarreled with her husband for keeping what she called a dirty bird about the house, saying that it only made extra work for her. now she was only too delighted to have some cause of complaint against the pet. she scolded and even cursed the poor little bird for her bad behavior, and not content with using these harsh, unfeeling words, in a fit of rage she seized the sparrow-who all this time had spread out her wings and bowed her head before the old woman, to show how sorry she was-and fetched the scissors and cut off the poor little bird's tongue.
Input Answer: Cut off the poor little bird's tongue.
Ground-truth Question What did the woman do to punish the bird?
Outputs ***BART-NarQA:*** What did the old woman do in her rage? ***BART-FairytaleQA:*** What did the old woman do after she seized her sparrow?
Input story section: "do not be sparing of the silver pieces in your pocket!" she cried after him as he went off.he went to the village, attended to everything, and came back. the woman tore the cloth apart, made a coat of it and put it on. no sooner had they walked a few miles before they could see a red cloud rising up in the south, like a flying bird."that is my mother," said the woman.in a moment the cloud was overhead. then the woman took the black tea-cups and threw them at it. seven she threw and seven fell to earth again. and then they could hear the mother in the cloud weeping and scolding, and thereupon the cloud disappeared.they went on for about four hours. then they heard a sound like the noise of silk being torn, and could see a cloud as black as ink, which was rushing up against the wind."alas, that is my father!" said the woman. "this is a matter of life and death, for he will not let us be! because of my love for you i will now have to disobey the holiest of laws!"
Input Answer: Took the black tea-cups and threw them at it.
Ground-truth Question What did the wife do when she saw her mother?
Outputs ***BART-NarQA:*** What did the woman do to try and kill her father? ***BART-FairytaleQA:*** What did the woman do after she saw her mother?

Table 12: Question Generation examples with event-related input answers by benchmark models.

C QG examples by benchmark models on event-based answers

Table 12 shows two QG examples that have an input of event-related ground-truth answers. We may notice that BART fine-tuned on NarrativeQA is able to generate questions that seem to be in a cor-

Category	Example QA Pair
Character	Q: *How did the man's daughter look?* A: *beautiful* Q: *Who were the brother and sister living with after their mom died?* A: *their stepmother*
Setting	Q: *Where did the man and his wife and two girls live?* A: *near the forest*
Action	Q: *What did the cook do after she opened the hamper?* A: *unpacked the vegetables* Q: *How did Johnny Town-Mouse and his friends treat Timmy Willie when they met him?* A: *Johnny Town-Mouse and his friends treat Timmy Willie poorly.*
Causal relationship	Q: *Why did the two mice come tumbling in, squeaking, and laughing?* A: *They were being chased by the cat.*
Outcome resolution	Q: *What happened to Timmy after he got in the hamper?* A: *The hamper takes him to the garden.*
Feeling	Q: *How did the princess feel in her new home?* A: *happy*
Prediction	Q: *How will the other animals treat the duckling?* A: *The other animals will look down on the duckling.*
Explicit	Q: *How did the girl feel when she saw the old woman's teeth?* A: *terrified* Context: *...but she had such great teeth that the girl was terrified...* Q: *What happened when the door of the stove was opened?* A: *The flames darted out of its mouth.* Context: *...when the door of the stove was opened, the flames darted out of its mouth. This is customary with all stoves...*
Implicit	Q: *What happened when the prince broke open one of the crow's eggs?* A1: *The prince found a beautiful palace inside.* A2: *There was a beautiful palace inside.* A3: *A little palace was inside and it grew until it covered as much ground as seven large barns.* Context: *The Swan Maiden lit in a great wide field, and there she told the prince to break open one of the crow's eggs. The prince did as she bade him, and what should he find but the most beautiful little palace, all of pure gold and silver. He set the palace on the ground, and it grew and grew and grew until it covered as much ground as seven large barns.*

Table 13: Example QA-pairs of FairytaleQA. We show one QA-pair for each narrative element as well as implicit and explicit.

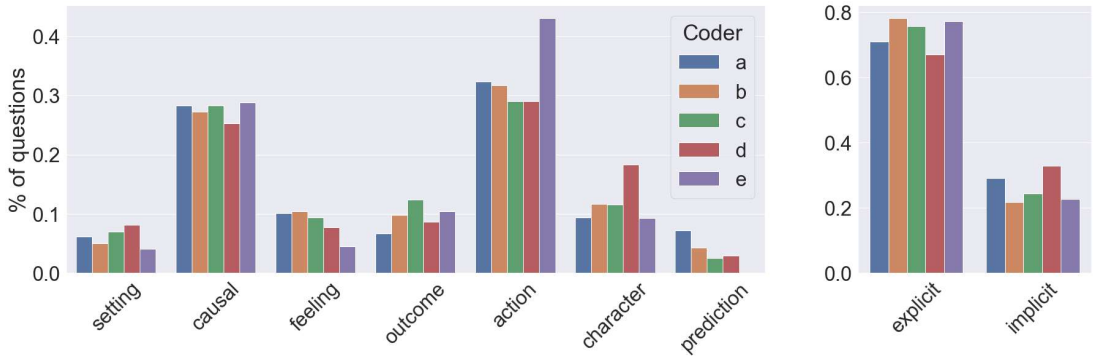

Figure 3: Percent of each question type by coder.

rect format but suffer from fact error, while BART fine-tuned on FairytaleQA is able to generate questions that are very similar to ground-truth questions and are semantically correct. Since the crowd workers only read the abstracts to create QA-pairs in NarrativeQA, in comparison, we ask our coders to read the complete story. This may lead to an issue with models fine-tuned on NarrativeQA where the evidence of the answer in the original text content is not detailed and obvious enough for QA-pairs in NarrativeQA so that the QG model fine-tuned on NarrativeQA is not ad good as models fine-tuned

on FairytaleQA in locating evidence.

D Example questions by category in FairytaleQA

Table 13 shows example QA-pairs for different annotations in FairytaleQA dataset. There is one example QA-pair for each narrative element as well as for implicit and explicit.

E Fine-tuning Parameters

For the QA task, we keep the following fine-tuning parameters consistent over different datasets: *learning rate* = $5e^{-6}$; *batch size* = 1; *epoch* = 1.
For the QG task, we keep the following fine-tuning parameters consistent over different datasets: *learning rate* = $5e^{-6}$; *batch size* = 1; *epoch* = 3.

F Proportion of Each Question Type

Figure 3 shows the percentage of each question type by coder.