KG + Narrative > LLM: Integrating a Commonsense Knowledge Graph with Children's Storybook Narratives

Anonymous ACL submission

Abstract

Structured knowledge such as Knowledge Graph (KG) has long been utilized by humans in real-world scenarios (e.g., clinical diagnosis and children's education) together with free-form narratives. Despite exceptional text generation ability, whether Large Language Models (LLMs) are adaptative to and well-performed in these specialized real-world tasks has been overlooked. **In the LLM era, is structured knowledge still useful for domain-specific tasks?** In this paper, we propose a new *interactive storytelling* task grounded in real-world needs: preschool teachers and parents educate children on real-world knowledge through questioning-answering (QA) beyond story narratives during storytelling. For this task, we 1) design an annotation framework to leverage established commonsense KG to enrich narrative QA, and 2) construct an expert-annotated `FairytaleCQA` dataset (5,868 QA-pairs) with external commonsense knowledge for evaluation. Our experiments show that: 1) expert-annotated structured knowledge can enhance LLMs' (e.g., GPT-4) performance; 2) our designed QAG pipeline can support a small fine-tuned LM to consistently outperform large LLMs on `FairytaleCQA`.

1 Introduction

Humans have spent considerable effort collecting and organizing structured knowledge, such as Knowledge Graphs (KGs), in various real-world tasks and scenarios (Vrandečić and Krötzsch, 2014; Lehmann et al., 2015). For example, in interactive storytelling, preschool teachers or parents commonly have structured knowledge in their minds and want to extend the story to associated real-world knowledge (Parish-Morris et al., 2013; Saracho, 2017) to enrich children's real-world perception, good moral qualities, etc. In another circumstance, clinicians employ structured rules and knowledge, such as medical protocols, when conducting clinical diagnosis for a patient (American Diabetes Association, 2011; ElSayed et al., 2023).

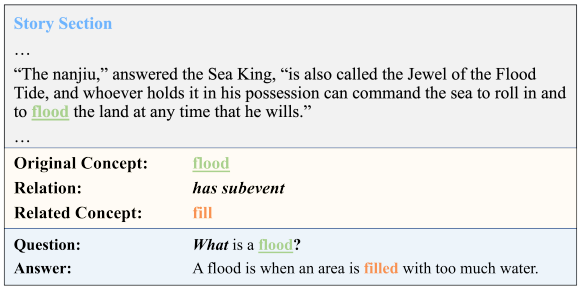

Figure 1: An example of `FairytaleCQA` dataset. In each story section, human educational experts select a concept word, link it to a desired external knowledge, and manually write an appropriate QA pair. Human annotators always prioritize educational appropriateness at each of the three steps.

In recent years, large language models (LLMs) such as GPT-3.5, GPT-4 (OpenAI, 2023), and Llama 2 (Touvron et al., 2023) have shown exceptional generation capability in various natural language generation (NLG) tasks (Robinson et al., 2022; Singhal et al., 2023). However, the performance of LLMs in real-world domain-specific tasks, where humans typically rely on structured knowledge (i.e., interactive storytelling), has been overlooked but holds significant importance.

In this paper, we propose a QA-pair generation (QAG) task grounded in real-world needs, where preschool teachers or parents want to extend story content to associated external commonsense knowledge during the *interactive storytelling* process. Despite existing AI-enabled storytelling systems (Shakeri et al., 2021; Zhang et al., 2022) have been increasingly utilized in supporting interactive storytelling activities, most are grounded in the story textual content (Xu et al., 2022), which does not faithfully facilitate parents need to incorporate external real-world knowledge during the storytelling process. Thus, these State-of-the-art (SOTA) systems have limited ability to generate QA-pairs associated with external knowledge.

To bridge the gap, we design and implement an annotation framework that retrieves and recommends structured commonsense knowledge from an external KG with carefully designed heuristics and recommendation algorithms. Leveraging our annotation framework, we then recruit 11 children's education experts to extend the fairytale stories from FAIRYTALEQA dataset and the resulting FairytaleCQA dataset [1] comprises 5,868 story-inspired QA-pairs associated with external commonsense knowledge.

We demonstrate the utility of structured knowledge with a comprehensive analysis of QAG experiments on different pipelines backed by SOTA LLMs, including GPT-3.5, GPT-4, and Llama 2 [2]. Specifically, we construct an end-to-end pipeline and a KG-assisted pipeline with expert-annotated structured knowledge as input to explore the usefulness of expert-annotated structured knowledge for LLMs. We carefully design the prompts with clear and informative instructions and compare the performance of robust SOTA LLMs in both zero-shot and few-shot In-Context Learning (ICL) settings. Automated evaluation and human evaluation on our FairytaleCQA dataset show that:

- **Human-prioritized knowledge (i.e., triplets) from external KG can elevate LLM performance in the domain-specific QAG task.**
- **Our carefully designed workflow can augment a fine-tuned small LM to outperform large LLMs (e.g., GPT-4) in real-world domain-specific tasks.**

2 Related Work

2.1 Structured Knowledge Source

Leveraging different structured external knowledge for constructing commonsense-related QA datasets (Talmor et al., 2018; Auer et al., 2023) has been widely explored and adopted. However, these datasets have limited relevance to children's education beyond story context. In addition, structured knowledge sources such as ATOMIC (Sap et al., 2019) and Wikidata (Vrandečić and Krötzsch, 2014) contain complex factual information, which might not be suitable for children's education. ConceptNet (Speer et al., 2017) is a vast graph widely used as an external knowledge source in NLP tasks (Bosselut et al., 2019; Xu et al., 2020). Knowledge in ConceptNet is represented in the simple triplet format of ($concept_1$, $relation$, $concept_2$) to support commonsense reasoning, aligning well with the need for children's education that the knowledge should be broad and not too tricky. Our work follows prior literature to use ConceptNet as our structured knowledge source to facilitate QA-pair annotation.

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
FairytaleCQA	278	5,868	Yes	Expert	Story books

Table 1: Properties of existing datasets focusing on children's education compared with our FairytaleCQA.

2.2 QA Datasets in the Educational Domain

General-purpose QA datasets, such as NarrativeQA (Kočiský et al., 2018) and SQuAD2.0 (Rajpurkar et al., 2018), primarily focus on crowdsourced QA-pairs grounded in texts, lacking the incorporation of external knowledge for enhanced comprehension and expertise in children's education. While QA datasets such as CommonsenseQA (Talmor et al., 2018) and SciQA (Auer et al., 2023)in the general domain contain commonsense, they usually lack appropriate context (e.g., fairytale stories) for QA-pairs to anchor on. Thus, these datasets are not appropriate to a specific scenario like children's education or a specific age group like children aged 3 to 6.

Targeting children's education, Zhao et al. (2023) propose StoryQA, a QA dataset containing out-of-context questions. Annotated by crowd workers with limited children's education knowledge and lacking structured external knowledge, this dataset potentially compromises the quality and consistency of generated QA-pairs for children's education. Experts-annotated QA datasets such as FAIRYTALEQA (Xu et al., 2022) and EduQG (Hadifar et al., 2023) center on story context, lacking out-of-context questions or external knowledge.

To meet parents' needs with an experts-labeled, large-scale QA dataset containing structured external knowledge, we propose FairytaleCQA. We summarize key properties of education-oriented QA datasets and FairytaleCQA in Table 1.

[1] We will release our dataset and code once our paper get accepted.

[2] We also experiment with 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 E.

2.3 QA-pair Annotation Frameworks

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, which support QA-pair annotation by providing source texts and allowing annotators to highlight a span of text as an answer to a question. Zhao et al. (2023) design a data collection user interface that allows annotators to type in answers in their own words. These aforementioned annotation frameworks are sufficient for document-grounded QA-pair annotation. However, without a structured form of external knowledge, annotators may have difficulty systematically incorporating external knowledge into anchored document text. Thus, an annotation framework that facilitates QA-pair annotations supported by coherent and structured external knowledge is essential.

2.4 QA-Pair Generation

Existing QAG methods could be broadly categorized into heuristics-based and neural network-based methods. Heuristics-based models (Yao, 2010; Labutov et al., 2015; Das et al., 2016) have more control over the generated QA-pairs, yet often lack diversity. Neural network-based methods (Zhou et al., 2018; Zhao et al., 2022) are more prevalent recently, with the rapid development of pre-trained language models (Devlin et al., 2019a; Liu et al., 2019). Yet, the generation qualities of neural network-based approaches highly depend on the training datasets, resulting in potential underperformance for domains requiring specific expertise, such as children's education.

Recent advances in LLMs (Chung et al., 2022; OpenAI, 2023) show exceptional natural language generation (NLG) capabilities. While conversational LLMs like GPT-4, and FLAN-T5 demonstrate superior zero-shot and few-shot in-context learning performance, their adaptability and performance in specialized domains, such as children's education, remain underexplored. We experiment with a series of QAG pipelines using SOTA LLMs to assess their performance thoroughly.

3 FairytaleCQA

FairytaleCQA aims to facilitate parents' storytelling process with structured knowledge. Our dataset consists of 5,868 QA-pairs annotated by children's education experts leveraging our specifically designed annotation framework. We present the core statistics of FairytaleCQA in Table 2 and show one example in Figure 1. Figure 2 illustrates the complete annotation process.

FairytaleCQA	Mean	St.D	Min	Max
# sections / story	14.7	9.2	2	60
# tokens per story	2196.7	1401.3	228	7577
# tokens / section	149.1	63.6	12	447
# questions / story	21.1	16.9	2	126
# questions / section	1.4	0.7	1	9
# tokens / question	5.4	1.7	3	19
# tokens / answer	4.9	2.3	1	20

Table 2: Core statistics of our FairytaleCQA dataset, which comprises 278 books and 5,868 QA-pairs.

3.1 Source Narrative

Plenty of excellent work (Xu et al., 2022; Zhao et al., 2023) has focused on creating high-quality text corpus for children's reading comprehension capabilities. Specifically, FAIRYTALEQA (Xu et al., 2022) comprises 278 classic fairytale stories from diverse origins, and all the stories have been evaluated as suitable for 10^{th}-grade children and younger. Afterward, the stories are parsed by children's 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. We build on prior work and take the story sections from FAIRYTALEQA as the source text for FairytaleCQA.

3.2 Annotation Framework

The ultimate goal of our annotation framework is to provide QA-pairs that **originate from the concepts in the stories and ask for associated external commonsense knowledge suitable for preschool children.** To better incorporate story texts with structured knowledge and facilitate experts' annotation process, we 1) develop carefully designed user interfaces, take the parent-children storytelling process into account, and 2) approach the annotation process by decomposing it into three steps:

1. **Concept Selection**: The first interface (Figure 5) displays one fairytale story section, and candidate concepts are highlighted in grey to select. Annotators need to identify a concept from the story that meets the following criteria: tier 1 or tier 2 (Beck et al., 2013) vocabulary and a concrete noun, verb, or adjective.
2. **Knowledge Matching**: In the second interface (Figure 6), annotators need to select commonsense knowledge based on the identified

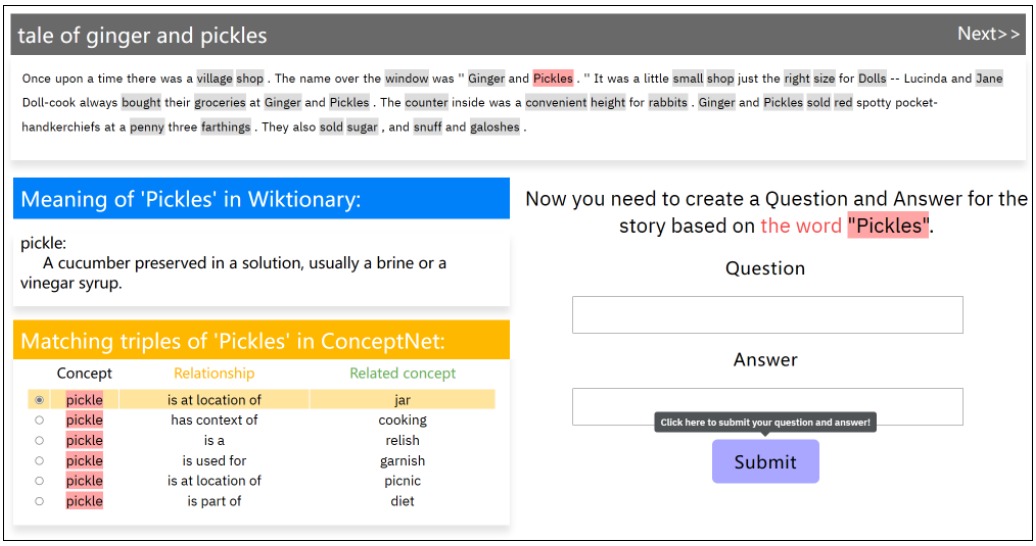

Figure 2: The user interface to facilitate our annotation task. The words highlighted in grey are candidate concepts. The blue block shows the Wiktionary explanation, and the yellow block lists our recommended triplets.

concept but goes beyond the story text.

3. **QA-Pair Creation**: The third interface (Figure 2) involves creating a QA-pair with either the question or answer containing the selected concept. The question should go beyond the stories' context and focus on the generated common-sense, fact-based knowledge.

To facilitate the annotation process by providing recommendations for external commonsense knowledge, we design our annotation framework by retrieving and recommending commonsense knowledge triplets from ConceptNet (Speer et al., 2017), a publicly available, large-scale commonsense Knowledge Graph.

We recruit 11 education experts experienced in preschool education for the annotation task. To ensure the created QA-pairs suit parents' real-world needs, experts are asked to mimic parents' habits during storytelling. Thus, for each story section, experts are asked to choose one or two concepts that are most beneficial for children's education from the text, according to parents' habit of asking questions in this storytelling scenario. When selecting commonsense triplets and creating QA-pairs, experts are asked to take children's cognitive and emotional levels into account and write QA-pairs that are most appropriate for 3-6-year-olds. Aligned with the user interface design demonstrated in Figure 2, we present the 3-step workflow of QA-pair annotation below, which follows Figure 3.

3.2.1 Step 1. Concept Selection

We develop a collection of heuristics to filter candidate concepts that meet the requirement for subtask 1 described in Section 3.2. First, we leverage the spaCy (Honnibal and Montani, 2017) to filter auxiliary words and punctuation[3] from the original text. Then, we use AllenNLP's semantic role labeling tool (Gardner et al., 2017) to tag the latent structure of each sentence in the story content. This process identifies and retains key elements represented by semantic roles , which are subsequently treated as potential candidate concepts.

3.2.2 Step 2. Knowledge Matching

Inspired by Xu et al. (2020)'s work of combining Wiktionary[4] and ConceptNet (Speer et al., 2017) for commonsense question answering, as well as filtering out weak relations in ConceptNet, we implement a knowledge matching module that can retrieve and rank the knowledge associated with each candidate concept in the source text.

More specifically, once the annotator selects a candidate concept, our knowledge matching module (1) retrieves a list of commonsense triplets, with the format of *(source concept, relation, target concept)* from ConceptNet as external knowledge; (2) filters out weak relations in ConceptNet, leaving 13 relation types for annotation (Complete relation list in Appendix B). In addition to the commonsense knowledge retrieval, we retrieve a sentence explanation from Wiktionary for each candidate concept to support expert annotations.

The second step in our knowledge matching module is to rank and select diverse and representa-

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

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

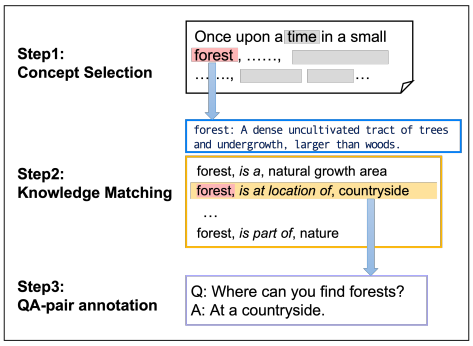

Figure 3: Workflow of the experts' annotation process. Experts need to select a concept first, then match it with the most suitable knowledge and finally create a QA-pair based on the selected knowledge.

tive triplets from all retrieved commonsense triplets associated with the selected concept. We use the concatenation of the relation and related concept in each triplet to calculate the average similarity between every other retrieved triplet using the Term Frequency-Inverse Document Frequency (TF-IDF).

We rank all retrieved triplets with $1 - \overline{s} + w$, where $\overline{s}$ denotes the similarity score and w denotes the weight of a triplet provided by ConceptNet, reflecting the combined influence and credibility of the triplet by summing up the weights coming from all the sources that support it. We recommend the top six ranked triplets to annotators to balance providing a sufficient selection and avoiding excessive distractions during the annotation task.

3.2.3 Step 3. QA-Pair Creation

Annotators need to create QA-pairs based on selected commonsense triplets. For each triplet, annotators are instructed to incorporate one concept in the question or answer and include the relation from the triplet in the resulting QA-pair.

3.3 Cross-Validation

The consistency between annotators' triplet selection and QA-pair creation is accessed through cross-validation. Details of our cross-validation are explained in Appendix A. Out of 100 randomly selected sections in the validation and test splits, 86% of the triplets that appear in the top-3 list are selected by both annotators, and 56% of the triplets are ranked top by the validator, indicating very high consistency between experts for triplet selection.

In addition, we evaluate the similarity of two QA-pairs (the question and answer are concatenated for the evaluation) created by two annotators based on the identical triplet with Rouge-L. The Rouge-L F1 score of QA-pair creation between annotators is 0.53, which shows a shared tendency among experts when it comes to selecting commonsense knowledge and creating a QA-pair that is both beneficial and appropriate for children's education. This observation reinforces the necessity of experts' annotation in constructing a high-quality QA dataset for children's education.

3.4 Statistics of FairytaleCQA

Figure 4 demonstrates the distribution of commonsense relations in the dataset, and Table 2 illustrates detailed statistics of the dataset. On average, each section is annotated with approximately 1.4 QA-pairs. In FairytaleCQA, the top 3 commonsense 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 commonsense relations. *is used for*, *is at location of* and *is capable of* each constitute 8.8%, 7.5%, and 5.2% of all commonsense relations. The proportion of other relations is less than 5%. The distribution of question types in FairytaleCQA is shown in Table 6. In FairytaleCQA, questions start with 'what', the most common question type, constituting 86.0%. Questions starting with 'why' and 'how' constitute about 7.2% and 2.4%, respectively.

According to experts' annotation, commonsense relation *is a* and 'what' questions have a much higher proportion than other relations and questions. Children aged 3-6, in the exploration stage and highly curious about the world (Chouinard et al., 2007; Jirout and Klahr, 2012), naturally tend to ask questions to satisfy their curiosity. Accordingly, parents are more inclined to use 'what' questions to inspire children's thinking and promote active knowledge acquisition (Yu et al., 2019). Consistent with parents' preferences, the fact that experts' annotated questions share a high consensus of 'what' questions is more in line with children's learning and cognitive characteristics.

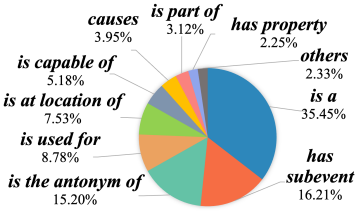

Figure 4: Distribution of commonsense relations annotated by experts in the FairytaleCQA dataset

4 Experiment

We investigate the utility of structured knowledge regarding the following research questions:

- **RQ1**: Is structured knowledge still useful in domain-specific tasks?
- **RQ2**: Can a much smaller model fine-tuned with KG-supported annotations beat generic LLMs in a specialized domain?

We approach RQ1 by conducting QA-pair generation (QAG) experiments with SOTA LLMs in an end-to-end (baseline) pipeline with carefully curated prompts and a KG-assisted pipeline, where we provide human-annotated triplets to LLMs. To further investigate RQ2 as well as demonstrate the usefulness of our dataset, we fine-tune a T5-Large model with our `FairytaleCQA` and comprehensively evaluate its QAG performance with LLMs across three carefully designed pipelines that simulate the human expert workflow.

The evaluation comprises six SOTA 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 H) with clear and informative instructions, including 13 relation types (Appendix B) in ConceptNet. The goal is to leverage LLMs to generate diverse triplets similar to those created by human education experts.

For both experiments, we utilize Rouge-L (Lin, 2004) to evaluate the quality of the concatenated QA-pairs between the generated ones and two expert-annotated ground-truths for each data, and report the averaged score across all test data. We perform experiments with GPT-3.5 and GPT-4 three times for each setting to calculate a robust and reliable averaged score. Additional scores of sentence similarity using Sentence Transformer (Reimers and Gurevych, 2019) are shown in Appendix E; however, we believe this metric can not faithfully represent the domain specialty with a generic evaluation model. As a result, We conduct a human evaluation to further evaluate the quality of QA-pairs generated by LLMs and experts' annotation from the educational perspective.

4.1 End2End QAG vs. KG-Assisted QAG (RQ1)

To investigate the utility of structured knowledge in this domain-specific QAG task, we carefully design two distinct QAG pipelines. For each LLM involved in this experiment (Llama 2, GPT-3.5, and GPT-4), we employ both **zero-shot** and **few-shot in-context learning (ICL)** (Wei et al., 2022a) approaches to thoroughly examine the QAG performance of SOTA LLMs for our specific QAG task, where we randomly sample examples from the validation split from `FairytaleCQA` as demonstrations for the few-shot ICL approaches. We fine-tune a T5-Large model for each pipeline to examine how a much smaller domain-specific model, supported by human-annotated triplets as additional input, performs compared to generic LLMs. The experiment settings and hyper-parameters can be found in Appendix D.

End-to-end QAG pipeline The system generates QA-pairs directly from a story section, serving as the baseline. To exploit LLMs' comprehensive generation ability and simulate experts' annotation process, we design two end-to-end variations:

1. *w/o triplets*: Directly generate QA-pairs from the input text (baseline).
2. *w/ triplets*: Generate a commonsense triplet alongside the QA-pair.

KG-assisted QAG pipeline We provide expert-annotated commonsense triplets for each story section as input guidance to examine the usefulness of structured knowledge in this specialized QAG task. The prompt inputs are shown in Table 18.

4.1.1 Experiment Results

We report the performance of the aforementioned LLMs with each proposed pipeline in Table 3 and report the complete results in Table 7 in Appendix, including LLMs that perform worse than GPT-4, such as Mistral-7B (Jiang et al., 2023). Examples of GPT-generated QA-pairs, as well as experts' annotations on the same section, can be found in Appendix G. Across all the end-to-end pipelines for each LLM, the 5-shot ICL pipeline consistently outperforms both zero-shot and 1-shot ICL pipelines.

For the end-to-end setting that asks LLMs to generate triplets along with QA-pairs, we can observe improvements on Llama 2 and GPT-3.5 with the ICL approach but also observe lower performance with GPT-4. This observation may imply that GPT-4 is already equipped with enough knowledge to generate QA-pairs, and the triplet generation requirement posts a negative effect on the QAG task. To explore GPT-4's strong generation capabilities under the end-to-end setting, we utilize the Chain-of-Thoughts (Wei et al., 2022b) prompting to facilitate this specialized QAG task on GPT-4. Nevertheless, the result in Table 3 does not witness an obvious improvement compared with the

Pipeline	Category	T5-Large fine-tuned (0.77B)	Llama 2 (7B)	GPT-3.5 (175B)	GPT-4 (1,760B)
End2End pipeline (w/o triplets)	zero-shot	0.332	0.213	0.194	0.277
	1-shot	-	0.192	0.239	0.272
	5-shot	-	0.241	0.262	0.287
End2End pipeline (w/ triplets)	zero-shot	0.279	0.177	0.220	0.243
	1-shot	-	0.206	0.252	0.251
	5-shot	-	0.269	0.264	0.248
	CoT	-	-	-	0.271
KG-assisted pipeline	10-shot	**0.510** (zero-shot)	**0.470**	**0.541**	**0.527**

Table 3: QAG performance of the end-to-end and KG-assisted pipelines with LLMs. LLMs are provided with structured knowledge annotated by experts in the KG-assisted pipeline.

ICL approach. It is worth noting that with the assistance of this structured knowledge, all LLMs as well as the domain-specific fine-tuned language model can far exceed the end-to-end pipeline in the QAG task, which justifies that **human-annotated structured knowledge is still useful in such real-world domain-specific tasks**.

4.2 Domain Fine-tuned T5-Large vs. LLMs (RQ2)

We further investigate the performance of a small model fine-tuned with domain-specific knowledge compared with generic LLMs without expert annotation. To establish robust LLM baselines and harness the full potential of their reasoning and natural language generation capabilities, we design a 2-step and 3-step pipeline in addition to the end-to-end pipeline by mimicking the experts' annotation workflow. For each multi-step pipeline, we also fine-tune a T5-Large model on `FairytaleCQA` for each step and utilize the model output for the previous step (e.g., generated triplets) as part of the input for the next model (e.g., generate QA-pairs given the story content and generated triplets).

2-step QAG pipeline The pipeline consists of two steps: (1) generates an external commonsense triplet given the story content, and (2) generates QA-pairs with the input of the generated triplets.

3-step QAG pipeline Mimicking the experts' annotation process, this pipeline comprises three steps: (1) selects a concept from the story first, (2) creates the corresponding commonsense triplet based on the selected concept, and (3) generates QA-pairs based on the generated triplet.

We select the best-performing LLMs in the end-to-end pipeline from the previous experiment, namely GPT-3.5, GPT-4 and Llama 2 for the 2-step QAG pipeline. Given the strong QAG capability of GPT-3.5 shown in the 2-step pipeline, we further conduct our 3-step pipeline through GPT-3.5 and strictly limit the number of QA-pairs generated in each section in the prompt (Table 19, 20).

Models	Category	End2End w/o triplets	End2End w/ triplets	2-step pipeline	3-step pipeline
T5-Large fine-tuned (0.77B)	zero-shot	**0.332**	**0.279**	**0.279**	**0.290**
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	-	-
	few-shot	0.241	0.269	0.263	-
GPT-3.5 (175B)	zero-shot	0.194	0.220	-	-
	few-shot	0.262	0.264	**0.279**	0.282
GPT-4 (1,760B)	zero-shot	0.277	0.243	-	-
	few-shot	0.287	0.251	0.271	-

Table 4: Rouge-L scores of generated QA-pairs using the T5-Large fine-tuned model and LLMs across end-to-end, 2-step and 3-step pipelines. **Bolded numbers** are global best performance within each setting.

4.2.1 Experiment Results

We present the models' performance in Table 4 and examples of generated results for each pipeline is shown in Appendix G. The system evaluation of the T5-Large model fine-tuned on our `FairytaleCQA` consistently outperforms generic LLMs across all pipelines by Rouge-L. This observation justifies that **a smaller language model assisted with domain expertise as well as structured knowledge can reliably perform better than generic LLMs in domain-specific scenarios.**

Comparing the models' performance across pipelines, the overall system's performance of the 2-step pipeline exhibits a slight enhancement compared to the end-to-end pipeline.

We attribute this to the challenge of creating commonsense triplets as properly and accurately as experts in the first step, as experts rely on structured external knowledge source ConceptNet to create QA-pairs. In other words, the domain experts exhibit much better "timing" of when and where to provide and incorporate structured knowledge, whereas generic LLMs fall short of this nuanced mental behavior in terms of domain-specific tasks.

The Rouge-L scores of the fine-tuned T5-Large and GPT-3.5 in the 3-step pipeline are both better than those of the 2-step pipeline, indicating that

Dimension	Human	T5-Large fine-tuned	GPT-4	p-value
Grammar Correctness	**4.893**	4.843	4.871	0.209/0.519
Answer Relevancy	**4.696**	4.329	4.379	<0.01
Contexual Consistency	**4.657**	4.639	4.529	<0.05
Educational Appropriateness	**4.493**	4.325	4.318	<0.01

Table 5: The human evaluation results of experts' annotation, GPT-4 and T5-Large fine-tuned on FairytaleCQA in the end-to-end pipeline setting.

using structured external knowledge like human annotators does assist the model in performing better on both identifying a concept and selecting a commonsense triplet. This result justifies the validity of our proposed annotation framework. By infusing structured knowledge with free-form narrative, the framework facilitates both domain-specific language models and LLMs. We believe this method of incorporating structured knowledge with free-form narrative can be applicable in similar tasks but of different specialized domains.

4.3 Human Evaluation

To comprehensively investigate the helpfulness of structured knowledge in the specific QAG task of children's education, we further conduct a human study to compare the generated QA-pairs.

More specifically, according to the superior performance of T5-Large fine-tuned on FairytaleCQA and GPT-4 with 5-shot ICL approach in an end-to-end pipeline, we select these two models along with experts' annotation for human evaluation. We randomly select ten story books from the test split of FairytaleCQA, and sample seven sections per book. For each section, there are three QA-pairs 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 education experts are asked to evaluate each QA-pair on the following four dimensions with a 5-point Likert scale:

1. *Grammar Correctness*: The QA-pair is in readable English grammar and words;
2. *Answer Relevancy*: The answer is correct corresponding to the question;
3. *Contextual Consistency*: The QA-pair originates from the story and goes beyond the context;
4. *Children's Educational Appropriateness*: The QA-pair is appropriate in young children's reading experience of interactive storytelling;

Table 5 illustrates the average scores in each dimension and the detailed paired sample *t-test* results are shown in Table 9 in the Appendix. We observe that experts-created QA-pairs outperform those generated by both GPT-4 and fine-tuned T5-Large model approaches on all four dimensions. Our paired sample *t-tests* shows that experts' annotation has significant differences in three out of four dimensions compared with models' generation. This justifies the utility of our FairytaleCQA.

For the *Contexual Conisistancy* dimension, in which we assess whether a QA-pair is both associated with story contexts and external commonsense knowledge, the fine-tuned T5-Large significantly outperformed GPT-4, behind experts' annotations.

For the *Children's Educational Appropriateness* dimension, the T5-Large model fine-tuned on FairytaleCQA also exhibits better performance than GPT-4. This result suggests that fine-tuned with KG-supported expert annotation, the T5-Large model can benefit from the assistance of structured knowledge as well as experts' domain-specificity.

Therefore, benefiting from experts' annotation assisted by structured knowledge, the fine-tuned T5-Large is capable of generating QA-pairs that 1) contain external structured knowledge, and 2) are appropriate for young children's interactive storytelling experience. The performance of T5-large fine-tuned on FairytaleCQA also proves that our proposed annotation framework can effectively infuse structured knowledge with free-form narrative, facilitating similar tasks in other specific domains.

5 Conclusion and Future Work

In summary, we present a real-world scenario where structured knowledge is needed to facilitate interactive storytelling. We collected a QA dataset, namely FairytaleCQA, for children's education by leveraging a novel annotation framework that facilitates scalable expert annotations using structured external knowledge. Our bi-fold experiments investigate the utility of structured knowledge and LLMs performance in domain-specific tasks.

One possible future work entails refining the structure of the QAG model structure, using LLMs to generate QA-pairs that align more closely with the actual needs of parents. Another future direction involves using FairytaleCQA and model to develop a human-AI education system, aiding parents and early educators in formulating questions during story readings, and addressing their language, knowledge, time, or motivation constraints.

6 Limitations

This work primarily focuses on investigating the usefulness of structured knowledge in domain-specific tasks. Our experiment with domain experts-annotated dataset solely utilizes a T5-Large model to generate QA-pairs. However, we are aware that the performance of other models such as BERT (Devlin et al., 2019b), BART (Lewis et al., 2019), etc. is to be further explored.

In this work, we try to comprehensively utilize LLMs generation capabilities in QAG; thus, we designed three QAG pipelines to investigate the performance of LLMs on these pipelines. Nevertheless, we can further experiment with more LLMs and explore more ICL approaches with each LLM. This is intended to enhance the generation of QA-pairs that are better suited for children's education. In addition, based on LLM's performance within each generation step, we can use the combination of well-performed LLMs to construct a robust QAG system in this scenario.

Besides, in the knowledge matching module of our annotation framework, we currently focus on commonsense representations involving two concepts and a relation. The incorporation of meta-paths connecting multiple concepts is an area that is still to be 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.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense Transformers for Automatic Knowledge Graph Construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

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.

Rubel Das, Antariksha Ray, Souvik Mondal, and Dipankar Das. 2016. A rule based question generation framework to deal with simple and complex sentences. *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 542–548. Conference Name: 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI) ISBN: 9781509020294 Place: Jaipur, India Publisher: IEEE.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

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.

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.

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.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA Reading Comprehension Challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328. Place: Cambridge, MA Publisher: MIT Press.

Igor Labutov, Sumit Basu, and Lucy Vanderwende. 2015. Deep Questions without Deep Understanding. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 889–898, Beijing, China. Association for Computational Linguistics.

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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]*.

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

Julia Parish-Morris, Neha Mahajan, Kathy Hirsh-Pasek, Roberta Michnick Golinkoff, and Molly Fuller Collins. 2013. Once Upon a Time: Parent–Child Dialogue and Storybook Reading in the Electronic Era. *Mind, Brain, and Education*, 7(3):200–211.

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.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

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.

Joshua Robinson, Christopher Michael Rytting, and David Wingate. 2022. Leveraging Large Language Models for Multiple Choice Question Answering. Publisher: arXiv Version Number: 3.

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.

Olivia N. Saracho. 2017. Parents' shared storybook reading – learning to read. *Early Child Development and Care*, 187(3-4):554–567.

Hanieh Shakeri, Carman Neustaedter, and Steve DiPaola. 2021. SAGA: Collaborative Storytelling with GPT-3. In *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '21, pages 163–166, New York, NY, USA. Association for Computing Machinery.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. 2023. Towards Expert-Level Medical Question Answering with Large Language Models. Publisher: arXiv Version Number: 1.

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).

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*.

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. 2022a. Chain of Thought Prompting Elicits Reasoning in Large Language Models. *ArXiv*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022b. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

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, 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.

Xuchen Yao. 2010. Question Generation with Minimal Recursion Semantics.

Yue Yu, Elizabeth Bonawitz, and Patrick Shafto. 2019. Pedagogical questions in parent–child conversations. *Child Development*, 90(1):147–161.

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, Dublin, Ireland. Association for Computational Linguistics.

Qingyu Zhou, Nan Yang, Furu Wei, Chuanqi Tan, Hangbo Bao, and Ming Zhou. 2018. Neural Question Generation from Text: A Preliminary Study. *Natural Language Processing and Chinese Computing*, 10619:662–671. Book Title: Natural Language Processing and Chinese Computing ISBN: 9783319736174 9783319736181 Place: Cham Publisher: Springer International Publishing.

Appendix

A Cross-Validation

To validate and ensure the quality of annotated QA-pairs across annotators and to assess the agreement of triplet selection and QA-pair creation between annotators, we implement additional user interfaces for the cross-validation process. We randomly selected 50 QA-pairs in each 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 $annotator_B$, correspondingly):

1. Shown in Figure 7, $annotator_A$ is provided the story section and the concept selected by $annotator_B$. For each selected concept, $annotator_A$ is asked to rank the top 3 triplets from the same recommended triplet list given to $annotator_B$, verifying the triplet selection agreement between annotators (Figure 8).
2. In the next step, $annotator_A$ is asked to create a QA-pair based on the word and triplet selected by $annotator_B$, evaluating the similarity of QA-pairs between annotators given the identical triplet (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 triplet, and $annotator_A$ is asked to write an answer to the question to cross-validate the question-answering agreement (Figure 10).

B 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*.

C Distribution of Question Type

The distribution of question type in `FairytaleCQA` is shown in Table 6.

D 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.

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 6: Distribution of question types in `FairytaleCQA`.

E 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 7 and 8.

Models	Categroy	End2End Pipeline w/o triplets		End2End Pipeline w/ triplets	
		Rouge-L	Sent Similarity	Rouge-L	Sent Similarity
T5-Large fine-tuned	zero-shot	**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
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	-	-	0.271	0.270

Table 7: Rouge-L and Sentence Similarity scores of LLMs on end-to-end pipeline. **Bolded numbers** are global best performance within each setting on each metrics.

F Human Evaluation Results

Table 9 demonstrates the paired sample t-Test results on our designed four dimensions. Three out of four dimensions exhibits significant difference.

G Examples of Generated QA-pairs

We randomly sample a section, and the generated QA-pairs of each pipeline can be found in Table 10, 11, 12, 13, 14. For the end-to-end pipeline, only the generation results of models with the best automatic evaluation results are demonstrated here.

Models	Categroy	2-step Pipeline		3-step Pipeline	
		Rouge-L	Sent Similarity	Rouge-L	Sent Similarity
T5-Large fine-tuned	zero-shot	**0.279**	0.263	**0.290**	**0.289**
Llama 2	10-shot	0.263	0.247	-	-
GPT-3.5	10-shot	**0.279**	**0.293**	0.282	0.247
GPT-4	10-shot	0.271	**0.293**	-	-

Table 8: Rouge-L and Sentence Similarity scores of LLMs on 2-step and 3-step pipelines. **Bolded numbers** are global best performance within each setting on each metrics.

H GPT Prompts

In order to utilize GPT's strong reasoning and generation capability as well as control GPT-generated questions as much as possible meets the needs of parents, we carefully design our prompts for GPT-3.5 and GPT-4.

For end-to-end 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 commonsense triplet and a QA-pair based on the triplet.

Table 15, 16 list our prompts for GPT in the two abovementioned approaches.

For 2-step pipeline, we first ask GPT to generate a commonsense triplet from a provided story section, and then we ask GPT to generate a QA-pair based on a triplet. Table 17, 18 show our prompts for the two steps, respectively.

We add an additional step before commonsense triplet generation in 3-step pipeline. For a provided story section, we ask GPT to identify a keyword in the text, then generate a commonsense triplet based on that keyword. Our prompts for the two steps are shown in Table 19 and 20. The final step is the same as the second step in 2-step pipeline.

I User Interface for Annotation System

We implement an annotation system to facilitate QA-pair annotation with associated external knowledge. Figure 5, 6 and 2 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.

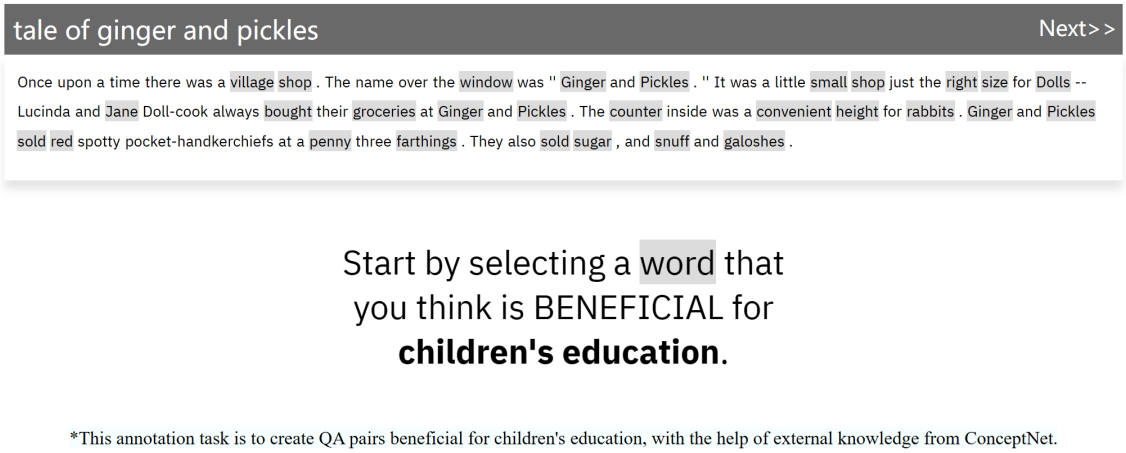

Figure 5: Annotation process1: Browse a displayed section, with candidate words highlighted in grey.

Dimension	Model	Mean	St.D	t	df	p-value
Grammar Correctness	Human	4.893	0.560			
	GPT-4	4.871	0.514	0.646	349	0.519
	T5-Large fine-tuned	4.842	0.585	1.259	349	0.209
Answer Relevancy**	Human	4.696	0.683			
	GPT-4	4.379	0.869	5.123	279	<0.01
	T5-Large fine-tuned	4.329	1.111	5.487	279	<0.01
Contexual Consistency*	Human	4.657	0.882			
	GPT-4	4.529	0.974	2.240	279	0.026
	T5-Large fine-tuned	4.639	0.972	5.487	279	0.729
Educational Appropriateness**	Human	4.493	0.892			
	GPT-4	4.318	2.974	3.113	279	<0.01
	T5-Large fine-tuned	4.325	0.972	2.937	279	<0.01

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

Table 9: The paired sample t Test result of human annotators in comparison of GPT-4 and T5-Large fine-tuned on `FairytaleCQA` in an end-to-end QAG setting.

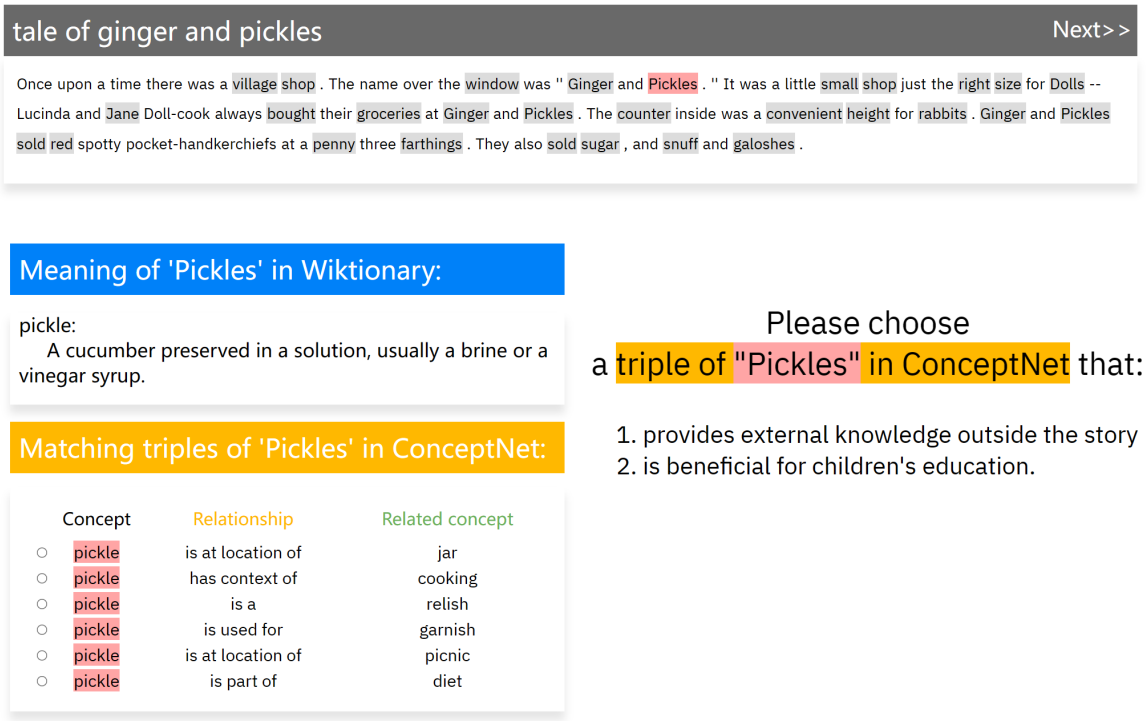

Figure 6: Annotation process2: After selecting a word (highlighted in red), related explanation in Wiktionary and candidate commonsense triplets in ConceptNet will display.

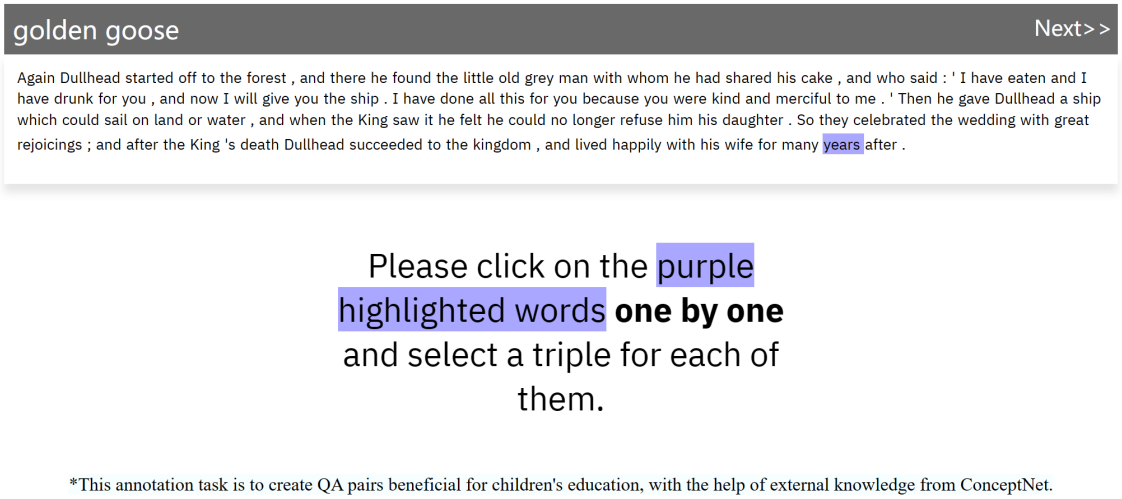

Figure 7: Cross-validation process1: Browse a displayed section, with candidate words highlighted in grey.

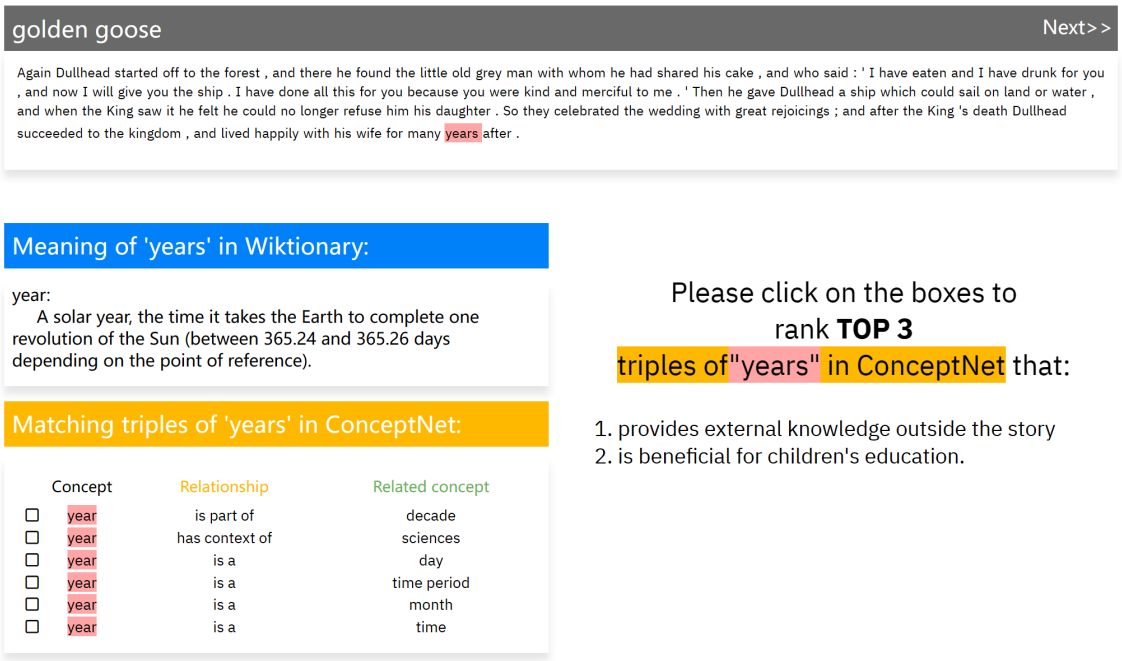

Figure 8: Cross-validation process2: Select a word annotated by others and rank the candidate triplets.

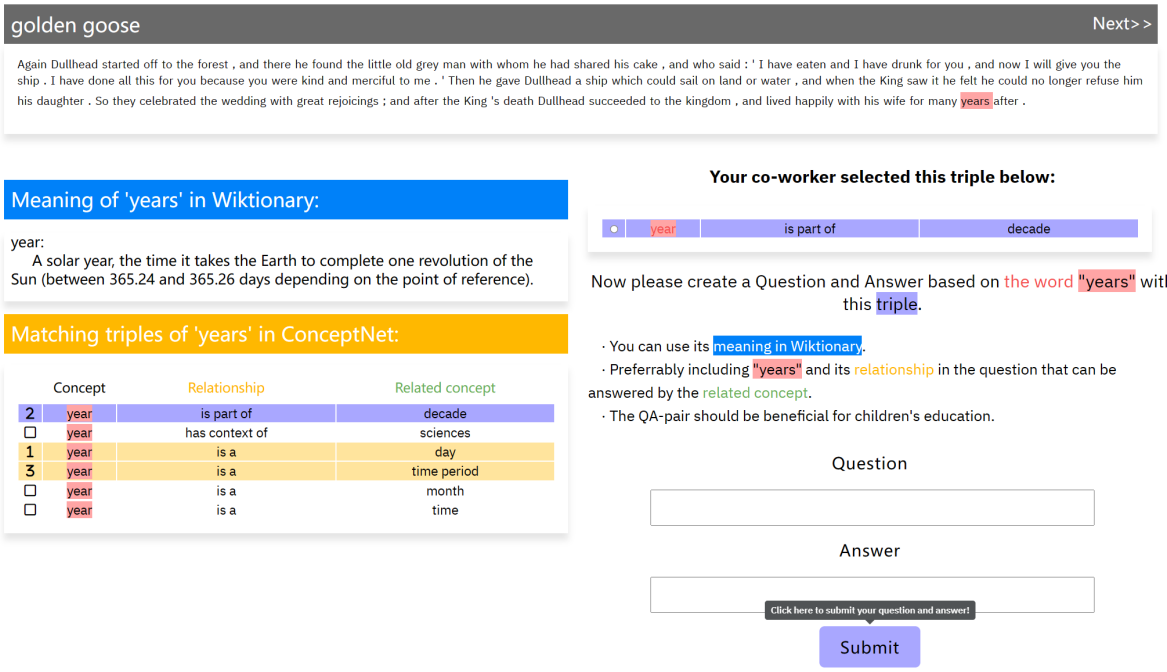

Figure 9: Cross-validation process3: After ranking top3 triplets, the triplet selected originally by the other annotator is displayed, the validator should create a QA-pair based on the original triplet.

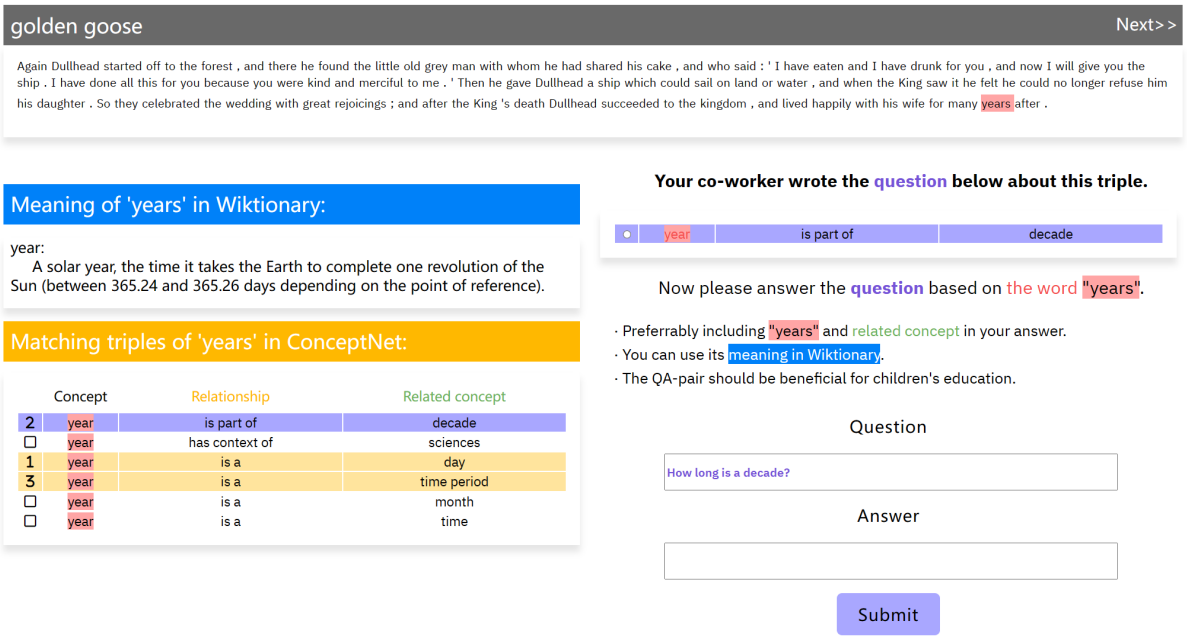

Figure 10: Cross-validation process4: Validator is asked to answer the question created by the other annotator using the triplet originally selected by the other annotator.

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.
...

Human annotated QA-pairs
Triplet: (bag, ***is used for***, carrying things) **Question:** ***What is*** a bag ***used for***? **Answer:** A bag is used for carrying things.
Triplet: (dagger, ***is a***, short sword) **Question:** ***What is*** a short sword called? **Answer:** A dagger.

Table 10: A randomly selected story section and two experts annotated QA-pairs.

End-to-end QAG pipeline (generate QA-pairs only)
GPT-4 5-shot **Question:** What is a dagger used for? **Answer:** Cutting.
T5-Large Fine-tuned **Question:** What is a sword? **Answer:** A sword is a long-bladed weapon.

Table 11: Generated QA-pairs using end-to-end QAG pipeline on GPT-4 and T5-Large without generating commonsense triplets. The QA-pairs are generated based on the story section in Table 10

End-to-end QAG pipeline (generate triplets and QA-pairs)
GPT-4 5-shot **Triplet:** (dagger, is used for, protection) **Question:** What can a dagger be used for? **Answer:** Protection.
T5-Large Fine-tuned **Triplet:** (dagger, is a, knife) **Question:** What is a dagger? **Answer:** A dagger is a knife.

Table 12: Generated QA-pairs and commonsense triplets using end-to-end QAG pipeline on GPT-4 and T5-Large. The QA-pairs are generated based on the story section in Table 10

2-step QAG pipeline
GPT-3.5 10-shot **Triplet:** (sword, is used for, fighting) **Question:** What is a sword used for? **Answer:** Fighting.
GPT-4 10-shot **Triplet:** (sword, is used for, protection) **Question:** What might someone use a sword for? **Answer:** Protection.
T5-Large Fine-tuned **Triplet:** (dagger, is a, knife) **Question:** What is a dagger? **Answer:** A dagger is a sharp knife.

Table 13: Generated QA-pairs and commonsense triplets using 2-step QAG pipeline on GPT-3.5, GPT-4 and T5-Large. The QA-pairs are generated based on the story section in Table 10

3-step QAG pipeline
GPT-3.5 1-shot **Triplet:** (lightning, is a, electrical discharge) **Question:** What is lightning? **Answer:** Electrical discharge.
T5-Large Fine-tuned **Triplet:** (bag, is used for, carrying things) **Question:** What do people use bags for? **Answer:** People use bags for carrying things.

Table 14: Generated QA-pairs and commonsense triplets using 3-step QAG pipeline on GPT-3.5 and T5-Large. The QA-pairs are generated based on the story section in Table 10

Prompt for GPT in end-to-end 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 common-sense, fact-based knowledge. The common-sense, 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 common-sense 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 15: Prompt for GPT in end-to-end QAG approach with generating commonsense triplet.

Prompt for GPT in end-to-end pipeline
generate triplets 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 common-sense relation based on the selected word. This common-sense relation should go beyond the context of the stories. For example, if your identified word is 'apple', your common-sense relation could be: apple grows on trees; apples are red. The common-sense, 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 common-sense 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 common-sense, 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.
5. After this, select one question-answer pair that you think best meet 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 common-sense knowledge and selected question-answer pair in the following format:
commonsense: (A, relation, B)
question: ...
answer: ...

⟨story ⟩:
{story1 for few-shot}

⟨response ⟩:
{response1 for few-shot}
... ...

⟨story ⟩:
{story for the current data}

⟨response ⟩:

Table 16: Prompt for GPT in end-to-end QAG approach with generating commonsense triplet.

Prompt for GPT in 2-step pipeline: Step 1

I need you to help generate commonsense knowledge for young children aged three to six. The commonsense knowledge you should write can be seen as a relation about two concepts. 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 common-sense, fact-based knowledge.
For example, if your identified word is 'apple', your commonsense relation could be: apple is a fruit; apple is used for eating.
The common-sense, 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 common-sense 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

Return the generated common-sense knowledge in the following format:

commonsense: (A, relation, B)

⟨story ⟩:
{story1 for few-shot}

⟨response ⟩:
{response1 for few-shot}
... ...

⟨story ⟩:
{story for the current data}

⟨response ⟩:

Table 17: Prompt for step 1 in GPT 2-step QAG approach.

Prompt for GPT in 2-step pipeline: Step 2

I need you to help generate a question and answer pair for young children aged three to six. I will provide you with a piece of commonsense knowledge. Please follow these steps:
1. Based on provided commonsense knowledge, generate a question and answer pair that either the question or the answer contains a concept in the commonsense knowledge.
The questions should be focused on commonsense, fact-based knowledge.
For example, given the commonsense knowledge of 'apple is used for eating', your question could be: what is apple used for?
Each question should have one single correct answer that would be the same regardless of the children's experiences. The answer should only be a concrete noun, verb, or adjective.
Return the generated question-answer pair in the following format:

question: ...
answer: ...

⟨commonsense knowledge ⟩:
{commonsense knowledge1 for few-shot}

⟨response ⟩:
{response1 for few-shot}
... ...

⟨commonsense knowledge ⟩:
{commonsense knowledge generated by GPT in Step 1 for the current data}

⟨response ⟩:

Table 18: Prompt for step 2 in GPT 2-step QAG approach.

Prompt for GPT in 3-step pipeline: Step 1

I need you to help identify a key word from a story text for young children aged three to six. The key word should be able to expand as commonsense knowledge. I will provide you with a short section of a story delimited by triple quotes, and candidate words in this section. Please follow these steps:
1. For all the candidate words in this section, identify a 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.
Return three identified key word in the following format:

key word:

⟨story ⟩:
{story1 for few-shot}
⟨candidate words ⟩:
{candidate words1 for few-shot}

⟨response ⟩:
{response1 for few-shot}
... ...

⟨story ⟩:
{story for the current data}
⟨candidate words ⟩:
{candidate words for the current data}

⟨response ⟩:

Table 19: Prompt for step 1 in GPT 3-step QAG approach.

Prompt for GPT in 3-step pipeline: Step 2

I need you to help generate commonsense knowledge based on a key word for young children aged three to six. The commonsense knowledge you should write can be seen as a relation about two concepts.
I will provide you with one key word identified in a story, and for each key word, I will provide you with six commonsense knowledge as candidate triples in the form of a triple such as 'A relation B', where A and B are two concepts and the key word can be either A or B. Please follow these steps:
1. Based on each selected word and candidate triples, choose one triple as a common-sense, fact-based knowledge that is best for children's education.
For example, if your key word is 'apple', your commonsense relation could be: (apple, is a, fruit); (apple, is used for,eating); (apple, is, sweet); (apple, has property, red); (apple, is at location of, trees); (apple, is used for, apple_pie).
Return one generated common-sense knowledge in the following format:

commonsense: (A, relation, B)
⟨key word⟩:
{key word1 for few-shot}
⟨candidate triples⟩:
{candidate triples1 for few-shot}

⟨response⟩:
{response1 for few-shot}
... ...

⟨key word⟩:
{key word for the current data}
⟨candidate triples⟩:
{candidate triples for the current data}

⟨response⟩:

Table 20: Prompt for step 2 in GPT 3-step QAG approach.