Exploring Domain Adaptation with LLMs for Real-World Augmented Question Answer Generation (RA-QAG) in Children Storytelling

Anonymous ACL submission

Abstract

In the real world, external domain-specific knowledge is commonly required, for instance, teachers often apply their expertise to ask preschoolers educational-crafted, story-inspired questions beyond the story content during interactive storytelling; however, existing storytelling systems could not effectively support such activity as the generated questions are mostly text-based. We formulate this type of common real-world application as a novel **Real-World Augmented QAG (RA-QAG)** task. This work aims to explore how well LLMs, equipped with various domain adaptation strategies (e.g., few-shot In-Context Learning, Chain-of-Thoughts, Retrieval-Augmented Generation), perform on the RA-QAG task in the context of children storytelling. We design and experiment with end-to-end and 2-Step QAG pipelines with different domain adaptation strategies to explore whether they can identify real-world knowledge and create QA pairs aligned with experts' annotation. Our automatic evaluation and human evaluation show that 1) RAG is a promising direction to approach real-world domain-specific tasks; 2) human experts still have more nuanced knowledge from which generic LLMs need to learn.

1 Introduction

Generic Large Language Models (LLMs) such as GPT-3.5, GPT-4 (OpenAI, 2023), and Llama 2 (Touvron et al., 2023) exhibit strong capabilities to generate various types of text solely using a snippet of narrative as input data, such as narrative question-answering (Shao et al., 2023a) and text summarization (Zhang et al., 2024). However, real-world scenarios are much more complicated: they commonly require additional domain-specific knowledge, which is often not present in the narrative but mastered by domain experts.

In this paper, we focus on a real-world scenario in the context of preschool children education: teachers often try to teach real-world knowledge while reading a storybook with a child through interactive question-and-answer activities (Xu et al., 2021). These teachers need to decide **where to ask** the question so that it is relevant to the story (i.e., anchored at a particular word in the narrative), **what to ask** so that the child can learn better (i.e., external knowledge goes beyond the story narrative), and **how to ask** the question (i.e., the style and difficulty) so that the question is engaging but not frustrating. The questions and answers crafted by human teachers while reading a storybook also need to align with children's cognitive development levels at different ages (Parish-Morris et al., 2013; Saracho, 2017; Xu et al., 2021). We formulate such type of real-world application as a novel narrative question-answer generation (QAG) task, namely **Real-World Augmented QAG (RA-QAG)**, that requires external domain knowledge to solve. The RA-QAG task for children storytelling, differs from traditional QAG tasks (Yao et al., 2021), demands models to 1) locate particular words in story narratives and link them to educationally appropriate knowledge, and 2) create children-centered and knowledgeable QA pairs.

Existing children storytelling systems (Shakeri et al., 2021; Zhang et al., 2022), despite showing effectiveness in supporting interactive storytelling, are mostly grounded in a story's textual content, thus leading to limited capability in the RA-QAG task (Yao et al., 2021). The recent advancement in large language models (LLMs), which encountered and learned profound world knowledge during the pre-training process, aroused significant attention in investigating LLMs' capabilities for real-world domain-specific tasks. Moreover, various types of generation strategies with the shared purpose of enhancing LLMs' domain-adaptation capabilities have emerged recently, including purely prompting-based strategies like few-shot In-Context Learning (ICL) (Brown et al., 2020) and Chain-of-

Thoughts (Wei et al., 2022a), as well as Retrieval Augmented Generation (RAG) (Lewis et al., 2020) based strategies, which retrieves domain-specific knowledge from external knowledge resources as additional guidance for LLMs. Nevertheless, how well the State-of-The-Art (SoTA) language models, equipped with a variety of domain adaptation strategies, perform on the RA-QAG task in the context of children education domain, remains underexplored. In addition, whether these models will come close to the performance of human domain experts, and if not, how close they will come, is also unknown but of significant practical implication for both the technical and educational communities.

Our primary contribution in this work aims to address the aforementioned "known unknowns" of LLMs' domain adaptation capabilities on the RA-QAG task of teacher-children storytelling we formulated using a recently published dataset, namely `FairytaleCQA` (Chen et al., 2023). We designed and experimented with different QAG strategies under **end-to-end** and **two-step** generation pipelines, where the latter one mimics the experts' annotation process to come up with a knowledge triple first before creating the QA pairs. The end-to-end pipeline comprises both traditional compact models fine-tuned on the training split of `FairytaleCQA` as well as LLMs supported by different prompting strategies (e.g., zero-shot, few-shot ICL, CoT), whereas the two-step pipeline consists of three strategies for Triple Selection: generated by LLM, trained retriever (RAG), and expert-annotated triples (Human). Our comprehensive benchmark experiment, along with in-depth analysis, reveals two critical findings:

- Fine-tuning models with expert annotations and leveraging RAG can enhance the pipeline's QAG quality.
- Experts' annotation remains more effective in elevating LLM performance on domain-specific tasks like interactive storytelling.

We further discuss the potential and limitations of RAG strategies for RA-QAG tasks in different real-world scenarios.

2 Related Work

2.1 Large Language Models for Domain Adaptation

Recent advancements in large language models (LLMs), including notable models like GPT-3.5, GPT-4 (OpenAI, 2023), and Llama (Touvron et al., 2023), have demonstrated exceptional capabilities in generating coherent and contextually relevant text. Many prompting techniques have been proposed recently to further enhance LLMs' task-solving and domain-adaptation capability without tuning the model parameters, such as few-shot In-Context Learning, Chain-of-Thought prompting, etc. Nevertheless, recent work discovered that LLMs' adaptability and performance in specialized domains off-the-shelf, such as in children education and mental health (Xu et al., 2023), is commonly compromised due to limited domain-specific knowledge (Cao et al., 2020; He et al., 2022).

Recently, RAG has emerged as a novel and promising domain-adaptation approach that retrieves external information as guidance to generate more up-to-date, accurate, and reliable responses (Lewis et al., 2020; Izacard et al., 2022) in various tasks (Izacard and Grave, 2021; Cai et al., 2019). The retrieval module in RAG aims to extract the most helpful external knowledge, which could be supported by a traditional trainable retriever model like BM25 (Robertson and Zaragoza, 2009) and BERT (Devlin et al., 2019), or advanced embedding models like BGE (Chen et al., 2024). In this work, we leverage BGE as the retriever model for RAG throughout the experiments.

2.2 The `FairytaleCQA` Dataset

General QA datasets such as NarrativeQA (Kočiský et al., 2018), SQuAD2.0 (Rajpurkar et al., 2018), CommonsenseQA (Talmor et al., 2018), and SciQA (Auer et al., 2023) either only consist of text-grounded, crowd-sourced QA pairs or fall short at considering children education appropriateness with incorporated knowledge, leading all these datasets less suitable for the QAG task augmented by real-world knowledge in the context of children education.

`FairytaleCQA` (Chen et al., 2023) is a recently published large-scale QA dataset annotated by children experts, specifically designed for children's interactive storytelling activities. This dataset contains $5,868$ QA pairs derived from children's fairytale stories and enriched with external real-world knowledge from ConceptNet (Speer et al., 2017), a wide-used knowledge graph of structured real-world knowledge. Such integration of story content with real-world knowledge appropriate for children education turns `FairytaleCQA` an ideal data resource for the tasks that require external domain knowledge to solve, and as a result, we utilize `FairytaleCQA` as the benchmark dataset for our proposed RA-QAG task.

3 Real-World-Augmented QA Pair Generation for Children Storytelling

In this section, we present the details of our experimental setup and methodology for the RA-QAG task we formulate from the real-world interactive storytelling scenario. We designed and experimented with different QAG strategies, including the utility of experts' domain knowledge annotation for RAG, with a particular focus on the following detailed research questions (RQ):

- **RQ1**: Can LLMs perform better by emulating human experts' QA pair creation process?
- **RQ2**: Can a compact model fine-tuned with experts' annotations outperform LLMs?
- **RQ3**: To what extent RAG can enhance LLMs' domain adaptability, and what is the gap remaining compared with human experts?

3.1 Dataset Preprocessing for Fine-tuning

`FairytaleCQA` contains 5,868 QA pairs from 278 children's fairytale books. Each QA pair in `FairytaleCQA` is grounded in a concept from the story text and a corresponding external knowledge triple from ConceptNet (Speer et al., 2017), which represent a real-world knowledge in the format of $(concept_1, relation, concept_2)$, annotated by children's educational experts.

The fine-tuning process of our retriever model in two-step QAG pipeline, as described in Section 3.2.2, and the compact model for end-to-end generation follows the original train/validation/split of `FairytaleCQA`, comprising $4,300/769/799$ QA pairs, accordingly. We leverage supervised contrastive training (Khosla et al., 2020) to fine-tune the retriever model by creating an equal amount of negative examples paired with positive examples – expert-annotated real-world knowledge triples for each story section. Negative examples are randomly sampled from the whole dataset excluding the expert annotations in the positive examples. The input for the retriever model becomes the concatenation of a story section and its corresponding positive and negative examples.

3.2 RA-QAG Pipelines

Catering to teachers' common practice during the interactive storytelling activity, our RA-QAG experiments aim to come up with the final artifacts of QA pairs that are associated with a particular concept in the story and related external knowledge, all of educational appropriateness for children education. We designed two types of QAG pipelines: 1) **end-to-end (E2E)** pipeline directly generates the QA pairs with no intermediate outputs, and 2) **two-step (2-step)** pipeline that simulates experts' annotation process for `FairytaleCQA`. Multiple generation strategies were further designed under each type of QAG pipeline. Figure 1 illustrates the structure of both pipelines, with additional illustrations of QAG strategies under the 2-Step pipeline.

3.2.1 End-to-end QAG Pipeline (E2E)

The end-to-end pipeline generates both the story-inspired real-world knowledge and a corresponding QA pair directly given the input of a story section. We follow the instructions provided to human experts to create `FairytaleCQA` as the basic prompt instructions, and collaborate with human educational experts to iteratively refine the prompt instructions for LLMs to incorporate educational guidelines as well as QAG instructions, as reported in Appendix A.2. The goal is to ask LLMs to generate diverse triples and corresponding QA pairs that are appropriate for interactive storytelling activity with children.

We leverage five robust LLMs for end-to-end generation, including GPT-4 (OpenAI, 2023), Llama 2 (Touvron et al., 2023), Mistral (Jiang et al., 2023), Alpaca (Taori et al., 2023) and FLAN-T5-XXL (Chung et al., 2022). For each LLM, we experiment with the following popular prompting strategies: zero-shot as baseline (denoted as ZS), few-shot In-Context Learning (denoted as FS(# OF SHOT), and Chain-of-Thoughts (denoted as COT). In addition, we fine-tune a traditional T5-Large model with expert-annotated triples and QA pairs for the end-to-end pipeline to compare the performance between domain-specific fine-tuned compact model and generic LLMs.

3.2.2 Two-Step QAG Pipeline (2-Step)

The 2-Step pipeline aims to mimic the expert annotation process in constructing `FairytaleCQA`, which is also aligned with existing works (Yao et al., 2021; Qu et al., 2021) for multi-step QAG generation. During the original annotation pro-

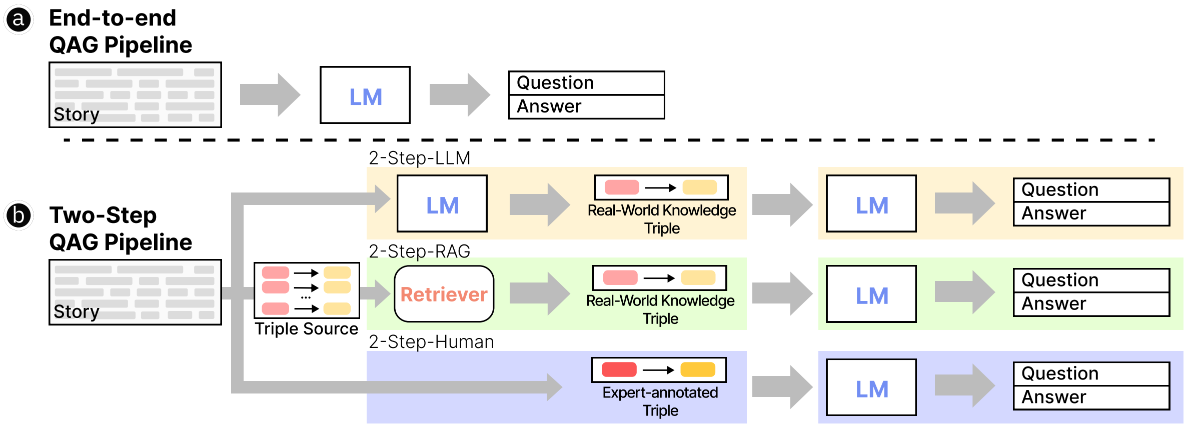

Figure 1: The structure of the end-to-end pipeline and the three QAG strategies under 2-Step pipeline, including the 2-STEP-LLM, 2-STEP-RAG and 2-STEP-HUMAN strategies.

cess, educational experts select a story-related real-world knowledge triple beyond the story narrative, assisted with a retrieval algorithm that recommends the most relevant real-world knowledge triples from ConceptNet. Afterward, the experts write a QA pair based on the selected knowledge triple. In both steps, the experts are explicitly instructed to consider the educational appropriateness of selecting triples as well as creating QA pairs.

Similarly, The 2-Step pipeline consists of the following steps: 1) generates an external real-world knowledge triple based on the story context, and 2) explicitly uses the generated triple as additional input to create the corresponding QA pair. We aim to investigate the effectiveness of RAG in supporting LLMs to retrieve external knowledge triples in the 2-Step QAG pipeline, and, as a result, we design three different knowledge triple generation strategies, including generated by LLMs directly, retrieved from ConceptNet via a trained retriever model, and directly using expert-annotated triples. Details of each strategy are illustrated below.

LLM Strategy (2-STEP-LLM) This is the basic strategy that asks the LLMs to only generate a real-world knowledge triple based on the input story in the first step, then feed the generated triple as additional input to the LLMs for the generation of a QA pair in the second step. For this strategy, we incorporated GPT-4 and Llama 2 as the LLM variations and also fine-tuned a T5-Large model for each step on `FairytaleCQA` as the compact model alternative.

Trained Retriever Strategy (2-STEP-RAG) This strategy represents the RAG approach for LLMs, where we attempt to mimic the two-step annotation process of human experts. Specifically, we follow the same process as reported in `FairytaleCQA` to locate associated external knowledge triples for concepts in stories. Firstly, we generate the list of candidate concepts from the story content. Then, we acquire the top six related knowledge triples from ConceptNet for every candidate concept as the external knowledge resource for the current story input. A retriever model was trained on the training split of `FairytaleCQA` to select the most relevant and helpful knowledge triple annotated by human experts, given the story content and the external knowledge resource. Once the retriever returns a triple, we ask the LLM to generate a corresponding QA pair, which is identical to the other 2-Step pipelines.

We incorporate two versions of the BGE model as the retriever: the original BGE, and the other one fine-tuned with `FairytaleCQA`. For fine-tuning the BGE embedding model, we process the data as described in Section 3.1, and leverage the BGE model to calculate the similarity between the embeddings of the story text and each suggested real-world knowledge triples.

Expert-Annotated Strategy (2-STEP-HUMAN) We also design and experiment with the 2-Step-Human strategy as the upper bound for 2-Step QAG pipelines by directly using the expert-annotated triples for the first step. For the second step of QA pair generation, we also trained a T5-Large model as the fine-tuned compact model variation. We aim to compare the performance of the aforementioned 2-Step QAG strategy with this expert knowledge-based strategy to investigate the gap remaining between human experts' knowledge and the RAG-enhanced LLMs.

Model	Prompting Strategy	Rouge-L (Triple)	Rouge-L (QA pair)
T5-Large fine-tuned (0.77B)	-	0.206	**0.279**
Llama 2 (7B)	ZS FS	0.154 <u>0.291</u>	0.177 <u>0.269</u>
GPT-4	ZS FS CoT	0.286 0.285 **0.295**	0.243 0.248 0.262

Table 1: QAG performance of LLMs and the fine-tuned T5-Large in the E2E QAG pipeline. We use 5-shot for both few-shot ICL methods. **Bolded numbers** are the best scores within each setting, and underlined numbers are the second-best scores within each setting.

4 Evaluation

Following the experiment setting described in Section 3, we conduct evaluations on both QAG pipelines. We carefully designed the prompt inputs by incorporating the instructions provided to the human experts for `FairytaleCQA`, and emphasized the educational appropriateness for generated QA pairs, as shown in Appendix A.2.

For automatic evaluation, we utilize **Rouge-L** (Lin, 2004) to evaluate the quality of the concatenated QA pairs between the generated ones and two ground truths annotated by experts, then report the average score across all test data. We also measure **Sentence-BERT** (SBERT) using Sentence Transformer (Reimers and Gurevych, 2019) and report the scores in Appendix A.1. We acknowledge that these similarity-based metrics cannot faithfully measure domain specificity, therefore, we conducted a **human evaluation** with education experts to further assess the quality of LLM-generated QA pairs from an educational perspective.

4.1 RQ1: E2E QAG vs. 2-Step QAG

We approach the RA-QAG task using two aforementioned QAG pipelines described in Section 3.2: end-to-end (E2E) and two-Step (2-Step). Our end-to-end pipeline 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). To thoroughly examine LLMs' performance in the QAG task for interactive storytelling of children, we employed various popular prompting approaches, including zero-shot, few-shot In-Context Learning (ICL)(Brown et al., 2020), and Chain-of-Thought (CoT)(Wei et al., 2022a).

4.1.1 Experiment Results and Analysis

We report the performance of the aforementioned LLMs with the end-to-end pipeline in Table 1, report the two-step pipeline results in Table 2, and report the complete results in Table 4 in Appendix A.1, including LLMs that perform worse than GPT-4, such as Alpaca and Mistral-7B (Jiang et al., 2023).

In the end-to-end pipeline, models with the 5-shot ICL approach consistently outperform those utilizing the zero-shot approach. To harness GPT-4' full potential under the end-to-end setting, we apply the Chain-of-Thoughts (Wei et al., 2022b) prompting strategy for this specialized QAG task, where we guide GPT-4 to identify real-world knowledge and create QA pairs like human experts. Table 1 illustrates that GPT-4 achieves superior performance in the end-to-end pipeline by asking it to "think step-by-step" (CoT), which simulates humans' thinking process.

Subsequently, we conduct a two-step QAG pipeline evaluation, where all language models are asked to locate and link a real-world knowledge triple from the story first, and then generate a corresponding QA pair. Overall, as shown in Table 2, the two-step QAG pipeline demonstrates superior performance compared to the end-to-end pipeline. This result justifies that by emulating human experts' real-world knowledge triple identification and QA pair creation process, LLMs can generate more educationally appropriate QA pairs.

However, despite the superior performance of the two-step QAG pipeline, we observed that the improvement is inconspicuous, particularly for the 2-STEP-LLM and 2-STEP-RAG strategies, where all models are not directly assisted by human expertise. Notably, both models in the first step of the two-step pipeline exhibit better performance than in the second step. We attribute this to two main challenges: 1) It is hard for models to create real-world knowledge triples as properly and accurately as human experts in the first step, as experts rely on structured external knowledge source ConceptNet to identify and associate real-world knowledge. 2) When generating QA pairs that integrate the real-world knowledge triple created in the first step, the quality of the QA pairs is affected by the appropriateness of the created triple. In addition, the second step would suffer from more loss

Strategy	Model: Step1-Triple	Rouge-L	Model: Step2-QA pair	Rouge-L
2-Step-LLM	T5-Large Fine-Tuned	0.331	T5-Large Fine-Tuned	0.279
	Llama 2	0.311	Llama 2	0.263
	GPT-4	0.290	GPT-4	0.269
2-Step-RAG	BGE	0.298	GPT-4	0.256
	BGE Fine-Tuned	0.328	GPT-4	0.278
2-Step-Human	Experts' Annotation (Ground Truth)	**1.000**	T5-Large Fine-Tuned	**0.510**
			Llama 2	0.413
			GPT-4	0.482

Table 2: The 2-Step QAG pipeline performance on both steps, including the 2-Step-LLM, 2-Step-RAG and 2-Step-Human strategies. For all LLMs involved in this setting, we used 5-shot ICL for the corresponding generation step. **Bolded numbers** are the best scores within each setting, and underlined numbers are the second-best scores within each setting.

in terms of using suitable vocabulary for 3-6-year-olds' comprehension as properly as human experts. 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.

4.2 RQ2: Domain-Specific Fine-tuned Models vs. LLMs

To investigate whether language models can learn from domain experts' knowledge, we compare the performance of a compact model fine-tuned with domain-specific knowledge retrieved by experts with generic LLMs without experts' annotation. Specifically, for both the end-to-end and 2-Step pipeline, we fine-tune a T5-Large model on `FairytaleCQA` to generate a real-world knowledge triple and QA pair simultaneously. For each step in the 2-Step-LLM strategy, we fine-tune a T5-Large model on `FairytaleCQA` and utilize the model output for the previous step (i.e., generated triples) as part of the input for the next model (i.e., generate QA pairs given the story content and generated triples). In addition, we fine-tune a BGE retriever to retrieve real-world knowledge triples based on story sections for the 2-Step-RAG strategy.

4.2.1 Experiment Results and Analysis

We present the models' performance in Table 1 and 2. The fine-tuned T5-Large system consistently outperforms generic LLMs across both the end-to-end pipeline and the 2-Step-LLM strategy by Rouge-L scores. In the 2-Step-RAG strategy, the fine-tuned BGE model also exhibits better performance when retrieving story-relate and educational-suitable triples compared to the original BGE model.

To thoroughly investigate the real-world knowledge triple identified in the first step of the 2-Step pipeline, we also investigate the relation distributions of real-world knowledge triples created by fine-tuned models and generic LLMs, in addition to QA pair evaluation. As illustrated in Figure 2, both the fine-tuned T5-Large and BGE models can create or retrieve real-world triples that are more closely aligned with expert annotations. In contrast, the relation distributions of models not fine-tuned with expert annotations, such as GPT-4, Llama 2, and the original BGE model, tend to be inconsistent with expert annotations. Also, LLMs generate a wider variety of triple types, many of which do not belong to the expert-annotated types (see the "others" column in Figure 2).

This observation justifies that **a smaller language model assisted with domain expertise (i.e., expert-annotated real-world knowledge) can reliably perform better than generic LLMs in domain-specific scenarios.**

4.3 RQ3: Retrieval-Augmented Models vs. Experts

We explore the potential of RAG compared with experts' annotation through two different strategies within the 2-Step pipeline. The first strategy is the 2-Step-RAG strategy. Here, we use a retriever to select a real-world knowledge triple relevant to a provided story section, and then employ a generator

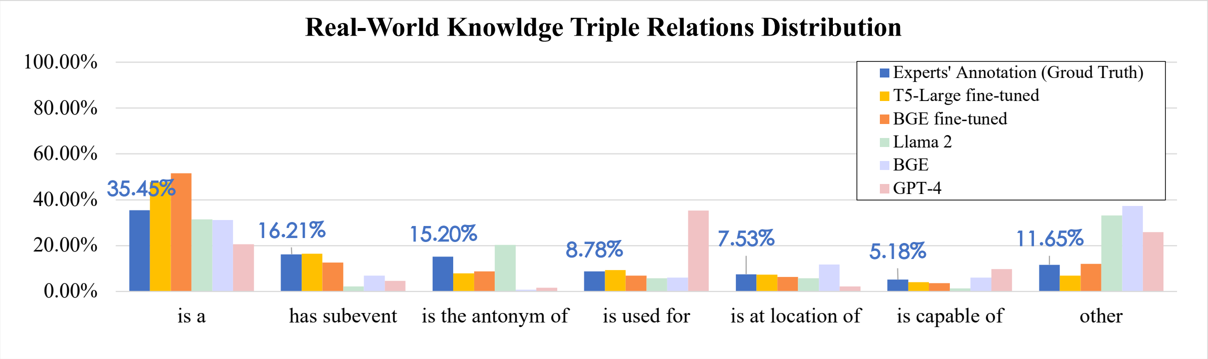

Figure 2: The distribution of triple relations in the 2-Step QAG pipeline, including experts' annotation, and triples created by fine-tuned T5-Large, fine-tuned BGE, Llama2, original BGE, and GPT-4.

Model	Grammar Correctness	Answer Relevancy	Contextual Consistency	Educational Appropriateness
T5-Large fine-tuned	4.867	4.478	**4.656**	**4.433**
GPT-4	**4.878**	**4.522**	4.578	4.389

Table 3: Human evaluation results of GPT-4 and fine-tuned T5-Large in 2-STEP-HUMAN. **Bolded numbers** are the best performance in each dimension.

to create a QA pair based on the retrieved triple. In the second strategy (i.e., 2-STEP-HUMAN), we provide language models with expert-annotated structured knowledge, guiding them to generate QA pairs based on the experts' annotations.

4.3.1 Experiment Results and Analysis

As shown in Table 2, the 2-STEP-RAG strategy, especially when fine-tuned with experts' annotation, performs better at retrieving more educationally appropriate triples compared to LLMs in the 2-STEP-LLM strategy where models are not assisted by human experts' annotations. This justifies that RAG is promising in terms of retrieving relevant knowledge for domain-specific tasks like interactive storytelling.

It is worth noting that by employing expert-annotated structured knowledge in the 2-STEP-HUMAN strategy, all LLMs as well as the domain-specific fine-tuned language model, can far exceed the end-to-end pipeline and 2-STEP-LLM and 2-STEP-RAG in the RA-QAG task. This proves that **domain expertise is still useful in such real-world domain-specific tasks. Despite RAG can improve the model performance to a certain extent on our QAG task, it cannot yet completely substitute the domain knowledge of human experts**.

4.4 Human Evaluation

To thoroughly assess the quality of LLM-generated QA pairs, as well as to comprehensively investigate the helpfulness of expert-annotated structured knowledge, we conduct a human study to compare the QA pairs generated by different models.

More specifically, according to the superior performance of fine-tuned T5-Large and GPT-4 in 2-STEP-HUMAN, we selected these two models for human evaluation. We recruit three education experts and randomly select 30 story sections of 16 books from the test split of `FairytaleCQA`. For each section, there are two QA pairs created based on the story narrative (experts' annotation, and QA pairs generated by GPT-4 and fine-tuned T5-Large), summing up to 60 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.

Considering teachers' practice in formulating questions and feedback during interactive storytelling (Xu et al., 2021; Zhang et al., 2022), we ask the experts to evaluate each QA pair on the following four dimensions with a 5-point Likert scale:

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

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

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

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

Table 3 illustrates the average scores in each dimension. We observe that GPT-4 performs better in the *Grammar Correctness* and *Answer Relevancy* dimensions, which is reasonable because LLMs like GPT-4 are trained on vast amounts of diverse corpora, making it easier for these models to generate more Grammatical correct text.

For the *Contextual Consistency* dimension, in which we assess whether a QA pair is both associated with the story and external real-world knowledge, the fine-tuned T5-Large outperformed GPT-4. For the *Children's Educational Appropriateness* dimension, the T5-Large model fine-tuned on `FairytaleCQA` also exhibits better performance.

This result suggests that fine-tuned with experts' annotation, the T5-Large model can generate QA pairs that 1) contain external structured knowledge, and 2) are appropriate for young children's interactive storytelling experience. Also, this result proves that our 2-Step pipeline can effectively infuse structured knowledge with free-form narrative, facilitating similar tasks in other specific domains.

4.5 Discussion

To approach the RA-QAG task we formulate, we construct comprehensive QAG pipelines and investigate the potential of various generation strategies in solving real-world tasks, as well as the effectiveness of human expertise.

The 2-STEP-RAG strategy, utilizing a fine-tuned retriever model, significantly enhances the performance of LLMs compared to 2-STEP-LLM. Notably, the BGE model only consists of 326 million parameters, whereas the T5-Large model consists of 770 million parameters. Fine-tuning a smaller retriever model like BGE can yield results almost identical to fine-tuning a larger LM, highlighting RAG as a more cost-effective method to improve LLM performance for real-world tasks. Comparing 2-STEP-RAG with 2-STEP-HUMAN, we also observe a notable improvement when models benefit from expert annotation. This underscores that **while RAG can impart domain knowledge to LLMs through fine-tuning on expert-annotated data, it does not obviate the need for human experts**.

By enabling LLMs to mimic teachers' practices in interactive storytelling activities and leveraging human-expert annotated knowledge for fine-tuning, all models' performance could be enhanced. This underscores the effectiveness of incorporating common human practices and leveraging human knowledge to improve model performance in real-world applications.

However, we also observe that the overall Rouge-L of QAG system evaluation is relatively low across all pipelines, even with GPT-4. **We attribute this to human experts' grasp of the timing.** During the creation of QA pairs, human experts strategically decide when to provide structured knowledge and what structured knowledge to incorporate. However, in the case of automatic QAG, the absence of this nuanced timing limits LLMs to provide appropriate structured knowledge for QAG. This illustrates the challenging nature of the QAG for interactive storytelling given SoTA language models, leaving significant space for future improvement.

5 Conclusion and Future Work

In this work, we focus on a common and critical real-world scenario: teachers attempting to impart real-world knowledge by posing story-inspired, educationally crafted questions and providing responsive feedback during interactive storytelling with preschool children. We formulate this real-world application into a novel QAG task, namely Real-world Augmented QAG (RA-QAG), and explore LLMs' performance when equipped with various domain adaptation strategies compared with human expertise. By employing few-shot ICL, Chain-of-Thoughts, and Retrieval-Augmented Generation, our QAG pipeline experiments demonstrate that: 1) RAG shows great potential for tackling real-world, domain-specific tasks; 2) Human experts still master domain expertise and intricate knowledge that generic LLMs need to learn from.

One future direction involves leveraging our pipeline designs to further develop human-AI collaborative educational systems, such as interactive storytelling systems, to better facilitate children's story-based learning of real-world knowledge, addressing parents' or teachers' practical constraints, such as limited time, expertise, and educational resources. In addition, we could further explore advanced QAG pipeline designs and generation strategies to enhance LLMs' domain adaptation ability in other real-world settings such as healthcare, law, and finance, to further investigate the RA-QAG task we propose in this work.

6 Limitations

This work primarily focuses on employing various generation strategies, including few-shot ICL, Chain-of-Thoughts, and Retrieval-Augmented Generation, to approach the RA-QAG task we formulate from the real-world interactive storytelling scenario, and investigate their effectiveness compared with human expertise. There are several limitations.

First, we experimented with the few-shot ICL, Chain-of-Thoughts, and Retrieval-Augmented Generation strategies; however, we are aware that there are more generatstrategies,gies as well as some instruction-finetuned LLMs, such as InstructGPT (Ouyang et al., 2022), can be further explored.

Second, for the 2-STEP-RAG strategy, we experiment with an original version of the BGE model and a fine-tuned one. We acknowledge that there exist many other Retriever models, such as LLM-Embedder (Zhang et al., 2023), and RAG approaches, such as Iterative Retrieval (Shao et al., 2023b), that could be implemented.

Third, we experiment with two different QAG pipeline designs. Although we investigate three variations in the 2-Step pipeline, more novel pipeline designs, such as multi-step generation pipelines, could be implemented to further explore their performance on our RA-QAG task.

References

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.

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.

Deng Cai, Yan Wang, Wei Bi, Zhaopeng Tu, Xiaojiang Liu, and Shuming Shi. 2019. Retrieval-guided dialogue response generation via a matching-to-generation framework. In *Conference on Empirical Methods in Natural Language Processing*.

Meng Cao, Yue Dong, Jiapeng Wu, and Jackie Chi Kit Cheung. 2020. Factual Error Correction for Abstractive Summarization Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6251–6258, Online. Association for Computational Linguistics.

Jiaju Chen, Yuxuan Lu, Shao Zhang, Bingsheng Yao, Yuanzhe Dong, Ying Xu, Yunyao Li, Qianwen Wang, Dakuo Wang, and Yuling Sun. 2023. FairytaleCQA: Integrating a Commonsense Knowledge Graph into Children's Storybook Narratives. *arXiv.org*.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation.

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.

Hangfeng He, Hongming Zhang, and Dan Roth. 2022. Rethinking with retrieval: Faithful large language model inference. *ArXiv*, abs/2301.00303.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane A. Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Few-shot learning with retrieval augmented language models. *ArXiv*, abs/2208.03299.

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.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised Contrastive Learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673. Curran Associates, Inc.

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.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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

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

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.

Fanyi Qu, Xin Jia, and Yunfang Wu. 2021. Asking questions like educational experts: Automatically generating question-answer pairs on real-world examination data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2583–2593, Online and Punta Cana, Dominican Republic. 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.

Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

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.

Zhenwei Shao, Zhou Yu, Meng Wang, and Jun Yu. 2023a. Prompting Large Language Models With Answer Heuristics for Knowledge-Based Visual Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14974–14983.

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023b. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9248–9274, Singapore. Association for Computational Linguistics.

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*.

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.

Xuhai Xu, Bingsheng Yao, Yu Dong, Hongfeng Yu, James A. Hendler, Anind K. Dey, and Dakuo Wang. 2023. Mental-llm. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8:1 – 32.

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

Bingsheng Yao, Dakuo Wang, Tongshuang 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*.

Peitian Zhang, Shitao Xiao, Zheng Liu, Zhicheng Dou, and Jian-Yun Nie. 2023. Retrieve anything to augment large language models. *arXiv preprint arXiv:2310.07554*.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2024. Benchmarking Large Language Models for News Summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

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.

A Appendix

A.1 Complete QAG pipeline Results

We demonstrate the complete performance of LLMs in our end-to-end QAG pipeline in Table 4, and the full results for 2-Step QAG is presented in Table 5.

Models	Prompting Strategy	Triple		QA pair	
		Rouge-L	SBERT	Rouge-L	SBERT
T5-Large fine-tuned	-	0.206	0.318	**0.279**	0.263
Alpaca	zero-shot	0.139	0.301	0.266	0.207
	1-shot	0.276	0.321	0.239	0.186
Mistral	zero-shot	0.209	0.348	0.209	0.229
	1-shot	0.240	0.363	0.231	0.241
	5-shot	0.280	0.372	0.257	0.251
Llama 2	zero-shot	0.154	0.340	0.177	0.225
	1-shot	0.200	0.367	0.206	0.237
	5-shot	0.291	0.370	0.269	0.253
Flan-T5-XXL	1-shot	0.275	0.375	0.194	0.209
GPT-3.5	zero-shot	0.219	0.373	0.220	0.252
	1-shot	0.245	0.386	0.252	0.271
	5-shot	0.274	0.384	0.264	0.266
	CoT			0.259	0.280
GPT-4	zero-shot	0.286	0.385	0.243	0.261
	1-shot	0.289	**0.413**	0.251	**0.292**
	5-shot	0.285	0.398	0.248	0.283
	CoT	**0.295**	0.404	0.262	**0.292**

Table 4: Rouge-L and SentenceBERT scores of LLMs in the end-to-end QAG pipeline. **Bolded numbers** are the best performance within each setting in each metric. Underlined numbers are the second-best scores within each setting.

A.2 GPT propmts

To thoroughly harness GPT's generation capabilities, we collaborated with educational experts, and iteratively designed and refined the prompts with clear and informative instructions.

For our QAG pipelines, there are three different variations:

- **End-to-end QAG setting**: Directly generate a real-world knowledge triple and QA pair from an input story section (Table 6).
- **Chain-of-Thought QAG approach**: Generate a real-world knowledge triple and QA pair by thinking step by step from an input story section (Table 7).
- **two-step QAG setting**: Generate a real-world knowledge triple in the first step from a story section, and generate a QA pair based on the generated triple in the second step (Table 8 and Table 9).

A.3 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.

Strategy	Model (Step1-Triple)	Rouge-L	SBERT	Model (Step2-QA pair)	Rouge-L	SBERT
2-Step-LLM	T5-Large fine-tuned	0.331	0.402	T5-Large fine-tuned	0.279	0.263
	Llama 2	0.311	0.353	Llama 2	0.263	0.247
	GPT-4	0.290	0.398	GPT-4	0.269	0.279
2-Step-RAG	BGE	0.298	0.395	GPT-4	0.256	0.268
	BGE fine-tuned	0.328	0.384	GPT-4	0.278	0.267
2-Step-Human	Experts' annotation (Ground Truth)	**1.000**	**1.000**	T5-Large fine-tuned	**0.510**	**0.834**
				Llama 2	0.413	0.690
				GPT-4	0.482	0.794

Table 5: The complete 2-Step QAG pipeline performance on both steps, including the 2-Step-LLM, 2-Step-RAG and 2-Step-Human strategies. For all LLMs involved in this setting, we used 5-shot ICL for the corresponding generation step. **Bolded numbers** are the best performance within each setting on each metric. Underlined numbers are the second-best scores within each setting on each metric.

Prompt for GPT in the end-to-end QAG pipeline

I need you to help generate a question and answer pair for young children aged three to six. I will provide you with a short section of a story delimited by triple quotes. Please follow these steps:

1. For each sentence, identify one key word that meets the following criteria: it is relatively complex, it is considered tier 1 or tier 2 vocabulary, and it is a concrete noun, verb, or adjective.

2. After this, you need to completely forget about the story that I gave you, remembering only the words you identified.

3. Based on each selected word, generate one real-world relation based on the selected word. This real-world relation should go beyond the context of the stories. For example, if your identified word is 'apple', your real-world relation could be: apple grows on trees; apples are red. The real-world, fact-based knowledge should be based on the selected word and is in the form of a triple such as 'A relation B', where A and B are two concepts and the selected word can be either A or B. You should use one of the following relations for the real-world knowledge:

causes
desires
has context of
has property
has subevent
is a
is at location of
is capable of
is created by
is made of
is part of
is the antonym of
is used for

4. After this, generate a question and answer pair based on the real-world, fact-based knowledge you generated. Either the question or the answer should contain that identified word. Each question should have one single correct answer that would be the same regardless of the children's experiences. The questions should be focused on real-world, fact-based knowledge and beneficial to educate children during storytelling.

5. After this, select one question-answer pair that you think best 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 real-world knowledge 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 6: Prompt for GPT in the QAG task with generating real-world knowledge triple and QA pairs directly from story.

Chain-of-Thoughts Prompt for GPT in the QAG pipeline

Q: Now, generate a question and answer pair containing real-world, fact-based knowledge associated with the following story for young children aged three to six.
<story>: ***{story1 for few-shot}***

A: Let's think step by step.
First, a key word can be identified from the story text, which 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.
The identified key word is: ***{concept word1 for few-shot}***
Then, based on the identified key word, one piece of real-world, fact-based knowledge can be generated 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.
The triple should use one of the following relations for the real-world knowledge:
causes
desires
has context of
has property
has subevent
is a
is at location of
is capable of
is created by
is made of
is part of
is the antonym of
is used for
The generated real-world knowledge is: ***{real-world knowledge for few-shot}***
Finally, a question and answer pair can be generated based on the generated real-world, fact-based knowledge.
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 generated question-answer pair is:
question: ***{question1 for few-shot}***
answer: ***{answer1 for few-shot}***
... ...

Q: Now, generate a question and answer pair containing real-world, fact-based knowledge associated with the following story for young children aged three to six:
<story>: ***{story for the current data}***

Table 7: Chain-of-Thoughts Prompt for GPT in the QAG task with generating real-world knowledge triple and QA pairs directly from story.

Prompt for GPT in two-step pipeline: Step 1

I need you to help generate real-world knowledge for young children aged three to six. The real-world 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 real-world, fact-based knowledge.
For example, if your identified word is ’apple’, your real-world relation could be: apple is a fruit; apple is used for eating.
The real-world, fact-based knowledge should be based on the selected word and is in the form of a triple such as ’A relation B’, where A and B are two concepts and the selected word can be either A or B. You should use one of the following relations for the real-world knowledge:
causes
desires
has context of
has property
has subevent
is a
is at location of
is capable of
is created by
is made of
is part of
is the antonym of
is used for
Return the generated real-world knowledge in the following format:

real-world knowledge triple: (A, relation, B)

⟨story ⟩:
{story1 for few-shot}

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

⟨story ⟩:
{story for the current data}

⟨response ⟩:

Table 8: Prompt for step 1 in GPT two-step QAG pipeline.

Prompt for GPT in two-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 real-world knowledge. Please follow these steps:
1. Based on provided real-world knowledge, generate a question and answer pair that either the question or the answer contains a concept in the real-world knowledge.
The questions should be focused on real-world, fact-based knowledge.
For example, given the real-world 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: ...
⟨story ⟩:
{story1 for few-shot}

⟨real-world knowledge triple ⟩:
{real-world knowledge triple1 for few-shot}

⟨response ⟩:
{response1 for few-shot}
... ...
⟨story ⟩:
{story for the current data}

⟨real-world knowledge triple⟩:
{real-world knowledge triple generated by GPT in Step 1 for the current data}

⟨response ⟩:

Table 9: Prompt for step 2 in GPT two-step QAG pipeline.