Improving Question Generation Quality for Educational LLMs: Evaluation Framework Construction and Fine-Tuning Optimization

Anonymous ACL submission

Abstract

The application of large language models (LLMs) in the field of education is becoming increasingly widespread, and question generation based on LLMs has begun to attract more attention. This is because it can save educators' time and enable personalized learning. However, existing studies mostly focus on the local rationality of the content generated by models, lacking a systematic comparison between the generated questions and human-crafted questions in terms of their overall characteristics. This work proposes an evaluation framework that covers both the content and form of questions, which comprehensively measures the gap between question generation by LLMs and that by humans, and puts forward a series of improvement methods targeting this gap. Specifically, we focus on seven dimensions to measure the differences between human-crafted and model-generated questions. Based on these findings, we propose a zero-shot method (Chain-of-Thought Prompting for Question Generation, CPQG) that operates without external knowledge bases. By combining CoT reasoning with prompt engineering, CPQG significantly enhances the model's intrinsic generation quality. Extensive experiments demonstrate that CPQG effectively narrows the gap between model-generated and human-crafted questions. Compared with the baseline, CPQG enables 7B-sized models to achieve a 10% average performance gain and even surpass GPT-4 in multiple dimensions.

1 Introduction

Large Language Models (LLMs) have achieved remarkable success in natural-language understanding (OpenAI and others, 2024; DeepSeek-AI et al., 2025), question answering (Zhang et al., 2023; Lu et al., 2022), and text generation (Li et al., 2024), sparking intense interest in their application to intelligent education (Dan et al., 2023; Zhang et al., 2025). Recently, several initiatives have developed education-oriented LLMs (ELLMs), such as EduChat (Dan et al., 2023), MathGPT[1], and MuduoLLM (Zhu et al., 2025). Building on these advancements, researchers have explored personalized education (Bhutoria, 2022; Maghsudi et al., 2021), tutoring systems (Gao et al., 2025; Wang and others, 2025), and learning content generation (Shao et al., 2024; Valentini et al., 2023). As a core capability, LLM-based question generation can significantly reduce educator workload while enabling adaptive learning experiences (Dan et al., 2023; Zhu et al., 2025).

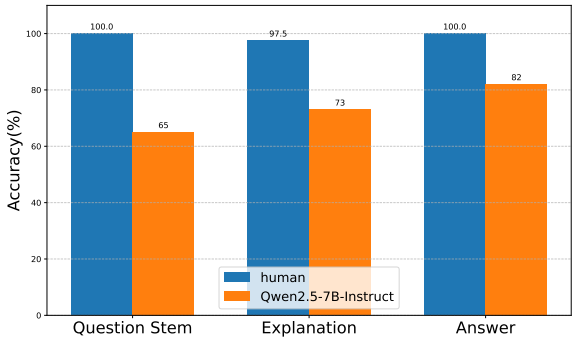

Figure 1: Results of quality comparison between model-generated questions and human-generated questions in three aspects (question stem, explanation, and answer).

Based on this, there has been a lot of research exploring question generation by LLMs. For example, researchers (Wang and others, 2025) attempted to guide the model in generating questions through prompt engineering. Some work tried to construct multiple-choice questions that are close to those generated by humans (Lee et al., 2024a; Scarlatos et al., 2024; Feng et al., 2024; Biancini et al., 2024). However, existing studies mostly focus on the local rationality (e.g., the fluency of questions) of the content generated by models, emphasizing the specific content of the questions, but lack a systematic

[1] https://math-gpt.org/

comparison between the generated questions and human-crafted questions in terms of their overall characteristics. This limitation makes it difficult for models to capture the overall logic of human-crafted questions, and thus unable to truly narrow the gap with manually generated questions. Figure 1 presents the comparison results between model-generated questions and manually generated questions across three common dimensions. This figure clearly indicates that there is a huge gap between current model-generated questions and manually generated ones.

For the above reasons, this paper supplements and improves a set of fine-grained evaluation frameworks for question generation by models, covering both content and form, based on the work of (Zhou et al., 2025), as shown in Figure 2. On this basis, a series of zero-shot methods to improve the quality of questions generated by models themselves are proposed to address the problems as shown in Figure 3, which significantly narrow the gap between model-generated questions and human-crafted ones. Specifically, we have improved the evaluation framework proposed by (Zhou et al., 2025), expanding the original five dimensions to seven. Detailed evaluation criteria for the supplementary dimensions are formulated based on the characteristics of human-crafted questions. A prompt dataset for question generation is constructed reversely from real questions, and LLMs are used as evaluation tools to conduct comprehensive evaluations on questions generated by humans, those generated by closed-source LLMs, and those generated by open-source models. Based on the evaluation results, we found that currently, the questions generated by models have significant flaws in three main dimensions, namely the question stem, explanations, and answers. In particular, the explanation aspect has become a shortcoming restricting the overall performance. Based on these three types of flaws, we have proposed a zero-shot method. This method endows the model with self-reflection ability through instruction fine-tuning combined with the latest long CoT (Chain of Thought) reasoning method, so as to enhance the quality of explanation generation. In addition, we have further subdivided the three dimensions of question stems, explanations, and answers, counted the specific errors occurring in them, and proposed targeted correction prompts to further improve the quality of question generation by the model. Experiments show that CPQG significantly narrows the gap between questions generated by models and those generated by humans.

Our contributions are summarized as follows:

- This paper proposes a comprehensive and fine-grained evaluation framework for questions generated by models, covering both content and form, which can accurately identify the differences between questions generated by large models and those generated by humans.
- A zero-shot method for improving the model's question generation ability is put forward, which systematically enhances the quality of the model's question generation by combining Chain of Thought (CoT) technology and prompt engineering.
- Experiments show that CPQG is significantly superior to other methods and surpasses GPT4 in multiple dimensions.

2 Related Work

2.1 Traditional Question Generation

Human-crafted questions design remain the dominant practice in education, yet creating well-targeted questions for every learner demands enormous time and effort from teachers. Early attempts at automation followed two main paths. One is rule-based systems (Heilman and Smith, 2010a; Chali and Hasan, 2015; Heilman and Smith, 2010b), which convert declarative sentences into interrogative ones, but rigid, hand-written rules restrict them to narrow domains. Neural approaches employ custom pre-trained models (Dong et al., 2019a; Du et al., 2017; Zhou et al., 2018), graph neural networks (Chen et al., 2020), transfer learning (Liao and Koh, 2020); however, generated questions often suffer from context-inconsistency issues.

2.2 LLMs-based Question Generation

Large models have attracted widespread attention and have gradually been applied to the field of education recently. Some work has proven that large models can generate high-quality question (Lee et al., 2024b; Doughty et al., 2024). Based on this, relevant attempts have been made to implement question generation on specialized educational large models (ELLMs). EduChat (Dan et al., 2023) requires users to submit reference questions, and based on these reference questions, ELLMs implement question generation. MuduoLLM (Zhu et al.,

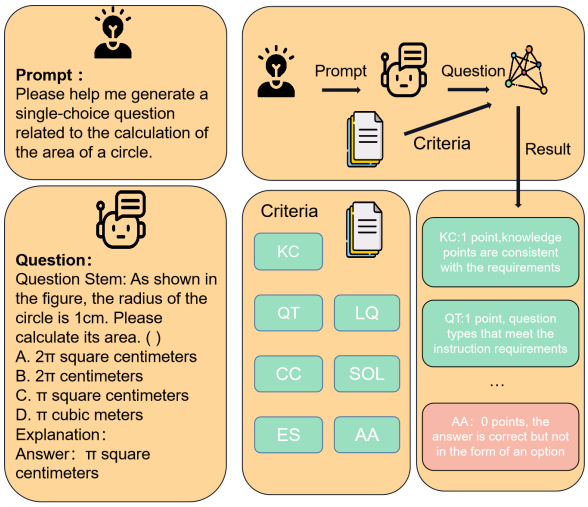

Figure 2: Evaluation framework, only abbreviations are shown in the figure, for specific content, refer to table 1.

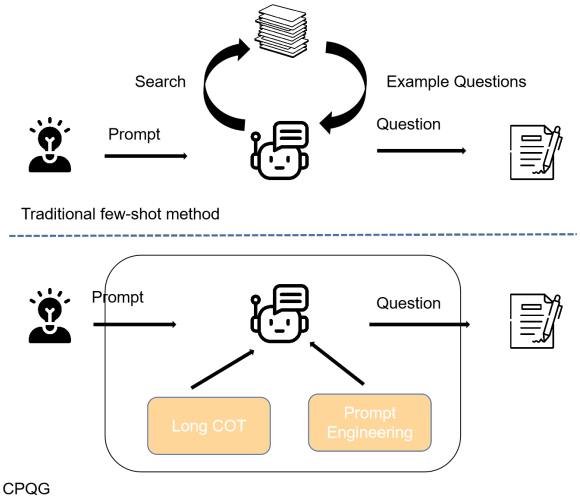

Figure 3: The comparison between our zero-shot method (CPQG) and few-shot method.

2025) extracts examples from the local database and then generates questions based on them. These methods are inherently limited by the quality of the provided example questions. Other works have begun to attempt to enhance the model's own generation capability. For example, one work has improved the quality of question generation through prompt engineering (Wang et al., 2025), and some work focuses specifically on multiple-choice questions, using fine-tuning to make multiple-choice questions more similar to those generated by humans (Lee et al., 2024a; Scarlatos et al., 2024; Feng et al., 2024; Biancini et al., 2024). However, these efforts only achieve making the generated questions more similar to human-crafted ones from a partial perspective, without improving the overall quality of the questions generated by the model. Our work aims to improve the model's own question generation capability. We identify the flaws in question generation from an overall perspective and propose a targeted method CPQG that combines instruction fine-tuning with prompt engineering, which has significantly enhanced the model's quality.

2.3 Evaluation Criteria

Generic criteria such as BLEU or BERTScore (Dong et al., 2019b; Liao and Koh, 2020) measure surface similarity to reference texts, yet they are ill-suited for assessing educational items. Custom criteria, such as asking users to provide answers to the questions via questionnaires (Lee et al., 2024a; Scarlatos et al., 2024; Feng et al., 2024; Biancini et al., 2024) are confined to multiple-choice questions and cannot be applied to all types of questions. Recognizing this limitation, (Zhou et al., 2025) proposed more universal criteria. However, their work has overlooked aspects such as the perspective of answers and the perspective of language fluency. Our evaluation criteria can comprehensively detect the quality of questions generated by the model. Compared with previous methods, they are more reasonable, more comprehensive and more universal.

2.4 Long CoT

Recent work has shown that chain-of-thought (CoT) prompting can markedly enhance model reasoning performance (Wei et al., 2023; Kojima et al., 2023; Zhang et al., 2022). long CoT, in particular, has been demonstrated to endow models with a self-reflective capability (DeepSeek-AI et al., 2025), and only a small volume of high-quality long CoT data is sufficient to yield substantial gains. More importantly, long CoT produces significantly higher-quality solutions to complex problems (DeepSeek-AI et al., 2025; Muennighoff et al., 2025; Moshkov et al., 2025). In this work, we use the dataset released by (Moshkov et al., 2025) to further strengthen our model's capacity for complex reasoning, and improved the quality of the model's explanation generation.

3 Problem Formulation

A complete question should at least include the question stem, explanation, and answer. For this purpose, we provide the following definitions for

question generation:

Definition (Question Generation Task, QGT) Let prompt P be the user-supplied input and model G a generator that, conditioned on P, outputs a triple

$$G(P) = \langle Q, E, A \rangle,$$

where

- Q is the question stem derived from the content of prompt P;
- E is the corresponding explanation generated from the stem Q;
- A is the answer produced from the explanation E.

4 Evaluation Framework

To evaluate the quality of question-generation produced by a model, we must first establish a comprehensive yet non-redundant set of criteria that can explicitly quantify the gap between generated questions and human-crafted questions. These criteria should be comprehensive enough with detailed scoring rules. Additionally, evaluating questions requires a set of prompts for question generation, and these prompts should avoid human intervention as much as possible. Finally, an objective evaluation subject is needed to accurately assess the questions.

To address these challenges, we construct our evaluation framework as follows:

1. **criteria** We extend the criteria proposed by (Zhou et al., 2025) to create a more comprehensive criteria.
2. **Data** We reverse-engineer prompts from human-crafted questions and use them as our evaluation dataset.
3. **Evaluator** We compare human assessment with model-based assessment and ultimately delegate the evaluation to LLMs.

We use these methods to evaluate the questions generated by humans, those generated by closed-source LLMs, and those generated by open-source models.

4.1 Evaluation Criteria

When we applied the evaluation protocol proposed by (Zhou et al., 2025), we identified three structural flaws:

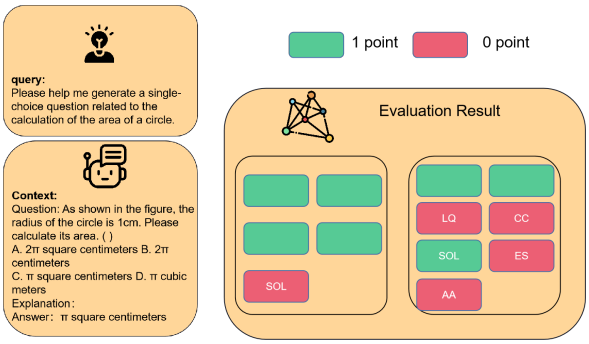

Figure 4: Different evaluation results for the same question under two different evaluation criteria, only abbreviations are shown in the figure; for specific content, please refer to Table 1.

- **Metric redundancy:** The same error may be penalized multiple times across different dimensions. For example, if a question contains a phrase as shown in the figure but no corresponding image is provided, this error will result in score deductions in multiple dimensions.
- **Missing perspectives:** For instance, this standard lacks an evaluation dimension for answers. If there is an error in the answer section, this standard will fail to detect it.
- **Dimension confusion:** For example, whether the language is fluent is categorized by this standard as whether the question is solvable, but language style and solvability are orthogonal and should not be confused.

Motivated by these observations, we revise the five-dimensional criteria of Zhou et al. (2025) into seven finer-grained criteria. The evaluation criteria are shown in Table 1.

Figure 4 presents a detailed comparison of the evaluation results for the same set of generated questions under both rubrics. In actual testing and evaluation, it shows the following results: when faced with the same question, (Zhou et al., 2025) standard only identified the presence of phrases like as shown in the figure, merely flagging them as factors that could affect solving the problem (SOL). In contrast, our standard detected multiple issues: missing content (missing explanation), an incorrect answer. It is worth noting that although the problem mentions "as shown in the figure" without providing the corresponding image, the relevant information is complete, meaning that this problem can be solved. However, it was deemed unsolvable

Table 1: Evaluation Criteria for Generated Questions.

Criteria	Description
Knowledge Coverage (KC)	Does the generated question accurately target and sufficiently cover the specified knowledge point(s)?
Question Type (QT)	Does the output strictly adhere to the requested question type?
Linguistic Quality (LQ)	Are the question, its explanation, and the answer free of grammatical, stylistic, or terminological issues?
Content Completeness (CC)	Are all expected components, namely the question statement, explanation, and answer present?
Solvability (SOL)	Is the question logically solvable given the information provided?
Explanation Soundness (ES)	Is the generated explanation both accurate and logically coherent?
Answer Accuracy (AA)	Is the final answer correct?

according to the standards of (Zhou et al., 2025), and we have corrected this issue in our standards. In addition to providing standards, we also provide sufficient examples for each dimension, detailed scoring protocols for each dimension are provided in Appendix.

4.2 Data Construction

To reduce the subjectivity of manually generated data, we used human-crafted questions as templates to reversely generate test data p for question generation. The procedure is as follows:

- **Initial prompt:** Manually compose a handful of seed prompts P_0, each explicitly tagged with the triplet knowledge point–question type–grade level.
- **Reverse generation:** Taking P_0 as a reference template, we submitted it together with human-generated questions (GSM8K) to LLMs (GPT-4o), and the model returned the predicted original prompt p_1.
- **Semantic filtering:** Low-quality prompts were removed from p_1, while prompts with rich language content were retained as p_2.

Repeat steps 2 and 3 until the content of the prompt is relatively rich. All the finally generated prompts

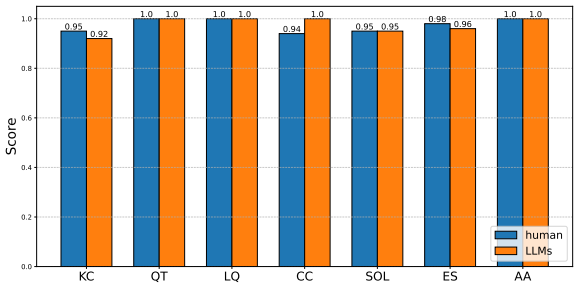

Figure 5: Survey results on the consistency between manual evaluation and large model evaluation

are retained, and 500 pieces of data are randomly selected from them as the test dataset t.

4.3 LLMs Evaluation

Previous studies relied on manual evaluation. Due to the lack of sufficiently detailed scoring criteria in the early stage, they had to depend on humans' subjective perception of quality. However, our criteria are sufficiently granular, with an adequate number of sample questions for each dimension, allowing the model to score the questions step by step. To verify the feasibility of this transition, we present a survey on the consistency between large model scoring and human scoring in Figure 5. Detailed information regarding the background of human evaluators, the specific evaluation process, and the raw consistency data is provided in Appendix D.

Figure 5 confirms that the models' judgments align almost perfectly with those of humans. This outcome rests on two factors: 1) the models possess near-human knowledge capabilities 2) The details of the standard scoring are well-developed, and the examples cover all scenarios. Large models can score directly according to the standards without making independent judgments.

Consequently, all subsequent experiments use the large model as the sole evaluator, with human review limited to spot-checks.

4.4 Criteria-based Evaluation Results

Because of probabilistic sampling, the same question may receive different scores even when evaluated by one LLM repeatedly, and systematic biases exist across different LLMs. To reduce random fluctuation, we adopt a method of multiple generations and multiple evaluations by the LLMs. The model is first asked to produce five independent questions on the test set t; each of these five outputs is then scored by three separate large models (DeepSeek-R1, GPT-4o, and Gemini-2.5-Pro). The

Table 2: Measurement results of questions generated by humans, closed-source LLMs, and open-source models.

Model	KC	QT	LQ	CC	SOL	ES	AA
Human-curated questions							
GSM8K	–	–	0.97	0.93	0.95	0.97	0.97
MATH	–	–	0.87	1.00	0.97	1.00	0.97
Gaokao	–	–	0.75	0.96	0.86	0.95	0.97
Closed-source LLMs							
GPT-4o	0.99	0.97	1.00	0.98	0.86	0.90	0.88
GPT-4	0.98	0.99	0.96	0.85	0.91	0.85	0.90
Gemini2.5-Pro	1.00	0.96	0.98	0.70	1.00	1.00	1.00
Deepseek-R1	0.99	0.96	0.98	0.90	1.00	0.98	0.934
Open-source models							
Qwen2.5-7B	0.94	0.93	0.89	0.87	0.82	0.64	0.73
Qwen3-8B	0.76	0.74	0.79	0.80	0.79	0.81	0.78
LLaMA2-8B	0.88	0.74	0.50	0.78	0.52	0.29	0.54

final criterion is the mean of all resulting scores. The instructions we use for generating questions are in Appendix.

We benchmarked mainstream large models such as DeepSeek-R1, GPT-4o, GPT-o4, and Gemini-2.5-Pro, as well as popular open-source models including Qwen2.5-7B-Instruct, Qwen3-8B, and Llama-8B. For the human-crafted questions, we used the widely adopted datasets GSM8K, MATH, and Gaokao. Table 2 summarizes the performance of all models under this unified evaluation framework.

Table 2 indicates that human-crafted questions consistently outperform models across all dimensions, validating our framework's ability to measure the performance gap. While LLMs approach human performance, open-source models like Qwen2.5-7B and Llama2.5-8B lag significantly, particularly in SOL, ES, and AA. Notably, Gemini2.5-Pro shows a lower CC (0.70) due to its tendency to merge explanations and answers, a formatting issue captured by our granular rubric despite its correct content. Other performance details are summarized in Table 2.

5 Methodology

To narrow the gap between the quality of questions generated by the model and those generated manually, we have implemented a zero-shot method that does not rely on external knowledge bases. Through instruction fine-tuning and prompt engineering, this method has significantly enhanced the model's question generation performance. Practice has demonstrated that our zero-shot method outperforms the commonly used few-shot methods.

5.1 Instruction Tuning

Table 2 shows that open-source small models perform poorly in three core components: Solvability (SOL), Explanation Soundness (ES), and Answer Accuracy (AA). Among them, the ES capability is at a relatively low level, which has become a key obstacle restricting the improvement of model performance. Therefore, this study takes enhancing the model's capability to generate explanation as one of its core goals.

The long Chain-of-Thought (CoT) enables the model to perform complex reasoning through step-by-step thinking and self-reflection. Given that question generation involves interlinked subtasks (stem, explanation, and answer generation) with high demands for logical consistency, applying long CoT is highly effective for minimizing errors and ensuring explanation quality.

Dataset The basic dataset used in this experiment is the OpenMathReasoning dataset, which was officially released by NVIDIA in 2025 (Moshkov et al., 2025). Its core feature is that it contains 3.2 million long COT samples, and the sample design focuses on improving the model's reasoning ability, thereby enhancing the quality of analysis generation. This dataset is highly aligned with the goals of this study. For detailed statistical information about the dataset and access methods, please refer to the original publication paper.

Furthermore, to simultaneously enhance the model's Answer Accuracy (AA) capability, based on the OpenMathReasoning dataset, the explanation (non-thinking processes) and corresponding answers are extracted from the chain-of-thought to form a result dataset (RD) containing the explanation-answer correspondence.

Separate Training First, the OpenMathReasoning dataset is used for single-turn dialogue fine-tuning of current SOTA models (e.g., Qwen2.5-7B-Instruct), with problem stem as inputs and chains of thought as outputs. The second round of fine-tuning focuses on improving the answer generation capability using RD dataset, where analyses serve as inputs and standardized formatted answers as outputs. According to (Longpre et al., 2023), we designed different prompt templates for the two rounds of fine-tuning respectively (see Appendix for details).

5.2 Prompt Engineering

We counted the errors in 1,000 model-generated questions across SOL, ES, and AA, and calculated the proportion of each type of error in the dataset. The results of the experiment are placed in the Appendix.

To further enhance generation quality, we apply prompt engineering (PE) with targeted instructions addressing identified errors. Specifically, we explicitly command the model to ensure answer uniqueness and provide step-by-step explanations: *"Guarantee a unique correct answer; proceed step-by-step without skipping or intermediate errors; ensure the final answer is logically derived and correctly formatted."* Our experiments confirm these instructions effectively mitigate common generation flaws.

6 Experiment

In this chapter, we use the Qwen2.5-7B-Instruct model as the baseline model to verify the effectiveness of our zero-shot method. To ensure fairness, the methods for evaluating the model's generation performance follow the framework we proposed earlier, and all experiments are conducted on a single A800.

6.1 Main Results

Table 3 summarizes the results. CPQG demonstrates substantial gains across almost all dimensions, notably increasing ES by 30% and AA by 21%, with an average improvement exceeding 10%. CPQG also rivals or surpasses GPT-4 in reasoning-heavy dimensions like SOL, ES, and AA.

Comparison Experiment Table 3 presents the best results of the few-shot method, and our zero-shot method is significantly superior to the few-shot method. For example, it exceeds the one-shot method by 18% in LQ and 12% in SOL. To further explore the few-shot method, we designed comparative experiments. Specifically, we generated a set of example questions for each question, then evaluated them respectively. The questions were divided into 6 groups according to their average scores: 0.5, 0.75, and 1.0. Then, the questions in each group were used as examples for one-shot, three-shot, and ten-shot experiments respectively. The experimental results are shown in Figure 6. First, all three few-shot methods are affected by the example questions: the higher the quality of the example questions, the higher the quality of the questions generated by the few-shot methods. The advantage of ten-shot over one-shot lies in that when the quality of example questions decreases, the quality of questions generated by the model decreases relatively less. However, this advantage becomes less obvious when the quality of example questions is very high. In summary, the few-shot method is easily affected by the quality of example questions, and the premise of using the few-shot method is the need to pre-construct high-quality reference texts. In contrast, our zero-shot method saves this time and achieves better results in terms of the quality of generated questions.

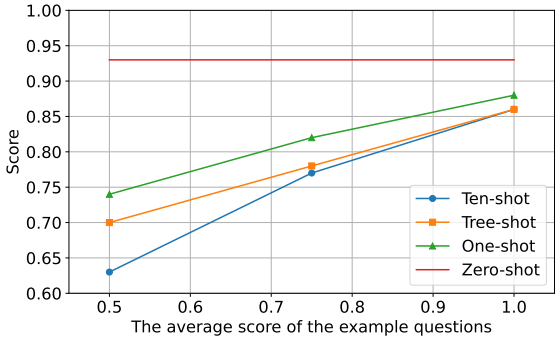

Figure 6: The performance of few-shot methods is influenced by the quality and quantity of example questions. The abscissa represents the average score of example questions, and the ordinate represents the average score of generated questions.

Ablation experiment We explored the importance of fine-tuning and specialized improvements. First, as observed from Table 3, when COT is removed, various metrics decline (with QT dropping by 18% in particular), which illustrates the importance of the COT method and proves that the COT method can overall improve the quality of question generation by the model. When result fine-tuning (RD) is removed, the AA metric also shows a decline, though the removal of CoT leads to the most significant drop in AA, emphasizing that logical reasoning is the foundation for accuracy. When individual improvements are removed, a downward trend is observed in the three dimensions of SOL, ES, and AA, which demonstrates the rationality of our hybrid improvement method.

QGT Fine Tuning To verify whether there is a better fine-tuning scheme for CoT, we attempted to directly fine-tune the model for the task of question generation, specifically fine-tuning QGT. First, we selected 5,000 high-quality questions

Table 3: The performance of the model in generating questions, where higher scores in all aspects indicate better performance.

Model	KC	QT	LQ	CC	SOL	ES	AA
Qwen2.5-7B-Instruct	0.94±0.03	0.93±0.02	0.89±0.04	0.87±0.05	0.82±0.03	0.64±0.06	0.73±0.04
GPT4	0.97±0.02	0.96±0.01	0.91±0.03	0.93±0.02	0.91±0.02	0.85±0.03	0.81±0.03
GPT4o	**0.99±0.01**	**0.97±0.02**	**0.99±0.01**	**0.98±0.01**	0.86±0.04	0.90±0.02	0.88±0.03
Comparative experiments							
One-shot	0.85±0.05	0.94±0.02	0.76±0.06	0.90±0.03	0.83±0.04	0.85±0.03	0.90±0.02
Ten-shot	0.91±0.04	0.88±0.03	0.83±0.05	0.95±0.02	0.86±0.03	0.85±0.04	0.90±0.03
QGT	0.28±0.08	0.75±0.07	0.25±0.09	0.85±0.04	0.50±0.08	0.65±0.07	0.55±0.08
CPQG							
CPQG	0.95±0.02	0.89±0.03	0.94±0.02	0.95±0.02	**0.95±0.01**	**0.94±0.02**	**0.94±0.02**
w/o CoT	0.82±0.05	0.71±0.06	0.71±0.07	0.90±0.03	0.85±0.04	0.81±0.05	0.89±0.03
w/o PE	0.95±0.03	0.90±0.02	0.85±0.04	0.90±0.03	**0.95±0.02**	0.89±0.03	0.91±0.02
w/o RD	0.89±0.04	0.85±0.03	0.93±0.02	0.93±0.02	0.93±0.03	0.93±0.02	0.90±0.03

from the GSM8K dataset, then reversely constructed prompts for question generation. With these prompts as input and the question stems, explanations, and answers as output, we fine-tuned the model. To ensure fairness, we excluded the part of data corresponding to the test set t. The experimental results are shown in Table 3. This method did not improve the question generation quality of the base model; instead, there was a significant decline. For example, SOL decreased by 32%. We speculate that this may be because the question generation task is relatively complex, involving three subtasks: generating question stems, generating explanations, and generating answers, which is quite different from pre-training tasks. As a result, the model may struggle to transfer to the question generation task.

6.2 Robustness Experiment

To test robustness, we introduced noise to the test dataset t by randomly replacing prompt words, question types, or knowledge points. As shown in Figure 7, performance remains stable across all dimensions, indicating strong robustness. This resilience likely stems from the long CoT process, which enables the model to self-reflect and correct for input noise during generation.

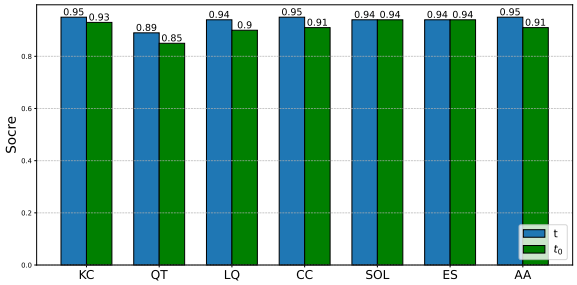

Figure 7: Performance of robustness, in the figure, t represents the test results of the dataset without added noise, while t_0 represents the test results after adding noise.

7 Conclusion

Currently, large models are widely applied in the education industry, and the generation of questions through large models is gradually attracting attention. To narrow the gap between questions generated by models and those created by humans, we have implemented a comprehensive and fine-grained test question evaluation framework. Based on this framework, we have identified potential issues in the questions currently generated by models. In addition, we have attempted to propose a zero-shot method (CPQG), which combines long CoT (Chain-of-Thought) technology and prompt engineering, significantly improving the model's generation quality. Experiments show that CPQG outperforms traditional few-shot methods and enables 7B-sized models to outperform GPT-4.

Limitations

Our study has several limitations that provide directions for future work. First, regarding the **subject matter limitation**: the current experiments and evaluations are primarily focused on the field of mathematics, particularly junior high school math-

ematics questions. In the future, we will further extend the CPQG method to other disciplines, such as humanities and languages, to verify its cross-domain generalizability.

Second, our **evaluation criteria** are currently defined from a relatively straightforward perspective. These criteria may not be directly applicable to more complex items, such as geometry problems that require spatial reasoning and auxiliary line construction. We plan to expand the evaluation framework to incorporate more diverse item types and deeper pedagogical reasoning.

Finally, our analysis revealed that we **overlooked the impact of question difficulty**. Different difficulty levels may significantly influence various evaluation metrics. In future work, we will incorporate difficulty as a controlled variable to achieve a more nuanced and comprehensive evaluation of model capabilities across varying levels of complexity.

References

Aditi Bhutoria. 2022. Personalized education and artificial intelligence in the united states, china, and india: A systematic review using a human-in-the-loop model. *Computers and Education: Artificial Intelligence*, 3:100068.

Giorgio Biancini, Alessio Ferrato, and Carla Limongelli. 2024. Multiple-choice question generation using large language models: Methodology and educator insights. In *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '24, page 584–590. ACM.

Yllias Chali and Sadid A. Hasan. 2015. Towards topic-to-question generation. *Computational Linguistics*, 41(1):1–20.

Yu Chen, Lingfei Wu, and Mohammed J. Zaki. 2020. Reinforcement learning based graph-to-sequence model for natural question generation. *Preprint*, arXiv:1908.04942.

Yuhao Dan, Zhikai Lei, and 1 others. 2023. Educhat: A large-scale language model-based chatbot system for intelligent education. *Preprint*, arXiv:2308.02773.

DeepSeek-AI, Daya Guo, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Li Dong, Nan Yang, and 1 others. 2019a. *Unified language model pre-training for natural language understanding and generation*. Curran Associates Inc., Red Hook, NY, USA.

Li Dong, Nan Yang, and 1 others. 2019b. Unified language model pre-training for natural language understanding and generation. *Preprint*, arXiv:1905.03197.

Jacob Doughty, Zipiao Wan, Anishka Bompelli, Jubahed Qayum, Taozhi Wang, Juran Zhang, Yujia Zheng, Aidan Doyle, Pragnya Sridhar, Arav Agarwal, Christopher Bogart, Eric Keylor, Can Kultur, Jaromir Savelka, and Majd Sakr. 2024. A comparative study of ai-generated (gpt-4) and human-crafted mcqs in programming education. In *Proceedings of the 26th Australasian Computing Education Conference*, ACE 2024, page 114–123. ACM.

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 (Volume 1: Long Papers)*, pages 1342–1352, Vancouver, Canada. Association for Computational Linguistics.

Wanyong Feng, Jaewook Lee, Hunter McNichols, and 1 others. 2024. Exploring automated distractor generation for math multiple-choice questions via large language models. *Preprint*, arXiv:2404.02124.

Weibo Gao, Qi Liu, and 1 others. 2025. Agent4edu: Generating learner response data by generative agents for intelligent education systems. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(22):23923–23932.

Michael Heilman and Noah A. Smith. 2010a. Good question! statistical ranking for question generation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 609–617, Los Angeles, California. Association for Computational Linguistics.

Michael Heilman and Noah A. Smith. 2010b. Good question! statistical ranking for question generation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, page 609–617, USA. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, and 1 others. 2023. Large language models are zero-shot reasoners. *Preprint*, arXiv:2205.11916.

Jaewook Lee, Digory Smith, Simon Woodhead, and Andrew Lan. 2024a. Math multiple choice question generation via human-large language model collaboration. *Preprint*, arXiv:2405.00864.

U Lee, H Jung, Y Jeon, and 1 others. 2024b. Few-shot is enough: exploring chatgpt prompt engineering method for automatic question generation in english education. *Education and Information Technologies*, 29(9):11483–11515.

Junyi Li, Tang, and 1 others. 2024. Pre-trained language models for text generation: A survey. *ACM Comput. Surv.*, 56(9).

Yin-Hsiang Liao and Jia-Ling Koh. 2020. Question generation through transfer learning. In *Trends in Artificial Intelligence Theory and Applications. Artificial Intelligence Practices: 33rd International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2020, Kitakyushu, Japan, September 22-25, 2020, Proceedings*, page 3–17, Berlin, Heidelberg. Springer-Verlag.

Shayne Longpre, Le Hou, and 1 others. 2023. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Pan Lu, Swaroop Mishra, and 1 others. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*, volume 35, pages 2507–2521. Curran Associates, Inc.

Setareh Maghsudi and 1 others. 2021. Personalized education in the artificial intelligence era: What to expect next. *IEEE Signal Processing Magazine*, 38(3):37–50.

Ivan Moshkov, Darragh Hanley, and 1 others. 2025. Aimo-2 winning solution: Building state-of-the-art mathematical reasoning models with openmathreasoning dataset. *Preprint*, arXiv:2504.16891.

Niklas Muennighoff, Zitong Yang, and 1 others. 2025. s1: Simple test-time scaling. *Preprint*, arXiv:2501.19393.

OpenAI and 1 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Alexander Scarlatos, Wanyong Feng, and 1 others. 2024. Improving automated distractor generation for math multiple-choice questions with overgenerate-and-rank. *Preprint*, arXiv:2405.05144.

Yijia Shao, Yucheng Jiang, and 1 others. 2024. Assisting in writing wikipedia-like articles from scratch with large language models. *Preprint*, arXiv:2402.14207.

Maria Valentini, Jennifer Weber, and 1 others. 2023. On the automatic generation and simplification of children's stories. *Preprint*, arXiv:2310.18502.

Lili Wang, Ruiyuan Song, Weitong Guo, and Hongwu Yang. 2025. Exploring prompt pattern for generative artificial intelligence in automatic question generation. *Interactive Learning Environments*, 33(3):2559–2584.

Tianfu Wang and 1 others. 2025. Llm-powered multi-agent framework for goal-oriented learning in intelligent tutoring system. *Preprint*, arXiv:2501.15749.

Jason Wei, Xuezhi Wang, Dale Schuurmans, and 1 others. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Xuan Zhang, Navid Rajabi, , and 1 others. 2023. Machine translation with large language models: Prompting, few-shot learning, and fine-tuning with QLoRA. In *Proceedings of the Eighth Conference on Machine Translation*, pages 468–481, Singapore. Association for Computational Linguistics.

Zheyuan Zhang, Daniel Zhang-Li, and 1 others. 2025. Simulating classroom education with LLM-empowered agents. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10364–10379, Albuquerque, New Mexico. Association for Computational Linguistics.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic chain of thought prompting in large language models. *Preprint*, arXiv:2210.03493.

Chengliang Zhou, Mei Wang, and Hua Huang. 2025. From answers to questions: Eqgbench for evaluating llms' educational question generation. https://github.com/clyde-zh/From_Answers_to_Questions_EQGBench_for_Evaluating_LLMs_Educational_Question_Generation. GitHub repository.

Qingyu Zhou, Nan Yang, and 1 others. 2018. Neural question generation from text: A preliminary study. In *Natural Language Processing and Chinese Computing*, pages 662–671, Cham. Springer International Publishing.

Qiannan Zhu, ZeChen Li, and 1 others. 2025. Muduollm: A high-performance llm for intelligent education solutions. https://huggingface.co/ERC-ITEA/MuduoLLM.

A Overview of Existing QG Evaluation Frameworks

In this section, we provide a detailed introduction to the existing Question Generation (QG) evaluation framework, specifically the one proposed by (Zhou et al., 2025). Their framework established a foundational set of five criteria for assessing the quality of educational questions:

- **Correctness:** Whether the question is factually accurate and consistent with the knowledge point.
- **Solvability:** Whether the question provides enough information for a student to derive the answer.
- **Linguistic Quality:** Assessing the fluency, grammar, and clarity of the question text.
- **Educational Value:** Whether the question is appropriate for the target grade level and learning objectives.
- **Clarity:** Whether the question is unambiguous and easy for students to understand.

While these criteria laid the groundwork for automated QG evaluation, our 7-dimensional framework extends this by explicitly separating Content Completeness, Explanation Soundness, and Answer Accuracy, which are often conflated or overlooked in earlier rubrics.

B Evaluation Criteria Details

This section is a supplement to the main text's chapter on evaluation criteria. Herein, we provide a detailed account of all the frequency measurement dimensions and specific scoring criteria of the evaluation standards, as well as the corresponding input instructions for large models.

B.1 Criteria

Knowledge-Point Coverage (1 vs. 0) 1: The question essentially covers the user-specified knowledge point(s); if the knowledge belongs to a non-mathematical discipline or exceeds junior-high level, the stem must supply the necessary explanation. 0: The question is from the wrong discipline, exceeds the required grade, or fails to address the specified knowledge.

Question-Type Accuracy (1 / 0.5 / 0) 1: The format perfectly matches the requested type (e.g. MCQ has at least four choices, fill-in-the-blank has underlines, open-ended questions contain no extraneous blanks). 0.5: The type is correct but the format is flawed (e.g. MCQ has fewer than four choices, fill-in-the-blank lacks underlines). 0: The type is wrong or the format is chaotic and unrecognizable.

Linguistic Clarity (1 vs. 0) 1: The stem, solution, and answer are clear and concise; no garbled text; mathematical expressions and units are correct; no typos or repetition. 0: Any garbled text, semantic errors, incorrect formulas, or wrong units.

Content Completeness (1 vs. 0) 1: The stem, solution, and answer are all present; no placeholders such as "see figure" without the actual figure. 0: Any part is missing or placeholders lack the actual material.

Solution Accuracy (1 vs. 0) 1: The solution is fully relevant, logically sound, step-by-step, and free of logical jumps. 0: The solution is irrelevant or contains errors that prevent reaching the answer.

Answer Correctness (1 vs. 0) 1: The final answer meets the question's requirement (e.g. MCQ returns the option label, not the value; open-ended returns the required result). 0: The answer does not match the solution or fails to satisfy the question.

Solvability (1 vs. 0) 1: All necessary information is provided; exactly one correct choice for single-choice MCQ; computations are feasible. 0: Information is insufficient or ambiguous; multiple or no correct answers.

B.2 Reverse-engineering prompts

The following is an example of a prompt for reverse-engineering real questions using a large model:

- I am a junior high school student and want to participate in a math contest. However, I am not particularly good at fraction calculations. Could you please generate a multiple-choice question to help me practice my calculation skills?

- I am a third-year junior high school teacher. Recently, when explaining mathematical multiplication in class, I found that students just memorize it by rote and cannot apply it. Please help me come up with a solution question that flexibly uses this knowledge point.

- The school's mathematics teaching seminar requires the design of tiered homework. It is necessary to prepare a question for the knowledge point of integer addition and subtraction, which is for the junior high school stage, with priority given to the form of solution questions.

C Case Study: Instruction and Evaluation Examples

This section provides illustrative examples of instructions used for question generation and evaluation.

C.1 Question Generation Examples

The following examples demonstrate the format used for generating educational questions and their corresponding multi-step explanations.

Prompt 1 (Example of generating a multiple-choice question on linear regression).

Response 1 (Detailed stem, step-by-step explanation, and unique answer).

C.2 Evaluation Examples

Below are schematic instructions and model-generated scoring reasons demonstrating how our 7-dimensional rubric is applied by LLMs.

Prompt 1 (System prompt for the LLM as an expert evaluator).

Response 1-7 (Granular scoring reasons for each dimension, ensuring transparency and logical consistency).

D Manual Evaluation Details

To ensure the reliability of our automated evaluation framework, we conducted a rigorous manual validation process. This section provides detailed information about the human evaluators, the evaluation procedure, and the consistency results.

Evaluator Background A total of 20 education experts were recruited for the manual evaluation. All evaluators hold at least a Master's degree in Education or a related subject area (e.g., Mathematics, Science), with an average of 8 years of teaching experience in middle or high schools. They were trained on our 7-dimensional rubric for 2 hours before starting the evaluation.

Evaluation Process We randomly sampled 200 questions from the model-generated outputs. Each question was independently scored by three human evaluators according to the same 7-dimensional rubric used for the LLM-as-judge. The evaluators were blind to the source of the questions (whether they were generated by our CPQG method, baseline models, or humans).

Consistency Analysis As shown in Figure 5, we calculated the agreement between the average human scores and the average LLM scores. The high alignment confirms that our granular rubric effectively minimizes subjective bias, allowing the LLM to serve as a reliable proxy for expert judgment. The specific inter-rater agreement (Cohen's Kappa) among human evaluators averaged 0.82 across all dimensions, indicating high reliability.

E Consistency Validation on Qwen2.5-7B

In this section, we conducted a consistency validation experiment using the Qwen2.5-7B model. Specifically, 100 questions generated by Qwen2.5-7B were randomly selected and submitted to both human experts (two senior graduate students in Education) and Large Language Models (LLMs) for evaluation based on our 7-dimensional rubric.

For the human evaluation, a voting mechanism was adopted: each question was independently assessed by the two students, and the final score was determined by their consensus or majority vote. The total number of correctly generated questions (those meeting all criteria) was counted as the final score. These results were then compared with the scores generated by the LLM evaluator. The high degree of alignment between the human voting results and the LLM scores is illustrated in Figure 5.