Multimodal Question Generation and Evaluation Using Large Language Models

Anonymous ACL submission

Abstract

To support the development of conversational agents for educational purposes, particularly those designed to engage children through interactive storytelling, there is a growing need for systems that can automatically generate relevant and pedagogically sound questions. Conversational agents can use such questions during interactive sessions to promote comprehension, reflection, and active participation. In this work, we develop an LLM-based pipeline that automates the generation of questions from story content, ensuring the appropriateness and clarity of questions to maximize children's learning outcomes. We use GPT-4o to generate interactive questions from stories based on various modality covering question types such as completion, recall, open-ended, and Wh questions. Our findings demonstrate the ability of the LLM to generate appropriate and contextually relevant questions, as well as its ability to align with human judgment in the evaluation of automatically generated questions.

1 Introduction

Question generation plays a vital role in educational settings, serving as a fundamental tool for assessing student understanding, promoting critical thinking, and facilitating active learning (Whitehurst et al., 1988; Zhang et al., 2022). Whether crafted by educators or generated automatically, well-designed questions can stimulate deeper engagement with content, encourage reflection, and provide valuable feedback on learning progress (Dietz Smith et al., 2024). The ability to generate contextually appropriate and pedagogically sound questions at scale has become increasingly important as educational systems seek to provide personalized and adaptive learning experiences. Automatic question generation (AQG) using large language models (LLMs) has emerged as a powerful solution to this challenge, offering scalable and personalized learning support. Recent advances in

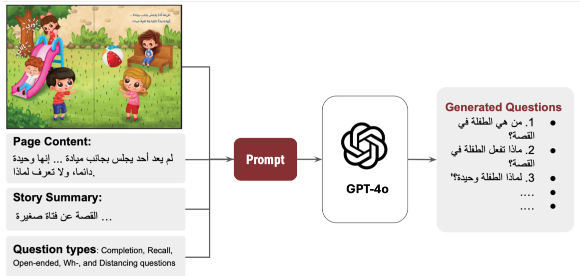

Figure 1: For each page, the story context (summary), the story page, textual content of the page , and a list of question types with their definitions are being concatenated with the prompt instructions and passed to the open AI's model. Then the model returns a list of questions and their corresponding question type. The model used was gpt-4-vision-preview

LLMs have significantly improved the efficiency and versatility of AQG, reducing the need for technical expertise and allowing educators to generate high-quality questions that can inspire student thinking and support self-assessment in online and offline learning environments (Yuan et al., 2022; Bulathwela et al., 2023; Jiang et al., 2024). In particular, AQG has shown promise in the domain of children's storytelling, where it can create questions with high cognitive demand that conversational agents use during reading sessions, fostering dialogic interaction between children and caregivers (Zhao et al., 2022; Lekshmi Narayanan et al., 2024).

While general-purpose models such as GPT-3 and GPT-4 have been successfully employed in educational AQG tasks (Lee et al., 2024; Yuan et al., 2022; Jiang et al., 2024), task-specific models such as MultiQG-TI and EduQG have also emerged. These specialized models leverage fine-tuning on specific datasets or incorporate multimodal inputs such as text and images to improve question quality and contextual relevance (Wang and Baraniuk, 2023; Bulathwela et al., 2023). In the Arabic context, fine-tuned transformer-based models have been used to develop end-to-end AQG systems trained on datasets such as Arabic-SQuAD and ARCD, achieving good quality as assessed by automatic evaluations (Alajmi et al., 2025; Lafkiar and En Nahnahi, 2025). Some approaches utilize LLMs not only to generate questions but also to evaluate and filter them based on relevance and difficulty (Yuan et al., 2022; Xiao et al., 2023). The types of questions generated through AQG systems are diverse, including multiple choice, open-ended, and closed questions, as well as more abstract categories such as prediction and concept questions (Lee et al., 2024; Jiang et al., 2024; Lekshmi Narayanan et al., 2024).

Despite significant progress, key challenges persist in the field. The evaluation of generated question quality remains particularly challenging, with common automatic evaluation metrics including ROUGE-L, BLEU, BERTScore, and cosine similarity (Zhao et al., 2022; Wang and Baraniuk, 2023; Lamsiyah et al., 2024), while manual evaluation often involves expert reviewers assessing fluency, relevance, and answerability (Cho et al., 2021; Alajmi et al., 2025). GPT models may perform inconsistently when generating yes/no questions or cloze-style multiple-choice items (Lee et al., 2024; Xiao et al., 2023). Furthermore, multimodal AQG systems face challenges related to contextual grounding and hallucination (Wang and Baraniuk, 2023). There is also a pressing need for more transparent and scalable evaluation frameworks and better integration of teacher-provided materials to fine-tune model outputs (Bulathwela et al., 2023; Xiao et al., 2023; Lekshmi Narayanan et al., 2024).

In this work, we develop an LLM-based pipeline that automates the generation of questions from Arabic children's story content. Our approach relies on multimodal question generation using LLMs. The LLM leverages dialogic reading strategies, specifically the CROWD framework, which encompasses several question types: Completion, Recall, Open-ended and Wh- questions (Zevenbergen and Whitehurst, 2003). The framework grounds the question generation process, ensuring that the LLM produces appropriate questions that support the goal of enhancing children's learning outcomes through interactive engagement. Our proposed pipeline aims to address some of these gaps by leveraging GPT-4 to generate pedagogically valuable questions from story content, tailored to support young learners. By focusing on question generation in narrative comprehension and diverse question modalities, our system contributes to the broader goal of enhancing educational interactions through LLMs.

We summarize our contribution as follows.

- We build a multimodal pipeline for question generation for illustrative stories following the CROWD framework.
- We design evaluation guidelines to assess the quality of the generated questions through human review.
- We develop a high-quality, scalable LLM-based evaluator, benchmarked against a batch with gold-standard human annotations, and find that it closely aligns with human judgments.

2 Data and Storybook Preprocessing

Our data set was constructed from 14 Arabic storybooks that cover a variety of age groups, all published by *We Love Reading*[1]. The textual content of each book was manually transcribed to ensure precision and consistency. For the visual modality, each double-page spread was semi-manually merged into a single panoramic image, thereby preserving the full illustration context and allowing precise alignment between text and visuals. This preprocessing step ensured that both modalities could be jointly leveraged for question generation.

3 Method

Our approach consists of using LLMs for question generation and question evaluation.

Question Generation In order to harness the potential of using LLMs to generate helpful and interactive questions for children, we refined the prompting strategy through iterative adjustments and human review of the generated outputs, with a particular focus on producing knowledge-based and educational questions. Questions are generated at the page level to ensure both local relevance and comprehensive coverage of the story. Each prompt provides the LLM with a holistic view of the page by including the summary of the story, textual content, and visual context (illustrative page of the story as images). Additionally, the prompt incorporates explicit instructions to adhere to the CROWD

[1]https://welovereading.org

framework of dialogic reading, which consists of five question types: completion, recall, open-ended, wh-question, and distancing. Each type focuses on specific aspects of learning and child engagement, such as fostering connections with personal experience, encouraging narrative recall, and assessing comprehension. Figure 1 illustrates an example of the constructed prompt.

Question Evaluation We measure the quality of the generated questions using an LLM-based and human-based evaluation. The evaluation consists of five questions and covers key aspects of clarity, appropriateness, and relevance, helping in the assessment of whether each question is well understood, contextually meaningful, and suitable for supporting children's learning outcomes. In addition to evaluating the question, we measure the effectiveness of incorporating various modality by asking about the modality contribution for the generated question, whether it is relevant to the image, the text, or both. For each evaluation question, the evaluator (human or LLM) is asked to give a score between 1-3 to indicate Yes / Partially / No. Full evaluation guidelines are presented in Appendix A.

We use 14 stories for our evaluation purposes, we randomly select one story from each age group (indicated by a star in Table 5). We unify the evaluation guidelines and questions for both the human evaluators and the LLM evaluator. For the evaluation prompt, we provide the context (i.e. story image and textual content) for each page along with the generated question and ask about the various evaluation dimensions.

To measure the agreement among evaluators, as well as between the LLM model and the majority vote of human evaluators, we used percentage agreement. In this metric, if all evaluators provided the same answer to every generated question, the percentage agreement would be 100%. This method is easy to interpret and accounts for the situation of no variance and no variability that might not be possible in other agreement metrics. The evaluation prompt is presented in Appendix B.

The model gpt-4o is used as the LLM evaluator, using the prompt described in Appendix B. For each generated question, the text and visual content of the corresponding page is appended to the evaluation prompt to ensure contextually grounded assessments. Human evaluations are performed by a native Arabic speaker. The evaluation instructions and structure mirror those provided to the LLM evaluator. In cases where the evaluator is uncertain, a default score of 2 (indicating uncertainty) is assigned.

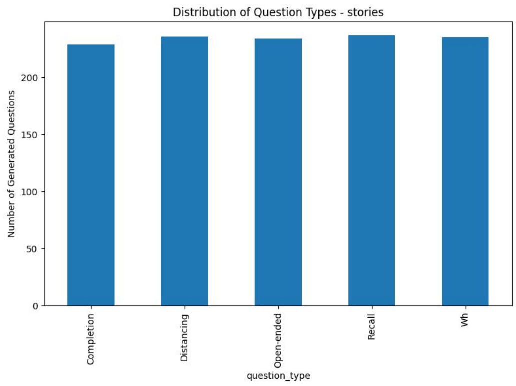

Figure 2: Distribution of generated question types across all stories

Story	Avg.Q per-Page	Total-Q
Amal	5.00	35
The Bridge To Dreamland	4.92	59
Questions in a Travel Bag	4.93	69
The Black Hen	4.93	69
The Open Faucet	4.88	78
"Um Hatta" the Cat	4.85	97
Why Did Electricity Run Away?	5.00	97
Something Really Strange	5.00	100
Word Cooker	5.00	100
Salma's Riddle	4.66	135
The Eid Gift	4.86	34
The Amazing Water Hero	5.00	60
I'd Like to Introduce You To	4.75	114
My Brother Hani	4.96	124
Average	**4.9**	**83.6**

Table 1: Average number of questions generated per page and total questions per story. Variation primarily due to differences in story length and content density.

4 Results

Question Generation We use gpt-4o to generate questions for stories and show the distribution of question types in Figure 2. The model produces a balanced set of questions in all CROWD categories. Each category is represented with a comparable frequency (approximately 230 questions per type), indicating that the model is not biased toward a particular form of questioning, but rather provides comprehensive coverage across diverse cognitive levels. We also examine the distribution of the generated questions in all stories. Table 1

Category (%)	LLM Eval.	Human Eval.
Both	84.6	70.1
Image	3.9	3.0
Irrelevant	4.4	0.0
Text	7.1	26.9

Table 2: Comparison of modality reliance between LLM and human evaluations.

Story	Q1	Q2	Q3	Q4	Q5
Amal	0.91	0.86	0.91	0.69	1.00
The Bridge To Dreamland	0.89	0.84	0.93	0.74	1.00
Questions in a Travel Bag	0.92	0.88	0.91	0.62	1.00
The Black Hen	0.96	0.96	0.97	0.83	1.00
The Open Faucet	0.96	0.91	0.88	0.74	1.00
"Um Hatta" the Cat	0.99	0.94	0.95	0.76	1.00
Why Did Electricity Run Away?	0.91	0.88	0.90	0.74	0.99
Something Really Strange	0.92	0.91	0.93	0.67	1.00
Word Cooker	0.93	0.90	0.98	0.82	0.99
Salma's Riddle	0.83	0.77	0.79	0.67	0.95
The Eid Gift	0.91	0.85	0.82	0.73	0.97
The Amazing Water Hero	0.90	0.85	0.90	0.68	1.00
I'd Like to Introduce You To	0.87	0.86	0.94	0.72	1.00
My Brother Hani	0.85	0.72	0.78	0.63	0.96
Average Agreement	0.91	0.87	0.90	0.72	0.99

Table 3: Agreement scores (Q1–Q5) by story.

shows the average number of questions generated per page and the total number of questions generated per story. The results show that the model maintains a stable density of approximately 4.8 to 5.0 questions per page across all stories, regardless of content. However, the total number of questions varies widely, ranging from 34 for shorter stories, such as "The Eid Gift", to 135 for longer stories, such as "Salma's Riddle". This indicates that the variation in the total questions is primarily driven by the length of the story rather than the inconsistency in the model generation process.

Question Evaluation We then evaluate the quality of the questions using both human and LLM evaluations. Agreement scores are calculated in all stories and evaluation questions. In general, to evaluate the alignment between automated and human assessment, we compare agreement scores between questions and stories, as well as modality reliance (text, image, or both).Table 2 shows that both LLM and human rating are mainly based on multimodal input, with LLM producing a higher proportion (84.6%) compared to human evaluation (70.1%). Table 3 summarizes the agreement scores in stories for the Q1-Q5 questions. Overall, agreement was

Evaluation Question	Human	LLM	Overlap
Is the question clear to a child?	0.98	0.99	0.98
Is the question relevant to the given image?	0.96	0.99	0.96
Is the question relevant to the page text?	0.96	0.98	0.96
Is the question about an important aspect (text+image)?	0.96	0.99	0.96
Is the question appropriate for a child?	0.98	1.00	0.98

Table 4: Comparison of human and LLM evaluations with percentage of "Yes" responses and their overlap.

consistently high ($\approx$ 0.85–1.0), with Q5 achieving the highest average score (0.99) in all stories. In contrast, Q4 showed the lowest agreement (0.72), indicating greater variability in responses. These results suggest that, while most question types yield stable agreement, certain prompts (e.g., Q4) may introduce interpretive differences across stories. As shown in Table 4, the human and LLM evaluations have near-perfect alignment in all five evaluation criteria (96–100%). The agreement is strongest for clarity (Q1) and appropriateness (Q5), both at 0.98 or higher, while the relevance to image, text, and integration (Q2-Q4) consistently scored 0.96. The overlap scores confirm that the model's judgments are highly consistent with human ratings.

5 Conclusion

This study presents a pipeline grounded in large language models (LLMs) for generating knowledge-based evaluation questions from children's stories, integrating both text and image modalities. Using `gpt-4o`, the system produced a balanced set of question types, completion, recall, open-ended, Wh, and distancing, with an average of 4.9 questions per page. The results indicated a significant concordance with human evaluations (96 to 100%), thus affirming the clarity, relevance, and suitability of the generated inquiries. The findings underscore the resilience of the methodology in a variety of narratives and its potential to facilitate social-emotional learning, as well as culturally relevant educational methodologies within early childhood environments. Future work will study the ability of conversational agents to use automatically generated questions to facilitate an interactive reading and learning session with children.

References

Anwar Alajmi, Haniah Altabaa, Sa'ed Abed, and Imtiaz Ahmad. 2025. Arabic question generation using transformers. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 24(3).

Sahan Bulathwela, Hamze Muse, and Emine Yilmaz. 2023. Scalable educational question generation with pre-trained language models. *Preprint*, arXiv:2305.07871.

Woon Sang Cho, Yizhe Zhang, Sudha Rao, Asli Celikyilmaz, Chenyan Xiong, Jianfeng Gao, Mengdi Wang, and Bill Dolan. 2021. Contrastive multi-document question generation. *Preprint*, arXiv:1911.03047.

Griffin Dietz Smith, Siddhartha Prasad, Matt J Davidson, Leah Findlater, and R Benjamin Shapiro. 2024. Contextq: Generated questions to support meaningful parent-child dialogue while co-reading. In *Proceedings of the 23rd Annual ACM Interaction Design and Children Conference*, pages 408–423.

Hang Jiang, Xiajie Zhang, Robert Mahari, Daniel Kessler, Eric Ma, Tal August, Irene Li, Alex Pentland, Yoon Kim, Deb Roy, and Jad Kabbara. 2024. Leveraging large language models for learning complex legal concepts through storytelling. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7194–7219, Bangkok, Thailand. Association for Computational Linguistics.

Said Lafkiar and Noureddine En Nahnahi. 2025. An end-to-end transformer-based model for arabic question generation. *Multimedia Tools and Applications*, 84(20):22009–22023.

Salima Lamsiyah, Abdelkader El Mahdaouy, Aria Nourbakhsh, and Christoph Schommer. 2024. Fine-tuning a large language model with reinforcement learning for educational question generation. In *Artificial Intelligence in Education*, pages 424–438, Cham. Springer Nature Switzerland.

U. Lee, H. Jung, Y. Jeon, and et al. 2024. Few-shot is enough: exploring chatgpt prompt engineering method for automatic question generation in english education. *Education and Information Technologies*, 29:11483–11515.

Arun Balajiee Lekshmi Narayanan, Ligia E. Gomez, Martha Michelle Soto Fernandez, Tri Nguyen, Chris Blais, Maria Adelaida Restrepo, and Arthur Glenberg. 2024. Genq: Automated question generation to support caregivers while reading stories with children. In *Proceedings of the XI Latin American Conference on Human Computer Interaction*, CLIHC '23, New York, NY, USA. Association for Computing Machinery.

Zichao Wang and Richard Baraniuk. 2023. MultiQG-TI: Towards question generation from multi-modal sources. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 682–691, Toronto, Canada. Association for Computational Linguistics.

Grover J Whitehurst, Francine L Falco, Christopher J Lonigan, Janet E Fischel, Barbara D DeBaryshe, Marta C Valdez-Menchaca, and Marie Caulfield. 1988. Accelerating language development through picture book reading. *Developmental psychology*, 24(4):552.

Changrong Xiao, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Lei Xia. 2023. Evaluating reading comprehension exercises generated by LLMs: A showcase of ChatGPT in education applications. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 610–625, Toronto, Canada. Association for Computational Linguistics.

Xingdi Yuan, Tong Wang, Yen-Hsiang Wang, Emery Fine, Rania Abdelghani, Pauline Lucas, Hélène Sauzéon, and Pierre-Yves Oudeyer. 2022. Selecting better samples from pre-trained llms: A case study on question generation. *Preprint*, arXiv:2209.11000.

Andrea A. Zevenbergen and Grover J. Whitehurst. 2003. Dialogic reading: A shared picture book reading intervention for preschoolers. In Anne van Kleeck, Steven A. Stahl, and Eurydice B. Bauer, editors, *On reading books to children: Parents and teachers*, pages 177–200. Lawrence Erlbaum Associates Publishers.

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. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–21.

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.

A Evaluation Guidelines

The annotator evaluation guidelines are shown in Table 6.

B Prompts

B.1 Question Generation

This appendix section gives example prompts for generating and evaluating the five types of questions designed for children aged 4 to 6. Words in all caps and square brackets were included verbatim

Split	Story	Age Group	Pages*
Train	Amal	7-10 YO	19/48
	The Bridge to Dreamland	5-8 YO	7/30
	Questions in a Travel Bag	4-6 YO	20/48
	The Black Hen	7-10 YO	24/56
	The Open Faucet	7-10 YO	16/28
	"Um Hatta" the Cat	4-6 YO	20/48
	Why Did Electricity Run Away?	4-6 YO	13/32
	Something Really Strange	4-6 YO	20/48
	Word Cooker	4-6 YO	20/48
	Salma's Riddle	4-6 YO	29/48
Test	The Eid Gift	3-7 YO	7/24
	The Amazing Water Hero	4-6 YO	12/25
	I Would Like to Introduce You To	7-10 YO	24/56
	My Brother Hani	4-6 YO	25/56

Table 5: Overview of Storybooks. Content pages refer to pages that include story content and merged pages.

as prompt variables. Words in parentheses were replaced with the relevant story text.

- **Completion Prompt**

The initial prompt is used to generate a candidate question. *Act as an early childhood reading instructor, generating completion prompts for children aged 4–6. The requirements are:*

- *The question should be based on repetition or rhyme in the story.*
- *It should be one sentence long and end with a blank for the child to complete.*
- *It must be grounded in the current sentence or phrase, without requiring broader story context.*
- *Example: I'll huff, and I'll puff, and I'll blow the house ____.*

(current page text) "With that context, generate a prompt of type 'completion' for the above text." Format your response in JSON using exactly the template below:

```
{
  "prompt": PROMPT
}
```

- **Recall Prompt**

Purpose: Ask about a past event that requires memory across pages. You are an expert in early childhood education. Generate a recall prompt suitable for a child aged 4–6. This prompt should ask about a thematically important event that occurred earlier in the story The requirements are:

- *The question should ask the child to recall a specific, thematically important event from the story.*
- *It must reference content that requires integrating across multiple pages or events.*
- *The question should begin with a wh-word (e.g., What, When, Who).*
- *Do NOT include compound or multi-part questions.*
- *Example: What did the lion do after the mouse helped him?*

(current page text) "With that context, generate a prompt of type 'recall' for the above text." Format your response in JSON using exactly the template below:

```
{
  "prompt": PROMPT
}
```

- **Open-ended Prompt**

The initial prompt is used to generate a candidate question. *You are a specialist in dialogic reading with young children. Create an open-ended question for a child aged 4–6. The requirements are:*

- *The question should invite speculation, prediction, or explanation related to characters, setting, or themes.*
- *Avoid simple factual recall or yes/no questions.*
- *The child should be encouraged to provide a thoughtful or imaginative response.*
- *Avoid asking about personal experiences.*
- *Example: What do you think the rabbit felt when he saw the trap?*

(current page text)

"With that context, generate a prompt of type 'open-ended' for the above text."

Format your response in JSON using exactly the template below:

```
{
  "prompt": PROMPT
}
```

- **Wh Prompt**

The initial prompt is used to generate a candidate question. *Act as a reading instructor for children. Based on the following story text, create a Wh-question for a child aged 4–6. The requirements are:*

Question	Answer Options
1- Is the question clear to a child? Ask yourself, if I was a child will I understand this question? Will I be able to comprehend it? Is the language simple enough to be understood by a child?	**1) It is not at all clear**: The question is ambiguous and is difficult for children to understand what is being asked. **2) It is somewhat clear**: The question is understandable but may have minor ambiguities. Choose it when uncertain. **3) It is clear**: The question is straightforward and unambiguous.
2 - Is the question relevant to the context of the given image? Given what is going on in the illustration of the image only, does the question make sense to be asked for this page?	**1) It is not at all related**: The question has no connection to the image. **2) It is somewhat related**: The question is indirectly related to the provided context. Choose it when uncertain. **3) It is related**: The question directly engages with the provided context.
3 - Is the question relevant to the context of the page's textual content? Only by referring to the text mentioned on the page and not the illustration.	**1) It is not at all related**: The question has no connection to the page's content. **2) It is somewhat related**: The question is indirectly related to the provided context. Choose it when uncertain. **3) It is related**: The question directly engages with the provided context.
4 - Is the question asking about an important aspect of the context (image and text)? Important aspects include: main events that support the storyline and are the core of the page content. This does NOT include, for example, details in illustrations that aren't relevant to the storyline or character development like "what is the person wearing?"	**1) Not at all important.** **2) It may be important.** **3) It is very important.**
5- Is the question appropriate for a child? Appropriate in terms of: Easy vocabulary. Does not include topics that could be frightening or too complex for the child (e.g., suicide/politics). Ask yourself, would I ask this question to a child?	**1) It is not appropriate**: The question contains content unsuitable for children. **2) It is somewhat appropriate**: The question may be suitable for children. Choose it when uncertain. **3) It is appropriate**: The question is suitable for children.

Table 6: Evaluation Questions used by Human evaluators and LLM

- *The question must start with What, Who, Where, Why, or How.*
- *Focus on descriptive details from the current page only (e.g., characters, actions, locations).*
- *Do not use multiple questions in one.*
- *Ensure the answer is directly supported by the text, without inference.*
- *Example: Where did the bear hide his food?*

(current page text)

"With that context, generate a prompt of type 'Wh' for the above text."

Format your response in JSON using exactly the template below:

```
{
  "prompt": PROMPT
}
```

- **Distancing Prompt**

The initial prompt is used to generate a candidate question.

You are a specialist in dialogic reading. Based on the following story excerpt, create a distancing question suitable for a 4–6 year-old child. The requirements are:

- *The question should encourage the child to connect the story to their own life experience.*
- *It must relate to the current story page but shift the frame of reference to the child's world.*

- *Use a wh-question or verb-based phrasing (e.g., Have you ever..., Can you remember...?).*

- *Ensure it cannot be answered with one word.*

- *Example: Have you ever had to help someone who was scared? What did you do?*

(current page text)
"With that context, generate a prompt of type 'distancing' for the above text."
Format your response in JSON using exactly the template below:

```
{
  "prompt": PROMPT
}
```

B.2 Question Evaluation

The evaluation prompt is used to assess the quality of the questions generated for children. Each evaluation considers both the page text and the illustration, but does not require explicit image description.

System instructions:

- *You are a helpful assistant tasked with evaluating educational questions for children.*

- *Each evaluation is based on a page of text and a corresponding illustration (image).*

- *Do not describe the image, only consider whether the question fits the context.*

- *Answer **only** in the specified JSON format, without explanation.*

Response format:

```
{
  "clarity": 1|2|3,
  "image_relevance": 1|2|3,
  "text_relevance": 1|2|3,
  "importance": 1|2|3,
  "appropriateness": 1|2|3
}
```

Evaluation criteria mapping:

- **Clarity** 1 = Not clear at all (ambiguous) 2 = Mostly clear (minor ambiguities, choose when uncertain) 3 = Very clear (straightforward and unambiguous)

- **Image relevance** 1 = Not related at all 2 = Somewhat related (indirect, choose when uncertain) 3 = Directly related to the image

- **Text relevance** 1 = Not related at all 2 = Somewhat related (indirect, choose when uncertain) 3 = Directly related to the text

- **Importance** 1 = Not important 2 = May be important 3 = Very important

- **Appropriateness** 1 = Not appropriate (unsuitable for children) 2 = Somewhat appropriate (uncertain) 3 = Appropriate for children