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

Jiaju Chen* East China Normal University
Yuxuan Lu Northeastern University
Shao Zhang* Shanghai Jiao Tong University
Bingsheng Yao Northeastern University
Yuanzhe Dong Stanford University
Ying Xu Harvard University
Yunyao Li Adobe
Qianwen Wang University of Minnesota Twin Cities
Dakuo Wang Northeastern University

Yuling Sun†
Lab of Artificial Intelligence for Education
East China Normal University

Abstract

Interactive story reading is common in early childhood education, where teachers expect to teach both language skills and real-world knowledge beyond the story. While many story reading systems have been developed for this activity, they often fail to infuse real-world knowledge into the conversation. This limitation can be attributed to the existing question-answering (QA) datasets used for children's education, upon which the systems are built, failing to capture the nuances of how education experts think when conducting interactive story reading activities. To bridge this gap, we design an annotation framework, empowered by existing knowledge graph to capture experts' annotations and thinking process, and leverage this framework to construct StorySparkQA dataset, which comprises $5,868$ expert-annotated QA pairs with real-world knowledge. We conduct automated and human expert evaluations across various QA pair generation settings to demonstrate that our StorySparkQA can effectively support models in generating QA pairs that target real-world knowledge beyond story content. StorySparkQA[1] is available at https://huggingface.co/datasets/NEU-HAI/StorySparkQA.

1 Introduction

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

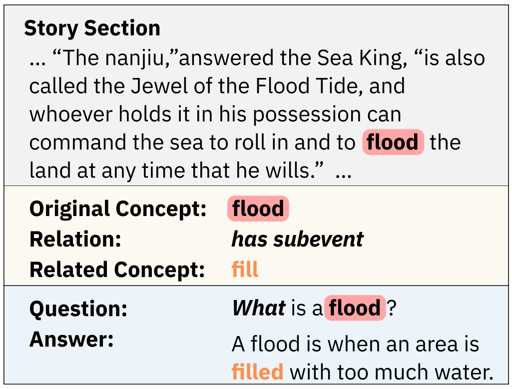

Figure 1: An example of StorySparkQA dataset. In each story section, educational experts select a concept word, link it to a desired external real-world knowledge, and write an appropriate QA pair. Additional data examples of StorySparkQA are presented in Appendix A.1.

Despite the benefits of interactive story reading, teachers often struggle to appropriately conduct such interactive story reading with children because of multi-facet difficulties (Golinkoff et al., 2019; Sun et al., 2024). Specifically, such interactive story reading needs teachers to identify the knowl-

*Work done when visiting Northeastern University.
† Corresponding Author: ylsun@cs.ecnu.edu.cn.
[1]The StorySparkQA dataset and annotation framework can also be found in: https://github.com/neuhai/StorySparkQA

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 17351–17370
November 12-16, 2024 ©2024 Association for Computational Linguistics

edge 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"**). In home settings, most parents also lack the educational expertise necessary to guide such educational conversations (Golinkoff et al., 2019; Sun et al., 2024). Meanwhile, today's parents often hardly maintain constant focus on 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)), have demonstrated utility in children's storytelling scenarios (Dietz et al., 2021). These systems primarily utilize the verbal communication interface and advanced language models to support natural conversation with humans (Mahmood et al., 2023; Chan et al., 2024; Yang et al., 2024; 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 in helping teaching real-world knowledge beyond the story narrative (Yao et al., 2021; Zhao et al., 2022), which is one of the main expectations of parents and teachers (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, to the best of our knowledge, no such data resources exist in the domain of children education. Further, the collection of such data resources requires annotators to recall a comprehensive and systematic 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`, 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 automated evaluation and human expert evaluations.

`StorySparkQA` can benefit different research directions in the children education, 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 and teachers expect. In addition, we believe our annotation framework has the potential to be generalized to domain-specific tasks analogous to the real world that require 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 6. 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.

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

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 triples: ($concept_1$, $relation$, $concept_2$). 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.

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, like the annotation toolkit for StoryQA (Zhao et al., 2023), support free-form input, 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 like BERT (Devlin et al., 2019) on QAG datasets for end-to-end generation was a prevalent approach, but such methodology heavily depends on the training data quality and lacks control of generated content, which is inappropriate for the children education domain. Existing works also attempted to design multi-step generation pipelines, which offer better control of the generated content (Yao et al., 2021; Wan et al., 2024).

Recent advancement in large language models (LLMs), such as GPT-3.5, GPT-4 (OpenAI, 2023), and Llama 3 (Dubey et al., 2024), supports free-form natural language input and output without the need for tuning model parameters. 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; Yao et al., 2024), Chain-of-Thought (i.e., ask models to think "step-by-step") (Wei et al., 2022), etc. The performance of these prompting methods on the general QAG task has also been evaluated (Ling and Afzaal, 2024; Lu et al., 2023). However, to what extent these prompting and modeling strategies are effective in **the QAG task for knowledge beyond the story content**, and whether a compact language model fine-tuned on domain-specific datasets performs better or worse than generic LLMs in the context of children education remains underexplored. For example, recent work demonstrates the unreliability of generic LLMs for the classification of mental health issues and the superiority of fine-tuning domain-specific language models with high-quality datasets (Xu et al., 2024). Our work attempts to step forward through comprehensive evaluation in Section 5.

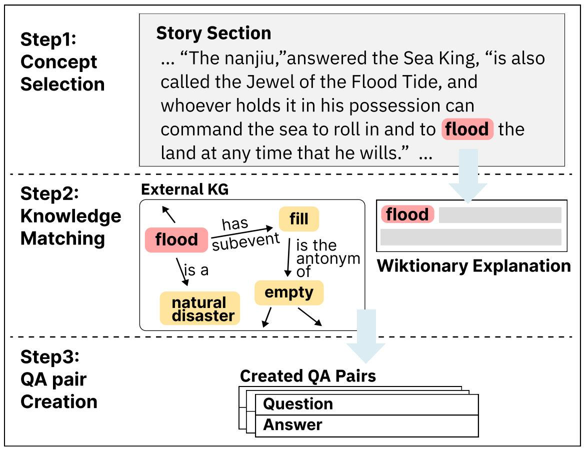

Figure 2: Workflow of the experts' annotation process. Experts need to select a concept first, then match it with the most suitable knowledge, and finally create a QA pair based on the selected knowledge.

3 Expert Annotation Framework

To better understand experts' mental procedures during annotating QA pairs enriched with external real-world knowledge, 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 (UI shown in Figure 5), experts identify an educationally suit-

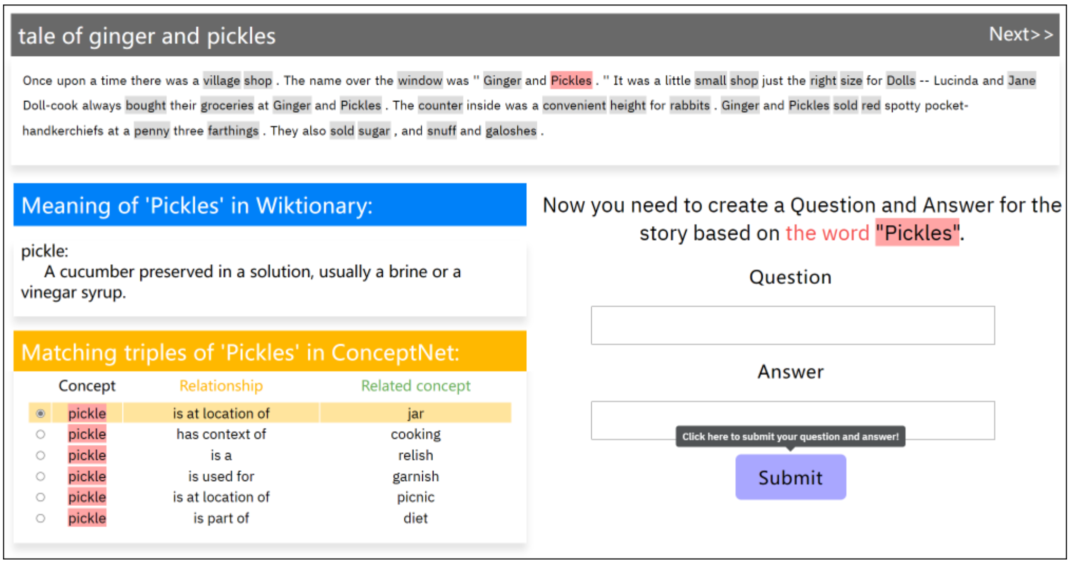

Figure 3: The user interface to facilitate our annotation task. The words highlighted in grey are candidate concepts. The blue block shows the Wiktionary explanation, and the yellow block lists our recommended triples.

able concept from the story text. 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.

Step 2. Knowledge Matching This step (UI shown in Figure 6) allows experts to select real-world knowledge based on the 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.3), 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) (Ramos et al., 2003).

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 annotation. We also retrieve the explanation for concepts from Wiktionary to better facilitate experts' annotations.

Step 3. QA pair Annotation This step enables annotators to create a QA pair based on the real-world knowledge triple they selected in step 2, and the corresponding UI is shown in Figure 3. 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 teachers' interactive story reading 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

[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'

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

StorySparkQA	Train 232 books with 4,300 QA pairs				Validation 23 books with 769 QA pairs				Test 23 books with 799 QA pairs			
	Mean	**SD**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**	**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.

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 education, FAIRYTALEQA (Xu et al., 2022) comprises 278 classic fairytale stories of various origins, and all the stories have been evaluated as suitable for 8$^{\text{th}}$-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.

4.2 Annotation Process

Following our annotation framework, we recruit 11 children education experts for the annotation task. The 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 by selecting concepts that are most beneficial for children's education from story text. The experts then proceed to select a real-world knowledge triple associated with the selected concept and create a QA pair based on the selected triple. In this process, experts are asked to consider children's cognitive and knowledge levels and write QA pairs that are most appropriate for 3- to 6-year-olds. We collect experts' mental procedures by recording their selected concepts, real-world knowledge triples, and created QA pairs during the annotation process.

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).

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 0.80, showing a shared

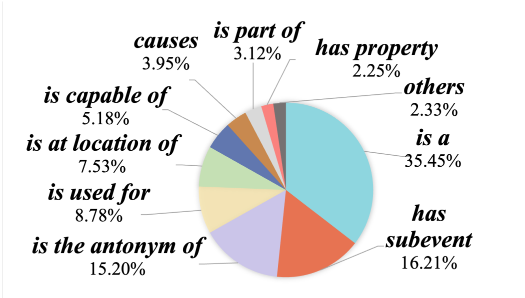

Figure 4: Distribution of real-world knowledge relations annotated by experts in the StorySparkQA dataset

tendency among experts when selecting real-world knowledge and creating a QA pair that is both beneficial and suitable for children education.

4.3 Statistics and Analysis of StorySparkQA

Statistics of StorySparkQA Figure 4 demonstrates the distribution of real-world knowledge relations in StorySparkQA, 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 7 in Appendix A.4. In StorySparkQA, questions starting with "what" are the most common type of question, constituting 86.0%. Questions starting with "why" and "how" constitute about 7.2% and 2.4%, respectively.

Analysis of StorySparkQA's Alignment with Real-World Needs According to experts' annotations, real-world knowledge relation *"is a"* and questions start with "what" have a much higher proportion than the others. Moreover, of all experts' annotations, the selected concept words are mostly nouns (65.06%) and adjectives (22.26%), which are easy for children to comprehend. Considering the cognitive development of children, especially those aged 3 to 6, they are typically in the stage of language development and exploration, full of curiosity about the world (Chouinard et al., 2007; Jirout and Klahr, 2012). It is therefore natural for them to ask questions as a way to satisfy this curiosity. Consequently, teachers are more inclined to use "what" questions with simple vocabulary to inspire children's thinking and encourage them to actively acquire knowledge (Taylor et al., 1994; Yu et al., 2019). Consistent with the actual habits of teachers, experts' annotated questions have a high consensus that "what" questions are more aligned with children's learning and cognitive characteristics.

Comparison with Existing Datasets Compared with existing QA datasets for children education (Xu et al., 2022; Zhao et al., 2023), StorySparkQA is unique mainly in the annotation process and data composition. Going beyond direct QA pair annotation utilized by most QA datasets (e.g., FAIRYTALEQA and StoryQA), StorySparkQA's annotation process provides step-by-step support with structured real-world knowledge to ease the effort that the experts need to craft children-appropriate and knowledgeable QA pairs. In addition, StorySparkQA includes not only QA pairs, but also corresponding story texts and expert-selected real-world knowledge triples, which reflects experts' mental procedures when creating these children-appropriated, knowledgeable QA pairs and leads to a more comprehensive dataset for children's knowledge expansion. Catering to teachers' practical needs for knowledge expansion in interactive story reading, StorySparkQA complements existing QA datasets, which often focus on narrative comprehension and commonsense question-answering. The key properties of StorySparkQA and related children education datasets are illustrated in Table 6 in Appendix A.2.

5 Benchmark Experiment

We benchmark the quality and usability of our StorySparkQA on the QAG task, which is required to meet the needs of teachers to guide children to learn some real-work knowledge during practical interactive story reading, as well as existing work of developing AI-assisted storytelling and reading 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 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 a story section as input and generating the QA pairs. To exploit LLMs' comprehensive generation ability, we design two variations to simulate experts' annotation process:

1. *QA pair generation*: Generate the QA pair.
2. *QA pair and triple generation*: Generate the associated real-world 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-7B (Taori et al., 2023), Mistral-7B (Jiang et al., 2023) and Llama 2(7B) (Touvron et al., 2023). We carefully design the prompt inputs (Appendix A.8) with clear and informative instructions, including 13 relation types (Appendix A.3) in ConceptNet. The goal is to leverage 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) approaches to thoroughly examine the QAG performance of these models with different prompting strategies. For GPT-3.5 and GPT-4, we also use **Chain-of-Thought (CoT)** (Wei et al., 2022) to further explore their QAG capabilities. Randomly sampled examples from the validation split are used 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. The experiment settings and hyper-parameters are reported in Appendix A.6.

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 data in the test split. For the setting that generates triples along with QA pairs, we evaluate the generated triples and QA pairs separately. The Rouge-L F1 score is chosen because it captures the sequential matching of texts, reflecting whether the generated QA pairs are textually similar to experts' annotation. However, we are aware that Rouge-L

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.

is not without limitations, such as a lack of semantic understanding. To address this, we incorporate **SBERT** using Sentence Transformer (Reimers and Gurevych, 2019) for automated evaluation and present the full results in Appendix A.7. We also conduct a human evaluation (see in Section 5.2) to further assess the quality of generated QA pairs. We perform experiments with GPT-3.5 and GPT-4 three times 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. Particularly, Alpaca's few-shot performance when generating QA pairs without triples falls behind the zero-shot setting. This is due to the repetitive and less diverse outputs generated by the model in the zero-shot setting (e.g., '*What is the relation between A and B? A is the antonym of B.*'). Such QA pairs are repetitive across the results, leading to a higher Rouge-L score as the words in it have a higher chance to match with the experts' annotations. Particularly, the relation *'is the antonym of'* occupies a significant proportion in the concatenation of QA pairs.

Remarkably, models using 5-shot demonstrations outperform those using 1-shot demonstrations. It is notable that GPT-3.5 achieves better performance in the few-shot setting compared to GPT-4. We believe this is caused by GPT-4's exceptional Natural Language Generation versatility, potentially resulting in longer and more grammatically complicated QA pairs that are unsuitable for 3- to 6-year-olds due to advanced vocabulary and knowledge comprehension challenges (Ouellette, 2006; Perfetti and Stafura, 2014). Meanwhile, 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 results do not indicate an improvement in QAG through the step of generating real-world knowledge triples. 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 in generating QA pairs enriched with real-world knowledge for children than conversational LLMs like GPT-3.5 and GPT-4 by Rouge-L. Additional analysis of the generated external knowledge types among experts' annotations, fine-tuned T5-Large, and GPT-4 is demonstrated in Appendix A.5.

5.2 Human Evaluation

To further compensate for the limitation of Rouge-L and **SBERT** as well as 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' annotations, 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 children's reading experience during interactive story reading.

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-test* results show that experts' annotations are significantly different in three out of four dimensions compared with models' generation. These justify `StorySparkQA`'s utility in catering to teachers' real-world needs in interactive story reading.

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 QA pairs with greater consistency in word usage. Therefore, compared with T5-Large, GPT-4 produces answers that connect more closely with the questions, resulting in greater coherence and accuracy between the questions and the answers.

In terms of *Contextual Consistency*, the fine-tuned T5-Large significantly outperformed GPT-4, behind experts' annotations. 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' annotations, 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 story reading activities.

5.3 Discussion

Comparing the best-performing SoTA LLMs in the QAG pipeline with the corresponding fine-tuned

Dimension	Model	Mean	SD	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

Table 3: The paired sample t-test result of children 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' annotations. * means p-value <0.05, and ** means p-value <0.01, both are statistically significant.

T5-Large, we can observe that the T5-Large can reliably generate QA pairs aligned more with experts' annotations 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 teachers' real-world educational expectations of interactive story reading: originating from the story and embodying educational-appropriate real-world knowledge. Worth mentioning that T5-Large only consists of 770 million parameters, whereas Alpaca-7B, Mistral-7B, and Llama 2 in our experiments consist of 7 billion parameters (10 times larger).

This observation justifies `StorySparkQA`'s utility in training a task-specific model that caters to teachers' real-world story reading 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 story reading.**

6 Conclusion and Future Work

In summary, we propose `StorySparkQA`, an expert-annotated, external-knowledge-enriched QA dataset for children education, by leveraging a novel annotation framework to facilitate scalable expert annotations through structured external knowledge. We demonstrate the effectiveness of `StorySparkQA` through an automated evaluation on various LLMs of generating QA pairs catering to teachers' needs and a human evaluation with children education experts.

One possible future work is refining the QAG pipeline structures and exploiting LLMs to generate QA pairs that align more closely with teachers' practical needs. Another future direction involves using `StorySparkQA` and language models to develop a human-AI collaborative education system (e.g., an interactive story reading system) (Wang et al., 2020, 2019), aiding parents and educators to formulate personalized questions during story readings, while addressing their language, knowledge, or time constraints. Also, fine-tuning LLMs (e.g., Llama 3) may lead to better performance on the QAG task, which offers a future direction to refine models' capabilities in real-world tasks like children's education.

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, however, this work is not without limitations. We could further explore LLMs' QAG capabilities with different models (e.g., GPT-4 and Llama 3), and domain-specific prompting methodologies, such as ICL with more demonstrations and RAG approaches (Edge et al., 2024) with multi-step generation pipelines. In addition, the size and scope limitations of expert annotations in our dataset may not be sufficient for developing NLP technologies that can be generalizable to similar scenarios. We call for future research to explore methods for scaling the data annotation process in real-world settings and to investigate strategies for efficiently optimizing or evaluating NLP technologies in low-resource scenarios where expert resources are scarce.

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.

Szeyi Chan, Shihan Fu, Jiachen Li, Bingsheng Yao, Smit Desai, Mirjana Prpa, and Dakuo Wang. 2024. Human and llm-based voice assistant interaction: An analytical framework for user verbal and nonverbal behaviors. *arXiv preprint arXiv:2408.16465*.

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.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

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.

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.

Jintao Ling and Muhammad Afzaal. 2024. Automatic question-answer pairs generation using pre-trained large language models in higher education. *Computers and Education: Artificial Intelligence*, 6:100252.

Yuxuan Lu, Bingsheng Yao, Shao Zhang, Yun Wang, Peng Zhang, Tun Lu, Toby Jia-Jun Li, and Dakuo Wang. 2023. Human Still Wins over LLM: An Empirical Study of Active Learning on Domain-Specific Annotation Tasks. *Preprint*, arXiv:2311.09825.

Amama Mahmood, Junxiang Wang, Bingsheng Yao, Dakuo Wang, and Chien-Ming Huang. 2023. Llm-powered conversational voice assistants: Interaction patterns, opportunities, challenges, and design guidelines. *arXiv preprint arXiv:2309.13879*.

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

Gene P Ouellette. 2006. What's meaning got to do with it: The role of vocabulary in word reading and reading comprehension. *Journal of educational psychology*, 98(3):554.

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.

Charles Perfetti and Joseph Stafura. 2014. Word knowledge in a theory of reading comprehension. *Scientific studies of Reading*, 18(1):22–37.

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.

Juan Ramos et al. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 29–48. Citeseer.

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, Jiaju Chen, Bingsheng Yao, Jiali Liu, 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.

Marjorie Taylor, Bonnie M Esbensen, and Robert T Bennett. 1994. Children's understanding of knowledge acquisition: The tendency for children to report that they have always known what they have just learned. *Child development*, 65(6):1581–1604.

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.

Yuwei Wan, Aswathy Ajith, Yixuan Liu, Ke Lu, Clara Grazian, Bram Hoex, Wenjie Zhang, Chunyu Kit, Tong Xie, and Ian Foster. 2024. Sciqag: A framework for auto-generated scientific question answering dataset with fine-grained evaluation. *arXiv preprint arXiv:2405.09939*.

Dakuo Wang, Elizabeth Churchill, Pattie Maes, Xiangmin Fan, Ben Shneiderman, Yuanchun Shi, and Qianying Wang. 2020. From human-human collaboration to human-ai collaboration: Designing ai systems that can work together with people. In *Extended abstracts of the 2020 CHI conference on human factors in computing systems*, pages 1–6.

Dakuo Wang, Justin D Weisz, Michael Muller, Parikshit Ram, Werner Geyer, Casey Dugan, Yla Tausczik, Horst Samulowitz, and Alexander Gray. 2019. Human-ai collaboration in data science: Exploring data scientists' perceptions of automated ai. *Proceedings of the ACM on human-computer interaction*, 3(CSCW):1–24.

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.

Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K Dey, and Dakuo Wang. 2024. Mental-llm: Leveraging large language models for mental health prediction via online text data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(1):1–32.

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.

Ziqi Yang, Xuhai Xu, Bingsheng Yao, Ethan Rogers, Shao Zhang, Stephen Intille, Nawar Shara, Guodong Gordon Gao, and Dakuo Wang. 2024. Talk2care: An llm-based voice assistant for communication between healthcare providers and older adults. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(2):1–35.

Bingsheng Yao, Guiming Chen, Ruishi Zou, Yuxuan Lu, Jiachen Li, Shao Zhang, Yisi Sang, Sijia Liu, James Hendler, and Dakuo Wang. 2024. More samples or more prompts? exploring effective few-shot in-context learning for llms with in-context sampling. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1772–1790.

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.

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*.

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

A Appendix

A.1 Sample Data of `StorySparkQA`

In Table 4 and 5, we present the sample data from `StorySparkQA`, which include expert-selected concepts, real-world knowledge triples, and created QA pairs.

Story Section:
At the time when the Tang dynasty reigned over the Middle Kingdom, there were master swordsmen of various kinds. Those who came first were the saints of the sword. They were able to take different shapes at will, and their swords were like strokes of lightning. ... They wore a hidden dagger at their side and carried a leather bag at their belt. By magic means they were able to turn human heads into flowing water. ...
Expert annotated QA pairs
Triple: (dagger, ***is a***, short sword) **Question:** ***What is*** a short sword called? **Answer:** A dagger.
Triple: (bag, ***is used for***, carrying things) **Question:** ***What is*** a bag ***used for***? **Answer:** A bag is used for carrying things.

Table 4: Example 1 of expert annotated data point in `StorySparkQA`.

A.2 Properties of Educational QA datasets

The key properties of educational QA datasets, including their number and type of sourcebooks, QA pairs, whether they contain external knowledge, annotators, annotation process, and data composition, are presented in Table 6.

A.3 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*.

Story Section:
... On hearing this the king walked to the window and stood for a few minutes with his back to the room, where the company of young men remained silent. Then he came back, his face white and stern. 'I tell you,' he said, 'and it is the solemn truth, that I would rather you had told me that the prince was dead, though he is my only son, than know that he would suffer such an injury without attempting to avenge it ...
Expert annotated QA pairs
Triple: (white, ***is a***, color) **Question:** ***What*** color is snow? **Answer:** A White.
Triple: (injury, ***is at location of***, hospital) **Question:** ***Where*** do you go if you get very hurt? **Answer:** You go to hospital if you get very hurt.

Table 5: Example 2 of expert annotated data point in `StorySparkQA`.

A.4 Distribution of Question Type

The distribution of question type in `StorySparkQA` is shown in Table 7.

A.5 Analysis of the Type of Generated External Knowledge

We calculated the real-world knowledge type distribution of experts' annotation and triples generated by T5-Large and GPT-4 with 5-shot ICL (the best-performing ones in automated evaluation). The results are shown in Table 8.

According to the result, the fine-tuned T5 model can generate real-world knowledge triples more aligned with experts' annotation, surpassing GPT-4, further proving that domain-specific fine-tuning provides the model with targeted knowledge and expertise that LLMs may lack.

A.6 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.

Dataset	# Books	# QA Pairs	External Knowledge	Annotator	Document Source	Annotation Process	Data Composition
StoryQA	148	38,703	Yes	Crowd-Sourced	Story books	Direct Annotation	1. Story Section 2. QA Pairs
FairytaleQA	278	10,580	No	Expert	Story books	Direct Annotation	1. Story Section 2. QA Pairs
EduQG	13	5,018	No	Expert	Text books	Direct Annotation	1. Source Documents 2. Questions and Answer Options
`StorySparkQA`	278	5,868	Yes	Expert	Story books	3-step Guided Annotation	1. Story Section 2. Real-world Knowledge Triples 3. QA Pairs

Table 6: Properties of existing datasets focusing on children education compared with our `StorySparkQA`.

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 7: Distribution of question types in `StorySparkQA`.

Relations	Experts' Annotation	T5-Large Fine-Tuned	GPT-4
is a	35.45%	47.67%	20.64%
has subevent	16.21%	16.44%	4.66%
is the antonym of	15.20%	7.95%	1.64%
is used for	8.78%	9.32%	35.25%
is at location of	7.53%	7.40%	2.19%
is capable of	5.18%	4.11%	9.77%
other	11.65%	6.85%	25.84%

Table 8: Comparison of real-world knowledge relation types across experts' annotations, fine-tuned T5-Large, and GPT-4.

A.7 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 9.

A.8 LLM Prompts

To utilize LLMs' strong reasoning and generation capability as well as control GPT-generated questions as much as possible to meet the needs of teachers, 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.

Models	Prompting Strategy	End2End Pipeline w/o Triples Rouge-L	End2End Pipeline w/o Triples SBERT	End2End Pipeline w/ Triples Rouge-L	End2End Pipeline w/ Triples SBERT
T5-Large Fine-Tuned (0.77B)	-	**0.332**	0.289	**0.279**	0.263
Alpaca (7B)	zero-shot	0.124	0.186	0.266	0.207
	1-shot	0.251	0.182	0.239	0.186
Mistral (7B)	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 (7B)	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 9: Rouge-L and SentenceBERT scores of LLMs in the QAG task. **Bolded numbers** are global best performance within each setting on each metric.

Table 10, 11 list our prompts for GPT in the two abovementioned approaches.

A.9 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.

Prompt for LLMs 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 story reading. 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: 1) causes, 2) desires, 3) has context of, 4) has property, 5) has subevent, 6) is a, 7) is at location of, 8) is capable of, 9) is created by, 10) is made of, 11) is part of, 12) is the antonym of, 13) 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 10: Prompt for LLMs in the QAG task with generating QA pairs directly from the story.

Prompt for LLMs 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: 1) causes, 2) desires, 3) has context of, 4) has property, 5) has subevent, 6) is a, 7) is at location of, 8) is capable of, 9) is created by, 10) is made of, 11) is part of, 12) is the antonym of, 13) 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 story reading. 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 11: Prompt for LLMs in the QAG task with generating real-world knowledge triple and QA pairs directly from the story.

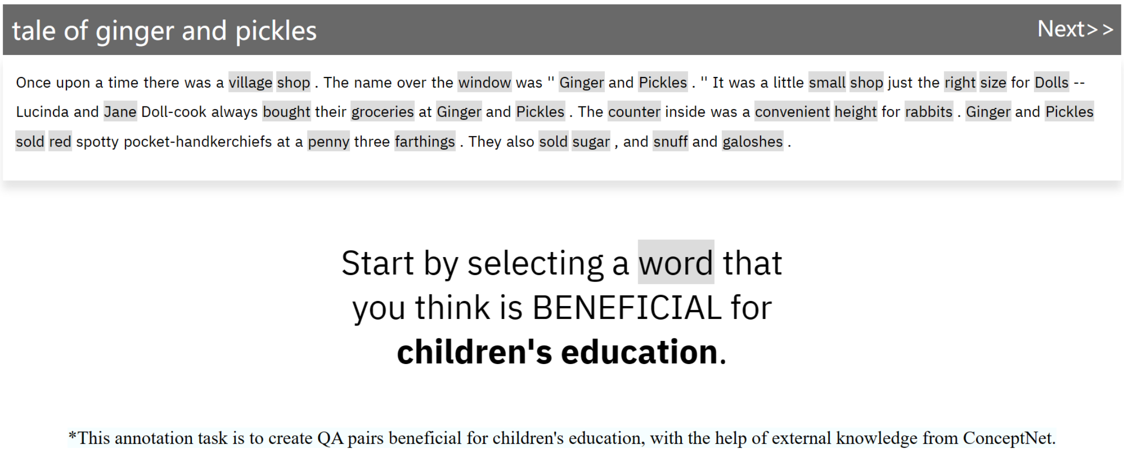

Figure 5: Annotation process 1: browse a displayed section, with candidate words highlighted in grey.

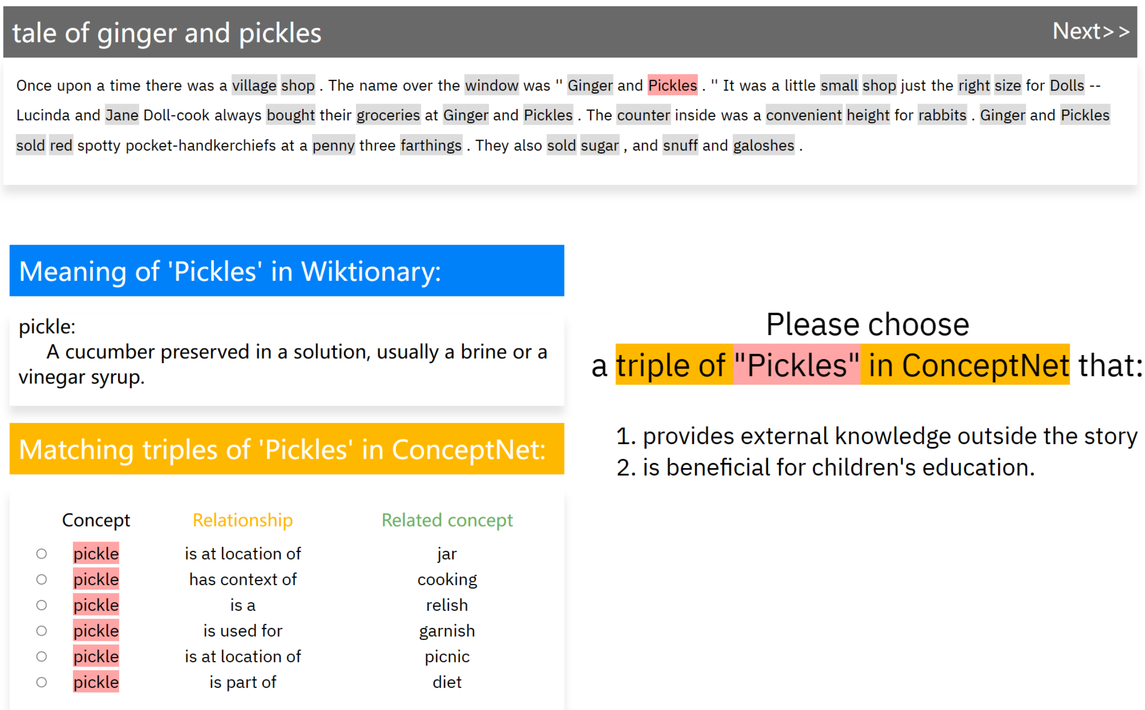

Figure 6: Annotation process 2: after selecting a word (highlighted in red), related explanation in Wiktionary and candidate real-world knowledge triples in ConceptNet will display.

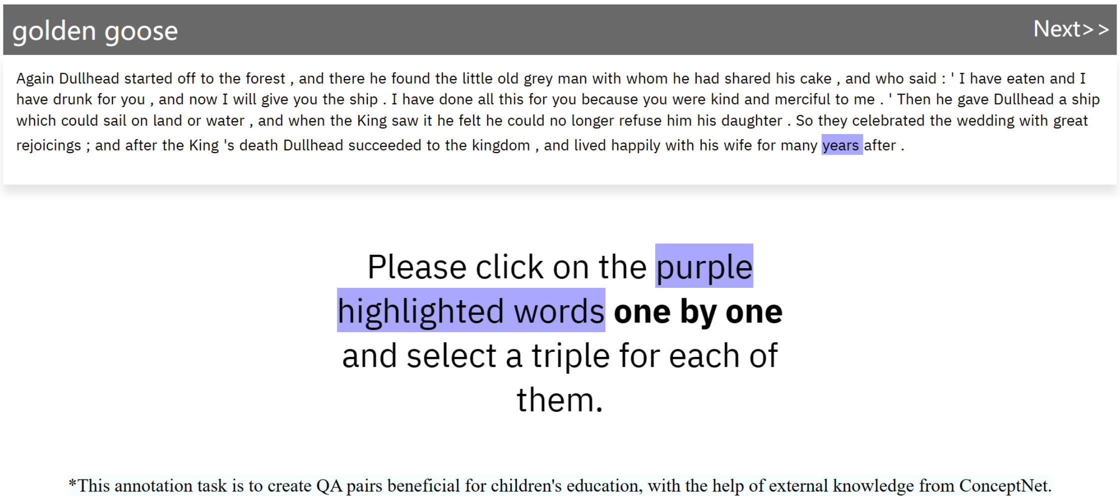

Figure 7: Cross-validation process 1: browse a displayed section, with candidate words highlighted in grey.

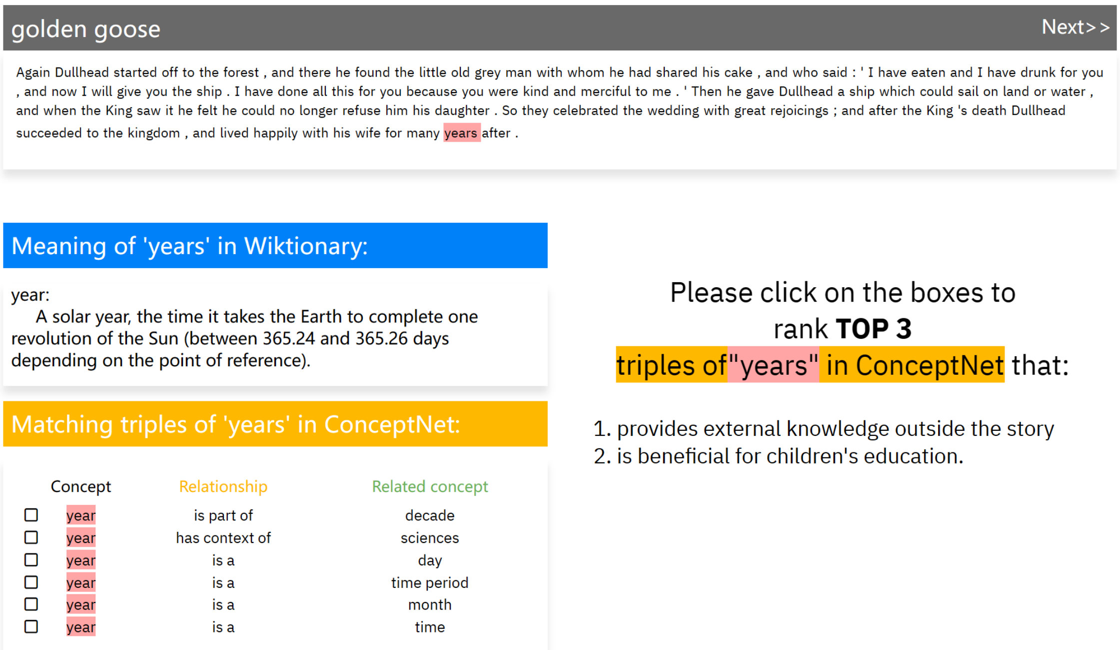

Figure 8: Cross-validation process 2: select a word annotated by others and rank the candidate triples.

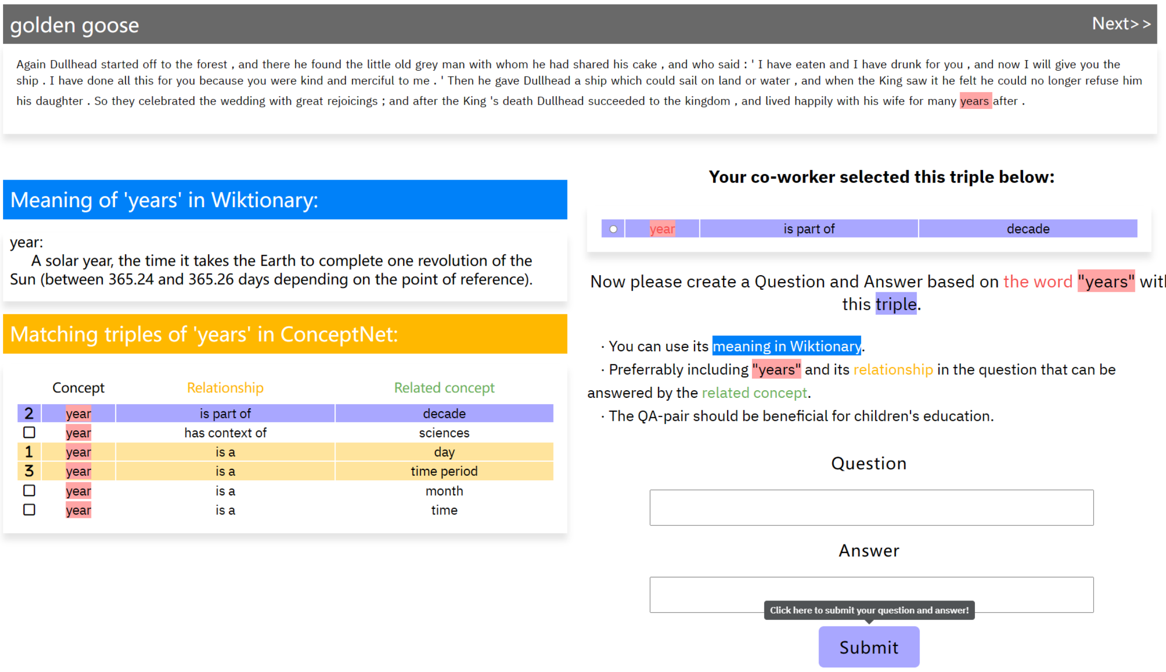

Figure 9: Cross-validation process 3: after ranking top3 triples, the triple selected originally by the other annotator is displayed, the validator should create a QA pair based on the original triple.

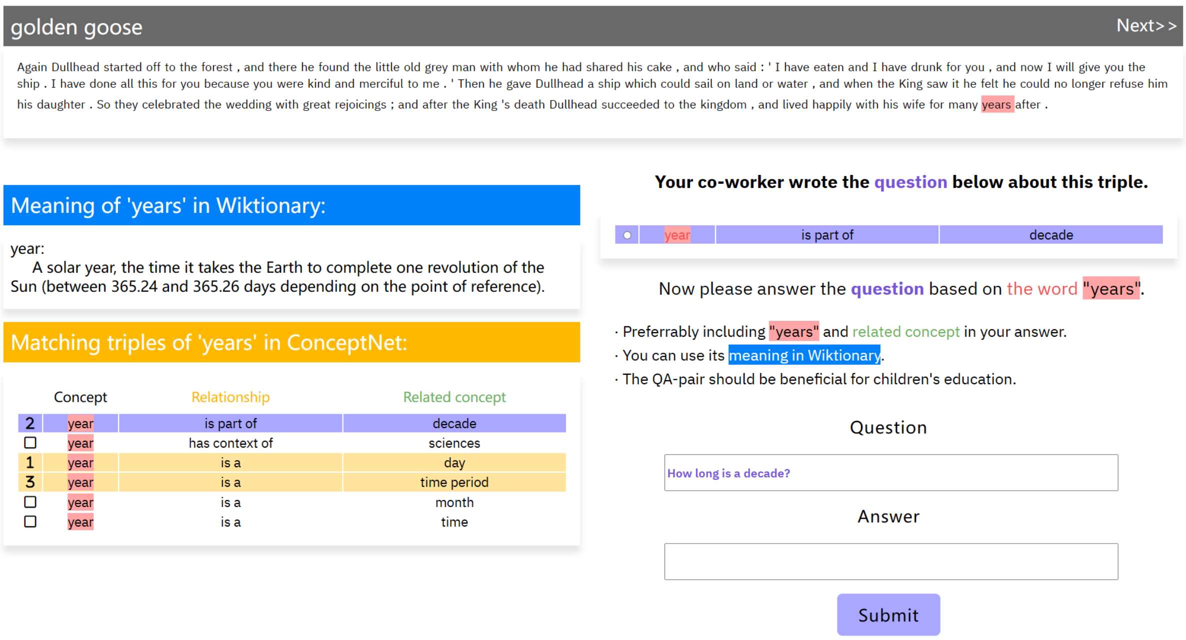

Figure 10: Cross-validation process 4: validator is asked to answer the question created by the other annotator using the triple originally selected by the other annotator.