INTERACTIVE LEARNING ENVIRONMENTS
2025, VOL. 33, NO. 3, 2559–2584
https://doi.org/10.1080/10494820.2024.2412082

Routledge
Taylor & Francis Group

Exploring prompt pattern for generative artificial intelligence in automatic question generation

Lili Wang [a], Ruiyuan Song [a], Weitong Guo [a,b] and Hongwu Yang [a,b]

[a]School of Educational Technology, Northwest Normal University, Lanzhou, People's Republic of China; [b]Key Laboratory of Education Digitalization of Gansu Province, Lanzhou, People's Republic of China

ABSTRACT

The construction of questions is an essential component in educational assessment and student learning processes. However, manually constructing questions is a complex task that requires not only professional training, substantial experience, and extensive resources from teachers but is also time-consuming. This article introduces an Automatic Question Generation (AQG) technology based on a prompt pattern to alleviate this burden and address the ongoing need for new questions in education. The essence of this method lies in constructing a prompt pattern grounded on a collective knowledge base derived from teachers, thereby enhancing the quality of the questions produced. Practical applications and expert evaluations demonstrate that integrating a prompt pattern with a collective knowledge base into Large Language Models (LLMs) results in high-quality questions with statistically significant results. These questions not only meet educational standards but also approach the quality of manually constructed questions by teachers in certain aspects. Our research further emphasizes the feasibility of AI-teacher collaboration in education.

ARTICLE HISTORY
Received 5 February 2024
Accepted 26 September 2024

KEYWORDS
Prompt pattern; generative artificial intelligence; collective knowledge base; automatic question generation; education

1. Introduction

Teaching through questions is an age-old practice that has been a cornerstone of education for centuries (Christenbury & Kelly, 1983). Good questions serve as catalysts for meaningful dialogue and empower students to construct their own understanding of the subject matter, promoting not just rote memorization but genuine comprehension and retention (Brookhart, 2014). Fink's (2013) practical research has shown that developing meaningful questions can prompt students to grapple with course content in meaningful ways, fostering deeper understanding and long-term retention. Therefore, the quality of question design is crucial to the effectiveness of teaching. Carefully crafted questions can bring new insights to students, spark in-depth discussions, and prompt comprehensive exploration of the reading topic (Killen & O'Toole, 2023). Conversely, poorly formulated questions can hinder learning by creating confusion, intimidating students, and limiting creative thinking (Tofade et al., 2013). However, designing a good question is time-consuming and labor-intensive. With the ongoing integration of artificial intelligence (AI) technology in the field of education, automatic question generation (AQG) technology has emerged. This technology not only addresses the limitations of traditional manual question design, such as slow speed and suboptimal

CONTACT Hongwu Yang yanghw@nwnu.edu.cn
This article has been corrected with minor changes. These changes do not impact the academic content of the article.

outcomes, but also promotes student autonomy in reading comprehension and facilitates deep learning practices (Liu, Zhang, et al., 2023a; Wang et al., 2024).

Currently, approaches utilizing templates, rules, and statistical methods have been proposed for AQG (Mulla & Gharpure, 2023). Previous studies have confirmed the potential of AQG in the field of education. However, integrating AQG into practical educational settings poses certain technical challenges. Firstly, most AQG research has been spearheaded by technical experts, with limited effective participation from educators (Kurdi et al., 2020). Secondly, existing AQG technologies, once developed, are difficult to adapt due to their heavy reliance on hand-crafted templates, training data, and technical architectures, often resulting in the generation of questions with subpar quality (Alsubait et al., 2016).

The advent of pre-trained Large Language Models (LLMs) has caused a qualitative leap in Generative Artificial Intelligence (AI), resulting in a transformative era in digital content creation and production (Al-Emran, 2024; Gupta et al., 2023). Notably, in December 2022, OpenAI released an LLM called ChatGPT (Chat Generative Pre-trained Transformer) to the public, attracting over a million subscribers within its inaugural week (Farrokhnia et al., 2023). The emergence of LLMs has revolutionized the AI field, providing unprecedented capabilities in natural language understanding and generation and offering the potential for educational change (Adeshola & Adepoju, 2023; Rahman & Watanobe, 2023; Rasul et al., 2023; Rospigliosi, 2023).

The applications of LLMs, such as ChatGPT, in education have recently made significant progress (Ding et al., 2021; Liu, Yuan, et al., 2023b). Research indicates that LLMs are capable of generating high-quality questions (Lee et al., 2023; Rospigliosi, 2023). This research evaluates the effectiveness of LLMs in question generation by providing prompt templates for expert assessment. However, the provided prompt templates are relatively simplistic, merely instructing LLMs to generate a specific type of question, without offering guidance to teachers on how to integrate their practical knowledge into LLMs for generating the necessary questions. According to White et al. (2023), informing LLMs about which information is important, the format of the information, and the expected output format can effectively enhance the quality of LLMs' output. LLMs, akin to students, require the inspiration of teachers' practical knowledge for their intelligent emergence (Jeon & Lee, 2023). Just as students enhance their abilities and expand their knowledge under the guidance and instruction of teachers, LLMs also need teachers to translate their expertise into prompts inputted into LLMs. This serves to train and optimize the LLMs' performance, enabling them to understand and generate knowledge in specific domains more accurately (Aithal & Aithal, 2023; Ausat et al., 2023). However, there is currently a lack of research on AQG based on prompt patterns guided by teachers' practical knowledge in this process of integrating new technology into teaching and learning.

Therefore, in this study, we designed, developed, and evaluated the quality of an AQG system driven by the prompt pattern. This system specifically focuses on incorporating a collective practical knowledge base of teachers. The research questions we propose are as follows:

RQ1: How can we design a collective practical knowledge base and integrate it with existing prompt patterns to form a new prompt pattern for the LLMs-based AQG system that can generate high-quality questions?

RQ2: To what extent can a prompt pattern with a generated question yield high validity and reliability?

The rest of the article is structured as follows. Firstly, we introduced the related work in Section 2. Secondly, we illustrated the prompt pattern in Section 3. Thirdly, we conducted the study using the development research method in Section 4 and showed the results in Section 5. We also discuss the results, implications, and limitations in Section 6. Finally, a brief conclusion was provided in Section 7.

2. Related work

2.1. Automatic question generation

The question can enable students to transition from passive learning to active learning, from merely acquiring knowledge to mastering how to learn, thereby promoting students' deep learning

(Douglas et al., 2012). Teachers can utilize questions to foster students' engagement with the text, bolster their reading comprehension and critical thinking skills, consequently enhancing their comprehension level and the quality of their thinking (Almeida, 2012; Robin, 2015). Therefore, articulating questions, a pivotal tactic for effective communication, is a crucial endeavor in the teaching-learning process.

Crafting effective questions is indeed a challenging endeavor. Educators must thoroughly grasp the subject matter and choose appropriate question styles and formats to assess student comprehension accurately. This process requires significant time and effort. Moreover, manual question generation may yield restricted variety and quality, potentially introducing biases in question selection, as it hinges on individual teachers' backgrounds and levels of expertise (Gorgun & Bulut, 2024). This underscores the necessity for AQG techniques in education.

Research on AQG has its roots in the 1970s. Over time, it has evolved from generating structured questions based on Knowledge Bases (KBs) to deriving unstructured questions from textual information. The significance of AQG research has grown in recent years, coinciding with parallel to the widespread adoption of ubiquitous learning (Gaebel et al., 2014; Goldbach & Hamza-Lup, 2017; Qayyum & Zawacki-Richter, 2018). AQG technology can provide teachers with valuable insights into the fundamental characteristics of effective question creation and evaluation. This represents a crucial step toward developing meta-cognitive knowledge about reading comprehension instruction.

Early research on AQG primarily adopted traditional rule-based methods, which were heavily reliant on handcrafted templates. These methods were often constrained to specific applications within closed domains, resulting in limited universality and scalability (Montenegro et al., 2012). Subsequent advancements in machine learning models introduced various methods such as pre-trained models (Dong et al., 2019), variational autoencoder (Wang et al., 2019), graph neural network (Chen et al., 2019), multi-task learning (Wang et al., 2017), reinforcement learning (Fan et al., 2018), and transfer learning (Liao & Koh, 2020). However, many AQG systems driven by these technologies sometimes struggle to generate contextually appropriate questions, often due to biases and imperfections in their training data (Liao & Koh, 2020). Moreover, they may lack the domain-specific expertise necessary to formulate relevant questions for specialized topics (Alsubait et al., 2012). The significant concern is that current research on AQG is primarily driven by technical personnel, leaving teachers unable to tailor questions to their specific requirements. To effectively harness AQG technology in education, exploring methods for systematically generating question types aligned with diverse educational needs is crucial. Additionally, it is important to conduct iterative design and validation of AQGs, as well as legitimize the entire process of AGQs and test the question generation with low-resource learning techniques, such as Prompt-based Learning, which can be crucial (Lee et al., 2023).

2.2. Applications of generative AI in educational contexts

Generative AI possesses the capability to comprehend and generate human language, predict the probability of word sequences, and produce new text based on given input (Chen et al., 2020). Typically, large language models (LLMs), which contain billions of parameters and utilize the Transformer architecture, are categorized as generative AI (Wu et al., 2023). These models are distinguished by their substantial scale, with parameter counts exceeding billions and training dataset sizes surpassing terabytes. Through exposure to extensive corpora, generative AI leverages its extensive parameter sets to identify and articulate the rules and logical relationships inherent in human language. As a result, it generates new texts that adhere to human linguistic conventions (Roumeliotis & Tselikas, 2023). Therefore, its utility spans across various domains. The robust situational learning capabilities of generative AI enable it to decompose complex problems into simpler components, employing multi-step reasoning to reach solutions (Mungoli, 2023; Savelka et al., 2023; Urhan et al., 2024).

The advancement of Generative AI has significantly augmented its utility in educational settings. In the field of education, Generative AI finds extensive application in areas such as assessment and evaluation (Chaudhry et al., 2023), plagiarism detection (Khalil, 2023), student performance prediction (Stojanov, 2023), intelligent assistance systems deployment (Pinto et al., 2023), learning environment management (Bahroun et al., 2023) and so on. For instance, Markel et al. (2023) introduced a novel AI tool named "GPTeach," specifically designed for teacher training. This tool enables aspiring educators to practice teaching with simulated students powered by GPT. Finnie-Ansley et al. (2022) introduced Codex, an LLM capable of annotating existing code in real-time, thereby facilitating students' rapid and precise comprehension of the codebase. Additionally, Codex can generate code based on students' textual descriptions and elucidate the underlying principles, assisting students in mastering the logic and techniques of code writing. Moreover, leveraging the profound language comprehension and dialogue generation abilities afforded by LLMs, Generative AI has enhanced its capabilities in essay correction, elevating both the dimension and sophistication of its intellectual engagement (Bouziane & Bouziane, 2024). Beyond fundamental corrections of spelling, grammar, and sentence construction, Generative AI can undertake sophisticated essay revisions, analyzing discourse structure and elements of writing style. After the correction phase, it can offer targeted advice on logical structure, vocabulary choices, and other refinements. Concurrently, it can generate multiple sample essays based on the given topic, serving as resources for teachers' essay instruction and references for student learning.

However, Generative AI is also controversial due to its tendency to generate false content, misleading information, and implicit biases. Studies (Savelka et al., 2023) show that the capabilities and limitations of GPT models for reasoning and analyzing code within educational settings have not been fully explored. Therefore, it has been highlighted that the issue of "garbage in, garbage out," akin to traditional computing devices, also persists in artificial intelligence, such as ChatGPT (Vidgen & Derczynski, 2020). With the assistance of generative AI, we can uncover new educational patterns, gain comprehensive insights, generate high-quality outputs, and receive personalized responses that release infinite possibilities from the vast amount of accessible data. However, to unlock this magical box, one must master the key – Prompt pattern.

2.3. Prompt pattern

In the context of advancements in LLMs, such as ChatGPT, Prompt Engineering emerges as a transformative practice, reshaping how these AI systems interpret and respond to textual queries (Wang, Shi, et al., 2023; Zhou et al., 2022). Prompt Engineering is a method trained or fine-tuned using specific prompts or cues to generate desired responses (Gu et al., 2023). Prompts serve as instructions fed to LLMs to enforce rules, automate processes, and ensure specific qualities and quantities of the generated output (White et al., 2023). By carefully designing prompts, researchers and developers can influence the output of the model, making it more specific, relevant, or contextually appropriate for a particular application or task (Haque et al., 2022). A well-constructed prompt acts as a conduit to direct the model's extensive knowledge and computational power into task-specific responses. Crafting this conduit requires an understanding of human psychology, linguistic nuances, and cultural contexts (Mungoli, 2023). These prompts' systematic design and optimization is essential to guide LLMs' responses, ensuring accuracy, relevance, and coherence (Sarkhel et al., 2022). Prompts not only enable LLMs to output content but also reshape the future of human-AI collaboration (Murungu, 2024).

The design of prompts encompasses various approaches, including established best practices for innovative research techniques. Each approach serves as a tailored input to define tasks, set constraints, provide examples, or specify the desired response format, thereby shaping the final outcome (Baidoo-Anu & Ansah 2023; Kaddour et al., 2023). Prompt patterns emerge as a crucial concept within this framework in Prompt Engineering. A prompt pattern refers to a reusable template or structure for crafting effective prompts. These patterns encapsulate proven strategies for

eliciting specific types of responses from LLMs, ensuring consistency and quality across various applications. For instance, a ''compare and contrast'' prompt pattern might follow a structure like: ''Compare [Topic A] and [Topic B] in terms of their [Aspect 1] and [Aspect 2].'' By leveraging such patterns, researchers and developers can streamline the prompt pattern creation process, enhance reproducibility, and facilitate the systematic exploration of LLM capabilities across different domains and tasks. At present, the design of prompt patterns mainly consists of the following key parts: Instruction, Context, Input Data, Output Indicator, and Example Code (Wei et al., 2022). These elements are essential in guiding LLMs to generate precise and relevant responses (Giray, 2023; Lu et al., 2021).

- **Instruction:** This pertains to the specific tasks or commands executed by LLMs, such as performing text classification tasks.
- **Context:** This refers to the external information or additional context provided to the LLMs to guide them in producing more refined outputs.
- **Input Data:** This provides the fundamental principles for the problem and an explanation of their significance.
- **Output Indicator:** This specifies the type or format of the desired output.
- **Example Code:** This illustrates how prompt patterns are applied in practical scenarios, offering a tangible demonstration of their implementation.

In summary, a substantial body of knowledge regarding effective prompts has been accumulated. These established prompt patterns provide reusable solutions for structuring interactions with LLMs. However, research has indicated that when LLMs are supplied with basic instructions lacking elaborative context, the generated outputs tend to be overly generalized (Luo et al., 2019). This phenomenon underscores an issue originating from the imprecision of the prompt, which may result in responses that, although informative, fail to adequately cater to the specific requirements or interests of the reader. A more effective prompt would narrow the focus to a particular aspect of software engineering or request detailed insights into how a specific trend is impacting the field. Such guidance would yield a more focused and pertinent response, affording a clearer and more concise understanding of the specified area in software engineering (Mungoli, 2023). Teachers, as guides and creators of knowledge, fulfill a pivotal role in devising these effective prompts, drawing on their expertise and experience. They resemble classroom instructors equipped to aid the ''student'' LLM in refining its learning and responding, thereby enhancing the quality and efficiency of LLM output. However, there is currently a lack of empirical research that melds educators' practical wisdom with prompt patterns to tackle AQG problems.

3. Prompt pattern for generative AI in education

The quality of outputs generated by generative AI is directly related to the quality of prompts provided by educators. As discussed in Subsection 2.3 of the prompt pattern, prompts given to LLMs can enhance interactions between educators and LLMs and help address various issues. An important contribution of this article is the design of a structured prompt pattern. The key is to create a pattern capable of summarizing and synthesizing the collective wisdom of teachers. This firstly involves translating teachers' practical knowledge into a knowledge base. The knowledge base is then input into LLMs as part of the prompt pattern process to facilitate the generation of more professional and accurate questions.

This study aims to generate questions that align with educators' pedagogical requirements by enhancing the ''Context'' component within the pattern framework detailed in Section 2.3, which comprises Instruction, Context, Input Data, Output Indicator, and Example code. We propose incorporating a teacher-based collective knowledge base, thereby enriching the generation process with expert insights. This enhancement is designed to assist LLMs in addressing various practical tasks in

an educational context, resulting in more structured and detailed outputs. In this context of question generation, the collective practical knowledge base refers to a systematized compilation of teachers' questioning strategies developed through years of classroom practice. This knowledge base is created by extracting and codifying educators' implicit knowledge when formulating questions. It encompasses a wide range of pedagogical strategies, best practices, and domain-specific understandings teachers have developed through years of classroom experience and professional development. The core of this collective knowledge base lies in our effort to transform teachers' tacit understanding of effective questioning into explicit, goal-oriented questioning strategies. By integrating this collective practical knowledge base into the prompt pattern, we aim to enhance the LLM's capacity to generate questions that not only accurately cover the subject matter but also align with established pedagogical practices. This approach bridges the gap between AI-generated questions and the nuanced, context-aware questioning techniques employed by experienced educators, thereby improving the quality and educational value of the generated questions. The schematic diagram of this pattern is illustrated in Figure 1.

Eager and Brunton (2023) provide guidance for crafting instructional text to guide producing high-quality outputs from LLMs in higher education. They recommend including six components in written prompts to facilitate effective prompt engineering: Verb, Focus, Context, Focus and Condition, Alignment, Constraints, and Limitations. Following this guidance, we have designed the configuration process of the collective knowledge base, with an emphasis on clear and precise instruction for prompt engineering. Initially, this process involves categorizing the knowledge itself and determining the structure and organization of the base. Subsequently, a precise and accurate definition of the knowledge is required to capture its essence and connotation, thus preventing ambiguity within the category. Furthermore, based on this foundation, providing a more explicit definition of the key features inherent in the knowledge is imperative. This ensures that LLMs can distinctly grasp the fundamental features of the knowledge conveyed, thereby enhancing the accuracy of their outputs. Building on the existing prompt patterns, a collective knowledge base forms the ultimate generative AI application patterns customized for the educational domain. In this pattern, the prompt categories encompass Role, Output Indicator, Type, Definition, Characteristic, and Example Code, each of which is summarized below.

Role: Specifying the user's role and conveying the idea that the LLMs must assume a specific role and provide outputs related to what such a persona would generate. This role can be expressed in

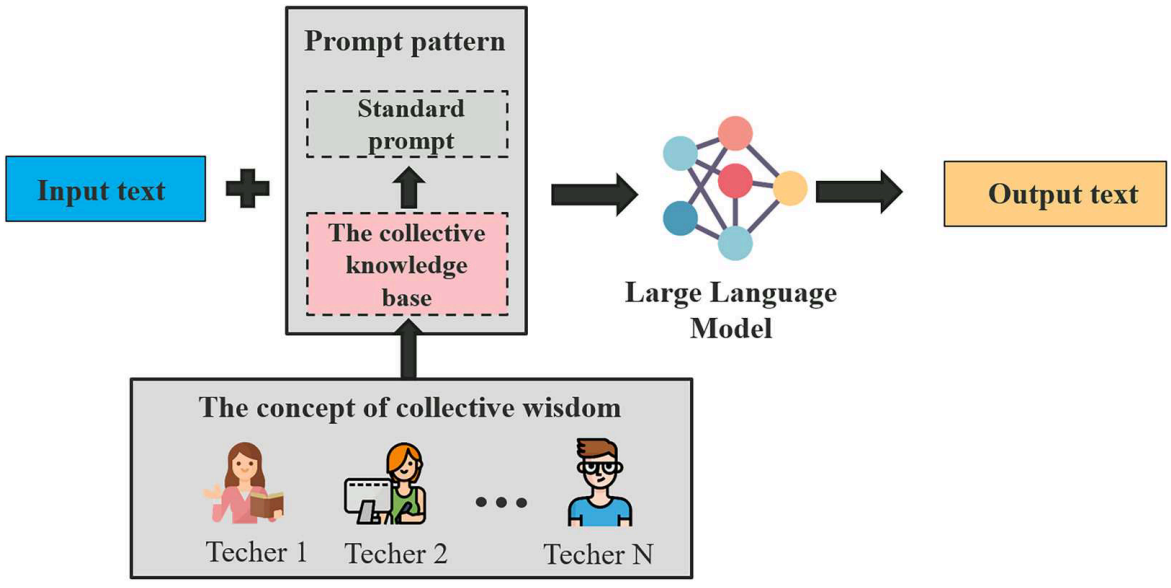

Figure 1. The prompt pattern for generative AI application in education.

several ways, including job descriptions, titles, fictional characters, historical figures, etc. When the model adopts a specific role, it draws upon relevant knowledge and mimics the communication style associated with that role (Zhang et al., 2023). The effectiveness of this technique largely depends on the training of the model and the chosen role's knowledge base. For example, specifying that LLMs should act as a middle school Chinese teacher and generate answers related to that role.

Output Indicator: Emphasizing the constraint or tailoring of the types, formats, structures, numbers, or other attributes of the output generated by LLMs to maintain consistency in output format and control the length and details of the generated questions. For instance, we can specify that LLMs can only generate 3 questions at a time in AQG tasks.

Type: Indicating the category of knowledge definitions in the collective knowledge base. LLMs can choose the appropriate interaction type from the collective knowledge base to ensure a swift achievement of the goal and refine the professionalism of the outputs. For example, suppose a teacher wishes to generate evaluation questions. In that case, he can define those question types, and their definition will guide the LLMs in retrieving the relevant information from the knowledge base. This ensures that the generated evaluation questions align with the demands of the teacher.

Definition: Referring to the specific definition of the type within the collective knowledge base, which could enable LLMs to comprehend intricate concepts and relationships in this category with enhanced professionalism and accuracy, thereby generating output content closely related to user needs, clear, and targeted. For example, if teachers need to generate "Evaluation Questions," providing a clear definition will help LLMs generate specific questions related to assessment rather than broad questions.

Characteristic: Representing features in the Definition category that consist of one or more elements. The purpose is to enable LLMs to comprehend the fundamental principles of specific outputs, clarify how to process inputs, make specific assumptions, or provide specific data, among other tasks. For example, Evaluation Questions have typical questioning characteristics, such as ① Do you agree? Why? ② Would it be better if … ? ③ Why do you think that? etc. When these features are input into LLMs, the LLMs will mimic and learn from these specified characteristics to generate more effective questions.

Example Code: Providing a specific practice case to guide LLMs in learning specific patterns, styles, or content, that enable LLMs to generate outputs that match or relate to the example. For instance, the specific example given for Evaluation Questions is "Do you agree with the author's statement of "Reading is good, reading many books, reading good books'? Why?" This allows LLMs to better understand the user's intentions through examples and apply these learned patterns to future generation tasks.

4. Method

This study employed the Development Research Methods to conduct our investigation. This methodological approach represents a pragmatic type of research that offers a way to test "theory" that has been only hypothesized and to validate practice that has been perpetuated essentially through unchallenged tradition (Richey & Klein, 2005). It differs from design-based research, as its purpose is to systematically explore the process of design, development, and evaluation, aimed at grounding the creation of both instructional and non-instructional products with empirical evidence. This type of method is particularly suitable when the problem focuses on emerging technologies (Seels & Richey, 2012).

Development-based research comes in two primary types: Type 1 studies may have an analysis phase, a design phase, a development phase, and an evaluation phase. Type 2 studies may have a model construction phase, a model implementation phase, and a model validation phase. According to the purpose of our research, we selected Type 1, which provides a structured blueprint for the entire process of research design, development, and evaluation, thereby establishing a

methodological foundation for evaluating the AQG system. This process is illustrated in Figure 2. In both the Methods and Results sections, we organized our research presentation around the design, development, and evaluation framework. Building upon this structure, we employed a mixed-methods approach to data analysis. This methodology facilitated the collection of participants' experiential feedback on our proposed AQG system through both quantitative questionnaires and qualitative open-ended discussions. Furthermore, we implemented a blind expert review process to assess the quality of the generated questions, thereby validating the efficacy of our proposed system. This comprehensive approach enabled us to triangulate our findings, providing a robust evaluation of the AQG system's performance and user acceptance.

4.1. Design

The focus of the research design phase is to obtain a collective knowledge base based on the questioning characteristics of the teachers' instructional designs, and then embed the collective knowledge base into the prompt pattern for Generative AI in education to generate questions. This stage mainly adopts coding and expert validation methods. Firstly, we downloaded the exemplary Chinese instructional designs of the primary school from The Smart Education of China Platform. Secondly, the first and second authors extracted the questions designed by teacher and questions characteristics from these instructional designs. Thirdly, based on the analysis of the above questions characteristics and combining with the practical needs of teachers in instructional practices, the two researchers classified questions into 11 types under the guidance of common question classification

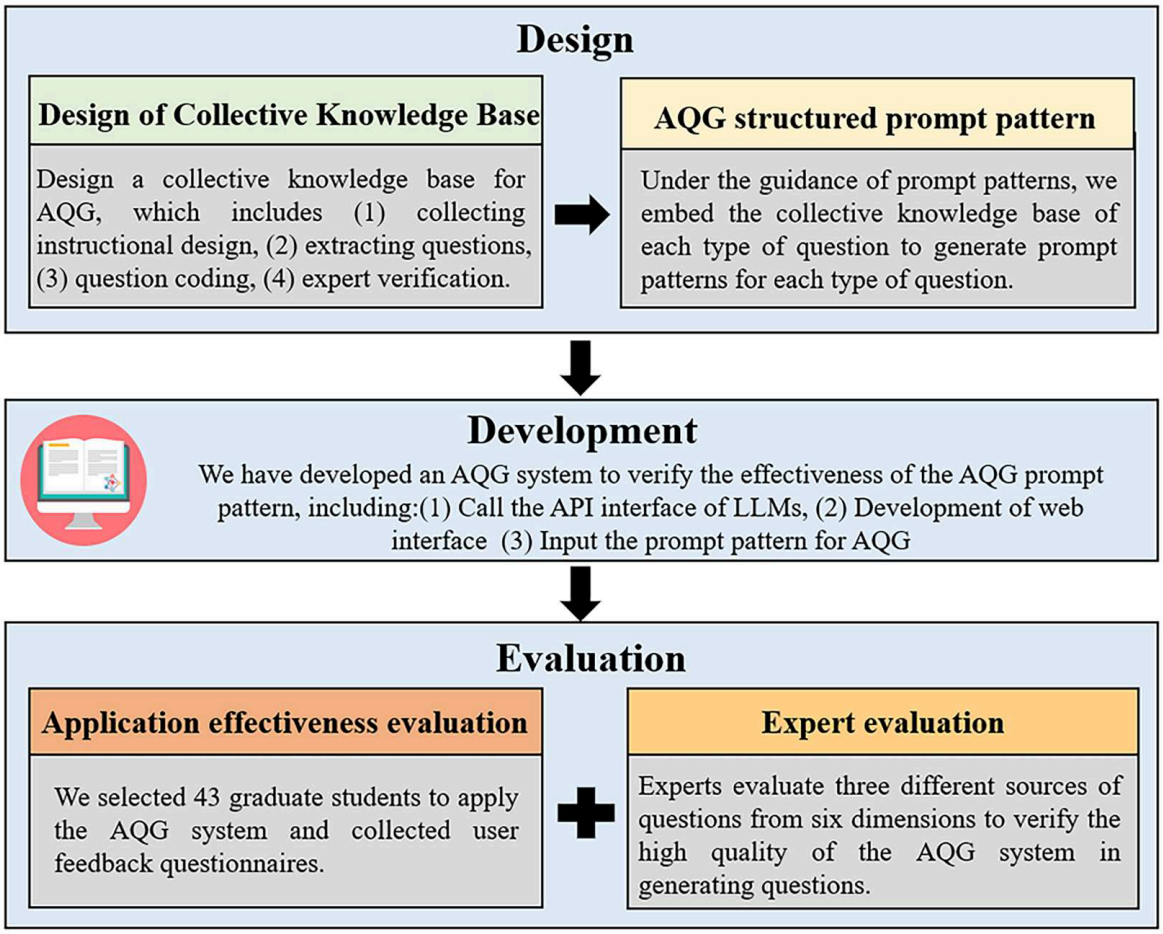

Figure 2. Research process diagram.

theories such as Bloom's Taxonomy (2020) (remembering, understanding, applying, analyzing, evaluating and creating), John Dewey's (2001) Five-Step Teaching Method involving posing difficult situations, defining or specifying the question, proposing various hypotheses to solve the difficulties, making inferences based on these hypotheses, and verifying or modifying the hypotheses, Halliday's (1998) common relationship types (classification, prediction, comparison, reasoning, and summary), as well as Scott' (1973) concepts of open-ended questions and closed-ended questions. Finally, the researchers independently encoded the characteristics represented by the questions. To facilitate a comprehensive analysis, the coders meticulously reviewed the characterization of each question multiple times to ascertain its appropriate category. Moreover, the coders-maintained blindness to each other's codes throughout the process.

After coding independently, we obtained a collective knowledge base that includes characteristics, definitions and types. To evaluate the effectiveness of our knowledge base, we invited three Chinese teachers with more than 10 years of teaching experience to review our defined collective knowledge base. Initially, we sent the defined collective knowledge base to the three teachers and explained our purpose to them. After a week, we conducted semi-structured interviews with each teacher for about half an hour to gather feedback and improve the clarity and accuracy of the knowledge base definitions. All three teachers hold senior professional titles, ensuring that the knowledge base we designed is practical and effective in educational settings.

4.2. Development

To verify the effectiveness of our proposed prompt pattern, we developed an AQG system with a user-friendly web interface using the Python programming language. We utilized APIs from OpenAI for this purpose. We inputted the AQG prompt pattern into the design page for operation. Additionally, we adjusted the output results through multiple rounds of debugging and configuring different parameters. Finally, we set the key parameters for LLMs to temperature = 0.5 and top-p = 0.8. Since the participants in the experiment were all Chinese, the experiment's design and the results' presentation were in Chinese. After inputting the text, the teacher can choose to manually input the text into the designated input area or directly upload a document for analysis. Subsequently, the teacher can generate the desired question type with a single click. At the same time, if a teacher needs to generate an answer to a question, they only need to click on the answer button to obtain the answer to the generated question. Moreover, teachers can also choose different LLMs, configure various parameters, and automatically generate questions according to their preferences.

4.3. Evaluation

This study employs a combination of questionnaire surveys and assessment reports on generated questions to validate the effectiveness of AQG based on generative AI prompt patterns. The evaluation phase endeavored to evaluate two key aspects with specific objectives. On one side, it aims to evaluate the validity of the questions generated by AQG systems with prompt patterns which have in practical applications. On the other side, it aims to evaluate whether the questions generated based on the AQG system have high quality.

4.3.1. Participants

We selected 43 third-year graduate students from normal universities to participate in the experiment. The reason for selecting these participants lies in their dual roles as both teachers and learners, enabling a comprehensive evaluation of the AQG system from both instructional and learning perspectives. This approach facilitates a more efficient assessment with a limited sample size. Furthermore, we included 12 teachers in the experimental process to conduct a more in-depth professional review of the AQG system. This additional step ensures a comprehensive evaluation

from experienced educators, contributing to a thorough assessment of the system's pedagogical effectiveness. Among these 12 teachers, eight are university lecturers or teaching assistants, with five having a background in computer science and three in education. This diverse composition allows for a comprehensive evaluation of the system from both technical and pedagogical perspectives. Additionally, the remaining four teachers are Chinese language teachers at the junior high school level, boasting rich teaching experience with service durations of 1.5, 4, 18, and 21 years. Their evaluations contribute valuable insights from the dimension of practical teaching experience.

4.3.2. Tools

The experimental tools comprise two components: research questionnaires and an assessment report. The research questionnaire was developed based on established measurement scales. The scales included a performance measurement scale, a technical acceptance measurement scale, and a platform reuse intention scale. Among these, the performance measurement scale reflects the effectiveness that learners expect when interacting with an AQG system during interactive learning. The scale was adapted from the perceived value scale (Zhai et al., 2017). The scale comprises three dimensions: perceived feedback, perceived value, and satisfaction. The perceived feedback dimension comprises 3 questions designed to measure participants' subjective perceptions of professionalism, accuracy, and consistency in category classification of questions generated by the system. The perceived value dimension includes 3 questions, primarily aimed at measuring which form of generated questions, those based on prompt patterns or non-prompt patterns, is more useful for teaching. The satisfaction dimension consists of 2 questions, primarily designed to measure participants' overall satisfaction with the AQG system. The technical acceptance measurement scale reflects learners' perceptions of the ease of use and interaction with the designed AQG system when they adopt a generative AI prompt pattern. This scale was adapted from the expected performance scale in studies on AI acceptance (Gansser & Reich, 2021). The scale includes two dimensions: practicability and operability, with 3 questions for each dimension. The reuse intention scale refers to learners' willingness to reuse a specific educational technology tool. This scale was adapted from the psychological needs research scale on human-computer collaborative reuse intention (Jahn et al., 2021). This scale also comprises 2 dimensions: autonomous demand and willingness to reuse, each with 3 questions. It's important to note that the autonomous demand dimension primarily measures participants' acceptance of this training session. All variables in the questionnaire were measured using a five-point Likert scale, ranging from "Strongly Agree," "Agree," "Neutral," "Disagree," to "Strongly Disagree." While modifying the measurement scale, efforts were made to ensure that the wording used in the scale was as clear and understandable as possible for the participants. Before formally distributing the survey questionnaire, industry experts were invited to review it to guarantee its scientific rigor and precision.

The design of an assessment report includes 40 reading passages, each accompanied by a variable number of questions ranging from 3 to 7, resulting in a total of 196 questions. These questions consist of 62 items directly generated by LLMs, 66 items generated by combining three questions from each category using the AQG system, and the remaining 68 questions selected from teachers' instructional designs. The researchers have already labeled the source of each question. However, before the assessment, all questions are shuffled, and assessors need to be made aware of the origin of each question during the assessment process. Ratings were collected on a 1-to-5 Likert scale to measure the extent of generated questions. Each question is assessed across six dimensions, namely:

- answerable, by looking at their context (Answerability);
- relevant to their context (Relevance);
- grammatically correct (Correctness);
- semantically sound (Soundness);

- well-posed and natural (Fluency);
- and, possessing pedagogical significance (Quality).

4.3.3. Procedure

The experiment consists of two components. The first part involves all participants engaging in learning on the platform and completing questionnaire assessments. The second part entails the selection of two teachers from the participating educators to evaluate the assessment report. The specific experimental design is outlined as follows.

Preceding the initiation of the experiment, a comprehensive presentation elucidating the platform's functionalities was delivered to all participants. Furthermore, a brief user manual was furnished to each participant. After the system demonstration, a concise meeting was convened to address any questions the participants had. We conducted two rounds of experiments to ensure that all participants were proficient in using the platform. In the first round of the experiment, participants were instructed to input a specified reading text. They were then given the freedom to choose any button to generate questions of various categories automatically. Participants were encouraged to discuss the effectiveness of the generated questions among themselves. In the second round of the experiment, participants were given the autonomy to select texts and generate questions of various categories. After completing both rounds of experiments, participants were asked to fill out feedback questionnaires. The researchers explained the purpose of the survey and assured participants of its anonymity to alleviate any concerns. Due to the non-uniform availability of participants for the experiment, it was conducted in three separate sessions. However, the procedures for each session remained consistent, with a duration of 80 min allocated for each experiment. Furthermore, we invited 5 teachers and 3 students who participated in the questionnaire to engage in open-ended discussions, aiming to glean deeper insights and solicit suggestions for refinement.

To assess the validity of our protocol and the generated questions, we sought feedback from teachers actively engaged in the field. This validation process ensured the practicality and accuracy of our tool in real-world educational scenarios. Therefore, we enlisted Chinese language teachers with teaching experience of 18 and 21 years, respectively, to evaluate the assessment report. It is noteworthy that these evaluators conducted their assessments without any prior exchange of information. Their task is to evaluate the extent of each generated question across six dimensions. They were encouraged to comment on questions they perceived as disagreement or strong disagreement, but specific mandatory requirements regarding whether to annotate and the length of annotations were not imposed. Moreover, they were also expected to answer the questions. The primary rationale behind this approach was to mitigate the potential impact of additional workload, which could influence teachers to choose agreement. Before the evaluations, we clearly communicated the scoring criteria, which gauged the protocol's ability to produce valuable questions for educators.

5. Results

5.1. Design

5.1.1. Design of collective knowledge base

To facilitate AQG supported by generative AI, it is imperative to understand the question design methodologies and rules that teachers use in education to obtain types, definitions, and characterizations for prompt patterns. We downloaded 318 excellent primary school Chinese instructional designs through The Smart Education of China Platform. Then, we selected 2653 questions from the instructional designs and extracted their corresponding characterization. We used case studies to analyze *Osmanthus Rain's prose and Reading Memories's narrative* in the fifth-grade Chinese textbooks, as presented in Tables 1 and 2.

Table 1. In-depth analysis of questioning strategies in Osmanthus Rain.

Specific questions in teacher instructional design	Characterization of the questions
1. It's the season when Osmanthus flowers open again. Do students like Osmanthus flowers? 2. Today, we are going to learn the text *Osmanthus Rain*. Students can guess what aspects of the story will be about Osmanthus.	This category is of the "What is the relationship between these facts?" type. ① What do you think will happen? Why? ② What does the title/picture tell you about the story/article? ③ What clues in the text helped you make a prediction? etc.
1. What is this article about? 2. How would you describe Osmanthus? 3. What is laurel really like in the field? Please use the description from the author's original article. 4. What kind of love affair did the Osmanthus flowers in my hometown have with my mother? Read the text aloud and find out the relevant words and phrases in the text.	The first level is "What are the facts?" type. Specific questions are characterized as ① What is it about? ② Who did what? ③ Where did it happen? ④ When did it happen? ⑤ Why did it happen? ⑥ How did it happen? etc. The second level is "What do you know about the facts?" Specific questions are characterized as ① What do we already know about … ? ② Can you give me an example? ③ How would you describe … … ? etc.
The author draws on the scenery to express the emotion. What feelings does the author imply by describing his love for Osmanthus flowers?	This category is of the "What is the relationship between these facts?" type. Specific questions are characterized as ① What does … … mean? ② What can you figure out the author didn't put in words? ③ What facts or ideas show … … ? ④ What are the implications of … ? ⑤ What statements support … … ? etc.
The fragrance of Osmanthus flowers is the same, and even the Osmanthus flowers on the hill in Hangzhou are more fragrant because there are more of them. We use our noses to distinguish the aroma, so what did my mother use to distinguish the aroma?	This category is of the " What is the relationship between these facts?" type. Specific questions are characterized as ① How would you compare/ contrast … ? ② How are … and … similar/ different? ③ What's the similarity/difference between … band … ? ④ How can you make a distinction between … and … ? ⑤ How can you identify the different parts? etc.
What is the author's favorite thing about Osmanthus? What joys did Osmanthus bring to the author's childhood?	This category is of the " What is the relationship between the facts?" type. Specific questions are characterized as ① What do you think is the reason for . … … ? ② What is the reason of … … ? ③ What is the result of … … ? etc.
What is the topic of this article? What is it mainly about?	This category is of the " What is the relationship between the facts?" type. Specific questions are characterized as ① What is the best way to summarize this article? ② What is the theme of … … ? What is the main idea of … … ? ③ How would you summarize … … ? ④ What conclusions can you draw from … … ? etc.
What does "Lingering Fragrance of Osmanthus Perfumes the Ten-Mile Radius" mean? What are the synonyms?	This category is of the "What are the facts?" type. Specific questions are characterized as ① Can you restate … … in your own words? ② Can you explain … … ? ③ Can you say it in a different way but with the same meaning? ④ Can you explain … in your own words? etc.

Two researchers independently encoded the characteristics of the questions. The Pearson correlation coefficient is 0.963, which suggests a high degree of concordance among the raters. A collective knowledge base include Characterizations, Definitions, and Types was obtained. Then, we invited three senior Chinese teachers to evaluate it who unanimously agreed that our collective knowledge base is valid, but disagreed on some characteristics. Specifically, in our initial collective knowledge base, "What are the facts?" was categorized as "Interactive Questions," but the two teachers believed that the questioning method in the "What are the facts?" category employed multiple strategies. For instance, ① What is it about? ② Who … ? ③ What … ? ④ Where … ? ⑤ When … ? ⑥ How … ? Another set of phrasings includes ① Can you restate … in your own words? ② Can you paraphrase … ? ③ Can you say it in a different way but with the same meaning? ④ Can you explain … in your own words? Therefore, under the "What are the facts?" category, questions are divided into two subcategories which is "Interpretive" and "Retrieval of Information" (1). Consequently, after considering the opinions of the teachers, we generated retrieved information questions (1) and (2) based on different questioning strategies and category relationships to address these varying explanations. The final classification of the questions is presented in Table 3.

Table 2. In-depth analysis of questioning strategies in *Memories of Reading*.

Specific questions in teacher instructional design	Characterization of the questions
1. What did the author learn about reading? Please use the description in the author's original text. 2. What can be said about "reading is great"?	The first level is "What are the facts?" type. Specific questions are characterized as ① What is it about? ② Who did what? ③ Where did it happen? ④ When did it happen? ⑤ Why did it happen? ⑥ How did it happen? etc. The second level is "What do you know about the facts?" Specific questions are characterized as ① What do we already know about … ? ② Can you give me an example? ③ How would you describe … … ? etc.
1. Please read the article. In what order does the author recount his reading experience? 2. Please sort out which stages of the author's reading experience.	This category is of the "What is the relationship between these facts?" type. Specific questions are characterized as ① Can you divide … into different categories? ② Can you divide … … into groups? etc.
1. Reading the passage, in what order does the writer describe his reading experience? 2. What are the stages of the author's reading experience?	This category is of the "What is the relationship between these facts?" type. Specific questions are characterized as ① Can you put … into different categories? ② Can you classify … into groups? etc.
1. How does the author describe works such as Journey to the West, The Canonization of the Gods, Water Margin and Dang Kou Zhi? Underline the key sentences. 2. What is the author's purpose in comparing Journey to the West with The Canonization of the Gods, and Water Margin with Dang Kou Zhi?	This category is of the " What is the relationship between the facts?" type. Specific questions are characterized as ① What is the best way to summarize this article? ② What is the theme of … … ? What is the main idea of … … ? ③ How would you summarize … … ? ④ What conclusions can you draw from … … ? etc.
By recalling his own reading experience, the author tells us that it is good to read, read many books and read good books. How does this inspire you?	This category is of the " What is the relationship between these facts?" type. Specific questions are characterized as ① What is the best way to summarize this article? ② What is the theme of … … ? What is the main idea of … … ? ③ How would you summarize … … ? ④ What conclusions can you draw from … … ? etc.
When the author remembers reading Romance of the Three Kingdoms as a young boy, he uses the words "He gritted his teeth tightly" and "unexpectedly," etc. Tell us what is so good about these words.	This category is of the "What are the facts?" type. Specific questions are characterized as ① Can you restate … … in your own words? ② Can you explain … … ? ③ Can you say it in a different way but with the same meaning? ④ Can you explain … in your own words? etc.
Do you agree with the author's statement "reading is good, reading many books, reading good books"? Why?	This category is of the "Do you agree or disagree with these facts?" type. Specific questions are characterized as ① Do you agree with … ? Why? ② Would it be better if … ? ③ Why do you think that … ? ④ What's your opinion of … ? ⑤ How would you improve/ prove … ? etc.
The author believes that "reading is good." Therefore, students combined their own experiences to talk about their reading sense.	This category is of the "What do the facts affect you?" type. Specific questions are characterized as ① How would you solve … ? ② What lessons did you learn that you can use in your own life? ③ Is there anything you can get from … ? ④ How did this story make you feel? ⑤ Does it give you any thoughts about … ?
1. After learning this text, we know that reading has many benefits. Students can draw up a reading plan for themselves. 2. There are very many benefits of reading. What efforts will you make to improve your reading habits?	This category is of the "How can you improve/create facts?" type. Specific questions are characterized as ① What is another way to look at … … ? ② How could you invent/ improve … … ? ③ How could you create a different … … ? ④ What changes would you make to solve … … ? ⑤ What would happen if … … etc.

5.1.2. AQG structured prompt pattern

With the guidance of experts, we adjust the questions' classification of characterization and form a collective knowledge base for each type of question, including question characterization, question definitions, and question types. Then we embed the collective knowledge base of each type of question into the prompt pattern to form a structured question prompt pattern. We illustrate our designed prompt pattern using Reasoning Questions as a case, as shown in Table 4. The design process for the other 10 types of questions is consistent with Reasoning Questions, requiring only modifications to the Type, Definition, Characteristics, and Example Code. Additionally, the role definition and the output indicator can be specified and modified according to practical needs.

Table 3. Categorized question knowledge base.

Characterization of the question	Definition of the question	Type of the question
What are the facts?	These are questions in which the reader uses their own words to paraphrase or interpret information from the reading text reasonably, reorganizing the language and making sure that all the important information is included.	Interpretive Questions
	Readers locate, identify, confirm, and extract valid information from a reading text. Retrieving information usually involves two levels of competence. The first level is the ability to directly extract surface information.	Retrieving Information Questions (1)
What do you think the facts are?	Searching for clues (such as titles, pictures, etc.) from the reading material to guess what will happen next.	Predictable Questions
What is the relationship between these facts?	These are questions in which the reader categorizes the information obtained from the reading text according to certain principles or ways.	Categorization Questions
	These are questions in which the reader draws conclusions or makes predictions based on the clues provided in the reading text and what they know.	Inferential Questions
	These are questions in which the reader discovers similarities and differences in things reflected in the factual information of the reading text.	Comparative and Contrastive Questions
	These are questions in which the reader deduces a logical relationship or determines a cause-and-effect relationship between one thing and another based on factual information in the reading text.	Reasoning Questions
	These are questions in which the reader uses their own words to process and summarize the reading text as a whole.	Summary Questions
What do you know about the facts?	Readers locate, identify, confirm, and extract valid information from a reading text. Retrieving information usually involves two levels of competence. The second level is initially memorizing or understanding the surface information.	Retrieving Information Questions (2)
Do you agree or disagree with these facts?	These are questions in which the reader uses evidence to reasonably and logically judge the author's point of view and suggest improvements.	Evaluative Questions
How do these facts affect you?	These are questions in which readers reflect on and use the knowledge they have gained from reading the text in relation to their own lives.	Application Questions
How could you improve or create facts?	These are questions in which the reader imagines and creates the content of the reading text.	Creative Questions

Table 4. A case of AQG prompt pattern.

Text content Input	Reading text	Romonosov is a famous Russian scientist. He was the son of a fisherman. When he was twelve years old, during the day, he followed his father to fish, and at night, he hid in a boarding house and read … …
Questioning strategy	Role definition	You're an elementary school Chinese teacher.
	Output indicator	3 questions
	Type of the question	This is Reasoning Questions.
	Definition of the question	These are questions in which the reader deduces a logical relationship or determines a cause-and-effect relationship between one thing and another based on factual information in the reading text. It belongs to the category of "What is the relationship between these facts?" questions.
	Characterization of the question	① What does … … mean? ② Can you see what the author didn't write? ③ What fact or opinion shows … … ? ④ What is the meaning of … … ? ⑤ What statements support … … ? etc.
	Example code	The author believes that "It is Good to Read," please combine your own experience and talk about your own feelings about reading.

5.2. Development

After development, we obtained an AQG system based on prompt pattern, as shown in Figure 3. In this system, after entering the text that needs to be read, the teacher clicks the question type button,

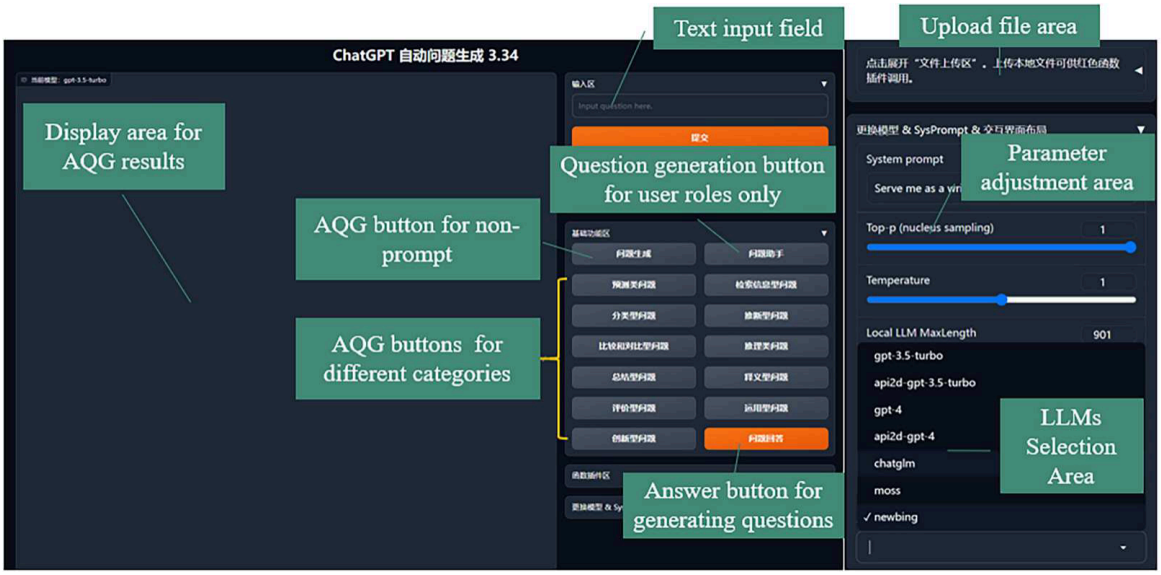

Figure 3. Automatic question generation system.

and the system can automatically generate three questions of this type. As shown in Figure 4, when the following text is entered, the system will generate three questions.

"Romonosov is a famous scientist in Russia. He is the son of a fisherman. When he was twelve years old, during the day, he would go fishing with his father, and at night, he would hide in a makeshift hut to read books. One day, Romonosov and his father were fishing at sea. Suddenly, a fierce wind blew, and his hat was blown onto the deck. Soon, huge waves surged in the sea, and the sails of the ship were blown off. The situation was critical. Romonosov climbed up the swaying mast, disregarding everything, and quickly fastened the fallen sails securely. The fishing boat stabilized again. After the storm passed, his father pulled him aside and said with a grin, "My child, I want to reward your bravery. Shall I buy you a deer skin coat?" Romonosov shook his head. "Then what do you want?" "I want a book, Dad, and nothing else." "Isn't a deer skin coat better than a book?" "Dad, I want a book that contains all knowledge. For example, why don't the stars in the sky fall down, and why does night turn into dawn?" Upon hearing this, his father and the sailors all widened their eyes in astonishment."

The system generates the following problem when the user clicks the "Predictive Question" button:

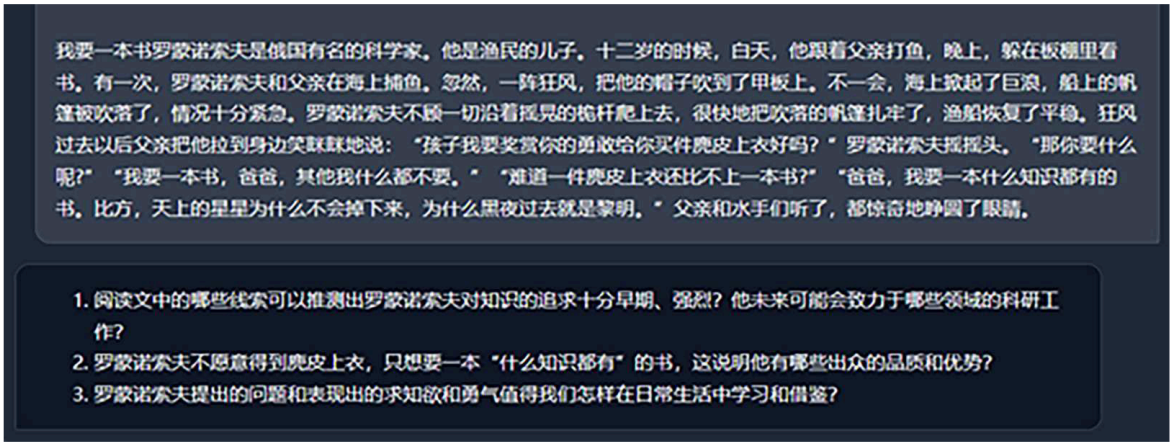

Figure 4. Predictive questions generate page.

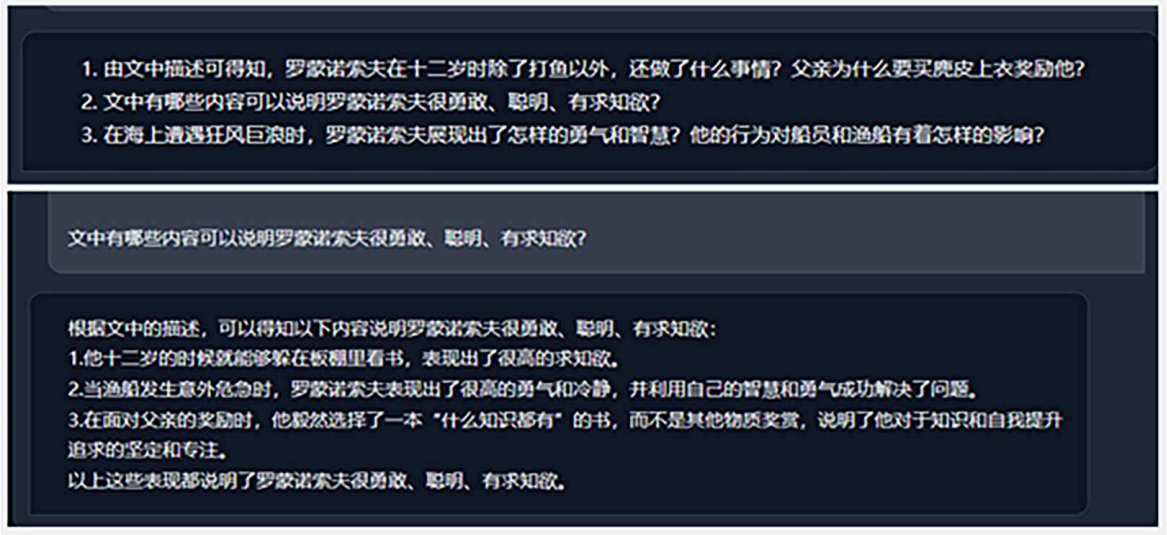

Figure 5. The generation and the answer to Retrieving Information Questions.

> "1. Which clues in the text indicate Romonosov's early and intense pursuit of knowledge? What areas of scientific research might he dedicate himself to in the future? 2. Romonosov refused to accept a deer skin coat and instead requested a book with 'all knowledge.' What outstanding qualities and strengths does this reveal about him? 3. How can we learn from and apply Romonosov's questions, curiosity, and courage in our everyday lives?"

Figure 5 presents the system-generated Retrieving Information Questions and its answer. The questions generated by the system are as follows:

> "1. Besides fishing, what else did Romonosov do at the age of twelve according to the text? Why did his father want to reward him with a deer skin coat? 2. What aspects in the text illustrate Romonosov's bravery, intelligence, and curiosity? 3. How did Romonosov demonstrate courage and wisdom when facing the fierce winds and huge waves at sea? What impact did his actions have on the crew and the fishing boat?"

For the question: "What content in the text illustrates Romonosov's bravery, intelligence, and curiosity?" the system generated the following answer:

> "According to the description in the text, we can learn from the following content that Romonosov is bravery, intelligence, and curiosity: 1. When he was twelve years old, he was able to hide in a makeshift hut to read books, showing an intense pursuit of knowledge. 2. When the fishing boat accidentally critical, Romonosov showed a high degree of courage and calm, and use his wisdom and courage to successfully solve the problem. 3. In the face of his father's reward, he resolutely chose a book of 'all knowledge', rather than other material rewards, which shows his firm and focused pursuit of knowledge and self-improvement. All of the above shows that Romonosov is bravery, intelligence, and curiosity."

5.3. Evaluation

5.3.1. Application effectiveness evaluation

We evaluated the application effectiveness of the AQG system in generating questions by analyzing the results of the questionnaire survey. The researchers conducted manual screening of the collected questionnaires, excluding those with missing items or providing uniform responses to all questions, resulting in 53 valid questionnaires. The response rate was 96.3%. Statistical analysis of the questionnaires was performed using SPSS software. The study utilized Principal Component Analysis to assess the reliability and validity of the questionnaire. The analysis results showed that Cronbach's α values were all greater than 0.7, indicating excellent internal consistency. We conducted a descriptive statistical analysis of the collected questionnaires. Firstly, we analyzed the results of perceived feedback in the perceived value dimension. We investigated whether questions generated based on prompt

patterns were more professional and accurate than the non-prompt method. The specific results are shown in Figure 6. Secondly, after conducting descriptive statistical analysis on the question categories in perceived feedback, we found that participants generally agreed with the questions generated within their respective categories. On average, 35% of participants strongly agreed with the categorization of various question types. Specifically, 70% of participants strongly agreed with the categorization of Information Retrieval Questions. However, 63% of participants disagreed with the consistency between the generated questions and the categories for Application-type Questions. Thirdly, through the analysis of perceived value, we found that, on average, over half of the participants indicated that they preferred the questions generated with prompt patterns and felt they could learn more from them. However, 75% of participants disagreed with the statement that the questions generated by our method took less time to generate. Regarding technical acceptance and the practicality of the AQG system, only an average of 48% of the selected participants agreed or strongly agreed. Particularly, in the aspect of whether the AQG system improved work efficiency, 37% of participants chose neutral or disagreed. However, it is worth noting that the satisfaction rate regarding the usability of the AQG system reached 90%. Regarding the intention to reuse, the average proportion of participants who agreed or strongly agreed on both autonomous demand and willingness to reuse was 69%.

5.3.2. Expert evaluation

We conducted statistical analysis on the judgment results of the question assessment reports of two teachers to verify the quality of the AQG system generated questions. Firstly, we conducted a Kappa test to assess the consistency of evaluation results between the two teachers. The specific results of the consistency test for each of the six dimensions under examination are presented in Table 5. The table reveals a moderate level of agreement between them. As there were inconsistencies in how each teacher assessed the questions, we also calculated the Intraclass Correlation Coefficient (ICC) for each dimension to measure the average agreement within groups. In this context, the obtained ICC ranged from 0.667 to 0.945, indicating robust results. Furthermore, we observed the annotations provided by the teachers. They identified key reasons for deeming questions as disagreeing, including unclear question directions, overly simplistic questions that fail to encourage critical thinking,

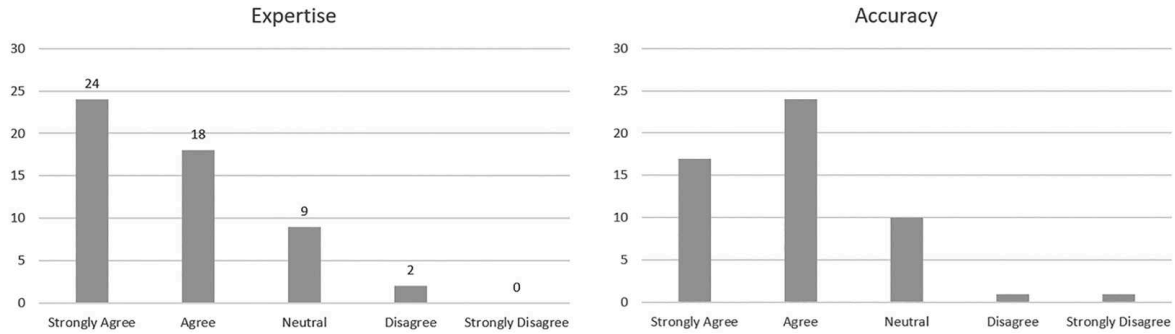

Figure 6. Experimental results of perceived value dimension.

Table 5. The consistency analysis of the assessment report.

Dimensions	Kappa	ICC
Answerability	0.552	0.945
Relevance	0.430	0.878
Correctness	0.458	0.826
Soundness	0.494	0.667
Fluency	0.567	0.667
Quality	0.519	0.874

questions that do not aid in student comprehension of perspectives, knowledge, phenomena, and values, or questions that do not reinforce previously learned concepts. Additionally, they highlighted those questions that were disconnected from the reading texts.

Subsequently, we statistically evaluated the assessment results of the assessment reports based on the original generation sources of the questions. Specifically, we compared questions directly generated by the ChatGPT with those generated by our designed prompt-based AQG system across the six dimensions developed for question evaluation. We conducted a count of labels in the assessment results to test for statistically significant differences, followed by normalization. The experimental results are illustrated in Figure 7. On the left side of the figure are the evaluation results of questions generated by ChatGPT without prompts, while on the right side are the evaluation results of questions generated by our designed AQG system. The evaluation was conducted across six dimensions, with the results for each dimension displayed according to the Likert 5-point scale, as illustrated in the legend of Figure 7.

From Figure 7, it is evident that questions generated by both the Non-prompt and AQG methods exhibit comparable quality across dimensions such as Relevance, Correctness, Soundness, and Fluency. An interesting observation arises in the Relevance dimension, where, despite both approaches demonstrating similar quality, the count of Strongly Agree responses is only half of that found in the Correctness, Soundness, and Fluency dimensions. This suggests that there is room for further improvement in enhancing the relevance of questions generated by both methods.

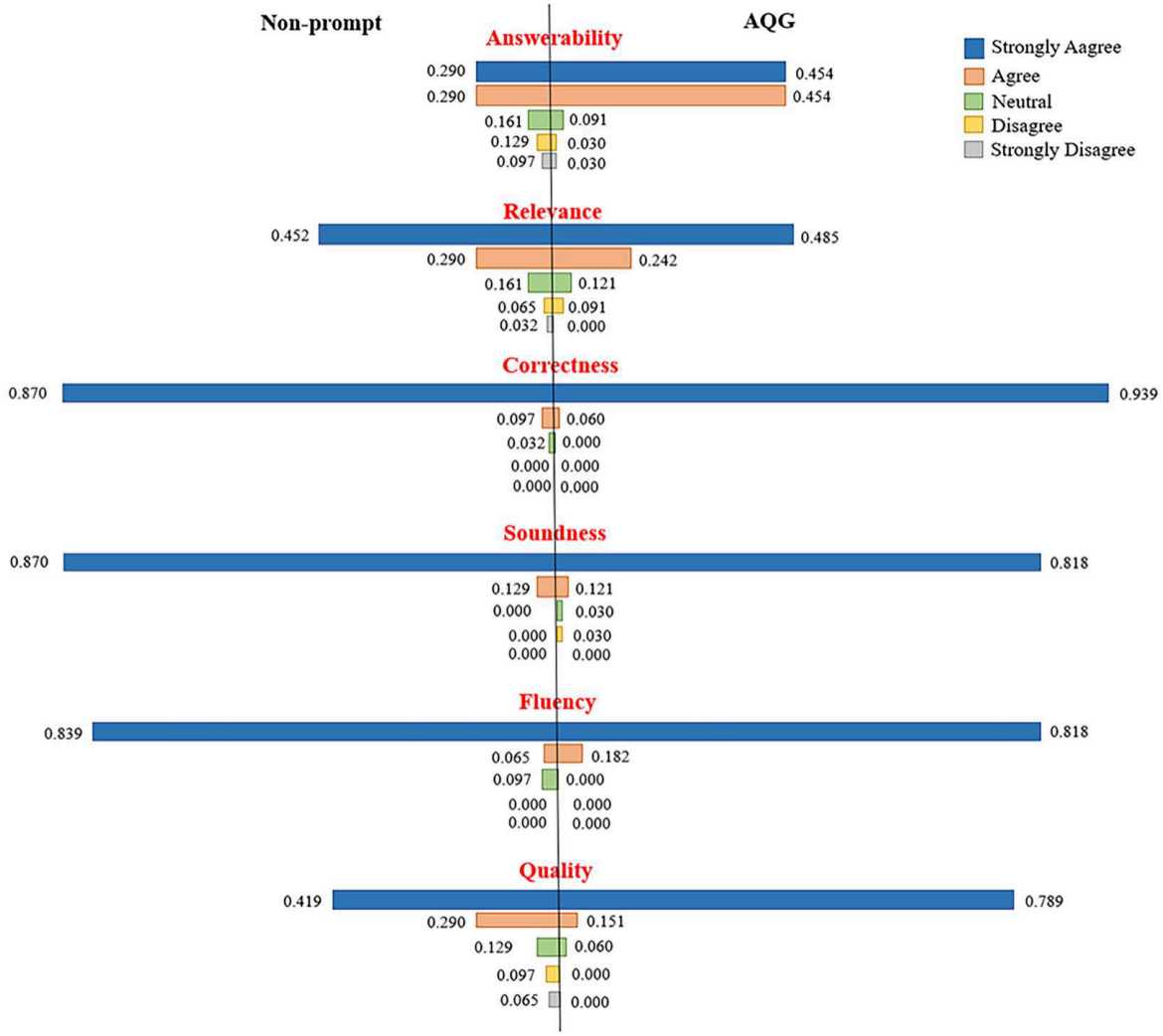

Figure 7. Extent evaluation results of questions generated by Non prompt and AQG systems.

Regarding Answerability and Quality dimensions, questions generated by the AQG method demonstrate a more substantial improvement (0.454 vs. 0.290, 0.789 vs. 0.419). Consequently, the system design, incorporating a prompt-based approach along with a collective knowledge base, has the potential to address issues associated with inaccurate and unprofessional outputs commonly observed in LLMs. The findings highlight the effectiveness of this approach in enhancing the overall quality of generated questions, particularly in terms of answerability and overall extent.

Similarly, we conducted a comparative analysis between questions generated by the AQG system utilizing prompt patterns and manually crafted questions. The detailed experimental outcomes are illustrated in Figure 8, where the evaluation results for manually generated questions are presented on the left, while those for questions generated by the AQG system are on the right. While human evaluators were expected to answer questions during the evaluation to ensure authenticity, such a setup, although not a substitute for assessing questions in a real classroom environment, provides a preliminary understanding of the relative assessment results between questions generated by the AQG system and those crafted manually.

Our observations reveal that questions generated by both humans and the AQG system demonstrate comparable quality across dimensions such as Correctness, Soundness, Fluency, and Quality. Particularly in the domains of Correctness and Fluency, there is no significant distinction between questions generated by AQG and those meticulously crafted by humans. However, AQG-generated questions exhibit a noticeable disadvantage in the Answerability and Relevance dimensions, with 3%

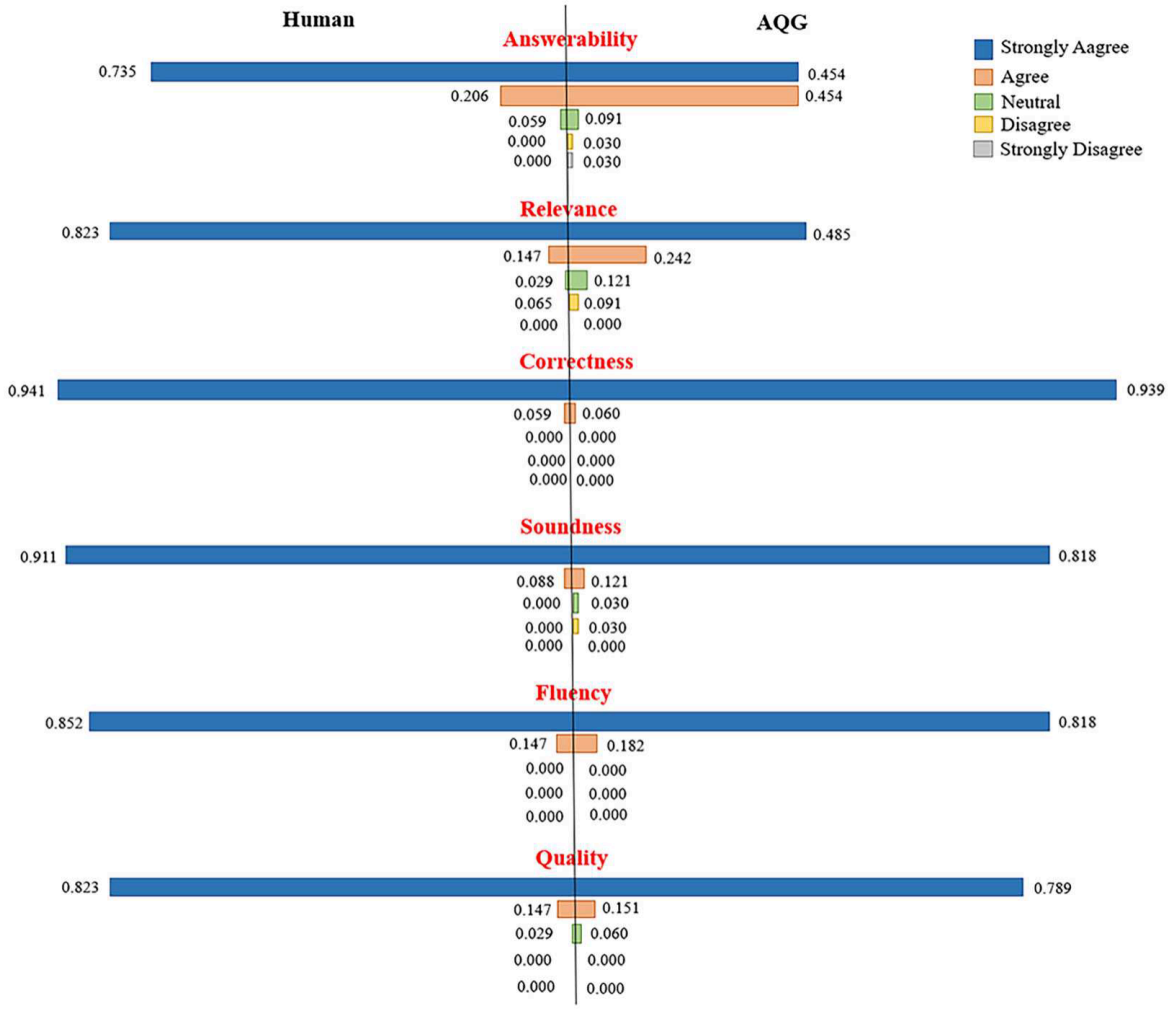

Figure 8. Extent evaluation results of questions generated by Human and AQG systems.

of the questions needing to be more complete. Despite the grammatical correctness, semantic coherence, and linguistic fluency exhibited by these questions, as well as the contextual relevance of the provided reading text, the questions posed cannot be answered based on the given reading text. This limitation may be attributed to the current state of the AQG system, which lacks advanced reasoning and logic capabilities in complex contexts, resulting in questions with limited depth. Nevertheless, when considering the overall assessment, questions generated by the AQG system and manually crafted questions exhibit a comparable extent.

6. Discussion, Implication, and Limitation

6.1. Discussion

6.1.1. RQ1. How can we design a collective practical knowledge base and integrate it with existing prompt patterns to form a new prompt pattern for LLMs-based AQG system that can generate high-quality questions?

Effective questions contribute to the development of students' critical thinking and creative thinking abilities, fostering their capacity for independent thought throughout the learning process. However, the task of crafting effective questions is not straightforward. Teachers need to thoroughly understand the material and select appropriate types and forms of questions to accurately assess students' comprehension. This meticulous process demands considerable time and effort. Furthermore, manual question creation is susceptible to limited variety and quality constraints, potentially introducing biases in question selection process based on individual teachers' backgrounds and expertise levels. This underscores the imperative for integrating AQG into education. Although there has been an automatic generation of questions based on natural language processing technology, this approach relies on technology-driven forces and couldn't be effectively adapted to educators.

To address these challenges, this study adopts a prompt engineering approach to optimize AQG of LLMs. The core of this method involves embedding a collective knowledge base into structured prompt patterns, rather than relying on traditional, simplistic prompt patterns. We meticulously analyzed the characteristics of questions designed by teachers in exemplary instructional designs. Guided by the question classification framework, we constructed a collective knowledge base encoding definitions and characteristics of each category of questions. We then evaluated its effectiveness through expert reviews. Subsequently, it was integrated as a component of the structured prompt pattern and fed into the LLMs. The collective knowledge base aligns with the learning process of LLMs, derived from collective intelligence, and covers a wide range of question types and difficulty levels in a balanced manner. It helps LLMs understand the questioning style of each question type, thereby enabling LLMs to generate high-quality questions. The results of the questionnaire survey and the evaluation report strongly demonstrate the effectiveness of the structured prompt pattern based on the collective knowledge base we designed for question generation. The AQG system we have designed is crucial for teachers to integrate AQG technology more effectively into practical applications, as they can effectively contribute their experiential knowledge about questioning. Additionally, the design process of the collective knowledge base can guide educational practitioners in designing structured prompt templates for similar types of research, enabling more effective application of LLMs to generate questions that better meet teaching needs.

6.1.2. RQ2. To what extent can a prompt pattern with generated question yield high validity and reliability?

Both the expert validation results and the questionnaire survey results demonstrate the validity and reliability of the new prompt-based AQG approach using ChatGPT. From the experimental results in Figure 8, we observe that the AQG system's ability to generate questions approaches human performance. In our analysis, the most significant challenges of the questions generated by the AQG

system are related to answerability and relevance. A typical example is in scenarios like the reading material "Tadpole Looks for Mom," where LLMs might pose unanswerable questions such as "Where did the mother place her son?" This is primarily attributed to ChatGPT's tendency to generate "hallucination" effects or seemingly plausible but incorrect answers, especially for complex tasks (Bang et al., 2023). OpenAI's own findings corroborated this observation (OpenAI, 2023), documenting the performance of ChatGPT across various tasks and noting its struggles to complex tasks that are also challenging for humans. However, when we input more explicit instructions from the knowledge base, LLMs generate higher quality questions, as evidenced by the results in Figure 7. We also found that LLMs had a higher accuracy in generating Information Retrieval Question, which may be attributed to the more consistent questioning style with the input text. This approach may be related to the training architecture of the LLMs. In the training of LLMs, when using contextual information without revealing label information, some randomly selected tokens in the sequence are masked with "[Mask]" to predict the relevant content (Wu et al., 2023). Therefore, the generation process of Information Retrieval Questions is associated with the training process of LLMs, leading to more significant experimental results. Hence, when using LLMs to generate questions in the future, it would be beneficial to guide LLMs from the technical architecture perspective to elicit questions effectively.

Indeed, while facing technical challenges, the prompt-based AQG method in this study shows significant promise in terms of validity and reliability. Despite some difficulties with Answerability and Relevance, the overall response indicates that the questions generated by AQG system can effectively meet most needs. This suggests that with further refinement and development, AQG system based on prompt patterns have the potential to become valuable tools in education.

6.2. Implication

This study offers a novel perspective on the ongoing discussion surrounding the integration of AQG as an AI tool in educational contexts. We explore the application of a prompt pattern for generative AI in education. The core idea involves synthesizing and consolidating the collective insights from educators' practical teaching experiences, transforming this knowledge into a well-organized foundation. LLMs can absorb this knowledge by inputting it into a structured prompt pattern, thereby aiding in the creation of more refined and precise content. Although we devised the AQG system as an illustration, the prompt pattern is adaptable to diverse text types. Teachers can infuse more human intelligence into LLMs following this design process, contributing to superior prompt engineering and tailored integration of LLMs within education. Currently, LLMs are not capable of, nor are they designed for, modeling the mind of the user (Wang, Wang, et al., 2023). They only train on a very large corpus of digital information, so human intelligence is needed to guide LLMs to produce higher quality questions. The outcomes of this study underscore the potential of synergistic collaborations between humans and AI in advancing educational studies. From the perspective of education, being a prompt engineer with domain specialization is one way to reimagine a supplementary and reconfigured role for the teacher. This may involve, at first, being the specialized educational prompt engineer in the supervised use of LLMs for learning purposes, but later, it involves imparting such prompt engineering skills to students, empowering them to work more autonomously and productively with the system for their own learning. It is consistent with the possibility of discussing the underlying concepts of how distributed cognition takes place when teachers engage with ChatGPT during human-AI collaborations (Kim et al., 2022; Ouyang & Jiao, 2021; Price & Flach, 2017).

6.3. Limitation and future research

Although empirical research results indicate the value of the prompt pattern proposed in our study for AQG, we acknowledge several limitations in our study that could potentially impact project implementation and research effectiveness. Firstly, our study only validated the prompt pattern

within the framework of qualitative analysis tasks, specifically AQG. This narrow scope may restrict its applicability in broader contexts. Secondly, the relatively small sample size utilized in our experiments could potentially compromise the assessment of the prompt pattern's overall effectiveness. Thirdly, the brevity of the experiments – limited to only two rounds of testing – could hinder a thorough validation of the real-time prompt pattern's efficacy. The extension would allow participants more time for assessment. Lastly, we did not individually scrutinize the impact of each element in the prompt pattern – Role, Output Indicator, Type, Definition, Characteristic, and Example Code – on the results.

In addressing the current limitations, our future research will focus on several key areas. Firstly, we will substantially increase the sample size to validate the applicability and effectiveness of the prompt pattern across various qualitative analysis tasks that require accessing collective wisdom of teacher. Secondly, as generative AI continues to be integrated into education, we will continuously refine and redesign the structure of the prompt pattern to keep pace with technological progress. This iterative process aims to make the pattern increasingly user-friendly and convenient for educators to implement in their teaching practices. Furthermore, we intend to validate each element in the prompt pattern against the results, incorporating practical knowledge bases to enhance the system's capabilities. This enhancement will easily and effectively facilitate educational practitioners in inputting their knowledge into the system. The system will autonomously encode this information, forming a powerful and practical knowledge base to guide the output of LLMs. This expansion will enable the system to address a wider range of educational scenarios, thereby enhancing its practicability and relevance.

7. Conclusion

This study realizes an AQG system based on prompt pattern that combines the collective knowledge base of the teacher community. Empirical application and expert evaluation have corroborated the efficacy of this methodology in generating high-quality questions. This contribution lies in offering an effective prompt pattern framework for qualitative analysis tasks requiring the collective intelligence of teachers. Although our research focuses on question generation, the pattern's versatility extends to various types of text data processing, including the categorizing common student errors, the comprehension of learners' learning patterns, and the formulation of differentiated instructional assignments. As a result, the study articulates a practical framework for applying generative AI in education, facilitating the integration of this technology into educational practices.

Competing Interests

My co-authors and I don't have any relevant financial or non-financial competing interests.

Funding

The research is supported by the research fund from the National Natural Science Foundation of China [grant numbers 62067008, 62267008].

Notes on contributors

Lili Wang is a doctoral candidate and assistant lecturer, specializing in the fields of intelligent education and generative artificial intelligence. Her research endeavors focus on exploring the synergies between advanced AI technologies and educational practices, aiming to innovate and improve learning methodologies. She can be reached at 1335502737@qq.com, and her ORCID iD is 0009-0005-0686-5181.

Ruiyuan Song is a Master's degree candidate, focusing on K-12 discourse comprehension in the field of education. Her research centers on advancing the understanding and improvement of reading skills within the context of K-12

education. For communication and collaboration, Ruiyuan can be contacted at 2930667320@qq.com, and her ORCID iD is 0009-0001-3817-0350.

Weitong Guo is an Associate Professor and Master's thesis advisor, specializing in the fields of intelligent education and artificial intelligence technology. As a dedicated academic, Dr. Guo's primary research focus revolves around the integration of artificial intelligence technologies into educational practices. For communication and collaboration, Dr. Guo can be reached at guowt@nwnu.edu.cn. Her ORCID iD is 0009-0006-9501-1010.

Hongwu Yang is a Professor and Ph.D. thesis advisor, specializing in the fields of intelligent education and artificial intelligence technology. With extensive experience, Dr. Yang's primary research focus is on advancing the integration of artificial intelligence technologies into educational practices. For inquiries and collaboration, Dr. Yang can be reached at yanghw@nwnu.edu.cn. His ORCID iD is 0000-0002-8939-3386.

ORCID

Lili Wang http://orcid.org/0009-0005-0686-5181
Ruiyuan Song http://orcid.org/0009-0001-3817-0350
Weitong Guo http://orcid.org/0009-0006-9501-1010
Hongwu Yang http://orcid.org/0000-0002-8939-3386

References

Adeshola, I., & Adepoju, A. P. (2023). The opportunities and challenges of ChatGPT in education. *Interactive Learning Environments*, 1–14. https://doi.org/10.1080/10494820.2023.2253858

Aithal, P., & Aithal, S. (2023). The changing role of higher education in the era of AI-based GPTs. *International Journal of Case Studies in Business, IT, and Education (IJCSBE)*, *7*(2), 183–197. http://doi.org/10.2139/ssrn.4609337

Al-Emran, M. (2024). Unleashing the role of ChatGPT in metaverse learning environments: Opportunities, challenges, and future research agendas. *Interactive Learning Environments*, 1–10. https://doi.org/10.1080/10494820.2024.2324326

Almeida, P. A. (2012). Can I ask a question? The importance of classroom questioning. *Procedia - Social and Behavioral Sciences*, *31*, 634–638. https://doi.org/10.1016/j.sbspro.2011.12.116

Alsubait, T., Parsia, B., & Sattler, U. (2012). Automatic generation of analogy questions for student assessment: An ontology-based approach. *Research in Learning Technology*, *20*(sup1). https://doi.org/10.3402/rlt.v20i0.19198

Alsubait, T., Parsia, B., & Sattler, U. (2016). Ontology-based multiple choice question generation. *KI - Künstliche Intelligenz*, *30*(2), 183–188. https://doi.org/10.1007/s13218-015-0405-9

Ausat, A. M. A., Massang, B., Efendi, M., Nofirman, N., & Riady, Y. (2023). Can Chat GPT replace the role of the teacher in the classroom: A fundamental analysis. *Journal on Education*, *5*(4), 16100–16106. https://doi.org/10.31004/joe.v5i4.2745

Bahroun, Z., Anane, C., Ahmed, V., & Zacca, A. (2023). Transforming education: A comprehensive review of generative artificial intelligence in educational settings through bibliometric and content analysis. *Sustainability*, *15*(17), 12983. https://doi.org/10.3390/su151712983

Baidoo-Anu, D., & Ansah, L. O. J. (2023). Education in the era of generative artificial intelligence (AI): understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI*, *7*(1), 52–62. https://doi.org/10.61969/jai.1337500

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., & Do, Q. V. (2023). A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *ArXiv./abs/2302.04023*. https://doi.org/10.48550/arXiv.2302.04023.

Bouziane, K., & Bouziane, A. (2024). Exploring the Role of AI in Essay Evaluation: A Comparative Analysis of ChatGPT and Human Corrections. *Researchsquare*. https://doi.org/10.21203/rs.3.rs-4139088/v1

Brookhart, S. M. (2014). *How to design questions and tasks to assess student thinking*. ASCD.

Chaudhry, I. S., Sarwary, S. A. M., El Refae, G. A., & Chabchoub, H. (2023). Time to revisit existing student's performance evaluation approach in higher education sector in a New Era of ChatGPT — A case study. *Cogent Education*, *10*(1), 2210461. https://doi.org/10.1080/2331186X.2023.2210461

Chen, Y., et al. (2019). Reinforcement learning based graph-to-sequence model for natural question generation. *arXiv preprint arXiv:1908.04942*.

Chen, X., Xie, H., Zou, D., & Hwang, G.-J. (2020). Application and theory gaps during the rise of artificial intelligence in education. *Computers and Education: Artificial Intelligence*, *1*, 100002. https://doi.org/10.1016/j.caeai.2020.100002

Christenbury, L., & Kelly, P. P. (1983). Questioning: A path to critical thinking.

Ding, N., Hu, S., Zhao, W., Chen, Y., Liu, Z., Zheng, H. T., & Sun, M. (2021). Openprompt: An open-source framework for prompt-learning. *arXiv preprint arXiv:.01998*. https://doi.org/10.48550/arXiv.2111.01998.

Dong, L., Yang, N., Wang, W., Wei, F., Liu, X., Wang, Y., Gao, J., Zhou, M., & Hon, H. W. (2019). Unified language model pre-training for natural language understanding and generation. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 32: Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS'19)* (Vol. 32, pp. 13042–13054).

Douglas, M., Wilson, J., & Ennis, S. (2012). Multiple-choice question tests: A convenient, flexible and effective learning tool? A case study. *Innovations in Education and Teaching International*, *49*(2), 111–121. https://doi.org/10.1080/14703297.2012.677596

Eager, B., & Brunton, R. (2023). Prompting higher education towards AI-augmented teaching and learning practice. *Journal of University Teaching and Learning Practice*, *20*(5), 02. https://doi.org/10.53761/1.20.5.02

Fan, Z., Wei, Z., Wang, S., Liu, Y., & Huang, X. J. (2018). A reinforcement learning framework for natural question generation using bi-discriminators. In E. M. Bender, L. Derczynski, & P. Isabelle (Eds.), *Proceedings of the 27th international conference on computational linguistics* (pp. 1763–1774). Association for Computational Linguistics.

Farrokhnia, M., Banihashem, S. K., Noroozi, O., & Wals, A. (2023). A SWOT analysis of ChatGPT: Implications for educational practice and research. *Innovations in Education and Teaching International*, 1–15.

Fink, L. D. (2013). *Creating significant learning experiences: An integrated approach to designing college courses*. John Wiley & Sons.

Finnie-Ansley, J., Denny, P., Becker, B. A., Luxton-Reilly, A., & Prather, J. (2022). The robots are coming: Exploring the implications of openai codex on introductory programming. In Institute of Electrical and Electronics Engineers (Eds.), *Proceedings of the 24th Australasian computing education conference* (pp. 10–19). Association for Computing Machinery.

Gaebel, M., Kupriyanova, V., Morais, R., & Colucci, E. (2014). E-learning in European higher education institutions: Results of a mapping survey conducted in October-December 2013. In European University Association (Eds.), *European University Association* (Vol. 92). EUA Publications.

Gansser, O. A., & Reich, C. S. (2021). A new acceptance model for artificial intelligence with extensions to UTAUT2: An empirical study in three segments of application. *Technology in Society*, *65*, 101535. https://doi.org/10.1016/j.techsoc.2021.101535

Giray, L. (2023). Prompt engineering with ChatGPT: A guide for academic writers. *Annals of Biomedical Engineering*, *51*(12), 2629–2633. https://doi.org/10.1007/s10439-023-03272-4

Goldbach, I., & Hamza-Lup, F. (2017). Survey on e-learning implementation in eastern-Europe spotlight on Romania. In Luca Andrea Ludovico, Università degli Studi di Milano, Italy Ahmed Mohamed Fahmy Yousef, Fayoum University, Egypt (Eds.), *ELML 2017: The ninth international conference on mobile, hybrid, and on-line learning. Hybrid, and on-line learning* (pp. 5–12). IARIA.

Gorgun, G., & Bulut, O. (2024). Exploring quality criteria and evaluation methods in automated question generation: A comprehensive survey. *Education and Information Technologies*, 1–32. https://doi.org/10.1007/s10639-024-12771-3

Gu, J., Han, Z., Chen, S., Beirami, A., He, B., Zhang, G., Liao, R., Qin, Y., Tresp, V., & Torr, P. (2023). A systematic survey of prompt engineering on vision-language foundation models. *arXiv preprint arXiv:2307.12980*. https://doi.org/10.48550/arXiv.2307.12980.

Gupta, M., Akiri, C., Aryal, K., Parker, E., & Praharaj, L. (2023). From ChatGPT to ThreatGPT: Impact of generative AI in cybersecurity and privacy. *IEEE Access*, *11*, 80218–80245. https://doi.org/10.1109/ACCESS.2023.3300381

Halliday, M. A. (1998). Things and relations. In J. R. Martin & R. Veel (Eds.), *Reading science: Critical and functional perspectives on discourses of science* (pp. 185–235). Routledge.

Haque, M. U., Dharmadasa, I., Sworna, Z. T., Rajapakse, R. N., & Ahmad, H. (2022). "I think this is the most disruptive technology": Exploring sentiments of ChatGPT early adopters using twitter data. *arXiv preprint arXiv:2212.05856*.

Jahn, K., Kordyaka, B., Machulska, A., Eiler, T. J., Gruenewald, A., Klucken, T., Brueck, R., Gethmann, C. F., & Niehaves, B. (2021). Individualized gamification elements: The impact of avatar and feedback design on reuse intention. *Computers in Human Behavior*, *119*, 106702. https://doi.org/10.1016/j.chb.2021.106702

Jeon, J., & Lee, S. (2023). Large language models in education: A focus on the complementary relationship between human teachers and ChatGPT. *Education and Information Technologies*, *28*(12), 15873–15892. https://doi.org/10.1007/s10639-023-11834-1

Kaddour, J., Harris, J., Mozes, M., Bradley, H., Raileanu, R., & McHardy, R. (2023). Challenges and applications of large language models. *arXiv preprint arXiv:.10169*. https://doi.org/10.48550/arXiv.2307.10169.

Khalil, M. E. E. (2023). Will ChatGPT get you caught? Rethinking of plagiarism detection. In P. Zaphiris & A. Ioannou (Eds.), *Learning and collaboration technologies. HCII 2023. Lecture notes in computer science* (Vol. 14040). Springer.

Killen, R., & O'Toole, M. (2023). *Effective teaching strategies 8e*. Cengage AU.

Kim, J., Lee, H., & Cho, Y. H. (2022). Learning design to support student-AI collaboration: Perspectives of leading teachers for AI in education. *Education and Information Technologies*, *27*(5), 6069–6104. https://doi.org/10.1007/s10639-021-10831-6

Kurdi, G., Leo, J., Parsia, B., Sattler, U., & Al-Emari, S. (2020). A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, *30*(1), 121–204. https://doi.org/10.1007/s40593-019-00186-y

Lee, U., Jung, H., Jeon, Y., Sohn, Y., Hwang, W., Moon, J., & Kim, H. (2023). Few-shot is enough: Exploring ChatGPT prompt engineering method for automatic question generation in English education. *Education and Information Technologies*, *29*(9), 11483–11515. https://doi.org/10.1007/s10639-023-12249-8

Liao, Y.-H., & Koh, J.-L. (2020). Question generation through transfer learning. In H. Fujita, P. Fournier-Viger, M. Ali, & J. Sasaki (Eds.), *International conference on industrial, engineering and other applications of applied intelligent systems* (pp. 3–17). Springer.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, *55*(9), 1–35. https://doi.org/10.1145/3560815

Liu, M., Zhang, J., Nyagoga, L. M., & Liu, L. (2023). Student-AI question Co-creation for enhancing Reading comprehension. *IEEE Transactions on Learning Technologies*, *17*, 815–826. https://doi.org/10.1109/TLT.2023.3333439

Lu, Y., Bartolo, M., Moore, A., Riedel, S., & Stenetorp, P. (2021). Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:.08786*. https://doi.org/10.48550/arXiv.2104.08786.

Luo, L., Ao, X., Song, Y., Li, J., Yang, X., He, Q., & Yu, D. (2019). Unsupervised neural aspect extraction with sememes. In Artificial Intelligence Journal Division (Eds.), *IJCAI* (pp. 5123–5129). International Joint Conferences on Artificial Intelligence Organization.

Markel, J. M., Opferman, S. G., Landay, J. A., & Piech, C. (2023). GPTeach: Interactive TA training with GPT-based students. In Institute of Electrical and Electronics Engineers (Eds.), *Proceedings of the tenth ACM conference on learning@ scale* (pp. 226–236). Association for Computing Machinery.

Montenegro, C. S., Engle, V. G., Acuba, M. G. J., & Ferrenal, A. M. A. (2012). Automated question generator for Tagalog informational texts using case markers. In Institute of Electrical and Electronics Engineers (Eds.), *TENCON 2012 IEEE region 10 conference* (pp. 1–5). IEEE.

Mulla, N., & Gharpure, P. (2023). Automatic question generation: A review of methodologies, datasets, evaluation metrics, and applications. *Progress in Artificial Intelligence*, *12*(1), 1–32. https://doi.org/10.1007/s13748-023-00295-9

Mungoli, N. (2023). Exploring the synergy of prompt engineering and reinforcement learning for enhanced control and responsiveness in chat GPT. *Journal of Electrical Electronics Engineering*, *2*(3), 201–205. https://doi.org/10.33140/JEEE

Murungu, R. (2024). Reimagining education in Africa: The transformative potential of prompt engineering. *OIDA International Journal of Sustainable Development*, *17*((01|1)), 47–62.

OpenAI. (2023). GPT-4 technical report. *arXiv:2303.08774*. https://doi.org/10.48550/arXiv.2303.08774.

Ouyang, F., & Jiao, P. (2021). Artificial intelligence in education: The three paradigms. *Computers and Education: Artificial Intelligence*, *2*, 100020. https://doi.org/10.1016/j.caeai.2021.100020

Pinto, A. S., Abreu, A., Costa, E., & Paiva, J. (2023). How machine learning (ML) is transforming higher education: A systematic literature review. *Journal of Information Systems Engineering and Management*, *8*(2), 21168, https://doi.org/10.55267/iadt.07.13227

Price, S., & Flach, P. A. (2017). Computational support for academic peer review: A perspective from artificial intelligence. *Communications of the ACM*, *60*(3), 70–79. https://doi.org/10.1145/2979672

Qayyum, A., & Zawacki-Richter, O. (2018). *Open and distance education in Australia, Europe and the Americas: National perspectives in a digital age*. Springer Nature.

Rahman, M. M., & Watanobe, Y. (2023). ChatGPT for education and research: Opportunities, threats, and strategies. *Applied Sciences*, *13*(9), 5783. https://doi.org/10.3390/app13095783

Rasul, T., Nair, S., Kalendra, D., Robin, M., de Oliveira Santini, F., Ladeira, W. J., Sun, M., Day, I., Rather, R. A., & Heathcote, L. (2023). The role of ChatGPT in higher education: Benefits, challenges, and future research directions. *Journal of Applied Learning & Teaching*, *6*(1), 1–16. https://doi.org/10.37074/jalt.2023.6.1.29

Richey, R. C., & Klein, J. D. (2005). Developmental research methods: Creating knowledge from instructional design and development practice. *Journal of Computing in Higher Education*, *16*(2), 23–38. https://doi.org/10.1007/BF02961473

Robin, B. R. (2015). The effective uses of digital storytelling as a teaching and learning tool. In J. Flood, S. B. Heath, & D. Lapp (Eds.), *Handbook of research on teaching literacy through the communicative and visual arts, volume II* (pp. 457–468). Routledge.

Rospigliosi, P. A. (2023). *Artificial intelligence in teaching and learning: what questions should we ask of ChatGPT?* (Vol. 31, pp. 1–3). Taylor & Francis.

Roumeliotis, K. I., & Tselikas, N. D. (2023). Chatgpt and open-AI models: A preliminary review. *Future Internet*, *15*(6), 192. https://doi.org/10.3390/fi15060192

Sarkhel, R., Huang, B., Lockard, C., & Shiralkar, P. (2022). Label-efficient self-training for attribute extraction from semi-structured web documents. *arXiv preprint arXiv:.13086*. https://doi.org/10.48550/arXiv.2208.13086

Savelka, J., Agarwal, A., Bogart, C., & Sakr, M. (2023). Large language models (GPT) struggle to answer multiple-choice questions about code. *arXiv preprint arXiv:.08033*. https://doi.org/10.48550/arXiv.2303.08033

Scott, J. L., Barnes, D., Britton, J., Rosen, H., & E., L. A. T. (1973). Language, the learner and the school. *The English Journal*, *62*(6), 934. https://doi.org/10.2307/813888.

Seels, B. B., & Richey, R. C. (2012). *Instructional technology: The definition and domains of the field*. IAP.

Stojanov, A. (2023). Learning with ChatGPT 3.5 as a more knowledgeable other: An autoethnographic study. *International Journal of Educational Technology in Higher Education*, *20*(1), 35. https://doi.org/10.1186/s41239-023-00404-7

Tofade, T., Elsner, J., & Haines, S. T. (2013). Best practice strategies for effective use of questions as a teaching tool. *American Journal of Pharmaceutical Education*, *77*(7), 155. https://doi.org/10.5688/ajpe777155

Urhan, S., Gençaslan, O., & Dost, Ş. (2024). An argumentation experience regarding concepts of calculus with ChatGPT. *Interactive Learning Environments*, 1–26. https://doi.org/10.1080/10494820.2024.2308093

Vidgen, B., & Derczynski, L. (2020). Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLoS One*, *15*(12), e0243300. https://doi.org/10.1371/journal.pone.0243300

Wang, T., Yuan, X., & Trischler, A. (2017). A joint model for question answering and question generation. *arXiv preprint arXiv:1706.01450*

Wang, W., Feng, S., Wang, D., & Zhang, Y. (2019). Answer-guided and semantic coherent question generation in open-domain conversation. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 5066–5076). Association for Computational Linguistics.

Wang, H. C., Chiang, Y.-H., & Chen, I.-F. (2024). A method for generating course test questions based on natural language processing and deep learning. *Education and Information Technologies*, *29*(7), 8843–8865. https://doi.org/10.1007/s10639-023-12159-9

Wang, J., Shi, E., Yu, S., Wu, Z., Ma, C., Dai, H., Yang, Q., Kang, Y., Wu, J., Hu, H., Yue, C., Zhang, H., Liu, Y., Pan, Y., Liu, Z., Sun, L., Li, X., Ge, B., Jiang, X., … Zhang, S. (2023). Prompt engineering for healthcare: Methodologies and applications. *arXiv preprint arXiv:2304.14670*. https://doi.org/10.48550/arXiv.2304.14670.

Wang, M., Wang, M., Xu, X., Yang, L., Cai, D., & Yin, M. (2023). Unleashing ChatGPT's power: A case study on optimizing information retrieval in flipped classrooms via prompt engineering. *IEEE Transactions on Learning Technologies*, *17*, 629–641. https://doi.org/10.1109/TLT.2023.3324714

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, *35*, 24824–24837.

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*. https://doi.org/10.48550/arXiv.2302.11382.

Williams, M. K. (2017). John Dewey in the 21st century. *Journal of Inquiry and Action in Education*, *9*(1), 7. https://digitalcommons.buffalostate.edu/jiae/vol9/iss1/7.

Wu, T., He, S., Liu, J., Sun, S., Liu, K., Han, Q.-L., & Tang, Y. (2023). A brief overview of ChatGPT: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica*, *10*(5), 1122–1136. https://doi.org/10.1109/JAS.2023.123618

Zhai, X., Gu, J., Liu, H., Liang, J. C., & Tsai, C. C. (2017). An experiential learning perspective on students' satisfaction model in a flipped classroom context. *Journal of Educational Technology & Society*, *20*(1), 198–210.

Zhang, Z., Gao, J., Dhaliwal, R. S., & Li, T. J.-J. (2023). Visar: A human-AI argumentative writing assistant with visual programming and rapid draft prototyping. In Institute of Electrical and Electronics Engineers (Eds.), *Proceedings of the 36th annual ACM symposium on user interface software and technology* (pp. 1–30). Association for Computing Machinery.

Zhou, Y., Muresanu, A. I., Han, Z., Paster, K., Pitis, S., Chan, H., & Ba, J. (2022). Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910*. https://doi.org/10.48550/arXiv.2211.01910.