Generating Contextualized Mathematics Multiple-Choice Questions Utilizing Large Language Models

Ruijia Li[1], Yiting Wang[3], Chanjin Zheng[1,2,4], Yuan-Hao Jiang[2,4], and Bo Jiang[2,4](✉)

[1] Faculty of Education, East China Normal University, Shanghai, China
[2] Lab of Artificial Intelligence for Education, East China Normal University, Shanghai, China
[3] Software Engineering Institute, East China Normal University, Shanghai, China
[4] Shanghai Institute of Artificial Intelligence for Education, East China Normal University, Shanghai, China
bjiang@deit.ecnu.edu.cn

Abstract. Applying mathematics to solve authentic question play important roles in math-ematics education. How to generate high-quality multiple-choice questions that have authentic context is a great challenge. By combining multiple iterations of large language model dialogues with auxiliary external tools and the LangChain framework, this work presents a novel method for automatically generating contextualized multiple-choice mathematics questions. To check the quality of generated questions, 30 questions were randomly selected and 13 human experts were invited to rate these questions. The survey result indicates that the questions produced by the proposed method exhibit a significantly higher quality compared to those generated directly by GPT4, and are already quite comparable in performance to questions that are meticulously crafted by humans across multiple dimensions. The code is available on the project home page: https://github.com/youzizzz1028/MCQ-generation-Chain.

Keywords: Automatic Question Generation · LangChain · ChatGPT · Core Literacy · Prompt Engineering

1 Introduction

1.1 Research Background

Creating multiple-choice questions (MCQs) manually requires considerable time and resources. Despite the provision of question banks, manual labor is still required for various tasks including question formatting, tagging, and integration. Moreover, an issue of inconsistent question quality persists. A critical aspect of learning system development is the investigation of methods to generate multiple-choice questions automatically while ensuring quality.

A. M. Olney et al. (Eds.): AIED 2024 Workshops, CCIS 2150, pp. 494–501, 2024.
https://doi.org/10.1007/978-3-031-64315-6_48

Another complexity arises from the additional demand for mathematics education. In mathematics, the Compulsory Education Mathematics Curriculum Standards released by Ministry of Education of China (2022 Edition) place significant emphasis on competency. Context is a crucial element in linking theoretical knowledge and practical skills to competency. It plays a vital role in connecting theory and practice, enabling the application of knowledge and abilities in a meaningful way. To enhance students' ability to apply their acquired knowledge to authentic situations, it is crucial to design math questions that incorporate authentic contexts. This presents a significant challenge to the ability of LLMs to utilize real-world knowledge.

The objective of this research is to investigate the potential application of large language models (LLMs) in deriving mathematical questions. Our method is to enhance the LLM's capability in item generation by incorporating code interpreters for logical operations and external knowledge bases for knowledge, with the aim of improving the quality and diversity of contextualized math questions [1].

1.2 Research Questions

This study aims to identify an efficient approach for the automated creation of context-specific multiple-choice questions (MCQs) in primary and middle schools. The objective of this study is to assess the quality of questions generated automatically, as well as those generated by our platform, questions crafted by human experts, and questions generated by ChatGPT. The research questions for this study are as follows:

1. What is the feasible approach, based on LLM, for generating automated multiple-choice questions in a contextualized manner?

2. What is the quality of questions generated using this solution?

2 Related Work

Kurdi [4] presents a classification and description of Automatic Question Generation (AQG) systems from seven dimensions: (a) the purpose and utilization of question generation, (b) the knowledge domain being examined, (c) the source of knowledge, (d) the generation method, (e) the type of questions, (f) the formatting of question structure, and (g) the evaluation of question quality. Cur-rent question generation technologies primarily aim to create quiz questions, support online learning platforms, or assist students in independent learning, with a focus on disciplines such as language, mathematics, humanities, and medicine. Classic methods of question generation primarily fall into three categories: (a) grammar-based, which identifies key concept words using techniques such as grammar trees and POS (Part-of-speech tagging), and generates questions by replacing these key concepts; (b) semantic-based, which primarily identifies the part of speech and dependencies between words using natural language processing techniques, and generates questions based on the extracted dependencies;

and (c) template-based, which involves the creation of preset question templates and the utilization of random number generation or keyword extraction to fill in the templates. Subsequently, with the emergence of deep learning technology, neural networks [7], reinforcement learning [2], and pre-trained language model shave been integrated into AQG systems. Additionally, some research has begun to focus on the utilization of LLMs for question generation [5].

Owing to the stringent requirements of mathematical precision and logical rigor, the utilization of connectionist artificial intelligence algorithms is not advisable for addressing mathematical challenges. Instead, it often necessitates the employment of symbolic methodologies for resolution [9]. Numerous technical solutions within AQG systems often fail to apply to the generation of mathematical problems. Presently, the most commonly used method remains the generation of random numbers based on predefined templates, which can result in challenges such as inflexible question formats, limited practical relevance, and limited scalability. Grévisse [3] delved into the quality of MCQs generated and discovered that despite the fluent language expression of GPT-generated questions, numerous issues persist, including excessively long question stems, absence of correct answers among the options, ambiguous expressions, and low relevance between the questions and knowledge points. Our experimental exploration concurs with these findings - the direct utilization of GPT models and prompts for question generation, particularly for MCQs, is challenging to achieve satisfactory out-comes. Therefore, it is imperative to introduce external knowledge bases as knowledge inputs and targeted teaching strategies. It is also necessary to introduce technologies such as code interpreters and PoT (Program-of-Thoughts) [1], adjust the parameters of LLM dialogue and control the question generation process.

3 Contextualized Mathematical Multiple-Choice Question Generation Methodology

The workflow of creating and generating multiple-choice questions is typically divided into the following six stages [3]: (a) Background Input (b) Preprocessing (c) Question Generation (d) Answer Generation (e) Distractor Generation (f) Formatting. Following this workflow, this study proposes a solution for automatically generating mathematical multiple-choice questions based on the LangChain framework, which integrates multi-round large language model dialogues and pluggable interfaces (such as SerpAPI, llm-math, etc.). This solution includes four core modules: Domain Agent, QG Question Generation Agent, Problem Solving Agent, and Option Generation Agent (Fig. 1). Each module is based on the GPT-4 model, and they are closed linked to generate mathematical multiple-choice questions with a certain quality assurance and real-world scenarios.

3.1 Domain Agent

This agent uses search engines to obtain quiz examples and problem-solving ideas highly related to specific knowledge point from extensive web resources, to

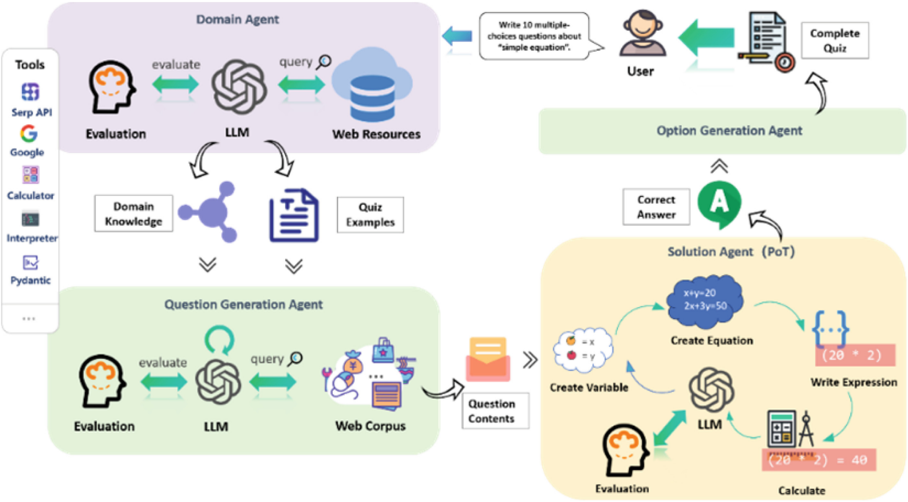


Fig. 1. Automated Solution for Producing Contextualized Multiple-Choice Questions

facilitate the establishment of domain foundation and minimize hallucinations and knowledge errors.

3.2 Question Generation Agent

This agent receives a range of parameters from preceding steps (such as domain knowledge, quiz examples, question quantity, difficulty level, etc.). By engaging in multiple iterations of experiments, a prompt is gradually crafted. Additionally, this agent utilizes SerpAPI as a supportive instrument to examine and validate the authenticity of the question context by leveraging extensive corpus data from the Internet. This approach ensures that the questions are grounded in reality, enriching contextualized characteristics.

3.3 Solution Agent

This agent designed for problem-solving utilizes llm-math interface and Python interpreter to conduct numerical computations. During analysis, the LLM provides Python expressions that can accurately resolve them. After gradual execution of program, the correct answer with natural language format is generated.

3.4 Option Generation Agent

This agent offers a standardized template of multiple-choice questions. Questions and corresponding correct answers are embedded in carefully designed prompt.

Then the agent generates distractors and randomly rearranges them, transforming the question's structure into MCQ format. Additionally, it incorporates formatting cues and coercion checks of data type, and returns a JSON formatted text as specified.

4 Quality Evaluation

4.1 Evaluation Methods

We collected evaluations and assessments from mathematics instructors on the quality of questions from different methods. 30 contextualized MCQs were gathered randomly from our platform, authentic exams, and ChatGPT4. We provide GPT-4 model with the same prompt used in our platform except format hint. Studies on the quality evaluation of contextualized MCQs have investigated the qualities that high-quality and fair MCQs should have [9]: the stem should be clearly expressed, focused on particular knowledge points, leading to a uniquely determined response, and employing positive phrases. The alternatives should be concise and straightforward, without the use of ambiguous adverbs or expressions like "all of the above are correct, all of the above are wrong", and with consistent length and language structure. Previous research has often used expert rating techniques to evaluate the quality of automatically produced questions [10]. Some studies merely gather survey responses for a single component of "question quality", but others collect expert views on dimensions such as answer-ability, knowledge relevance, and question complexity [4].

Based on the preceding research, this study argues that assessing the quality of questions may be divided into three dimensions: rationality of the stem, rationality of options, and contextual appropriateness. In addition, this research looks at experts' perspectives on the producing potential to see whether there are subjective distinctions between produced questions and authentic questions. The following Table 1 provides detailed explanation and definition of each dimension.

Table 1. Definitions of Research Dimensions and Scoring Standards

Dimension	Definition and Scoring Standards
Rationality of the Stem	The expression should be clear and concise; closely related to knowledge points; lead to a uniquely determined answer; use positive expressions, such as "the following is correct..." rather than "the following is incorrect...".
Rationality of options	Contain the only correct answer; be concise and clear, avoid the use of vague adverbs (such as "often", "occasionally"); avoid the expression of "all of the above are correct" or "all of the above are wrong"; have consistent length and grammatical structure.
Contextual appropriateness	It conforms to real life; the expression is clear, and the structure is simple; it meets the cognitive level of students; the contextual background is integrated with the knowledge learned.

4.2 Evaluation Results

The Cronbach's alpha value of the consistency among scorers is 0.898, indicating that the scoring criteria of the scorers are consistent and their understanding and scoring methods for the questions are similar. Independent sample T-tests were conducted on the questions generated by this platform (Ours) and the authentic questions (AQ), as well as on the questions generated by this platform and directly generated by GPT4. The results are shown in the following Tables 2 and 3.

Table 2. Comparison of Evaluation Results between Ours and AQ

Dimension	Source	AVG	S.D	p-value
Contextual appropriateness	AQ	4.417	0.4387	0.439
	Ours	4.267	0.4924	
Rationality of options	AQ	4.55	0.2611	1.000
	Ours	4.55	0.3680	
Rationality of Stem	AQ	4.592	0.2353	0.457
	Ours	4.508	0.2999	
Producing Potential	AQ	2.039	0.4226	0.084
	Ours	2.433	0.6243	

Table 3. Comparison of Evaluation Results between Ours and GPT4

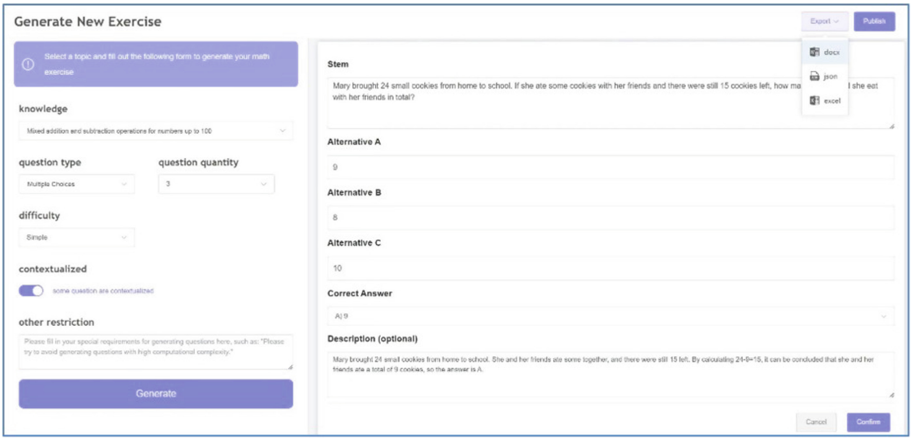
Dimension	Source	AVG	S.D	p-value
Contextual appropriateness	GPT4	3.983	0.3817	0.236
	Ours	4.267	0.4924	
Rationality of options	GPT4	4.050	0.6058	0.043
	Ours	4.55	0.3680	
Rationality of Stem	GPT4	4.017	0.3920	0.009
	Ours	4.508	0.2999	
Producing Potential	GPT4	3.417	0.5231	0.004
	Ours	2.433	0.6243	

Upon analysis of the data, it is evident that the questions produced by this platform exhibit minimal variance compared to authentic questions across various metrics. Notably, when compared to questions directly generated by GPT4, the platform's questions exhibit statistical significance in three dimensions. While the contextual appropriateness does not reach statistical significance, its numerical value surpasses GPT4-generated questions. This indicates that the technological approach employed by our platform enhances the scientific and accuracy of the questions based on GPT4, without significant deviation from real-world questions. Therefore, it is anticipated that this platform will be suitable for practical applications in the future.

5 Discussion and Conclusion

This research presents a technique for automatically creating mathematical MCQs with context using the LangChain framework. It combines multi-round LLM dialogue with plug-gable external tools. Questions created by this solution may be guaranteed to a certain degree, with a much greater quality than those directly generated by GPT4, and already comparable to questions prepared by human experts in some respects. By using the semantic expression capabilities of LLM and external searchable knowledge base, the produced questions are more closely connected with real-life scenarios, in accordance with the goals of developing and evaluating the fundamental mathematical literacy skills.

Various research has indicated that ChatGPT should advise and collaborate on educational exams to reduce teachers' workload rather than replace them [11]. This study intends to improve question quality and recognize the need of sensitivity in education. It suggests using machine-generated exercises as teaching materials rather of giving them to students. Our platform design gives educators easy tools for creating and managing questions and banks (Fig. 2). It may also output files in docx, excel, JSON, and other formats, simplifying conventional exams and enabling compatibility with other educational platforms.



Generate New Exercise

Select a topic and fill out the following form to generate your math exercise

knowledge: Mixed addition and subtraction operations for numbers up to 100

question type: Multiple Choices

question quantity: 3

difficulty: Simple

contextualized: some question are contextualized

other restriction: Please fill in your special requirements for generating questions here, such as: "Please try to avoid generating questions with high computational complexity."

Stem: Mary brought 24 small cookies from home to school. If she ate some cookies with her friends and there were still 15 cookies left, how ma l she eat with her friends in total?

Alternative A: 9

Alternative B: 8

Alternative C: 10

Correct Answer: A) 9

Description (optional): Mary brought 24 small cookies from home to school. She and her friends ate some together, and there were still 15 left. By calculating 24-9=15, it can be concluded that she and her friends ate a total of 9 cookies, so the answer is A.

Fig. 2. Automated Solution for Producing Contextualized Multiple-Choice Questions

Nevertheless, there are still some limitations in this method. Initially, the sup-port for various question kinds is still incomplete yet, this solution specifically focuses on MCQs in the field of mathematics. However, it does not yet provide any ways for creating fill-in-the-blank questions or true/false questions. Further-more, the approach used to manage the complexity of question generated is quite rudimentary. The system merely modifies the difficulty of questions by modifying prompt and depending on the LLMs' comprehension, without using exact quantitative techniques. Currently, the produced questions are restricted to numerical and algebraic concepts and do not provide support for geometry questions or statistics chart questions in many formats. For future work, researchers could attempt to migrate the solution proposed by this study to other subjects and question types. The researchers could also optimize on the control of difficulty and the source of the input knowledge points, or embed more precise retrieval-enhanced algorithms to improve the quality of the generated questions.

Acknowledgements. This work was partially supported by the National Natural Science Foundation of China under Grant 61977058, and Natural Science Foundation of Shanghai under Grant 23ZR1418500.

References

1. Chen, W.: Program of thoughts prompting: disentangling computation from reasoning for numerical reasoning tasks. arXiv preprint arXiv:2211.12588 (2022)
2. Chen, Y.: Reinforcement learning based graph-to-sequence model for natural question generation. arXiv preprint arXiv:1908.04942 (2019)
3. Grévisse, C.: Comparative quality analysis of GPT-based multiple choice question generation. Int. Conf. Appl. Inform., 435–447. Springer Nature Switzerland (2023). https://doi.org/10.1007/978-3-031-46813-1_29
4. Kurdi, G.: A systematic review of automatic question generation for educational purposes. Int. J. Artif. Intell. Educ. **30**, 121–204 (2020)
5. Lee, U.: Few-shot is enough: exploring ChatGPT prompt engineering method for automatic question generation in English education. Educ. Inf. Technol., 1–33 (2023). https://doi.org/10.1007/s10639-023-12249-8
6. Siddiq, F.: Taking a future perspective by learning from the past-a systematic review of assessment instruments that aim to measure primary and secondary school students' ICT literacy. Educ. Res. Rev. **19**, 58–84 (2016)
7. Zhou, Q., Yang, N., Wei, F., Tan, C., Bao, H., Zhou, M.: Neural question generation from text: a preliminary study. In: Huang, X., Jiang, J., Zhao, D., Feng, Y., Hong, Yu. (eds.) NLPCC 2017. LNCS (LNAI), vol. 10619, pp. 662–671. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-73618-1_56
8. Zong, M.: Solving math word problems concerning systems of equations with GPT-3. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 15972–15979 (2023)
9. Boland, R.J., Lester, N.A., Williams, E.: Writing multiple-choice questions. Acad. Psychiatry **34**(4), 310–316 (2010). https://doi.org/10.1176/appi.ap.34.4.310
10. Das, B., Majumder, M., Phadikar, S., Sekh, A.A.: Automatic question generation and answer assessment: a survey. Res. Pract. Technol. Enhanced Learn. **16**(1), 1–15 (2021). https://doi.org/10.1186/s41039-021-00151-1
11. Lo, C.K.: What is the impact of ChatGPT on education? A rapid review of the literature. Educ. Sci. **13**(4), 410 (2023). https://doi.org/10.3390/educsci13040410