StorySparkQA: Expert-Annotated QA Pairs with Real-World Knowledge for Children's Story-based Learning

Anonymous ACL submission

Abstract

Interactive storytelling between parents and children is a common activity in the real world, in which parents expect to teach children both language skills and real-world knowledge beyond the story narratives. While increasing AI-assisted storytelling systems have been developed and used in children's story-based interaction and learning scenarios, existing systems often fall short of generating real-world knowledge infused conversation to meet parents' practical expectation of interactive storytelling, with the foremost reason of existing question-answering (QA) datasets these systems build on focusing mainly on the knowledge answerable within the story content. To bridge this gap, we designed an annotation framework empowered by real-world knowledge graph to facilitate experts' annotations while collecting their mental procedures. Further, we leveraged this annotation framework to build StorySparkQA, a dataset of $5,868$ expert-annotated QA pairs with real-world knowledge beyond story context. A comprehensive benchmarking experiment, including both automated and human expert evaluation within various QA pair generation (QAG) settings, demonstrates the usability of our StorySparkQA on the story-based knowledgeable QAG task. Worth mentioning that a traditional compact model fine-tuned on StorySparkQA can reliably outperform robust LLMs. This further highlights the complexity of such real-world tasks.

1 Introduction

Interactive storytelling is a common parent-child activity, where parents often sit together with preschool children, read storybooks, and proactively engage in question-answering (QA) conversations with them (Wright, 1995; Isbell et al., 2004). Typically, such guided conversations are based on but beyond the story narratives (Kotaman, 2013), with parents' expectations of guiding children to learn real-world knowledge and improving their historical, cultural and emotional awareness (Sun et al., 2024). This story-based immersive interaction has been proven effective in better supporting preschoolers' knowledge learning (Zhang et al., 2024), enhancing their reading comprehension capabilities (Xu et al., 2021), etc.

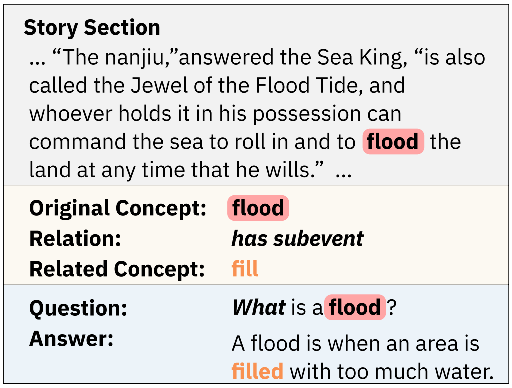

Figure 1: An example of StorySparkQA dataset. In each story section, educational experts select a concept word, link it to a desired external real-world knowledge, and write an appropriate QA pair.

Nevertheless, parents often struggle with appropriately conducting such interactive storytelling with children because of multi-facet difficulties (Golinkoff et al., 2019; Sun et al., 2024). Specifically, such interactive storytelling needs parents to identify the knowledge concept of interest during storytelling, formulate the real-world knowledge piece they want to teach in mind (**"what to ask"**), then ask an engaging question (**"how to ask"**) to children at the appropriate time (**"when to ask"**). Yet, most parents lack the necessary educational expertise and language skills to guide such educational conversations (Golinkoff et al., 2019; Sun et al., 2024). Also, parents in contemporary society often hardly maintain high concentration to accompany their children due to the need to deal with other work and family chores at the same time (Zhang et al., 2022; Sun et al., 2024).

Recently, AI-assisted storytelling systems (e.g.

StoryBuddy (Zhang et al., 2022), TaleMate (Vargas-Diaz et al., 2023), MatheMyths(Zhang et al., 2024)), backed by advanced language models that can drive the natural conversation with humans, have demonstrated effectiveness in children's storytelling scenarios (Dietz et al., 2021). Nevertheless, existing AI-assisted storytelling systems are not without limitations. Particularly, building on top of data resources with mostly extractive QA pairs (e.g., FairytaleQA (Xu et al., 2022)) – where the answers can be found directly in the story narrative – these systems fall short at helping parents teach real-world knowledge beyond the story narrative (Yao et al., 2021), which actually are one main expectation of parents (Sun et al., 2024).

We believe a promising approach to bridge this gap is to effectively and exhaustively collect education experts' knowledge, including their step-by-step thinking process as well as the appropriate QA pairs as final artifacts, nevertheless, no such data resources exist to the best of our knowledge in children's education domain. Further, the collection of such data resources requires annotators to recall a comprehensive and systematical external knowledge range for a given story text, which is challenging even for education experts (Berry et al., 2016). As a result, this work aims to facilitate experts' large-coverage knowledge collection and data annotation, and build an expert-labeled, large-scale QA dataset to support story-based educational QA generation with tri-fold contributions:

- We designed an annotation framework empowered by ConceptNet(Speer et al., 2017), a knowledge graph (KG) of structured real-world knowledge, to facilitate education experts creating appropriate story-based educational QA pairs, while collecting experts' mental procedures during data annotation.

- Based on the proposed annotation framework, we build `StorySparkQA`[1], an expert-labeled QA dataset consisting of $5,868$ story-based QA pairs infused with real-world knowledge.

- We demonstrate the utility of our `StorySparkQA` on the QA pair generation (QAG) task, benchmarked with a set of popular language models (fine-tuned T5-Large (Raffel et al., 2020), zero-shot, few-shot, and Chain-of-Thought with GPT-4 (OpenAI, 2023), Llama 2 (Touvron et al., 2023), etc[2]), through both automated evaluation and human expert evaluations.

`StorySparkQA` can benefit different research aspects in children's education domain, particularly in better understanding domain experts' thinking process, and training models to generate story-based QA pairs infused with real-world knowledge, with the ultimate goal of broadening children's knowledge scope beyond story narratives that parents expect. In addition, we believe our annotation framework possesses the potential to be generalized in analogous real-world domain-specific tasks requiring structured external knowledge (Vrandečić and Krötzsch, 2014; Lehmann et al., 2015), such that clinicians use structured guidelines and knowledge for diagnosing (ElSayed et al., 2023; American Diabetes Association, 2011).

2 Related Work

2.1 Children Education and Real-World Knowledge Resources

Existing datasets in the education domain (e.g., StoryQA (Zhao et al., 2023), FAIRYTALEQA (Xu et al., 2022), and EduQG (Hadifar et al., 2023)) mostly comprise QA-pairs grounded in the story, lacking real-world knowledge beyond the story. We present key properties of related children education datasets in Table 4. On the other hand, general-purpose datasets like CommonsenseQA (Talmor et al., 2018) and SciQA (Auer et al., 2023) integrate crowd-sourced commonsense with narratives, but lack educational appropriateness aligned with children's knowledge level.

Many popular real-world knowledge resources, such as ATOMIC (Sap et al., 2019) and Wikidata (Vrandečić and Krötzsch, 2014), are too complicated for children's knowledge level. A more appropriate option is ConceptNet (Speer et al., 2017), a very large-scale knowledge graph for real-world concepts and relations stored in triplets: ($concept_1$, $relation$, $concept_2$). The simplicity of triple representations makes ConceptNet suitable for children education, as demonstrated in prior literature (Xu et al., 2020), thus, our work also leverages ConceptNet to support experts' annotation process.

[1]We will release our dataset and code upon acceptance.

[2]We also experiment with GPT-3.5, Flan-T5-XXL (Chung et al., 2022), Alpaca (Taori et al., 2023) and Mistral-7B (Jiang et al., 2023) and report the results in Appendix A.5.

2.2 QA pair Annotation Frameworks

Some existing annotation frameworks (such as Potato (Pei et al., 2022) and Piaf (Keraron et al., 2020)) mostly focus on facilitating extractive QA pairs grounded in the text, that is, providing source texts and allowing annotators to highlight a span of text as an answer to a question. Some others (such as (Zhao et al., 2023)) support free-form input, that is, that is, allowing annotators to type in answers in their own words through the data collection user interface. In either type, existing annotation frameworks can't support story-based external knowledge collection and story data annotation effectively, in which annotators are required to recall comprehensive and systematical real-world knowledge for a given story text. Our study bridges this gap by proposing an external knowledge-empowered annotation framework.

2.3 QA Pair Generation (QAG)

Fine-tuning traditional pre-trained language models (e.g., BERT (Devlin et al., 2019) and GPT) on QAG datasets for end-to-end generation was a prevalent approach, but such methodology heavily depends on the training data quality and lack control of generated content, which is inappropriate for the children education domain. Existing works also attempted to design multi-step generation pipelines, which offers better control of the generated content.

The recent advancement in large language models (LLMs), such as GPT-3.5, GPT-4 (OpenAI, 2023), and Llama 2 (Touvron et al., 2023), supports free-form natural language input and output without the need for tuning model parameters. Also, many prompting strategies were developed to further enhance models' task-solving and domain-adaptation capabilities, including few-shot in-context learning (i.e., add a few examples in input) (Brown et al., 2020), Chain-of-Thought (i.e., ask models to think "step-by-step") (Wei et al., 2022), etc. However, to what extent these disparate prompting and modeling strategies are effective in the QAG task for knowledge beyond the story content remains under-explored, and this work attempts to step forward through the comprehensive evaluation in Section 5.

3 Expert Annotation Framework

To facilitate a better understanding of education experts' thinking process during the data annotation process for story-related QA pairs with knowledge

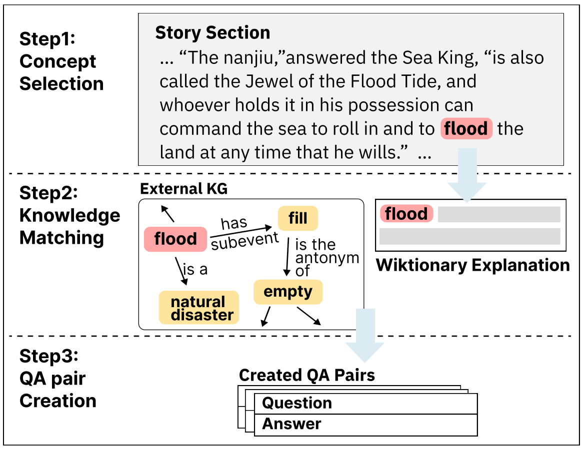

Figure 2: Workflow of the experts' annotation process. Experts need to select a concept first, then match it with the most suitable knowledge, and finally create a QA pair based on the selected knowledge.

beyond the story content, we proposed a three-step QA pair annotation framework with interactive user interfaces (UI). Particularly, considering the challenges facing annotators in recalling the comprehensive and systematical external knowledge for a given story text (Berry et al., 2016), our framework incorporates ConceptNet, a large-scale real-world Knowledge Graph, to support experts' large-coverage knowledge collection. The workflow of our annotation framework is shown in Figure 2.

Step 1. Concept Selection In this step, experts identify an educationally appropriate concept from the story content. We develop a collection of heuristics to filter candidate concepts that are tier 1 or tier 2 [3] vocabulary and a concrete noun, verb, or adjective. First, we leverage the spaCy (Honnibal and Montani, 2017) English model to filter auxiliary words and punctuation [4] from the original story text. Then, we use AllenNLP's (Gardner et al., 2017) semantic role labeling tool to tag the latent structure of each sentence in the story context. This process identifies and retains key elements represented by semantic roles, including agents, goals, and results, which are subsequently treated as potential candidate concepts. We design the UI, shown in Figure 5, to display one story section and allow experts to select highlighted candidate concepts in grey.

Step 2. Knowledge Matching This step allows experts to select real-world knowledge based on the

[3] Tier 1 words are common and basic words. Tier 2 contains high-frequency words of various domains (Beck et al., 2013).

[4] tagged by 'auxiliary', 'adposition', 'determiner', 'particle', 'punctuation', 'symbol', and 'other'

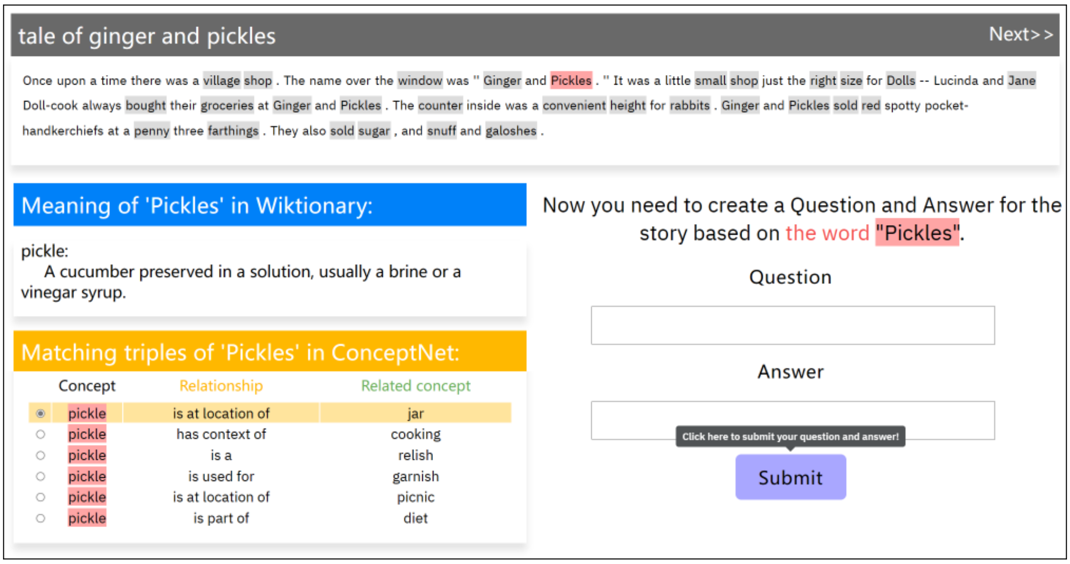

Figure 3: The user interface to facilitate our annotation task. The words highlighted in grey are candidate concepts. The blue block shows the Wiktionary explanation, and the yellow block lists our recommended triples.

concept selected previously. Inspired by Xu et al. (2020)'s work of combining and filtering knowledge from Wiktionary [5] and ConceptNet (Speer et al., 2017) for commonsense question answering, we implement a knowledge matching module that can retrieve and rank external knowledge associated with each concept selected by the experts.

Specifically, once experts select a candidate concept, our knowledge matching module (1) retrieves a list of real-world knowledge triples, with the format of $(source\ concept,\ relation,\ target\ concept)$ from ConceptNet; (2) filters out weak relations in ConceptNet (complete relation list in Appendix A.2), and (3) rank knowledge triples by concatenating concepts and relationships, and calculating the average similarity between every other triple with the Term Frequency-Inverse Document Frequency (TF-IDF).

We rank all retrieved triples with $1 - \overline{s} + w$, where $\overline{s}$ denotes the similarity score and w denotes the weight of a triple provided by ConceptNet, reflecting the combined influence and credibility of the triple by summing up the weights coming from all the sources that support it. The top six ranked triples are shown to annotators to balance between providing a sufficient selection and avoiding excessive distractions during the annotation task. We also retrieve the explanation for expert-selected concepts from Wiktionary to better facilitate experts' annotations. The UI is shown in Figure 6.

[5] https://www.wiktionary.org/

Step 3. QA pair Annotation We develop the third UI (Figure 3), enabling annotators to create a QA pair based on the triple they selected in step 2. In this step, experts are instructed to incorporate one concept in the question or answer and include the relation from the triple in the resulting QA pair.

4 StorySparkQA

StorySparkQA aims to facilitate parents' storytelling process with appropriate real-world knowledge: **practical, factual, everyday information that helps preschoolers understand the world around them.** Our dataset consists of $5,868$ QA pairs annotated by children education experts leveraging our designed annotation framework. We present the core statistics of StorySparkQA in Table 1 and show one example in Figure 1.

4.1 Source Narrative

Among the existing story-based datasets for children's education, FAIRYTALEQA (Xu et al., 2022) comprises 278 classic fairytale stories of various origins, and all the stories have been evaluated as suitable for 10^{th}-grade children and younger. The original stories were parsed by education experts into shorter sections of around 150 words, which leads the FAIRYTALEQA dataset to a unique and high-quality text corpus for children's reading comprehension. As a result, we take the story sections from FAIRYTALEQA as the source text for our StorySparkQA dataset.

StorySparkQA	Train 232 books with 4, 300 QA pairs				Validation 23 books with 769 QA pairs				Test 23 books with 799 QA pairs			
	Mean	**St.D**	**Min**	**Max**	**Mean**	**St.D**	**Min**	**Max**	**Mean**	**St.D**	**Min**	**Max**
# sections / 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 / section	149.6	64.8	12	447	147.8	56.7	33	298	145.8	58.6	24	290
# questions / story	18.5	14.5	2	126	33.4	22.1	4	115	34.7	21.1	8	90
# questions / section	1.3	0.6	1	9	2.1	0.3	2	3	2.1	0.3	2	3
# tokens / question	5.2	2.0	3	19	5.9	1.6	3	13	6.0	1.7	3	13
# tokens / answer	5.4	3.7	1	20	3.8	2.3	1	12	3.8	2.3	1	12

Table 1: Core statistics of our StorySparkQA dataset, which has 278 books and 5, 868 QA pairs.

4.2 Annotation Process

Following our annotation framework, we recruit 11 education experts for the annotation task. The education experts all have a minimum of 3 years of practical experience (e.g., kindergarten teachers) in learning science and possess relevant educational backgrounds. For each story section, experts are asked to first identify a concept from the story. The selected concepts should be considered the most beneficial for children's education from the text by educational experts. The experts then proceed to select a real-world knowledge triple and create a QA pair based on the selected triple. In this process, experts are asked to take into account children's cognitive and knowledge levels and write QA pairs that are most appropriate for 3-6-year-olds.

4.2.1 Cross-Validation

To ensure the quality and consistency of annotated QA pairs among annotators, as well as to evaluate agreement in selecting triples and creating QA pairs between annotators, we designed additional cross-validation procedures with corresponding UIs. We randomly selected 50 QA pairs in both the test and validation split (100 QA pairs in total) and two annotators were asked to cross-validate each other's annotation (denoted by $annotator_A$ and $annotator_B$, accordingly):

1. Shown in Figure 7, $annotator_A$ is provided with the story section and the concept selected by $annotator_B$. For each selected concept, $annotator_A$ is asked to rank the top 3 triples from the same recommended triple list given to $annotator_B$, verifying the triple selection agreement between annotators (Figure 8).

2. In the next step, $annotator_A$ is asked to create an QA pair based on the word and triple selected by $annotator_B$, evaluating the similarity of QA pairs between annotators given the identical triple (Figure 9).

3. After submitting the QA pair in Step 2, $annotator_A$ is provided with the question created by $annotator_B$ based on the same triple, and $annotator_A$ is asked to write an answer to the question to cross-validate the question-answering agreement (Figure 10).

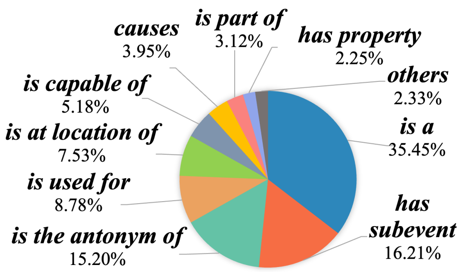

Figure 4: Distribution of real-world knowledge relations annotated by experts in the StorySparkQA dataset

Of the 100 randomly selected sections in the validation and test splits, 86% of the triples that appear in the top-3 list are selected by both annotators and 56% of the triples are ranked top by the validator, indicating a very high consistency between experts for triple selection.

In addition, we evaluate the similarity of the concatenated QA pairs created by each of the annotators based on the same triple with Rouge-L (Lin, 2004) and SBERT (Reimers and Gurevych, 2019) scores. The Rouge-L F1 score of QA pair creation between annotators is 0.53, and the SBERT score is 79.7%. The results show a shared tendency among experts when selecting real-world knowledge and creating a QA pair that is both beneficial and suitable for children's education.

4.3 Statistics of StorySparkQA

Figure 4 demonstrates the distribution of real-world knowledge relations in the dataset, and Table 1 illustrates detailed statistics of the dataset. On average, each section is annotated with approximately 1.4 QA pairs. In StorySparkQA, the top 3 real-world knowledge relations selected by experts are *is a*, *has subevent* and *is the antonym of*, respectively constituting 35.5%, 16.2% and 15.2% of all real-world knowledge relations. The distribution of question types in StorySparkQA is shown in Table 5 in Appendix A.3. In StorySparkQA, questions start with 'what', the most common type of question, which constitutes 86.0%. Questions starting with 'why' and 'how' constitute about 7.2% and 2.4%, respectively.

According to experts' annotation, real-world knowledge relation *is a* and questions start with 'what' have a much higher proportion than the others. Considering the characteristics of cognitive development of children, especially in the age group of 3-6 years, children are usually in the exploration stage and full of curiosity about the world (Chouinard et al., 2007; Jirout and Klahr, 2012), thus it is normal for them to ask questions to satisfy their curiosity. Consequently, parents are more inclined to use 'what' questions to inspire children's thinking and encourage them to actively acquire knowledge (Yu et al., 2019). Consistent with the actual habits of parents and teachers, experts' annotated questions have a high consensus that 'what' questions are more in line with children's learning and cognitive characteristics.

5 Benchmark Experiment

We benchmark the quality and usability of our StorySparkQA on the QA pair generation (QAG) task, which is required to meet the needs of parents in guiding children to learn some real-work knowledge during the real-world storytelling, as well as existing work of developing AI-assisted storytelling systems (Yao et al., 2021; Dietz et al., 2021; Zhang et al., 2022). We conduct an **automated evaluation**, reported in Section 5.1 to measure the semantic similarity of generated QA pairs with experts-annotated QA pairs, benchmarked with a T5-Large model fine-tuned on StorySparkQA and a set of robust LLMs. Considering the limitation of automated evaluation in evaluating the educational appropriateness of generated QA pairs, we further conduct a **human evaluation**, reported in Section 5.2, with children's education experts.

5.1 Automated Evaluation

We now elaborate on the settings and results of our QAG experiments with various language models, through which to demonstrate the usability of StorySparkQA.

5.1.1 Experiment Settings

The QAG task involves taking the input of a story section and generating the QA pairs. To exploit LLMs' comprehensive generation ability, we design two variations to simulate experts' annotation process:

1. *w/o triples*: Generate the QA pair alone.

2. *w/ triples*: Generate the associated knowledge triple alongside the QA pair.

The automatic evaluation comprises six popular LLMs: GPT-3.5, GPT-4 (OpenAI, 2023), FLAN-T5-XXL (Chung et al., 2022), Alpaca (Taori et al., 2023), Mistral (Jiang et al., 2023) and Llama 2 (Touvron et al., 2023). We carefully design the prompt inputs (Appendix A.6) with clear and informative instructions, including 13 relation types (Appendix A.2) in ConceptNet. The goal is leveraging LLMs to generate diverse triples similar to those created by human education experts.

For each LLM involved in this experiment (GPT-3.5, GPT-4, FLAN-T5-XXL, Alpaca, Mistral, and Llama 2), we employ **zero-shot**, **few-shot in-context learning (ICL)** (Brown et al., 2020) and Chain-of-Thought (Wei et al., 2022) approaches to thoroughly examine the QAG performance of these models with different prompting strategies. We randomly sample examples from the validation split as demonstrations for the few-shot ICL approaches. We also fine-tune a T5-Large model to examine how a much smaller domain-specific model, supported by expert-annotated triples as additional input, performs compared to generic LLMs. We report the experiment settings and hyper-parameters in Appendix A.4.

We utilize **Rouge-L** (Lin, 2004) to evaluate the quality of concatenated QA pairs between the generated ones and two expert-annotated ground truths of each data, and report the averaged score across all test data. Additional scores of **SBERT** using Sentence Transformer (Reimers and Gurevych, 2019) are shown in Appendix A.5. We perform experiments with GPT-3.5 and GPT-4 three times

Model	Prompting Strategy	QAG w/o triples	QAG w/ triples
T5-Large fine-tuned (0.77B)	-	**0.332**	**0.279**
Alpaca (7B)	zero-shot	0.124	0.266
	few-shot	0.251	0.239
Mistral (7B)	zero-shot	0.229	0.209
	few-shot	0.267	0.257
Llama 2 (7B)	zero-shot	0.213	0.177
	1-shot	0.192	0.206
	5-shot	0.241	0.269
GPT-3.5	zero-shot	0.194	0.220
	1-shot	0.239	0.252
	5-shot	0.262	0.264
	CoT	-	0.259
GPT-4	zero-shot	0.277	0.243
	1-shot	0.272	0.251
	5-shot	0.287	0.248
	CoT	-	0.262

Table 2: QAG performance of LLMs with different prompting strategies and the fine-tuned T5-Large model. **Bolded numbers** are the best scores within each setting.

for each setting to calculate a robust and reliable average score.

5.1.2 Results and Analysis

In table 2, we show the zero-shot, few-shot ICL, and CoT performances on all models in both settings of the QAG task.

Generally, zero-shot QAG performance on these models falls short of the few-shot ICL QAG performance. Remarkably, models using 5-shot demonstrations outperform those using 1-shot demonstrations. Models employing the Chain-of-Thought prompting method do not imply an obvious improvement compared to the few-shot ICL QAG performance. For the setting of generating triples along with QA pairs (*w/ triples*), the automatic evaluation results do not indicate an improvement in QAG through the step of generating knowledge triples in the real world. We attribute this to the potential complexity of the task that asks LLMs to generate real-world knowledge triples and corresponding QA pairs simultaneously. It is worth noting that T5-Large fine-tuned on our `StorySparkQA` has a relatively better performance than conversational LLMs like GPT-3.5 and GPT-4 by Rouge-L.

5.2 Human Evaluation

To thoroughly assess the quality and usability of LLM-generated QA pairs, particularly in terms of educational appropriateness, we conducted a human study with four education experts to compare expert-annotated QA pairs and those generated by fine-tuned T5-Large and GPT-4 with 5-shot ICL, the best-performing ones in automated evaluation.

We randomly select ten story books from the test split of `StorySparkQA`, and sample seven sections per book. For each section, three QA pairs are created based on the story narrative (experts' annotation, and QA pairs generated by GPT-4 and fine-tuned T5-Large), summing up 210 QA pairs for the human evaluation. QA pairs are randomized for each section, and the sources are omitted to the human subjects for a fair evaluation.

Four experts evaluate each QA pair on the following four dimensions with a 5-point Likert scale:

1. *Grammar Correctness*: The QA pair uses comprehensible English Grammar;

2. *Answer Relevancy*: The answer is correct and corresponds to a question;

3. *Contextual Consistency*: The QA pair originates from the story and goes beyond the story's immediate context;

4. *Children's Educational Appropriateness*: The QA pair is appropriate for young children's reading experience during interactive storytelling;

5.2.1 Results and Analysis

Table 3 illustrates the average scores of each dimension and paired sample *t-test* results. We observe that expert-created QA pairs outperform those generated by models in all four dimensions. The paired sample *t-tests* results show that experts' annotation has significant differences in three out of four dimensions compared with models' generation. These justify the utility of our `StorySparkQA` in catering to parents' real-world needs in interactive storytelling.

In terms of *Grammar Correctness* and *Answer Relevancy*, GPT-4 achieves better performance than the fine-tuned T5-Large. We believe it to be reasonable because LLMs such as GPT-4 are trained on vast amounts of corpora, enabling them to generate text with higher grammatical accuracy.

Dimension	Model	Mean	St.D	t	df	p-value
Grammar Correctness	Human	4.893	0.560			
	T5-Large fine-tuned	4.842	0.585	1.259	279	0.209
	GPT-4	**4.871**	0.514	0.646	279	0.519
Answer Relevancy**	Human	4.696	0.683			
	T5-Large fine-tuned	4.329	1.111	5.487	279	<0.01
	GPT-4	**4.379**	0.869	5.123	279	<0.01
Contextual Consistency*	Human	4.657	0.882			
	T5-Large fine-tuned	**4.639**	0.972	5.487	279	0.729
	GPT-4	4.529	0.974	2.240	279	0.026
Educational Appropriateness**	Human	4.493	0.892			
	T5-Large fine-tuned	**4.325**	0.972	2.937	279	<0.01
	GPT-4	4.318	2.974	3.113	279	<0.01

Note: * denotes p-value <0.05, ** denotes p-value <0.01

Table 3: The paired sample t-test result of children's education experts in comparison of GPT-4 and T5-Large fine-tuned on `StorySparkQA` in the QAG task. **Bolded numbers** are the best scores within each dimension excluding human experts' annotation.

In terms of *Contextual Consistency*, the fine-tuned T5-Large significantly outperformed GPT-4, behind experts' annotation. A similar result could be found in *Children's Educational Appropriateness*, wherein the T5-Large model fine-tuned on `StorySparkQA` also exhibits better performance.

These results suggest that fine-tuned with experts' annotation, the T5-Large model can generate QA pairs that 1) contain external structured knowledge connected to the story narrative, and 2) are appropriate for young children to learn during the interactive storytelling activities.

5.3 Discussion

Comparing the best-performing SoTA LLMs in the QAG pipeline with the corresponding fine-tuned T5-Large, we can observe that the T5-Large can reliably generate QA pairs aligned more with experts' annotation in terms of Rouge-L score according to system evaluation, regardless of whether generating QA pairs along real-world knowledge triples. Drawing from the results of our human evaluation, the fine-tuned T5-Large exhibits better capabilities in generating QA pairs that suit parents' real-world educational expectations of interactive storytelling: originating from the story and embodying educational-appropriate real-world knowledge. Worth mentioning that T5-Large only consists of 770 million parameters, whereas Alpaca, Mistral and Llama in our experiments consist of 7 billion parameters (10 times larger).

This observation justifies the utility of `StorySparkQA` in training a task-specific model that caters to parents' real-world storytelling needs on the one hand, and **demonstrates the usefulness of combining structured real-world knowledge and free-form narratives in domain-specific tasks such as interactive storytelling.**

6 Conclusion and Future Work

In summary, we propose a QA dataset for children's education, named `StorySparkQA`, by leveraging a novel annotation framework that facilitates scalable expert annotations through structured external knowledge. `StorySparkQA` integrates external knowledge into children's story-based learning texts. We demonstrate the utility of `StorySparkQA` through an automated evaluation on various LLMs of generating QA pairs catering to parents' needs as well as a human evaluation with children's education experts.

One possible future work is refining the structure of QAG pipelines and exploiting LLMs for generating QA pairs that align more closely with parents' real-world needs. Another future direction involves using `StorySparkQA` and language models to develop a human-AI collaborative education system (e.g., an interactive storytelling system), aiding parents and educators to formulate personalized questions during story readings, while addressing their language, knowledge, skill, or time constraints.

7 Limitations

This work primarily focuses on constructing an expert-annotated, large-scale QA dataset consisting of story-based QA pairs associated with real-world knowledge beyond the story narrative. There are several limitations.

First, in the benchmark experiment, despite we have employed various prompting strategies to harness LLMs' generation potential, more prompting methods, e.g., 10-shot ICL for GPT-4 and Llama 2, could be further explored.

Second, we implement and evaluate one QAG pipeline in the end-to-end setting. Although we experiment with two different variations, we acknowledge that more novel pipeline designs, such as multi-step generation pipelines, could be implemented to further explore `StorySparkQA`'s utility.

Third, our experiment with a fine-tuned language model solely utilizes a T5-Large model to generate QA pairs. We recognize that the performance of other models, such as BERT (Devlin et al., 2019), BART (Lewis et al., 2019), etc., as well as some instruction-finetuned LLMs, such as InstructGPT (Ouyang et al., 2022), can be further explored.

Additionally, in the knowledge matching module of the proposed annotation framework, we currently focus on knowledge represented in the triplet of two concepts and a relation. The incorporation of meta-paths connecting multiple concepts is under-explored.

References

American Diabetes Association. 2011. Diagnosis and Classification of Diabetes Mellitus. *Diabetes Care*, 34(Supplement_1):S62–S69.

Sören Auer, Dante A. C. Barone, Cassiano Bartz, Eduardo G. Cortes, Mohamad Yaser Jaradeh, Oliver Karras, Manolis Koubarakis, Dmitry Mouromtsev, Dmitrii Pliukhin, Daniil Radyush, Ivan Shilin, Markus Stocker, and Eleni Tsalapati. 2023. The sciqa scientific question answering benchmark for scholarly knowledge. *Scientific Reports*, 13(1):7240.

Isabel L Beck, Margaret G McKeown, and Linda Kucan. 2013. *Bringing words to life: Robust vocabulary instruction*. Guilford Press.

Amanda Berry, Fien Depaepe, and Jan Van Driel. 2016. Pedagogical content knowledge in teacher education. *International Handbook of Teacher Education: Volume 1*, pages 347–386.

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. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, pages 1877–1901, Red Hook, NY, USA. Curran Associates Inc.

Michelle M. Chouinard, P. L. Harris, and Michael P. Maratsos. 2007. Children's Questions: A Mechanism for Cognitive Development. *Monographs of the Society for Research in Child Development*, 72(1):i–129. Publisher: [Society for Research in Child Development, Wiley].

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling Instruction-Finetuned Language Models. *Preprint*, arxiv:2210.11416.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North*, pages 4171–4186.

Griffin Dietz, Jimmy K Le, Nadin Tamer, Jenny Han, Hyowon Gweon, Elizabeth L Murnane, and James A. Landay. 2021. StoryCoder: Teaching Computational Thinking Concepts Through Storytelling in a Voice-Guided App for Children. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15. Conference Name: CHI '21: CHI Conference on Human Factors in Computing Systems ISBN: 9781450380966 Place: Yokohama Japan Publisher: ACM.

Nuha A. ElSayed, Grazia Aleppo, Vanita R. Aroda, Raveendhara R. Bannuru, Florence M. Brown, Dennis Bruemmer, Billy S. Collins, Marisa E. Hilliard, Diana Isaacs, Eric L. Johnson, Scott Kahan, Kamlesh Khunti, Jose Leon, Sarah K. Lyons, Mary Lou Perry, Priya Prahalad, Richard E. Pratley, Jane Jeffrie Seley, Robert C. Stanton, Robert A. Gabbay, and null on behalf of the American Diabetes Association. 2023. 2. Classification and Diagnosis of Diabetes: Standards of Care in Diabetes-2023. *Diabetes Care*, 46(Suppl 1):S19–S40.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. 2017. Allennlp: A deep semantic natural language processing platform.

Roberta Michnick Golinkoff, Erika Hoff, Meredith L. Rowe, Catherine S. Tamis-LeMonda, and Kathy Hirsh-Pasek. 2019. Language Matters: Denying the Existence of the 30-Million-Word Gap Has Serious Consequences. *Child Development*, 90(3):985–992.

Amir Hadifar, Semere Kiros Bitew, Johannes Deleu, Chris Develder, and Thomas Demeester. 2023. EduQG: A Multi-Format Multiple-Choice Dataset for the Educational Domain. *IEEE Access*, 11:20885–20896.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Rebecca Isbell, Joseph Sobol, Liane Lindauer, and April Lowrance. 2004. The Effects of Storytelling and Story Reading on the Oral Language Complexity and Story Comprehension of Young Children. *Early Childhood Education Journal*, 32(3):157–163.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. Publisher: arXiv Version Number: 1.

Jamie Jirout and David Klahr. 2012. Children's scientific curiosity: In search of an operational definition of an elusive concept. *Developmental Review*, 32(2):125–160.

Rachel Keraron, Guillaume Lancrenon, Mathilde Bras, Frédéric Allary, Gilles Moyse, Thomas Scialom, Edmundo-Pavel Soriano-Morales, and Jacopo Staiano. 2020. Project PIAF: Building a Native French Question-Answering Dataset. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5481–5490, Marseille, France. European Language Resources Association.

Huseyin Kotaman. 2013. Impacts of Dialogical Storybook Reading on Young Children's Reading Attitudes and Vocabulary Development. *Reading Improvement*, 50(4):199–204. Publisher: Project Innovation, Inc ERIC Number: EJ1023501.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. 2015. DBpedia – A large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, 6(2):167–195. Publisher: IOS Press.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdel rahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Annual Meeting of the Association for Computational Linguistics*.

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.

OpenAI. 2023. GPT-4 Technical Report. *ArXiv*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. 2022. Training language models to follow instructions with human feedback. *ArXiv*, abs/2203.02155.

Jiaxin Pei, Aparna Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Apostolos Dedeloudis, Jackson Sargent, and David Jurgens. 2022. POTATO: The Portable Text Annotation Tool. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 327–337, Abu Dhabi, UAE. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):3027–3035.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

Yuling Sun, Jiali Liu, Bingsheng Yao, Jiaju Chen, Dakuo Wang, Xiaojuan Ma, Yuxuan Lu, Ying Xu, and Liang He. 2024. Exploring Parent's Needs for Children-Centered AI to Support Preschoolers' Storytelling and Reading Activities. *Preprint*, arxiv:2401.13804.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, A. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *ArXiv*.

Daniel Vargas-Diaz, Sulakna Karunaratna, Jisun Kim, Sang Won Lee, and Koeun Choi. 2023. TaleMate: Collaborating with Voice Agents for Parent-Child Joint Reading Experiences. In *Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23 Adjunct, pages 1–3, New York, NY, USA. Association for Computing Machinery.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, E. Chi, F. Xia, Quoc Le, and Denny Zhou. 2022. Chain of Thought Prompting Elicits Reasoning in Large Language Models. *ArXiv*.

Andrew Wright. 1995. *Storytelling with Children*. Oxford University Press. Google-Books-ID: IuQOKN63TCwC.

Yichong Xu, Chenguang Zhu, Ruochen Xu, Yang Liu, Michael Zeng, and Xuedong Huang. 2020. Fusing Context Into Knowledge Graph for Commonsense Reasoning. *ArXiv*.

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.

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. 2022. 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 Sherry Wu, T. Hoang, Branda Sun, Toby Jia-Jun Li, Mo Yu, and Ying Xu. 2021. It is AI's Turn to Ask Human a Question: Question and Answer Pair Generation for Children Storybooks in FairytaleQA Dataset. *ArXiv*.

Yue Yu, Elizabeth Bonawitz, and Patrick Shafto. 2019. Pedagogical Questions in Parent–Child Conversations. *Child Development*, 90(1):147–161. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cdev.12850.

Chao Zhang, Xuechen Liu, Katherine Ziska, Soobin Jeon, Chi-Lin Yu, and Ying Xu. 2024. Mathemyths: leveraging large language models to teach mathematical language through child-ai co-creative storytelling. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–23.

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 Chatbot for Parent-Child Interactive Storytelling with Flexible Parental Involvement. *CHI Conference on Human Factors in Computing Systems*, pages 1–21. Conference Name: CHI '22: CHI Conference on Human Factors in Computing Systems ISBN: 9781450391573 Place: New Orleans LA USA Publisher: ACM.

Sanqiang Zhao, Seokhwan Kim, Yang Liu, Robinson Piramuthu, and Dilek Hakkani-Tür. 2023. Storyqa: Story grounded question answering dataset. In *AAAI 2023 Workshop on Knowledge Augmented Methods for NLP*.

A Appendix

A.1 Properties of Educational QA datasets

Dataset	# books	# QA pairs	External Knowledge	Annotator	Document Source
StoryQA	148	38,703	Yes	Crowd-Sourced	Story books
FAIRYTALEQA	278	10,580	No	Expert	Story books
EduQG	13	5,018	No	Expert	Text books
StorySparkQA	278	5,868	Yes	Expert	Story books

Table 4: Properties of existing datasets focusing on children's education compared with our StorySparkQA.

A.2 ConceptNet Relations

We follow Xu et al. (2020)'s work to filter out weak relations in ConceptNet, and our ranking algorithm uses the following 13 relations in our annotation framework as well as GPT prompts: *causes*, *desires*, *has context of*, *has property*, *has subevent*, *is a*, *is at location of*, *is capable of*, *is created by*, *is made of*, *is part of*, *is the antonym of*, *is used for*.

A.3 Distribution of Question Type

The distribution of question type in StorySparkQA is shown in Table 5.

Interrogative	**Train split**	**Val split**	**Test split**	**Total percentage (%)**
what	3779	628	641	86.01
why	227	93	105	7.24
who	76	10	14	1.70
where	41	3	7	0.87
when	20	12	8	0.68
how	112	13	15	2.39
other	42	10	9	1.04

Table 5: Distribution of question types in StorySparkQA.

A.4 Hyper-parameters and Experiment Settings

We conducted our experiments on Google Colab with A100. Following common practice when fine-tuning the T5-Large model, we use the learning rate of 1e-4 and train our model on 3 epochs.

A.5 Complete QAG Pipeline Results

We demonstrate the complete performance of LLMs in our QAG pipeline using both zero-shot and few-shot ICL approaches in Table 6.

Models	**Prompting Strategy**	**End2End Pipeline w/o triples**		**End2End Pipeline w/ triples**	
		Rouge-L	SBERT	Rouge-L	SBERT
T5-Large fine-tuned	-	**0.332**	0.289	**0.279**	0.263
Alpaca	zero-shot	0.124	0.186	0.266	0.207
	1-shot	0.251	0.182	0.239	0.186
Mistral	zero-shot	0.229	0.237	0.209	0.229
	1-shot	0.227	0.237	0.231	0.241
	5-shot	0.267	0.241	0.257	0.251
Llama 2	zero-shot	0.213	0.234	0.177	0.225
	1-shot	0.192	0.217	0.206	0.237
	5-shot	0.241	0.240	0.269	0.253
Flan-T5-XXL	1-shot	0.264	0.246	0.194	0.209
GPT-3.5	zero-shot	0.194	0.233	0.220	0.252
	1-shot	0.239	0.262	0.252	0.271
	5-shot	0.262	0.279	0.264	0.266
	CoT	-	-	0.259	0.280
GPT-4	zero-shot	0.277	0.252	0.243	0.261
	1-shot	0.272	0.279	0.251	**0.292**
	5-shot	0.287	**0.311**	0.248	0.283
	CoT	-	-	0.262	**0.292**

Table 6: Rouge-L and SentenceBERT scores of LLMs in the QAG task. **Bolded numbers** are global best performance within each setting on each metric.

A.6 GPT Prompts

To utilize GPT's strong reasoning and generation capability as well as control GPT-generated questions as much as possible to meet the needs of parents, we carefully design our prompts.

For the QAG pipeline, there are two variations based on the system: (1) Directly generate a QA pair based on a provided story section. (2) From a story section, generate a real-world knowledge triple and a QA pair simultaneously.

Table 7, 8 list our prompts for GPT in the two abovementioned approaches.

A.7 User Interface for Annotation System

We implement an annotation system to facilitate QA pair annotation with associated external knowledge. Figure 5, 6 and 3 show the annotation interface for human experts.

We also conduct cross-validation to assess the agreement among annotators. Figure 7, 8, 9 and 10 demonstrate user interfaces for each step to support the cross-validation process.

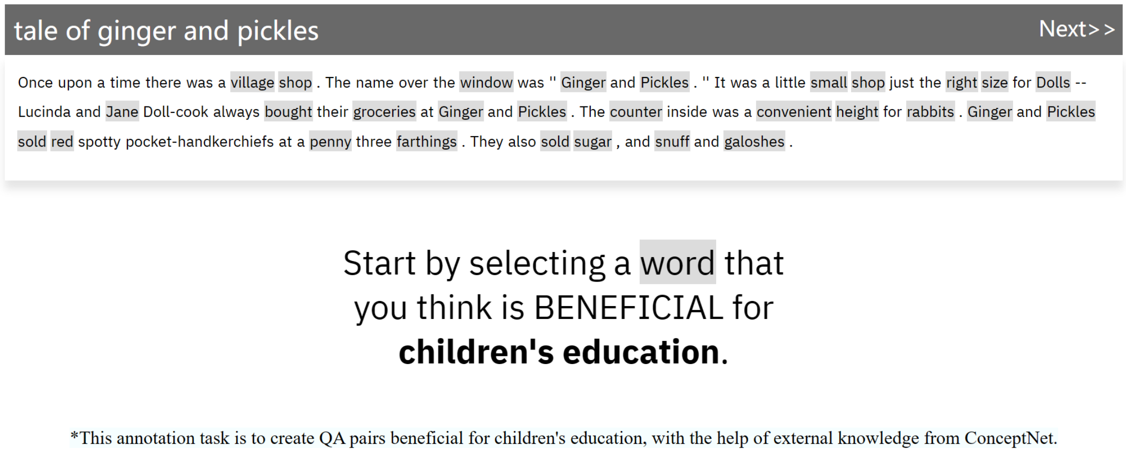

Figure 5: Annotation process1: Browse a displayed section, with candidate words highlighted in grey.

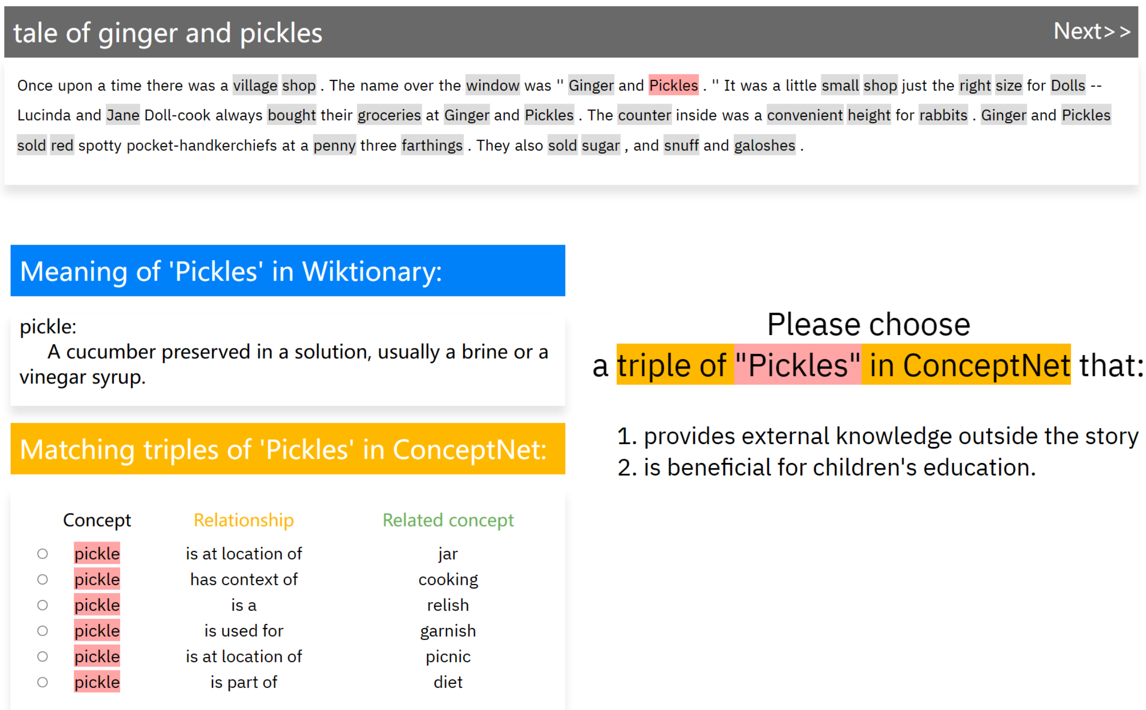

Figure 6: Annotation process2: After selecting a word (highlighted in red), related explanation in Wiktionary and candidate real-world knowledge triples in ConceptNet will display.

Prompt for GPT in the QAG pipeline (generate QA pairs only)
I need you to help generate a question and answer pair for young children aged three to six. I will provide you with a short section of a story delimited by triple quotes. Please follow these steps: 1. For each sentence, identify one key word that meets the following criteria: it is relatively complex, it is considered tier 1 or tier 2 vocabulary, and it is a concrete noun, verb, or adjective. 2. After this, you need to completely forget about the story that I gave you, remembering only the words you identified. 3. Based on each selected word, generate a question and answer pair that either the question or the answer contains that word. For example, if your identified word is 'apple', your question could be: where do apples grow? what do apples taste like? what color are apples? These questions should go beyond the context of the stories. Each question should have one single correct answer that would be the same regardless of the children's experiences. The questions should be focused on real-world, fact-based knowledge and beneficial to educate children during storytelling. The real-world, fact-based knowledge should be based on the selected word and is in the form of a triple such as A relation B, where A and B are two concepts and the selected word can be either A or B. You should use one of the following relations for the real-world knowledge: causes desires has context of has property has subevent is a is at location of is capable of is created by is made of is part of is the antonym of is used for 4. After this, select one question-answer pair that you think best meets my criteria. Please note that the question should be answerable without reading the story. The answer should only be a concrete noun, verb, or adjective. Return the selected question-answer pair in the following format: question: ... answer: ... ⟨story⟩: ***{story1 for few-shot}*** ⟨response⟩: ***{response1 for few-shot}*** ... ... ⟨story⟩: ***{story for the current data}*** ⟨response⟩:

Table 7: Prompt for GPT in the QAG task with generating QA pairs directly from the story.

Prompt for GPT in the QAG pipeline (generate triples and QA pairs)
I need you to help generate a question and answer pair for young children aged three to six. I will provide you with a short section of a story delimited by triple quotes. Please follow these steps: 1. For each sentence, identify one key word that meets the following criteria: it is relatively complex, it is considered tier 1 or tier 2 vocabulary, and it is a concrete noun, verb, or adjective. 2. After this, you need to completely forget about the story that I gave you, remembering only the words you identified. 3. Based on each selected word, generate one real-world relation based on the selected word. This real-world relation should go beyond the context of the stories. For example, if your identified word is 'apple', your real-world relation could be: apple grows on trees; apples are red. The real-world, fact-based knowledge should be based on the selected word and is in the form of a triple such as 'A relation B', where A and B are two concepts and the selected word can be either A or B. You should use one of the following relations for the real-world knowledge: causes desires has context of has property has subevent is a is at location of is capable of is created by is made of is part of is the antonym of is used for 4. After this, generate a question and answer pair based on the real-world, fact-based knowledge you generated. Either the question or the answer should contain that identified word. Each question should have one single correct answer that would be the same regardless of the children's experiences. The questions should be focused on real-world, fact-based knowledge and beneficial to educate children during storytelling. 5. After this, select one question-answer pair that you think best meets my criteria. Please note that the question should be answerable without reading the story. The answer should only be a concrete noun, verb, or adjective. Return the generated real-world knowledge triple and selected question-answer pair in the following format: real-world knowledge triple: (A, relation, B) question: ... answer: ... ⟨story⟩: ***{story1 for few-shot}*** ⟨response⟩: ***{response1 for few-shot}*** ... ... ⟨story⟩: ***{story for the current data}*** ⟨response⟩:

Table 8: Prompt for GPT in the QAG task with generating real-world knowledge triple and QA pairs directly from the story.

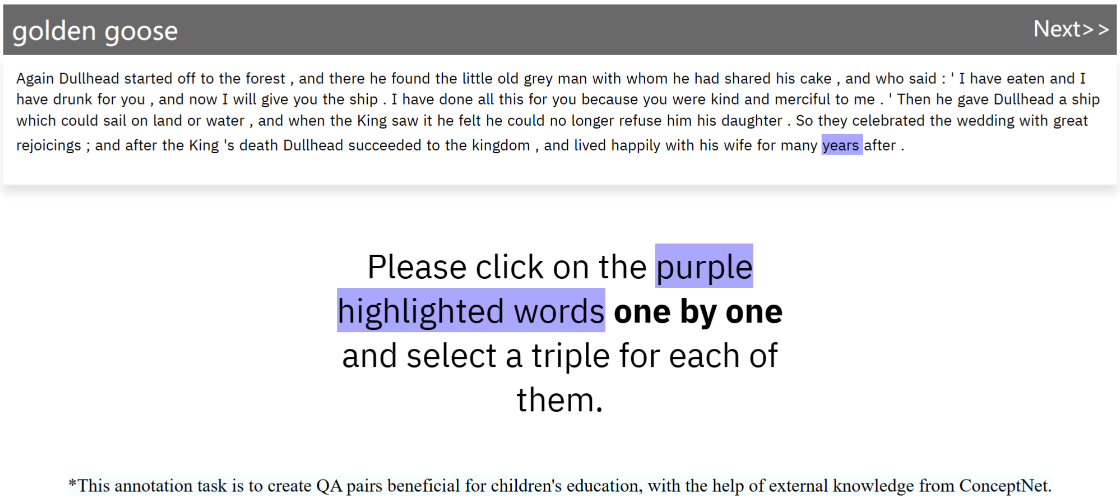

Figure 7: Cross-validation process1: Browse a displayed section, with candidate words highlighted in grey.

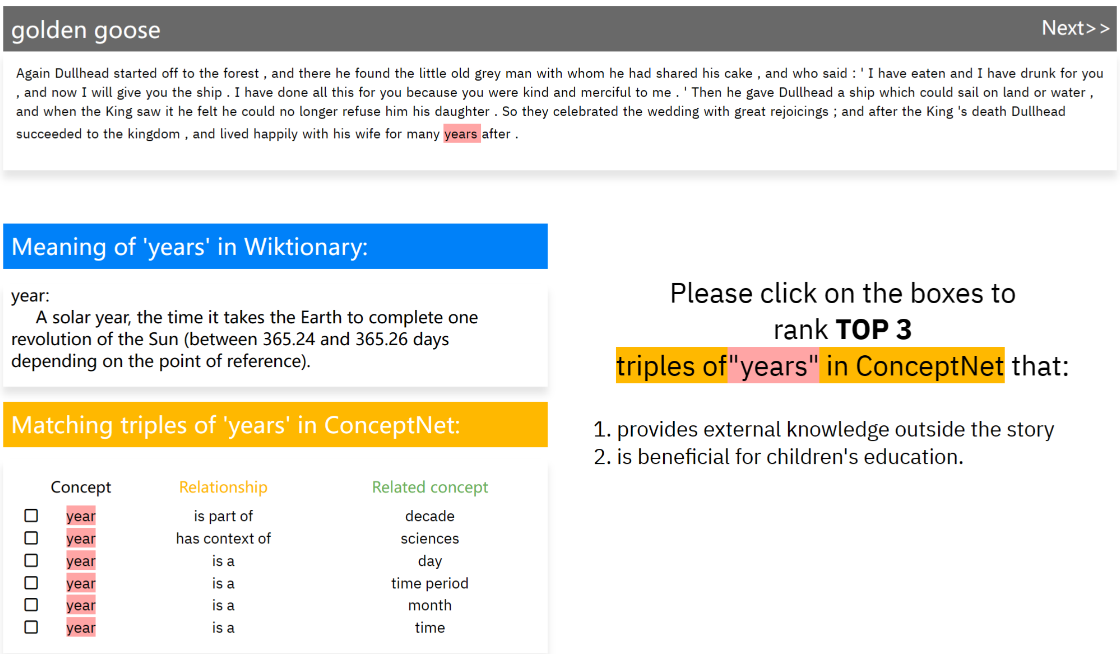

Figure 8: Cross-validation process2: Select a word annotated by others and rank the candidate triples.

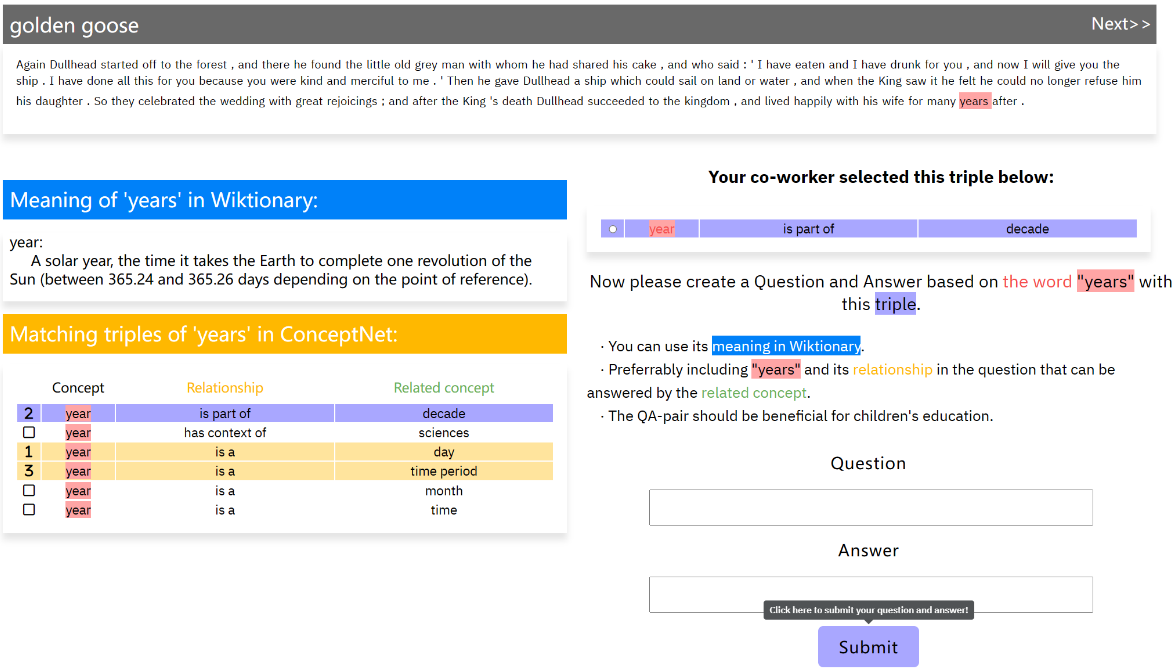

Figure 9: Cross-validation process3: After ranking top3 triples, the triple selected originally by the other annotator is displayed, the validator should create a QA pair based on the original triple.

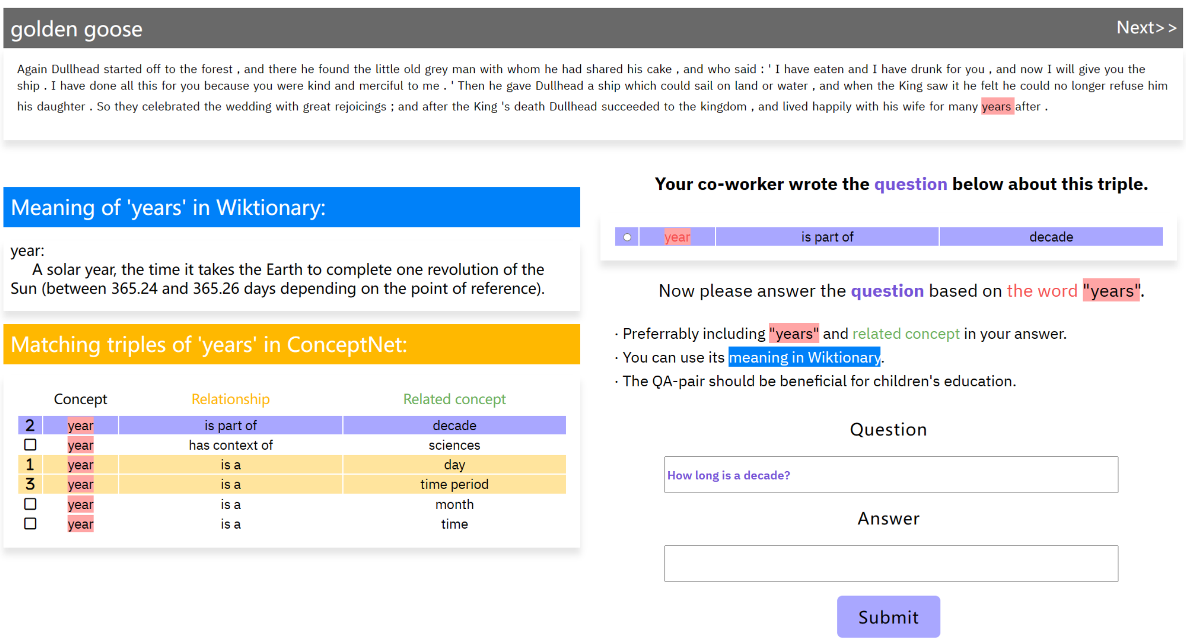

Figure 10: Cross-validation process4: Validator is asked to answer the question created by the other annotator using the triple originally selected by the other annotator.