Few-shot Question Generation for Reading Comprehension

Yin Poon,
John S. Y. Lee
Dept. of Linguistics and Translation
City University of Hong Kong
{yinpoon2,jsylee}@cityu.edu.hk

Yu Yan Lam, Wing Lam Suen,
Elsie Li Chen Ong, Samuel Kai Wah Chu
School of Nursing and Health Studies
Hong Kong Metropolitan University
{yuylam,wlsuen,
eong,skwchu}@hkmu.edu.hk

Abstract

According to the internationally recognized PIRLS (Progress in International Reading Literacy Study) assessment standards, reading comprehension questions should require not only information retrieval, but also higher-order processes such as inferencing, interpreting and evaluation. However, these kinds of questions are often not available in large quantities for training question generation models. This paper investigates whether pre-trained Large Language Models (LLMs) can produce higher-order questions. Human assessment on a Chinese dataset shows that few-shot LLM prompting generates more usable and higher-order questions than two competitive neural baselines.

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 comprehension questions should require not only information retrieval, but also higher-order processes such as inferencing, interpreting and evaluation (Mullis and Martin, 2019). However, existing QG benchmarks such as SQuAD (Rajpurkar et al., 2016) mostly focus on factoid short-answer questions. There is therefore a dearth of publicly available training data for the more challenging types of questions (Mulla and Gharpure, 2023) — those requiring inference, synthesis and critique — especially for languages other than English.

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). The italicized processes are those required by *higher-order* questions.

This paper investigates the generation of these higher-order questions with few or no training samples. Our contribution is two-fold. First, we report the first QG evaluation based on PIRLS, an internationally recognized standard for reading comprehension assessment, and demonstrate a high level of human agreement on PIRLS question type classification (Table 1). Second, in experiments on a Chinese dataset, we show that existing QG neural models generate predominantly information-retrieval questions, while few-shot prompting of a Large Language Model (LLM) can generate higher proportions of higher-order questions. The LLM-based approach can therefore produce a balanced set of questions that is desirable in the education setting with minimal supervision.

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 ef-

Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10), pages 21–27
August 16, 2024 ©2024 Association for Computational Linguistics

<table>
<tr><td colspan="2">Excerpt of input passage (in Chinese):
太阳和地球虽然相距1.5亿公里，但它却会提供光和热。除此以外，它还会给地球带来意想不到的“礼物”呢！其实太阳的表面常常发生爆炸，在最活跃的时候，更会把表面的物质抛射出去，形成太阳风暴。当太阳风暴经过地球时，不但会损毁人造卫星，干扰无线电通讯，...
Even though the Sun is 150 million kilometers away from Earth, it provides light and heat. Besides, it also gives a surprising ‘gift’ to Earth! There are frequent explosions on the surface of the Sun ... forming solar storms. When a solar storm passes by the Earth, it not only destroys satellites and interfere with wireless communication, ...</td></tr>
<tr><td>Type</td><td>Example Question</td></tr>
<tr><td>Retrieval: word-match</td><td>太阳和地球虽然相距一亿五千万公里，但它却会提供什么？
Even though the Sun is 150 million kilometers away from Earth, What does it provide?</td></tr>
<tr><td>Retrieval: paraphrase</td><td>文章提到太阳和地球之间的距离是多少?
What is the distance between the sun and the Earth, as mentioned in the passage?</td></tr>
<tr><td>Inferencing</td><td>根据文章，太阳爆炸造成的“太阳风暴”会对地球造成哪些影响?
How is the Earth affected by the solar storms caused by explosions on the Sun?</td></tr>
<tr><td>Integrating</td><td>文章中提到太阳常常发生爆炸会带来什么「礼物」？
According to the passage, what ‘gift’ is brought by the frequent explosions at the Sun?</td></tr>
<tr><td>Evaluation</td><td>作者认为太阳的影响对地球有什么优势和缺陷
What does the author think are the Sun’s positive and negative impact on the Earth?</td></tr>
</table>

Table 2: Example input passage and output questions of each PIRLS question type (Section 3.2)

fective (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 encoder-decoder LMs produces high-quality questions (Ushio et al., 2022), but has not been evaluated in terms of question type.

There have been a few QG studies on LLMs in the education setting. 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 similar approach with InstructGPT achieved an adherence rate between 67% and 69% for generating 9 question types (Elkins et al., 2023). A fine-tuned version of ChatGPT was able to generate questions that are competitive with human ones in terms of readability, correctness, coherence and engagement (Xiao et al., 2023). It remains unknown how these approaches compare to off-the-shelf neural QG models in terms of generating higher-order questions.

3 Evaluation metric

To accurately evaluate the utility and nature of the generated questions, manual assessment is necessary since automatic methods cannot yet reliably determine usability and PIRLS question types.

3.1 Usability

The human assessor assesses 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.

A question classified as one of the first two categories is said to be “***usable***”. Only usable questions are further analyzed on their question type.

3.2 PIRLS question type

According to the International Association for the Evaluation of Educational Achievement, a reading comprehension question should address one of four 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 can be found in Table 2.[1] A question classified as Inferencing, Integrating or Evaluation is considered as "***higher order***". For pedagogical purposes, a well-balanced set of questions should include not only Retrieval questions but also higher-order ones (Mullis and Martin, 2019).

4 Approach

We adopted the answer-agnostic setting for QG, since the target answer is not always found within the input text. The input is a Chinese text without any specified answer span.

4.1 Baseline: pipeline model

We used the DuReader pipeline QG model (Li et al., 2021), a publicly available QG system for Chinese. It performs two subtasks in sequence: answer generation[2] using an extractor trained in the Universal IE framework (Lu et al., 2022)[3]; followed by question generation[4] with a base model fine-tuned with UNIMO (Li et al., 2021).[5]

4.2 Baseline: Seq2seq model

A seq2seq model, trained directly to generate a question-answer pair from a passage, serves as a second baseline. It has been found to be robust in comparison with the pipeline and multitask approach, and computationally less intensive (Ushio et al., 2023).[6] We used the Chinese version of their publicly available end-to-end QAG model.[7]

Model	Unusable	Usable w/ minor rev.	Usable wo/ rev.
Zero-shot	31.5%	6.0%	62.5%
Few-shot	22.0%	7.0%	**71.0%**
Pipeline	46.5%	**18.5%**	35.0%
Seq2seq	**54.0%**	11.0%	35.0%

Table 3: Evaluation results on usability

4.3 LLM: Zero-shot

We used the Chinese version of Stanford Alpaca (Cui et al., 2023)[8], a LLaMA Model that can comprehend and execute instructions (Touvron et al., 2023).[9] We are not aware of any published research on prompt engineering for Chinese QG. Six candidate prompts, with varied keywords on inference, reasoning, and word usage were informally evaluated on a small set of passages randomly taken from Chinese-language public examinations.[10] As shown in Table 7 (Appendix B), the following prompt produced the largest number of usable and non-word-matching questions:

> 基于给定的文章，生成一个需要推断
> 的简答题。你的输出应该包含一个简
> 答问题和这个问题的对应的答案。
> 文章:<input>

[Translation: "Based on the given passage, generate a short-answer question that requires deduction. Your output should include a question and its answer. Passage: <input>]

4.4 LLM: Few-shot

In the few-shot approach, the prompt above is accompanied with N sample pairs of input passage and question, according to the template in Table 8 (Appendix B). We set $N = 5$, with all five sample passage-question pairs taken from the public examination papers mentioned above.

5 Dataset

Our evaluation data was drawn from the dev set of DuReader_robust (Tang et al., 2021), a widely used Chinese Q&A dataset[11]. Due to its filtering step, the pipeline model in Section 4.1 may not

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

[2] https://github.com/PaddlePaddle/PaddleNLP/tree/develop/applications/question_answering/unsupervised_qa

[3] We used the extractor `uie-base-answer-extractor` and the filter `uie-base-qa-filter`

[4] https://github.com/PaddlePaddle/Research/tree/master/NLP/UNIMO

[5] `unimo-text-1.0-question-generation`

[6] https://github.com/asahi417/lm-question-generation/

[7] `mt5-small-zhquad-qag`

[8] `Chinese-Alpaca-2-7B`

[9] https://github.com/tatsu-lab/stanford_alpaca

[10] https://www.hkeaa.edu.hk/en/sa_tsa/

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

Model	Unusable	Retrieval	Higher-order			Total
			Inferencing	Integrating	Evaluation	higher-order
Zero-shot	31.5%	39.0%	15.5%	9.0%	**5.0%**	29.5%
Few-shot	22.0%	**46.5%**	**16.5%**	**13.5%**	1.5%	**31.5%**
Pipeline	46.5%	45.5%	6.0%	2.0%	0%	8.0%
Seq2seq	**54.0%**	39.5%	4.5%	2.0%	0%	6.5%

Table 4: Evaluation results on PIRLS question types (first 5 columns add to 100%)

generate any question for some passages. Our test set consists of the first 200 passages for which the pipeline model successfully produced an output.

Two human assessors, both native speakers of Chinese with a Bachelor's degree, independently evaluated the questions generated for each of these 200 passages in terms of their usability (Section 3.1) and question type (Section 3.2). A third assessor, a native speaker of Chinese with a Master's degree, adjudicated in case of disagreement.

6 Agreement

The two assessors agreed 85.0% of the time in the 3-way classification on usability (Section 3.1), leading to a Kappa of 0.739, a "substantial" level of agreement (Landis and Koch, 1977).

In terms of question types, the two human assessors agreed in 93.5% of the cases, yielding a Kappa of 0.861, at the "Almost perfect" level of agreement (Landis and Koch, 1977) The most common disagreement (19 cases) is between Retrieval and Inferencing, typically in judging whether a paraphrase deviates sufficiently from the original expression to require inference. The two assessors also disagreed in 9 cases on whether the answer must be derived from different parts of the passage (Integrating) or from just a single sentence (Inferencing).

7 Results

7.1 Usability

The LLM-based approaches attained higher usability rates (Table 3). Among questions generated by zero-shot prompting, 62.5% can be used without revision. Few-shot prompting, with only five example passage-question pairs, produced a significant boost, with 71% ready for use without revision. The pipeline and Seq2seq neural models yielded substantially more unusable questions and fewer questions that are immediately ready (35.0%). The amount of unlabeled language data used in training — an order of magnitude larger in LLMs than the neural models — likely contributed to the grammaticality and fluency of the generated questions.

	Retrieval	Infer.	Integr.	Eval.
Retrieval	334	13	3	0
Infer.	6	66	1	0
Integr.	1	8	47	0
Eval.	0	0	0	13

Table 5: Confusion matrix of the two human annotators on PIRLS question types

7.2 PIRLS question types

Both neural QG models produced very limited number of higher-order questions, likely because there were few such questions in the training samples. Despite the lack of such samples, zero-shot LLM produces substantially more higher-order questions (29.5%), and few-shot prompting further increases the proportion (31.5%) (Table 4). It appears that Alpaca was able to learn the characteristics of higher-order questions even with only five samples.

8 Conclusion

Higher-order questions are important for assessment in reading comprehension. However, there is a lack of publicly available datasets of these challenging questions in languages other than English. This paper has presented the first study on automatic question generation (QG) for reading comprehension based on PIRLS, assuming no or minimal supervision. Experiments on Chinese passages show that zero-shot LLM produces more usable and more higher-order questions than two competitive off-the-shelf neural QG models, and few-shot prompting further improves the performance.

In future work, we plan to investigate tailored prompts for producing the different PIRLS question types, and to construct a Chinese dataset of higher-order questions for fine-tuning an LLM.

Limitations

The evaluation has focused on the quality of the questions, but cannot show their pedagogical impact on the students. 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.

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

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and Effective Text Encoding for Chinese LLaMA and Alpaca. In *arXiv:2304.08177.*

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).*

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.

Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. 2021. UNIMO: Towards Unified-Modal Understanding and Generation via Cross-Modal Contrastive Learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, page 2592–2607.

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.

Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2022. Unified Structure Generation for Universal Information Extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, page 5755–5772.

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.

N. Mulla and P. Gharpure. 2023. Automatic question generation: a review of methodologies, datasets, evaluation metrics, and applications. *Progress in Artificial Intelligence*, 12:1–32.

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

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.

Hongxuan Tang, Hongyu Li, Jing Liu, Yu Hong, Hua Wu, and Haifeng Wang. 2021. DuReader_robust: A Chinese Dataset Towards Evaluating Robustness and Generalization of Machine Reading Comprehension in Real-World Applications. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, page 955–963.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and efficient foundation language models. In *https://arxiv.org/abs/2302.13971*.

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.

Asahi Ushio, Fernando Alva-Manchego, and Jose Camacho-Collados. 2023. An Empirical Comparison of LM-based Question and Answer Generation Methods. In *Findings of the Association for Computational Linguistics: ACL 2023*, page 14262–14272.

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: Prompt selection and implementation

Table 6 lists the six prompts that were evaluated. The top of Table 7 shows zero-shot evaluation results on a set of 42 passages randomly chosen from public examinations on the Chinese-language subject in Hong Kong.[12] “Creative” refers to the parameter values {temperature=0.8, top_p=1}. Prompt #3 was found to produce the highest proportion of usable questions and questions that are not word-matching in nature.

The bottom of Table 7 shows the tuning of the temperature and top_p values. “Conservative” refers to the values {temperature=0.5, top_p=0.5}; “Less Creative” refers to the values {temperature=0.6, top_p=0.9}. We empirically set the temperature and top_p values at 0.6 and 0.9 in the rest of the experiments since they produced more usable and non-word-matching questions than the other values.

The few-shot template is shown in Table 8.

[12] https://www.hkeaa.edu.hk/en/sa_tsa/

ID	Prompt (in Chinese)	Keywords
0	基于给定的文章，你需要提炼出一个答案，并以此答案为基础 构建一个问题。你的输出应该包含问题和答案。 文章:{input}	none
1	基于给定的文章，提炼出一个答案，然后根据这个答案创造一个 需要推理的问题。确保你的输出包含这个需要推理的问题和 对应的答案。 文章:{input}	reasoning
2	基于给定的文章，提炼出一个答案，然后根据这个答案生成一个 新的简答题，也就是说，新的简答题需要使用与上下文不同的 词语来表达相同的含义。你的输出应该包含那个简答问题和 对应的答案。输出格式如下所示: 问题: 答案: 文章:{input}	vocabulary
3	基于给定的文章，生成一个需要推断的简答题。你的输出应该 包含一个简答问题和这个问题的对应的答案。 文章:{input}	deduction
4	请根据文章内容，生成一个需要推理的简答题。你的输出格式 应如下所示: 问题: 答案: 文章:{input}	reasoning
5	根据文章，生成一个需要推断的问题。问题措辞需要与上下文 不会完全一样。你的输出应该包含问题和答案。 文章:{input}	deduction; vocabulary

Table 6: Candidate prompts (in Chinese) for LLM-based question generation with keywords specifying deduction (*tuiduan*), reasoning (*tuili*), and varied vocabulary (keywords are underlined in this table for clarity but not in the experiments)

ID	Parameters	% Usable	% Non-word-matching
0	Creative	47.62	40.48
1		57.14	57.14
2		59.52	45.24
3		**66.67**	**61.9**
4		61.9	59.52
5		54.76	52.38
3	Conservative	**73.81**	61.9
3	Less Creative	**73.81**	**66.67**
3	Creative	66.67	61.9

Table 7: Evaluation results for prompt selection and parameter tuning (the prompt corresponding to each ID can be found in Table 6)

文章: {example passage 1}
简答题: {example question 1}
答案: {example answer 1}
...
文章: {example passage 5}
简答题: {example answer 5}
答案: {example question 5}

基于给定的文章，生成一个需要
推断的简答题。你的输出应该包含
一个简答问题和这个问题的对应的答案。
文章: <input>
简答题:
答案:

Table 8: Prompt template for few-shot question generation [Translation: “Based on the given passage, generate a short-answer question that requires inference. Your output should include a question and its answer. Passage: <input>]