PIRLS Category-specific Question Generation for Reading Comprehension

Yin Poon[1], **Qiong Wang**[2], **John S. Y. Lee**[2], **Yu Yan Lam**[1],
Samuel Kai Wah Chu[1]

[1]School of Nursing and Health Studies, Hong Kong Metropolitan University,
{ypoon, yuylam, skwchu}@hkmu.edu.hk
[2]Department of Linguistics and Translation, City University of Hong Kong,
{wang.qiong, jsylee}@cityu.edu.hk

Abstract

According to the internationally recognized PIRLS (Progress in International Reading Literacy Study) assessment standards, reading comprehension questions should encompass all four comprehension processes: retrieval, inferencing, integrating and evaluation. This paper investigates whether Large Language Models can produce high-quality questions for each of these categories. Human assessment on a Chinese dataset shows that GPT-4o can generate usable and category-specific questions, ranging from 74% to 90% accuracy depending on the category.

1 Introduction

Given the importance of asking questions for effective learning (Dillon, 2006; Etemadzadeh et al., 2013; Kurdi et al., 2020), there has been extensive effort in developing automatic Question Generation (QG) models to produce high-quality questions for reading materials in educational systems (Heilman and Smith, 2010; Lindberg et al., 2013). Through automatic creation of pedagogical and assessment material, QG benefits teachers by reducing their workload. It also levels the playing field for students, providing them with instant and free access to questions for review and practice.

According to PIRLS (Progress in International Reading Literacy Study), reading requires four comprehension processes: retrieval, inferencing, integrating and evaluation (Mullis and Martin, 2019) as described in Table 1. A balanced set of questions, involving all four processes, is therefore needed to assess reading comprehension. However, existing QG benchmarks such as SQuAD (Rajpurkar et al., 2016) mostly focus on factoid short-answer questions.

Process	Description
Retrieval	Focus on and Retrieve Explicitly Stated Information
Inferencing	Make Straightforward Inferences
Integrating	Interpret and Integrate Ideas and Information
Evaluation	Evaluate and Critique Content and Textual Elements

Table 1: Comprehension processes in reading according to PIRLS (Mullis and Martin, 2019)

This paper investigates question generation of the four PIRLS categories with Large Language Models (LLMs) using zero-shot, few-shot and fine-tuning approaches. Our contribution is two-fold. In this first attempt of QG based on PIRLS, an internationally recognized standard for reading comprehension assessment, we show that GPT-4o can generate high-quality questions with category-specific prompts. Second, we contribute a dataset of Chinese passages and questions, annotated with PIRLS categories, as a benchmark for future research.[1]

2 Previous work

Early QG approaches mostly relied on heuristics, linguistic templates and rules (Labutov et al., 2015; Mostow et al., 2016). With the availability of large-scale datasets, QG began to be formulated as a sequence-to-sequence generation task. An encoder-decoder architecture with a global attention mechanism was found to be effective (Du et al., 2017; Kim et al., 2019), but can be further improved with transformer-based approaches (Scialom et al., 2019), and fully fine-tuned language models (LM) (Xiao et al., 2021). Answer-agnostic QG can be performed via joint Question and Answer Generation (QAG) (Lewis et al., 2021). A QAG model based on fine-tuning

[1] Code and data for this paper are available at `https://github.com/pypoon/PIRLS-QG-ZH`

<table>
<tr><td colspan="2">Excerpt of input passage (in Chinese):
传统的「英式奶茶」采用名贵锡兰红茶，加入牛奶和糖冲泡，饮用时会配以蛋糕。...
「港式奶茶」的对象是一般市民，食肆会选用较廉价的茶叶和淡奶，以降低成本。...
此外，为配合华人喜欢喝浓茶的习惯，「港式奶茶」茶味普遍较浓。...
The traditional "British milk tea" is made from posh Ceylon black tea, added with milk and sugar, and served with cake. ... "Hong Kong-style milk tea" is aimed at the general public, and restaurants will use cheaper tea leaves and evaporated milk to reduce costs. ... In addition, to match the Chinese habit of drinking strong tea, "Hong Kong-style milk tea" generally has a stronger tea flavor. ...</td></tr>
<tr><td>Type</td><td>Example Question</td></tr>
<tr><td>Retrieval</td><td>食肆如何降低奶茶的制作成本?
How can restaurants reduce the cost of making milk tea?</td></tr>
<tr><td>Inferencing</td><td>「英式奶茶」的目标客户群是哪些人?
Who are the target customers of "British milk tea"?</td></tr>
<tr><td>Integrating</td><td>「英式奶茶」和「港式奶茶」有什么区别?
What is the difference between "British milk tea" and "Hong Kong-style milk" tea?</td></tr>
<tr><td>Evaluation</td><td>作者先介绍「英式奶茶」,再介绍「港式奶茶」。作者为什么这样安排?
The author first introduces "British milk tea" and then "Hong Kong-style milk tea".
Why did the author arrange it this way?</td></tr>
</table>

Table 2: Example input passage and output questions of each PIRLS question type (Section 4)

encoder-decoder LMs produces high-quality questions (Ushio et al., 2022), but has not been evaluated in terms of question type. The most recent research has adopted LLMs. On a textbook dataset, few-shot prompting with GPT-3 was able to generate human-like questions ready for classroom use (Wang et al., 2022). A fine-tuned version of ChatGPT was able to generate questions that are competitive with human ones (Xiao et al., 2023).

Type-specific QG enables the user to request questions that suit their purposes. Controllable question generation has mainly focused on difficulty (Uto et al., 2023) and content (Li and Zhang, 2024), such as action, feeling, or setting. While Cao and Wang (2021) attempted QG according to a question topology (Olney et al., 2012), their approach was primarily template-based. In a study most closely related to ours, Elkins et al. (2023) used InstructGPT to generate six kinds of questions in Bloom's taxonomy (Krathwohl, 2002). Experimental results on Wikipedia passages on various disciplines showed that accuracy varied widely, from 36.1% to 91.7% across different categories. Since neither Olney's or Bloom's Taxonomy is designed for grade-school reading comprehension, this project will adopt the PIRLS framework. Further, we report the effect of fine-tuning LLMs and contribute a dataset in Chinese, which has more limited resources for QG.

3 Dataset

Existing reading comprehension datasets in Chinese, such as the Delta Reading Comprehension Dataset[2] and DuReader[3], are primarily drawn from newspapers, Wikipedia and user logs. Further, the questions are not annotated with their categories. We therefore constructed new datasets using Chinese-language pedagogical materials:

Training set The fine-tuning data consists of 804 manually composed questions about 72 passages taken from published Chinese story books. The average passage length is 1,131 Chinese characters. There are a total of 201 questions for each PIRLS category; 181 questions of these were used for training, and the remaining 20 for validation.

Test set The test set consists of 50 passages from a public reading comprehension assessment[4], with 25 passages from Grade 3, and 25 passages from Grade 6. The average passage length is 648 Chinese characters.

[2] https://github.com/DRCKnowledgeTeam/DRCD

[3] https://github.com/baidu/DuReader

[4] Downloaded from the website of the Territory-wide System Assessment (TSA) https://www.bca.hkeaa.edu.hk/web/TSA/en/PriPaperSchema.html.

4 Annotation Scheme

According to the International Association for the Evaluation of Educational Achievement, a reading comprehension question should address the following comprehension processes, as defined in the PIRLS standards (Table 1):

Retrieval The answer is explicitly given in a text span in the passage.

Inferencing Answering the question requires inferences about ideas or information that is not explicitly stated.

Integrating Answering the question "requires comprehension of the entire text, or at least significant portions of it." (Mullis and Martin, 2019)

Evaluation The answer "involves a judgement about some aspect of the text", and is not necessarily found in the passage.

Example questions for each category can be found in Table 2.[5]

5 Approach

The input is a Chinese text, without any specified answer span. We used two LLMs — GPT-4o[6] and *LLaMa-3* (Cui and Yao, 2024)[7] to generate questions[8] for the text, using the following prompts (see prompts in Table 6):

Zero-shot For each of the four PIRLS category, a different prompt describing the requirements of the category is used.

Generic Unlike the zero-shot approach, the prompt does not specify the question category. This serves to gauge the effectiveness of the description of PIRLS categories used in the zero-shot prompt.

Few-shot The PIRLS category-specific prompt used in zero-shot above is accompanied with an input passage and N sample questions, according to the template in Table 8 (Appendix B). We set $N = 5$, with a sample passage and five questions taken from the training set.

Fine-tuned We fine-tuned[9] LLaMa-3 on the training set (Section 3), using the PIRLS category-specific prompts shown in Table 6.

For each passage in the test set, a question was generated from each prompt type described above.

Model	Unusable	Usable	
		minor rev.	wo/ rev.
Llama-3 (generic)	4%	24%	72%
Llama-3 (zero-shot)	4%	17.5%	78.5%
Llama-3 (few-shot)	14%	15%	71%
Llama-3 (fine-tuned)	15%	26.5%	58.5%
GPT-4o (generic)	2%	10%	88%
GPT-4o (zero-shot)	**0%**	**4%**	**96%**

Table 3: Evaluation results on usability using the scale defined in Section 6

6 Evaluation set-up

Four assessors, all native Chinese speakers with a bachelor's degree, annotated each generated question on its *usability* and *PIRLS category*. The order of the questions was randomized to avoid bias. Each question was independently evaluated by two of the assessors. In case of disagreement, a PIRLS expert with a Master's degree in Education, adjudicated the decision.

First, the assessors rated the quality of the question on the following three-point scale:

Usable without revision The question can be used as is: it is grammatical, fluent, and relevant for the input passage.

Usable with minor revision The question is relevant for the input passage, but requires improvement in its linguistic quality, e.g., correction of grammatical errors, better vocabulary choice or phrasing.

Unusable The question is irrelevant for the passage, or cannot be understood.

[5] The Chinese passage is taken from a Chinese-language public examination at https://www.hkeaa.edu.hk/en/sa_tsa/

[6] https://openai.com/index/hello-gpt-4o/

[7] Chinese 8B Instruct-v1, downloaded from https://huggingface.co/hfl/llama-3-chinese-8b-instruct

[8] max_tokens=200; temperature=0.6; top_p=0.9 for both LLMs

[9] The fine-tuning was performed for 1 epoch using the following hyperparameters: learning rate=1e-4; lora_rank=64; lora_alpha=128; lora_dropout=0.05; batch_size = 1; gradient_accumulation_steps=8; max_seq_length=3303. On an A100 GPU, the training took 4 minutes and 34 seconds.

Model	PIRLS category				Average
	Retrieval	Inferencing	Integrating	Evaluation	
Llama-3 (generic)	56%	32%	8%	0%	24%
Llama-3 (zero-shot)	78%	40%	22%	20%	40%
Llama-3 (few-shot)	82%	26%	10%	4%	30.5%
Llama-3 (fine-tuned)	68%	42%	10%	34%	38.5%
GPT-4o (generic)	54%	32%	12%	0%	24.5%
GPT-4o (zero-shot)	**86%**	**74%**	**78%**	**90%**	82%

Table 4: Accuracy in question category (denominator includes unusable questions)

Category	Retrieval	Infer.	Integr.	Eval.
Retrieval	**43**	6	1	0
Infer.	8	**37**	3	2
Integr.	0	3	**39**	8
Eval.	0	0	5	**45**

Table 5: Confusion matrix of the PIRLS category of the questions generated by GPT-4o (zero-shot)

Then, the usable questions (either without revision or with minor revision) were classified in terms of PIRLS question type (Section 4).

7 Results

7.1 Question Usability

Inter-annotator agreement. The four assessors agreed on 90% of questions on the usable vs. unusable classification, leading to a 0.499 weighted Kappa score, a "moderate" level of agreement (Landis and Koch, 1977).

Usability. Using the generic prompt, only 72% of the questions generated by Llama-3 were usable without revision (Table 3). The category-specific zero-shot prompt, which supplied more detailed requirements on the questions to be generated, increased the proportion of directly usable questions to 78.5%. Providing examples through few-shot and fine-tuning, however, resulted in more unusable questions. Our human evaluators reported that the model was led to overly prefer the wording in the given samples, even if it results in unnatural questions.

On GPT-4o, the category-specific prompts also led to gains in usability over the generic one. Overall, GPT-4o attained substantially superior performance, with a vast majority of the generated questions (96%) assessed as directly usable.

7.2 Question category

Inter-annotator agreement. Excluding the unusable questions, the assessors agreed on 55.17% of the generated questions on the 4-way classification of PIRLS category. This yielded a 0.494 weighted kappa score, a "moderate" level of agreement (Landis and Koch, 1977).

Accuracy in category. As expected, the generic prompt, which gave no specific instruction on question category, led to the lowest accuracy for both Llama-3 (24%) and GPT-4o (24.5%). Both models would be hardly useful for teachers looking for higher-order questions that require inferencing, integrating or evaluation, since they produced mostly 'retrieval'-type questions (56% and 54%, respectively). The category-specific (zero-shot) prompts improved the accuracy across all categories, raising the average accuracy to 40% for Llama-3 and 82% for GPT-4o. This result suggests that both models were able to understand the instructions in the prompt.

On Llama-3, the few-shot approach improved the generation of 'retrieval' questions to 82%. The five samples, however, appeared to be insufficient for the higher-order categories, resulting in lower accuracy. With larger quantity of training data for these higher-order categories, the fine-tuned model offered better performance for 'Inferencing' and 'Evaluation'.

The GPT-4o zero-shot approach achieved the best performance across all categories, with an average of 82% accuracy. As shown in the confusion matrix (Table 5), most errors were within one category above or below the target in the PIRLS scale.

8 Conclusion

A variety of question types, targeting various comprehension processes, is necessary for assessing reading comprehension. This paper has presented the first study on automatic question generation for reading comprehension based on the four categories in the PIRLS framework. Experiments on

Chinese passages show that zero-shot GPT-4o can produce questions belonging to the target category at 74% to 90% accuracy, outperforming both the zero-shot and fine-tuned LLaMA-3 model.

This research has focused on assisting teachers in designing a variety of question types, to test students' skills in reading comprehension. In future work, we plan to extend the experiment to the quality of the answers, to further automate the test design process. We also plan to deploy the automatically generated questions in real-world classrooms to measure their pedagogical impact on students.

Limitations and Ethics Consideration

At the time of system deployment, users should be clearly informed that the automatically generated questions should be viewed only as a first draft, to minimize the risk that the teacher may fail to edit an unusable question and pass it to students.

Considering the high cost of using few-shot generation, we did not test GPT-4o on few-shot prompts in this paper. Typically, generating integrating and evaluation questions requires a full text or several passages. Our focus was on finding a cost-effective approach to generate reading comprehension questions. Therefore, we suggest that future research explore the few-shot prompts in GPT-4o.

Acknowledgements

We gratefully acknowledge support from the e-Learning Ancillary Facilities Programme of the Quality Education Fund (Project "Knowledge Overlord - A Self-sustaining AI Game-based Online Platform to Enhance Student's Literacy Ability and 21st Century Skills"); and from a Teaching Development Grant at City University of Hong Kong (project 6000834).

References

Shuyang Cao and Lu Wang. 2021. Controllable Open-ended Question Generation with A New Question Type Ontology. In *Proc. 59th Annual Meeting of the Association for Computational Linguistics*, page 6424–6439.

Y. Cui and X. Yao. 2024. Rethinking LLM Language Adaptation: A Case Study on Chinese Mixtral. In *arXiv preprint arXiv:2403.01851.*

James T. Dillon. 2006. Effect of questions in education and other enterprises. In *Rethinking schooling*, page 145–174. Routledge.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to Ask: Neural Question Generation for Reading Comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL).*

Sabina Elkins, Ekaterina Kochmar, Iulian Serban, and Jackie C. K. Cheung. 2023. How Useful Are Educational Questions Generated by Large Language Models? *AIED 2023, CCIS*, 1831:536–542.

Atika Etemadzadeh, Samira Seifi, and Hamid Roohbakhsh Far. 2013. The role of questioning technique in developing thinking skills: The ongoing effect on writing skill. *Procedia-Social and Behavioral Sciences*, 70:1024–1031.

Michael Heilman and Noah A. Smith. 2010. Good Question! Statistical Ranking for Question Generation. In *Proc. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL (HLT-NAACL)*, page 609–617.

Yanghoon Kim, Hwanhee Lee, Joongbo Shin, and Kyomin Jung. 2019. Improving Neural Question Generation Using Answer Separation. In *Proc. 33rd AAAI Conference on Artificial Intelligence (AAAI-19).*

D. R. Krathwohl. 2002. A revision of Bloom's taxonomy: An overview. *Theory into practice*, 41(4):212–218.

Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler, and Salam Al-Emari. 2020. A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30:121–204.

I. Labutov, S. Basu, and L. Vanderwende. 2015. Deep questions without deep understanding. In *Proc. ACL.*

J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33:159–174.

Patrick Lewis, Yuxiang Wu, Linqing Liu, Pasquale Minervini, Heinrich Küttler, Aleksandra Piktus, Pontus Stenetorp, and Sebastian Riedel. 2021. PAQ: 65 million probably-asked questions and what you can do with them. *Transactions of the Association for Computational Linguistics*, 9:1098–1115.

Kunze Li and Yu Zhang. 2024. Planning first, question second: An llm-guided method for controllable question generation. In *indings of the Association for Computational Linguistics ACL 2024*, page 4715–4729.

David Lindberg, Fred Popowich, John Nesbit, and Phil Winne. 2013. Generating natural language questions to support learning on-line. In *Proceedings of the 14th European Workshop on Natural Language Generation*, page 105–114.

Jack Mostow, Yi ting Huang, Hyeju Jang, Anders Weinstein, Joe Valeri, and Donna Gates. 2016. Developing, evaluating, and refining an automatic generator of diagnostic multiple choice cloze questions to assess children's comprehension while reading. *Natural Language Engineering*, 23(2):245–294.

Ina V. S. Mullis and Michael O. Martin. 2019. *PIRLS 2021 Assessment Frameworks*. International Association for the Evaluation of Educational Achievement.

Andrew M. Olney, Arthur C. Graesser, and Natalie K. Person. 2012. Question generation from concept maps. *Dialogue and Discourse*, 3(2):75–99.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, page 2383–2392.

Thomas Scialom, Benjamin Piwowarski, and Jacopo Staiano. 2019. Self-Attention Architectures for Answer-Agnostic Neural Question Generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, page 6027–6032.

Asahi Ushio, Fernando Alva-Manchego, and Jose Camacho-Collados. 2022. Generative Language Models for Paragraph-Level Question Generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, page 670–688.

Masaki Uto, Yuto Tomikawa, and Ayaka Suzuki. 2023. Difficulty-Controllable Neural Question Generation for Reading Comprehension using Item Response Theory. In *Proc. 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, page 119–129.

Z. Wang, J. Valdez, D. Basu Mallick, and R. G. Baraniuk. 2022. Towards Human-Like Educational Question Generation with Large Language Models. *Artificial Intelligence in Education. AIED 2022. Lecture Notes in Computer Science*, 13355.

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 *Proc. 18th Workshop on Innovative Use of NLP for Building Educational Applications*, page 610–625.

Dongling Xiao, Han Zhang, Yukun Li, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. Ernie-gen: an enhanced multi-flow pre-training and fine-tuning framework for natural language generation. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, page 3997–4003.

A Appendix: Instruction to Human Assessors

The human assessors gave consent to the data collection and were informed that the results would remain anonymous. They were shown the following instructions:

```
<passage>
<question>
```

1. Is the question understandable and relevant for the passage?

2. Does the language quality of the question need to be improved?

3. If the answer to #1 is "Yes", choose one of the categories for the question:

   - Retrieval (Focus on and Retrieve Explicitly Stated Information)
   - Inferencing (Make Straightforward Inferences)
   - Integrating (Interpret and Integrate Ideas and Information)
   - Evaluation (Evaluate and Critique Content Textual Elements)

B Appendix: Few-shot prompt template

The prompts are shown in Table 6, and their English translation in Table 7. The few-shot template is shown in Table 8.

Type	Prompt (in Chinese)
System prompt	你是一個能幹的閱讀理解問題生成器，始終遵循給定的説明和要求來生成問題。
Generic prompt	基於所提供的文章，請創作一個簡答題，並提供對應的答案。 文章:{input passage}
Retrieval questions (PIRLS level 1)	基於所提供的文章，請創作一個屬於PIRLS第一層次的簡答題，並提供對應的答案。這個問題應著重於檢索文本中明確表述的信息，也就是資訊檢索型的問題。此類問題要求考生識別和回憶文本中明確提到的信息，如事件的順序、角色的特徵或進行比較等。 文章:{input passage}
Inferencing questions (PIRLS level 2)	基於所提供的文章，請創作一個屬於PIRLS第二層次的簡答題，並提供對應的答案。這個問題應鼓勵考生從文本中進行直接推理，進一步超越單純的信息提取，也就是需要進行簡單推理的問題。這類問題需要考生進行直接推理，例如理解因果關係或推測未明確陳述但可以從文本邏輯推導出的結果。 文章:{input passage}
Integrating questions (PIRLS level 3)	基於所提供的文章，請創作一個屬於PIRLS第三層次的簡答題，並提供對應的答案。這個問題應促使考生解釋想法並整合文本不同部分信息，也就是需要進行解釋及整合的問題。這類問題需要考生全面理解並能夠從文本的不同部分綜合信息，如解釋角色的感受和行為，並整合文本中的想法和信息。 文章:{input passage}
Evaluation questions (PIRLS level 4)	基於所提供的文章，請創作一個屬於PIRLS第四層次的簡答題，並提供對應的答案。這個問題應需要考生批判性地檢視和評估文本內容、語言和文本元素，也就是評鑒型的問題。這類問題是最高層次的問題，問題挑戰考生批判性地評估文本的內容、語言和文本元素，如對價值、期望和接受度作出判斷，或考慮他們如果處於某個角色的位置會如何反應。 文章:{input passage}

Table 6: LLM prompts for generating questions for each PIRLS category

Type	Prompt (in English)
System prompt	You are a capable reading comprehension question generator, always following the given instructions and requirements to generate questions.
Generic prompt	Based on the given passage, create a short-answer question and provide a corresponding answer. article:{input passage}
Retrieval questions (PIRLS level 1)	Based on the article provided, please create a short answer question belonging to PIRLS level 1 and provide the corresponding answer. This question should focus on retrieving information explicitly stated in the text, i.e. an information retrieval type question. This kind of question requires candidates to identify and recall information explicitly mentioned in the text, such as the sequence of events, character traits, or making comparisons. article:{input passage}
Inferencing questions (PIRLS level 2)	Based on the article provided, please create a short answer question belonging to PIRLS level 2 and provide the corresponding answer. This question should encourage candidates to make straightforward inferences from the article, moving further beyond information retrieval, i.e. a question requiring simple inferences. This type of question requires candidates to make straightforward inferences, such as understanding cause and effect relationships or inferring consequences that are not explicitly stated but can be logically deduced from the text. article:{input passage}
Integrating questions (PIRLS level 3)	Based on the article provided, please create a short answer question belonging to the PIRLS level 3 and provide the corresponding answer. This question should prompt the candidate to interpret ideas and integrate information from different parts of the text, i.e. a question that requires interpretation and integration. This type of question requires candidates to have a comprehensive understanding and be able to integrate information from different parts of the text, such as explaining a character's feelings and actions, and integrating ideas and information across the text. article:{input passage}
Evaluation questions (PIRLS level 4)	Based on the article provided, please create a short answer question belonging to PIRLS level 4 and provide the corresponding answer. This question should require candidates to critically examine and evaluate the text content, language, and textual elements, i.e. an evaluative question. This type of question is the highest-level question that challenges candidates to critically evaluate a text content, language, and textual elements, such as making judgments about value, desirability, and acceptability or considering how they would react if they were in a character's position. article:{input passage}

Table 7: LLM prompts for generating questions for each PIRLS category (English translation)

{category-specific prompt}
範例文章及相應的範例問題(請參考範例來創作問題):
{範例文章:{example passage}
PIRLS第{required level}層次範例問題1:{example question-answer pair 1}
...
PIRLS第{required level}層次範例問題5:{example question-answer pair 5}}
文章: {input passage}

Table 8: Prompt template for few-shot question generation