Education Reform and Development, 2024, Volume 6, Issue 12
https://ojs.bbwpublisher.com/index.php/ERD
Online ISSN: 2652-5372
Print ISSN: 2652-5364

The Practice of Large Language Models in Automated Question Generation: A Case Study of ChatGLM in High School Information Technology Curriculum

Yanxin Chen[1,2], Ling He[1,2*]

[1]Jiangxi Science and Technology Normal University, Nanchang 330038, China
[2]VR Perception and Interaction Key Laboratory, Nanchang 330038, China

***Corresponding author:** Ling He, lynlynhe126@126.com

Abstract: With the advancement of information technology, the field of education is undergoing transformation. In the teaching of high school information technology, scientific and efficient question formulation is crucial for enhancing the quality of education. Traditional methods of question formulation rely on the experience of teachers, leading to issues such as inconsistent difficulty levels and incomplete coverage of knowledge points. Large language models (LLMs), such as ChatGLM, leverage natural language processing and deep learning technologies to automatically generate questions that align with the curriculum, thereby improving scientific accuracy and precision, enhancing diversity and innovation, and catering to students' personalized needs. Additionally, LLMs can enhance the efficiency of question formulation and reduce the burden on teachers. This paper explores the application value of large language models in the formulation of questions for high school information technology, through empirical research comparing the performance of human-generated and ChatGLM-generated questions in terms of accuracy, relevance, clarity, and willingness. The study selected two chapters, "Data and Information" and "Fundamentals of Algorithms," and employed both human and ChatGLM-generated questions, inviting teachers to evaluate them. Through data analysis and statistical testing, we reveal the advantages and limitations of large language models in educational question formulation, providing insights for the intelligent development of educational assessment systems.

Keywords: LLMs; ChatGLM; Question generation; High school information technology

Online publication: January 15, 2025

1. The value of combining large language models with high school information technology question-setting

1.1. Enhancing the scientific and precision of question-setting

In high school information technology education, the scientific nature and precision of question-setting directly

impact students' learning outcomes and teachers' teaching quality [1]. Traditional question-setting methods often rely on teachers' experience and subjective judgment, which can lead to inconsistencies in question difficulty and incomplete coverage of knowledge points. The introduction of large language models can significantly enhance the scientific and precision of question-setting.

(1) Large language models can analyze a vast amount of teaching data and students' learning behaviors to automatically generate questions that align with the curriculum standards. For instance, the model can create programming questions that cover different difficulty levels based on the curriculum standards and knowledge distribution, ensuring that each student practices at a level suitable for their abilities.

(2) Large language models can assist teachers in quickly assessing the quality of questions through intelligent question-analysis tools [2,3]. For example, the model can analyze the difficulty coefficient of questions, the coverage of knowledge points, and the alignment with students' actual learning situations, thereby providing teachers with scientific recommendations for question-setting.

(3) Large language models can dynamically adjust question content based on students' learning data, ensuring that each student is challenged along the most suitable learning path. Through these intelligent question-setting tools, teachers can more efficiently design scientifically and precisely tailored questions, enhancing teaching effectiveness.

1.2. Increasing the diversity and creativity of questions

Traditional information technology questions are often limited to fixed question types and formats, making it difficult to stimulate students' creative thinking and practical abilities. The introduction of large language models can significantly enhance the diversity and creativity of questions.

(1) Large language models can generate various forms of questions using generative technology [4,5]. For example, the model can create programming challenge questions, project design questions, virtual experiment questions, allowing students to practice and explore in different contexts.

(2) Large language models can assist teachers in designing innovative questions through intelligent question-generation tools [6]. For instance, the model can generate programming tasks related to real-life scenarios based on students' interests and learning needs, such as designing a smart home control system or developing an online learning platform. These innovative questions not only stimulate students' interest in learning but also help them better understand the practical applications of information technology in real life.

(3) Large language models can simulate real-life scenarios to create challenging questions. For example, the model can design a virtual programming competition where students must complete complex programming tasks within a limited time, thereby enhancing their practical skills and innovative awareness [7]. Through these diverse and creative questions, students can continuously challenge themselves during their learning process, fostering creative thinking and problem-solving abilities [8].

1.3. Promoting the alignment of questions with students' individualized needs

Each student's learning abilities and interests vary, and traditional question-setting methods often fail to meet students' individualized needs. The introduction of large language models can significantly promote the alignment of questions with students' individualized needs [9,10].

(1) Large language models can analyze students' learning data to automatically generate personalized questions suitable for different students. For example, the model can create programming questions of

varying difficulty and type based on students' programming skills and learning progress, ensuring that each student practices along a suitable learning path.

(2) Large language models can assist students in selecting the most appropriate questions through intelligent question-recommendation systems. For instance, the model can recommend programming tasks related to cutting-edge technologies such as artificial intelligence and data analysis based on students' interests and learning goals, helping them expand their knowledge and enhance their practical abilities.

(3) Large language models can provide real-time feedback to help students adjust their learning strategies. For example, the model can offer feedback and suggestions based on students' performance, helping them identify weak areas in their learning and providing targeted practice questions. Through these personalized question-setting tools, students can continuously optimize their learning paths during the learning process, improving their learning outcomes.

1.4. Improving the efficiency and operability of question-setting

In traditional question-setting processes, teachers often spend a significant amount of time and effort designing, proofreading, and evaluating questions, which not only increases teachers' workload but also leads to low question-setting efficiency. The introduction of large language models can significantly improve the efficiency and operability of question-setting [11].

(1) Large language models can quickly generate a large number of high-quality questions through automated question-generation tools. For example, teachers only need to input curriculum standards and knowledge point requirements, and the model can generate a series of questions that meet the requirements in a short time, greatly reducing teachers' workload.

(2) Large language models can assist teachers in quickly revising and optimizing questions through intelligent question-editing tools. For instance, the model can automatically detect grammatical errors, logical flaws, or deviations in knowledge points within questions and provide modification suggestions, thereby improving the quality and accuracy of questions.

(3) Large language models can help teachers better organize and manage question resources through question-management platforms. For example, the model can automatically categorize and archive different types of questions, making it easier for teachers to quickly find and use them during teaching. Through these efficient question-setting tools, teachers can focus more on teaching design and student guidance, thereby enhancing overall teaching efficiency.

1.5. Summary

The integration of large language models into high school information technology question-setting not only elevates the scientific rigor and precision of questions but also enriches their diversity and creativity, aligns them with students' individualized needs, and substantially enhances the efficiency and practicality of question-setting. These advancements contribute to the modernization and intelligent evolution of high school information technology education, offering students a more enriching learning experience and providing teachers with more effective instructional support.

2. Methods and steps for integrating large language models with high school information technology

2.1. Collecting high school information technology questions

The content of the propositions is derived from the "Data and Information" (Data & Info) and "Algorithm Basics" (Algo) chapters in the compulsory textbook for the first year of high school [12]. In the "Data and Information" chapter, the basic concepts, characteristics of data, and the process of transforming data into information are detailed, along with the significant role of information in modern society. This chapter spans 15 pages with a total of approximately 20,000 words. Following this, the "Algorithm Basics" chapter, which is 16 pages long with about 22,000 words, provides an in-depth yet accessible introduction to the fundamental principles, classifications, design methods of algorithms, and their applications in computer science and other fields.

The proposition methods are divided into two categories: manual proposition and ChatGLM proposition. Manual propositions are primarily based on post-class exercises and past years' selected questions, aiming to consolidate students' classroom knowledge and assess their understanding. On the other hand, ChatGLM propositions utilize artificial intelligence technology [13], generating questions through the natural language processing model ChatGLM. These propositions dynamically adjust the difficulty and type of questions based on students' learning progress and ability levels, providing a personalized learning experience. In the innovative practice of ChatGLM propositions, the study adopts a Prompt engineering strategy that combines role-playing, tasks, and pseudocode (**Table 1**). This strategy ensures the accuracy and pertinence of conveying the intention and requirements of the propositions to the ChatGLM model through the structured expression of pseudocode, enhancing the scientific nature of the propositions. By this strategy, this study guides the ChatGLM model to deeply understand the educational objectives and students' needs, generating questions that meet educational standards and are highly personalized. In terms of the number of propositions, each chapter has 10 questions, totaling 20 questions for both chapters, thus equating to 20 manual propositions and 20 ChatGLM propositions.

Table 1. Prompt details.

Role: You are a high school information technology teacher, possessing highly specialized subject knowledge and pedagogical knowledge, and you can understand Chinese accurately and without error.
Task: Create a written exam question for students.
Requirements: Closely align with the content of the textbook.
Number of questions: 1.
Textbook content: {textbook_content}
Output content: Output the content in the format of pseudocode logic.
[
{"Seq":XX, Question":"XXX"}
]

2.2. Evaluation of human-generated questions vs. questions generated by large language models

The study invited five teachers with extensive teaching experience in the field of high school information technology to conduct a comprehensive and in-depth evaluation of a series of questions. The evaluation process covered several key dimensions, including the hitting of question knowledge points (the degree to which the

tested knowledge points align with the content of the textbook), the fitting (the degree to which the questions reflect core competencies [14,15]), clarity (whether the description of the questions is clear and unambiguous), and willing to use (whether teachers are willing to use the questions in their teaching). To ensure the objectivity and accuracy of the evaluation results, each dimension was scored on a scale of 1 to 5.

2.3. Analysis of results from human-generated and large language model-generated questions

The study first examined the accuracy performance of manual propositions and ChatGLM propositions across different chapters and analyzed the differences between the two through multiple evaluation metrics. By analyzing **Figure 1**, it was found that the average accuracy of manual propositions in the Data & Info chapter was 4.36 with a standard deviation of 0.56, and in the Algo chapter, it was 4.28 with a standard deviation of 0.61. In comparison, the average accuracy of ChatGLM propositions in the Data & Info chapter was 4.18 with a standard deviation of 0.60, and in the Algo chapter, it was 4.16 with a standard deviation of 0.55. These data indicate that the average accuracy of manual propositions was slightly higher than that of ChatGLM propositions in both chapters, although the difference was not substantial. The comparison of standard deviations shows that manual propositions were more stable in the Data & Info chapter, while ChatGLM propositions were more consistent in the Algo chapter.

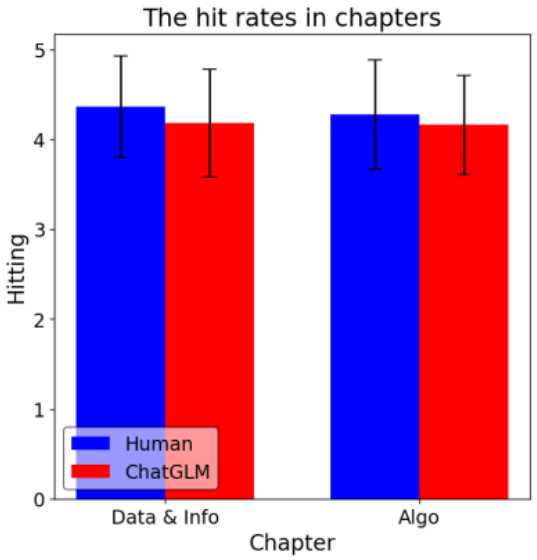

Figure 1. Hitting in chapters.

Further analysis of the data in **Table 2** and **Figure 2** evaluated the performance of Human and ChatGLM on the three indicators of Fitting, Clarity, and Willingness. On the Fitting indicator, the average score for Human was 4.11 with a standard deviation of 0.567, while the average score for ChatGLM was slightly lower at 4.10 with a standard deviation of 0.461. This indicates that the performance of both was very close on this indicator. On the Clarity indicator, the average score for Human was 4.03 with a standard deviation of 0.502, while the ChatGLM model showed a higher average score of 4.14 with a standard deviation of 0.569, indicating that ChatGLM performed slightly better than Human in terms of Clarity. On the Willing indicator, the average score for manual proposers was 3.05 with a standard deviation of 0.539, while the average score for the ChatGLM model was significantly higher, reaching 3.98 with a standard deviation of 0.568. This result clearly demonstrates

that ChatGLM significantly outperformed Human on the Willing indicator.

Table 2. Statistics from different methods

	Method	Mean	SD
Hitting	Human	4.32	0.584
	GLM	4.17	0.570
Fitting	Human	4.11	0.567
	GLM	4.10	0.461
Clarity	Human	4.03	0.502
	GLM	4.14	0.569
Willing	Human	3.05	0.539
	GLM	3.98	0.568

*Note: GLM = ChatGLM

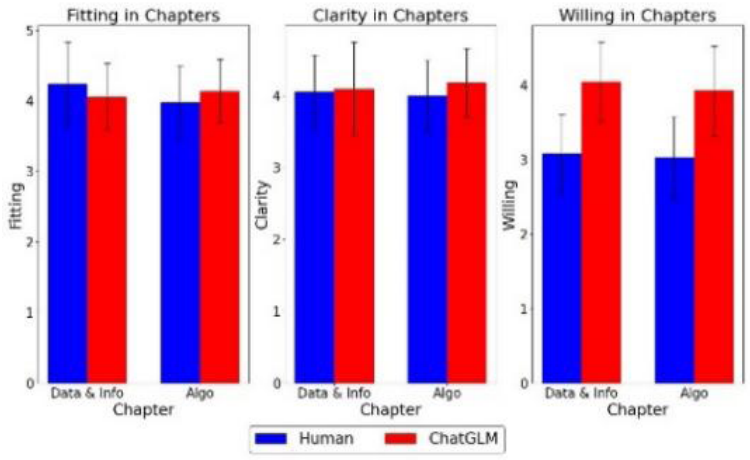

Figure 2. Fitting, Clarity and Willing.

Table 3. One-way ANOVA (Welch's) results for the effect of sources

	df1	df2	*p*
Hitting	1	198	0.067
Fitting	1	190	0.891
Clarity	1	195	0.149
Willing	1	197	<0.001

To explore the impact of different proposition methods on the variables of hit rate, fit, clarity, and willingness, the study employed a one-way analysis of variance (ANOVA) using Welch's method. The results in **Table 3** show that the *p*-values for Hitting ($p = 0.067$), Fitting ($p = 0.891$), and Clarity ($p = 0.149$) did not reach significance levels ($p < 0.05$), indicating that the mean differences in these variables between different proposition methods were not significant. However, the *p*-value for Willing was less than 0.001, significantly

lower than the conventional threshold, and the average Willing score for ChatGLM (3.98) was significantly higher than that for Human (3.05), strongly suggesting that there were significant mean differences in Willing between different proposition methods. In summary, only the Willing variable showed significant mean differences between different proposition methods, while the hit rate, fit, and clarity variables did not show such differences. Through interviews, the teachers who participated in the scoring indicated that they were willing to adopt ChatGLM because it could alleviate the burden of teaching.

3. Conclusion

This study conducted a detailed data analysis and statistical testing to compare the performance of manual propositions and ChatGLM propositions across multiple evaluation metrics, including accuracy, fit, clarity, and willingness. The results indicated that although manual propositions had a slightly higher average accuracy than ChatGLM propositions in certain chapters, the differences were not statistically significant. In the comparison of standard deviations, manual propositions exhibited greater stability in the Data & Info chapter, while ChatGLM propositions demonstrated more consistency in the Algo chapter. On the Fitting indicator, the performance of manual propositions and ChatGLM propositions was remarkably close, indicating similar levels of alignment between the proposition content and teaching objectives. However, on the Clarity indicator, ChatGLM propositions outperformed manual propositions slightly, likely due to the linguistic generation advantages of ChatGLM, which provided clearer and more accurate expressions. The most significant difference was observed in the Willing indicator, where the average score of ChatGLM propositions was significantly higher than that of manual propositions, suggesting that teachers and evaluators were more inclined to accept and use propositions generated by ChatGLM. This result may reflect the potential value of ChatGLM in reducing the workload of teachers and enhancing the efficiency of proposition generation. Through one-way analysis of variance (ANOVA) using Welch's method, it was found that only the Willing variable showed significant mean differences between different proposition methods, while the Hitting, Fitting and Clarity variables did not exhibit such differences. This result further supports the advantage of ChatGLM in enhancing the acceptance of propositions.

In summary, ChatGLM demonstrates significant potential in improving the clarity of propositions and enhancing teachers' willingness to accept them. Although manual propositions still hold certain advantages in some aspects, the integration and optimization of ChatGLM could bring revolutionary changes to the educational assessment system. Future research could explore further optimizations of the ChatGLM model to enhance the clarity and teachers' willingness to use propositions while maintaining high levels of fit and accuracy. Considering the continuous technological advancements, future educational assessment systems are likely to integrate more AI tools similar to ChatGLM to improve teaching efficiency and assessment quality.

Funding

Jiangxi Provincial Education Science Planning Project, "Research on Factors Influencing Learning Engagement in Gamified Learning Environments" (Project No.: 22YB130)

Disclosure statement

The authors declare no conflict of interest.

References

[1] Song R, 2023, What Can Linguistics Contribute to the Improvement of Large-Scale Language Models? Language Strategy Research, 8(4): 53.

[2] Chen Z, Lang W, An H, et al., 2024, A Proposition Intelligent Assistance System Based on Large Language Models. Telecommunications Express, 2024(4): 1–6.

[3] Zhang C, Du L, Zhu X, et al., 2023, Research on Educational Question-Answering Systems Based on Large Language Models. Journal of Beijing University of Posts and Telecommunications (Social Sciences Edition), 25(6): 79–88.

[4] He L, 2024, The Application of Large Language Models in Language Assessment. Foreign Language Teaching and Research, 2024(6): 10–22.

[5] Wang C, 2024, A Paradigm Comparison Between Large Language Models and Generative Linguistics. Journal of Nanjing University of Science and Technology (Social Sciences Edition), 37(4): 61–69.

[6] Yu H, 2010, On Model Reconstruction and Divergent Thinking from the Perspective of College Entrance Examination Proposition. Physics Bulletin, 2010(3): 4.

[7] Xiang L, Li G, Li H, 2024, A Reliability Code Automatic Generation Method Based on Knowledge Graphs and GPT Models. Chinese Journal of Computational Mechanics, 41(2): 217–225.

[8] Liu J, 2023, Exploring Guidance Methods in the Process of Creative Thinking and Design Thinking. National Common Language Teaching and Research, 2023(7): 1–3.

[9] Yang M, Shi L, Su Z, et al., 2024, Research on User Personalized Demand Service Matching Model Based on AIGC. Packaging Engineering, 45(20): 109–119 + 182.

[10] Lan H, 2020, Research on Personalized Cross-Language Query Expansion Based on User Interest Models. Information Systems Engineering, 2020(3): 3.

[11] Xiao F, 2024, "Speech Acts" in Human-Machine Collaboration in the Era of Large Models. Jianghan Forum, 2024(10): 49–56.

[12] Li F, Xiong Z, 2017, Core Literacy-Oriented Information Technology Curriculum: The "Data and Computing" Module. China Educational Technology, 2017(1): 32–37.

[13] Li S, 2024, Reflections on the Reform of Electromagnetic Fields and Waves Teaching Assisted by the Large Language Model ChatGLM. Creative Education Studies, 12(5): 5.

[14] Editorial Department, 2018, The Ministry of Education Issues the General High School Curriculum Plan and Curriculum Standards (2017 Edition). China Ethnic Education, 2018(2): 3.

[15] Su Q, 2023, Analysis of Instructional Design Strategies for High School Information Technology Courses from the Perspective of Core Literacy. Computer Campus, 2023: 6957–6958.