Education and Lifelong Development Research

Vol. 1, No. 2, p. 72-82, 2024

Original article

Heuristic question sequence generation based on retrieval augmentation

Zhihao Yang[1], Zhengzhou Zhu[1]*

[1]*School of Software & Microelectronics, Peking University, Beijing 100871, P. R. China*

Keywords:
Heuristic education
question sequence generation
personalized knowledge path
large language model
retrieval-augmented generation

Cited as:
Yang, Z., Zhu, Z. (2024). Heuristic question sequence generation based on retrieval augmentation. Education and Lifelong Development Research, 1(2): 72-82.
https://doi.org/10.46690/elder.2024.02.03

Abstract:
Traditional education and many Intelligent Tutoring Systems (ITSs) still focus on one-way indoctrination and lack the cultivation of students' independent learning abilities. This article innovatively applies artificial intelligence to Socratic Method, guiding students to solve problems independently through a series of questions rather than direct answers. The main contributions include: (1) A personalized knowledge path planning algorithm is designed, using Q-matrix and student test records to update students' knowledge mastery. Combined with the graph database, Dijkstra algorithm is used to construct knowledge path. (2) We utilized the retrieved knowledge path and modified Least-to-Most Prompting to guide GLM-4 to generate an orderly and controllable question sequence. We also design an interactive algorithm to help students think about the answers by interacting with them. (3) A heuristic question sequence generation system is implemented with the objective of promoting students' self-learning through chapter testing and question answering. Experiment and user study show that the efforts made by this paper in retrieval augmented generation have a positive effect on improving the impact of question sequence generation.

1. Introduction

In the history of education, Socratic Method represents a highly influential educational philosophy that emphasizes guiding students to self-discovery and self-education through dialogue and questioning. The core of this method is that teachers do not impart knowledge directly, but stimulate students' curiosity and critical thinking by asking a series of questions, prompting them to explore the nature of the problem in depth.

Under the background of modern education, the demand for personalized learning is growing rapidly. Teachers' limited energy is difficult to support the teaching of all students in the class in accordance with their aptitude. Therefore, providing personalized feedback to different students' doubts has gradually become one of the research focuses in the field of Artificial Intelligence in Education (AIED). Compared with recommending relevant teaching content to students to passively learn, asking students back questions is more conducive to promoting students' independent thinking and inspiring students to solve new problems through knowledge recall and search (Connor-Greene, 2000). Currently, few works consider inspiring students to think independently as the development goal of ITSs. For doubts raised by students, the existing system does not have the ability to ask questions back, and it cannot allow students to solve their original doubts by answering questions step by step.

Designing such an algorithm requires ensuring that the question sequence can effectively inspire learners' thinking and understanding, which involves the assessment of the learner's knowledge level, the recognition of knowledge path, and the generation of question sequences. The capacity of Large Language Models (LLMs) to generate text with great power opens up new possibilities for question sequence creation. In particular, the Chain-of-Thought (CoT) prompting approach significantly enhances the logical reasoning abilities of LLMs, and coincides with the generation of progressive question sequences explored in this paper. However, relying

Yandy Scientific Press

*Corresponding author.
E-mail address: zhihaoyang@stu.pku.edu.cn (Z. Yang); zhuzz@pku.edu.cn (Z. Zhu).

Received April 28, 2024; revised May 31, 2024; accepted June 15, 2024; available online June 20, 2024.

solely on LLMs to generate questions inevitably leads to the hallucination, which is unacceptable in the education field. Therefore, how to generate high-quality and controllable question sequences using LLMs becomes one of the issues to be solved in this paper. Knowledge graphs can express the knowledge structure of a course according to logical relationships. By retrieving the graphs, it is possible to provide a priori knowledge for models, which could potentially address the above issue. By assessing the knowledge acquisition level of students and providing structured knowledge representation for LLMs, it is anticipated that the system will be able to generate high-quality question sequences.

Based on the above background, the issue that this paper seeks to address is: to generate personalized heuristic question sequences for different questions raised by different students (hereinafter referred to as target questions), and to inspire students to form solutions to the target problems by letting students gradually think about sub-problems, which can cultivate students' independent learning ability. Our research questions are deconstructed into the following key points:

1) **What is the basis for generating the question sequence?** The algorithm needs to start from the knowledge points that students have mastered, search for personalized knowledge paths that can reach their target questions, and generate corresponding heuristic question sequences based on this path to build a scaffolding between mastered and unmastered knowledge.
2) **How to generate personalized question sequences for different students' learning status?** Different students have different knowledge absorption conditions, and there are differences in their mastery of course content. Therefore, it is necessary to design knowledge level assessment algorithms that are accurate and support real-time updates.
3) **How to guide students to answer questions step by step?** This paper needs to study how to generate a logically progressive question sequence and improve the controllability of the question form. In addition, the sequence of questions requires students to solve them sequentially, which requires the algorithm to have the ability to conduct multiple rounds of dialogue with students.

2. Related works

Asking students questions is more helpful to improve learning results than letting students study textbooks(Connor-Greene, 2000). Currently, most intelligent human-computer interaction systems on the market are conversational and require question generation capabilities to identify user needs (Mulla & Gharpure, 2023). As for education, Scaffolding Instruction is helpful to cultivate students' metacognitive ability. This method breaks down complex problems into several simple problems, and guides students to gradually build new knowledge schema on the basis of the established knowledge schema by setting up scaffolding. However, to integrate scaffolding teaching method into human-computer interaction system, it is necessary to study the intelligent generation of question sequences that can inspire students' thinking according to students' doubts.

Question generation is related to whether the algorithm can effectively promote students' independent thinking. However, at present, relevant researches focus on generating a single question based on facts or answers, and there is a lack of exploration of the generation mechanism of multiple related questions. The powerful text generation capability of LLMs brings new possibilities for the generation of problem sequences. In recent years, researchers have initiated a new research field, prompt engineering, with the objective of investigating the capacity of LLMs to learn based on context. This endeavor aims to identify more suitable input texts that can guide the model to generate the desired output or accomplish a specific task, which is a significant reference for this paper. Therefore, this section will focus on the problem generation algorithm and prompt engineering literature review.

2.1 Question generation

In the field of intelligent education, many ITSs rely on hand-crafted rules by experts to generate feedback (St-Hilaire et al., 2022), or use template-based methods to create questions, or generate static questions (Chen et al., 2018). As the rules or templates are handcrafted, the scalability and coverage of this approach are very limited, and designing templates requires a lot of manpower and resources. In order to make the questions more relevant to the learning of different students, a recent work by Srivastava and Goodman (Srivastava & Goodman, 2021) proposed a difficulty-controlled model to generate personalized questions based on the level of students. Lately, there have also been studies on the application of pre-trained language models to language translation education. Kulshreshtha et al. (2022) proposed a method to automatically generate personalized feedback. By combining causality analysis and Transformer based on text similarity, the method can identify the correct and wrong parts or missing parts in students' answers, and ask questions in natural language to guide students to find the correct answers. Bulathwela et al. (Bulathwela, Muse, & Yilmaz, 2023) used S2ORC (Lo et al., 2020), SQuAD 1.1 (Rajpurkar et al., 2016) and SciQ (Welbl, Liu, & Gardner, 2017) to conduct multi-stage pre-training and fine-tuning of T5 model (Raffel et al., 2020) to generate questions of science courses. This study proved that training at each stage contributes to the quality of question generation of language model.

In essence, the above-mentioned question generation algorithms generate simple questions with single sentences by direct mapping. However, in real teaching scenarios, teachers and students are highly interactive. When students need to solve their own problems, teachers will constantly provide heuristic feedback for students' questions or answers. In order for ITS to be able to give similar feedback like a human teacher, we must ensure that it can interact with students like a human teacher (Alkhatlan & Kalita, 2019), that is, generate a heuristic sequence of questions based on student questions to guide students to find answers step by step. Although the generation of question sequences is a relatively

underdeveloped area in national and international research, the stepwise reasoning of LLMs and the mounting of the knowledge base is helpful for contributing to this field.

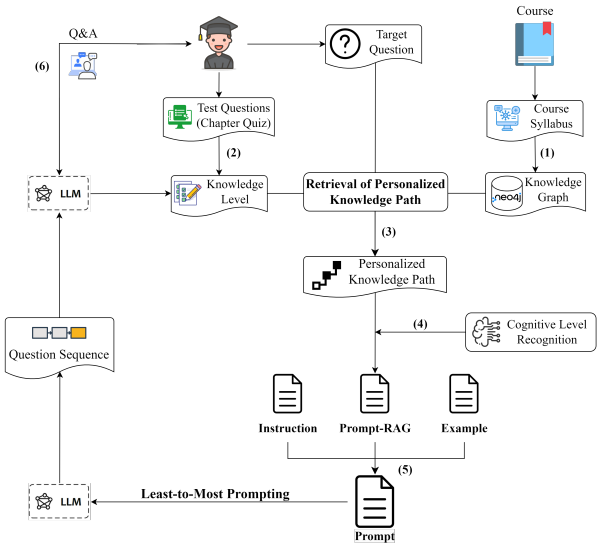

Fig. 1. Flow chart of heuristic problem sequence generation algorithm.

2.2 Prompt engineering on LLMs

With the sharp increase in the parameters of deep learning models, context-based learning capabilities such as GPT-4 and GLM-4 emerged. People can directly apply the pre-trained model to new tasks instead of fine-tuning it (Li et al., 2023; Yang et al., 2022). Therefore, Prompt Engineering has attracted the attention of scholars at home and abroad.

Many prompt schemes have been proposed in recent years, and these schemes have their own advantages and limitations in different application scenarios. Among them, the Chain-of-Thought (CoT) prompting (Chowdhery, 2023; Wei et al., 2022) has taken an important step in narrowing the gap between human intelligence and machine intelligence. Inspired by CoT, Liang et al. (2023) define the Question Generation over Knowledge Bases (KBQG) task as a reasoning problem, in which the generation of complete question is divided into the generation of a series of sub-problems, and similar logical forms are selected from the unlabeled data pool based on the vector similarity of logical forms, so as to guide the LLM question generation.

However, if the question to be solved by LLMs is more complex than the example shown in the prompt, the CoT prompting often performs poorly. Zhou et al. proposed the Least-to-Most prompting (Zhou et al., 2022), which decomposed the complex problem into a series of simpler sub-problems and then solved these sub-problems in order. The answer inference for each sub-problem depends on all the sub-problems and their answers in the preceding order. According to the experimental results reported by the authors, this method can solve more difficult problems than prompt examples in three tasks: symbolic manipulation, compositional generalization and math reasoning.

Question generation in education is a knowledge-intensive task, and the content generation needs the guidance of professional domain knowledge. Compared with fine-tuning LLMs that require a large amount of expert annotation data and expensive computing resources, retrieval-augmented generation (Gao et al., 2023) builds a retrieval system based on AI models to access external knowledge sources to construct hints. This approach respects the facts, produces traceable answers, helps alleviate the "hallucination" problem, and appeals to a large number of educational practitioners who prefer "plug and play." Wang et al. (2022) explored the impact of various factors on the generation of educational questions by prompting LLMs. These factors included the structure, source, number, and text length of examples. To assess the influence of these factors on the generation of questions in a series of generation scenarios, the researchers employed index calculation and manual inspection. The objective is to identify a set of strategies that are most likely to prompt high-quality questions.

3. Method

3.1 Overview

To solve the research questions proposed in this paper, implement the heuristic problem sequence generation algorithm, and be able to effectively interact with students, this chapter needs to complete the following work in order. The overview of the algorithm design is shown in Fig. 1.

1) The question sequence that the algorithm presents to students must adhere to a set of rules and align with the knowledge structure and logic. Therefore, this paper needs to construct a knowledge graph according to the course syllabus as a domain knowledge base to ensure the accuracy and rationality of the heuristic path.
2) It is essential that the generated heuristic questions adapt to the current learning state of students, necessitating the real-time assessment of their knowledge levels. Consequently, this study needs to develop a system to realize the chapter test function, model the implication relationship between the test questions and the knowledge points, and deduce the knowledge point set that the students have mastered by combining it with the test record.
3) When a student asks a question, the algorithm must be able to identify the starting point and the end point of the heuristic path in the course knowledge graph according to the student's knowledge level, and plan out a logical progressive knowledge path as the basis for the language model to generate the question sequence.
4) The labels and attributes of the knowledge on the path are different, so the reasonable ways of asking students questions are also different. Therefore, the algorithm needs to give different cognitive levels according to different knowledge descriptions, subsequently regulating the manner of question formulation.
5) Based on the information such as personalized knowledge path and question asking method planned by students, this paper needs a standardized process to construct

prompt context and guide the LLM to generate question sequence.

6) The algorithm must possess the capacity for multi-round dialogue, in which students are posed a question in turn within the question sequence. Therefore, students can be guided to finally solve their doubts by gradually answering sub-questions.

3.2 Construction of course knowledge graph

For the "what to retrieve" problem of retrieval augmentation generation, research conducted both domestically and internationally has evolved from simple token retrieval (Khandelwal et al., 2019) and entity retrieval (Nishikawa et al., 2022) to more complex structures, such as chunk retrieval (Ram et al., 2023) and knowledge graph retrieval (Kang et al., 2023). Sequeda et al. (Sequeda, Allemang, & Jacob, 2023) used the knowledge graph as additional information related to user requests to enrich the prompt of the LLMs and applied it to the question answering task. They found that attaching the knowledge graph could significantly improve the accuracy of the LLMs' responses. Inspired by the above studies, this paper chooses to construct a knowledge graph based on course content as an external retrieval library to provide domain knowledge as a reference for LLMs. The following content takes "Software Engineering" as an example. This paper presents the knowledge structure of the course as a knowledge graph, which is stored in the Neo4j graph database. The graph comprises three kinds of nodes: *chapter*, unit, and knowledge point. The relationships between nodes are defined by *successor* and subordinate, in order to ensure that the starting point of the retrieved knowledge path is upstream of the end point.

3.3 Knowledge level assessment

The objective of this section is to elucidate the methodology for modeling the knowledge schema established by students through their test-making records and for representing the knowledge points that need to be mastered to answer each test question. Most cognitive diagnosis algorithms employ Q-matrix to model the specific skills tested by each item, thereby enabling the model to provide more detailed diagnosis results and to reveal the mastery of students in specific learning fields. Inspired by these efforts, we select test questions and label each question with which knowledge points from the course syllabus it examined to construct Q-matrix. Specifically, the Q-matrix is a Boolean matrix with J rows and K columns consisting of 0 and 1, where J is the number of test questions, K is the number of knowledge points, and the element in row j and column k is defined as formula (1).

$$q_{jk} = \begin{cases} 0, \text{question j doesn}'\text{t test knowledge k} \\ 1, \text{question j tests knowledge k} \end{cases} \tag{1}$$

Every time a student completes a test, the records will be generated, which is used to generate a student-test matrix Y. The elements in row i and column j of the matrix are defined as Formula (2).

$$y_{ij} = \begin{cases} 0, \text{student i gets question j wrong} \\ 1, \text{student i gets question j right} \end{cases} \tag{2}$$

Similarly, the element of row i and column k in the student-knowledge matrix A is defined as Formula (3).

$$a_{ik} = \begin{cases} 0, \text{student i hasn}'\text{t mastered j knowledge k} \\ 1, \text{student i has mastered j knowledge k} \end{cases} \tag{3}$$

Inspired by the Deterministic Inputs, Noisy "And" gate model (DINA) (De La Torre, 2009) that calculates each element of the Y-matrix from the Q-matrix and the A-matrix, In this paper, the calculation method of element is defined as formula (4).

$$a_{ik} = \prod_{j=1}^{J} = y_{ij}^{q_{jk}} \tag{4}$$

Where J is the number of test questions. Each time a student completes a chapter test, the system will update the student's mastered knowledge to ensure that the student's knowledge level can be tracked in real time.

3.4 Personalized knowledge path retrieval algorithm

According to the knowledge level assessment algorithm, the students' established knowledge schema can be recorded, which can be used as the search scope of the starting point of their personalized knowledge path. The problems to be solved in this section are as follows: First, according to the target questions raised by the students, identify a specific knowledge point within the knowledge graph as the end point of the personalized knowledge path; Secondly, develop a knowledge path and construct a new knowledge schema for students, drawing upon the existing knowledge schema established by them. The language model is guided to generate a sequence of questions according to the path, and the students are inspired to solve the problem through the progressive questioning of multiple questions.

To fully represent the semantic information of knowledge points and support the accurate matching between the target question and knowledge points, inspired by the research conducted by (Abu-Rasheed et al, 2024), all knowledge points (including the labels and attributes) in the knowledge graph are represented as vectors in multidimensional space, and the knowledge graph is retrieved by the target question.

Since the retrieval between the question and knowledge point is sentence-level semantic matching, considering the time cost of the retrieval algorithm, the matching model chooses a two-tower structure (Reimers & Gurevych, 2019). includes two feature extraction models based on BERT (Devlin et al., 2018) with shared parameters. The feature extraction model takes the question or knowledge point as input, and the pooling layer maps them into the 768-dimension vector.

The pre-training task of BERT does not include optimizing the representation of the input at the sentence level. Therefore, loading the pre-training parameters of BERT-Base-Chinese is insufficient to complete the accurate sentence-level matching problem, let alone solve the question-knowledge

retrieval problem in education. This paper takes the approach of pre-training and two-stage fine-tuning to gradually adapt the language model to the specific task in this paper. General Chinese text matching data (Chinese-SNLI, Chinese-MnLI and OCNLI (Hu et al., 2020)), uestion generation data in education (LearningQ) (Chen et al., 2018), and specific course question data (questions designed by teachers in software engineering, knowledge points examined by the questions and corresponding descriptions were manually marked, and a total of 209 question-knowledge points were obtained) were trained to ensure that the model gradually learns domain knowledge. It is worth noting that to maintain the consistency of training paradigms in multiple stages of the model, as well as the consistency between training and reasoning stages, cosine similarity is selected in this paper to measure the similarity between problem Q and knowledge point K, as shown in formula (5).

$$sim(Q,K) = cos(Q,K) = \frac{Q^T K}{\|Q\|\|K\|} \tag{5}$$

Contrastive learning was chosen as the framework to train the matching model, and the optimization goal follows (Chen et al, 2020). Let (q_i,k_i) be the i-th problem-knowledge pairing in a training batch, and (Q_i,K_i) be their vector representation, then for each (q_i,k_i), the loss function is designed as formula (6).

$$loss_{qk} = -log \frac{e^{sim(Q,K)/\tau}}{\sum_{j=1}^{N} e^{sim(Q,K)/\tau}} \tag{6}$$

Where N is the amount of data in a training batch, and τ is the temperature.

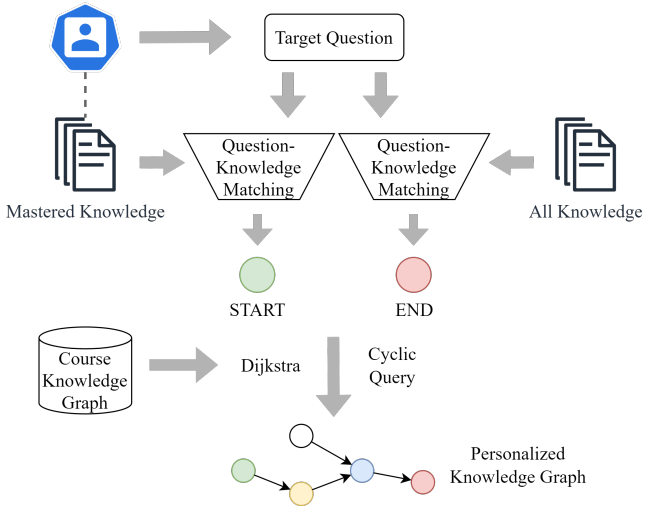

Fig. 2. Personalized knowledge path retrieval process.

The question-knowledge matching model is used to retrieve the starting point and end point in the knowledge graph. Then Dijkstra, a graph search algorithm that solves the single-source shortest path problem, is used to find the knowledge path from start point to end point. Moreover, given that multiple nodes may belong to a single node and that a single node may depend on multiple precursor nodes, it is not possible to express a complete knowledge path by means of a single shortest path. Consequently, cyclic query is employed to retrieve the sibling nodes of a node, that is, other nodes that are jointly subordinate to the next node or that jointly succeed the next node. The retrieval workflow is shown in Fig 2.

3.5 Bloom cognitive hierarchy recognition model

In educational scenarios, students' responses to questions often require higher-order cognitive skills (such as applying, analyzing, etc.) (Chen et al., 2018). Therefore, this paper introduces the taxonomy revised by Bloom (Conklin, 2005) and considers six cognitive levels of knowledge to provide cognitive level information for knowledge description on the personalized knowledge path, so that the generated questions are more consistent with the properties of knowledge.

Considering that some knowledge descriptions may contain two or more cognitive levels, for example, the corresponding knowledge goal of "concepts and technologies of association between objects" is "to explain the association between objects and related concepts, to implement technologies, and to use simple association concepts to solve complex association problems", which includes two cognitive levels of "understanding" and "applying". Therefore, inspired by SpanEmo (Alhuzali & Ananiadou, 2021), this paper designs a multi-label classification of cognitive level recognition, which includes six cognitive levels and one "unrelated".

Let $(S_i,Y_i)_{i=1}^N$ be a set of N pieces of data with 7 category labels, where S_i represents the knowledge description of the input model and $y_i \in \{0,1\}^7$ represents the collection of S_i labels. As shown in Fig. 3, both the tag set and the knowledge description are passed to the BERT encoder, and the vector H_i for each knowledge is obtained from formula (7).

$$H_i = BERT([CLS]+C+[SEP]+S_i) \tag{7}$$

Where, [CLS] and [SEP] are specific tokens in the input of BERT model, represents the label vector of the cognitive level, $H_i \in R^{seq\_len \times dim}$, seq_len is the length of the input text, dim is the dimension of token representation vector. This design not only allows BERT to learn the attention information between different cognitive levels and knowledge description texts, but also to learn a vector representation for each cognitive level.

A feed-forward network is further introduced to compute a prediction score $\hat{y}$ for each knowledge description, and then compute the cross-entropy with the real label y as an optimization target for model training.

In order to facilitate the gradual adaptation of the model to the cognitive level recognition task, this paper pre-loads BERT-Base-Chinese for the BERT encoder, then uses LearningQ and teaching objective data of Software Engineering to finetune the model in a sequential manner. This approach enables the model to more effectively learn the specific knowledge associated with individual courses, while simultaneously acquiring the broader knowledge base associated with the general educational field. For the knowledge points and triples in the personalized knowledge path, the algorithm inputs them into the cognitive level recognition model to predict the cognitive level with the highest score to ask the knowledge, and further

enhance the controllability of question generation.

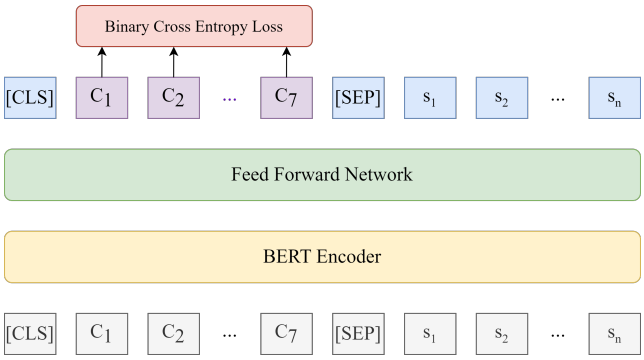

Fig. 3. Cognitive hierarchy recognition model structure diagram.

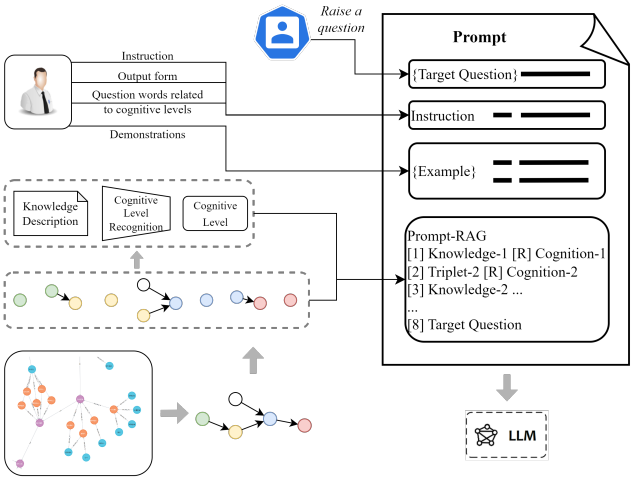

Fig. 4. Overview of prompt construction.

3.6 Prompt construction

A detailed instruction for question sequence generation is first defined. Secondly, for the target questions raised by the students, the personalized knowledge path is retrieved in the course knowledge graph according to the scheme described in 3.4. The description of each knowledge point is composed of its label and attribute. Assuming that the length of the retrieved knowledge path is n, there are $2n$ serialized knowledge points, including original knowledge points on n shortest paths, $n-1$ triplets in the knowledge path (including the main path and bypass) and the target questions raised by the student. Based on this part, the prompt can control the number of sub-questions in the question sequence. For these $2n$ knowledge descriptions, the cognitive level is given respectively according to 3.5, and the common question words (see Table 1) corresponding to each cognitive level (Conklin, 2005) are are added into the instruction, which controls the question formulation.

Table 1. Statistics of the 40 questions proposed by students.

Cognitive level	Representative question words
Remembering	recognize, recall
Understanding	Interpret, summarize, exemplify
Applying	execute, implement, use
Analyzing	differentiate, organize
Evaluating	check, judge
Creating	produce, design

To sum up, this section summarizes the prompt method which makes the question sequence path-manageable, number-controllable and questioning-reliable. The prompt construction scheme is shown in Fig. 4.

3.7 Q&A interactive algorithm

The Least-to-Most Prompting is highly related to the idea of this paper, including two stages: problem decomposition and sequential answering of sub-questions. However, this method only injects human thinking into the prompt, and the generation and answer of sub-questions are completely done by the model independently, which is easy to cause error accumulation. In addition, the special use of LLMs as chatbots also provides students with the possibility to participate in multi-round dialogue, similar to the discussion between students and teachers in daily learning. If humans can engage in the reasoning process of the LLM, they can identify and correct errors in the model's output timely. Furthermore, the model can assess the merits of human responses, thereby facilitating mutual guidance and collaborative exploration between humans and AI. This approach helps to promote the effective solution of the problem. The same is true in the education field. Students should keep thinking and output while receiving new knowledge, and students who can maintain frequent communication with teachers usually have a stronger grasp of knowledge.

In summary, based on the Least-to-Most Prompting, this paper makes adaptive improvements to the Q&A interaction algorithm, allowing students to participate in the reasoning process of the LLM, so that students can learn in the process of interacting with AI. The comparison diagram between the improved prompt scheme and Least-to-Most is shown in Fig. 5.

The aforementioned process serves as a framework through which the LLM can identify models that have demonstrated proficiency in in-context learning and logical reasoning.

GLM-4 has been demonstrated proficiency in comprehending complex contexts and logical relationships, and it matches or even surpasses GPT-4 on multiple tasks (GLM et al., 2024). Moreover, a substantial quantity of Chinese data was used to train the GLM-4 (Yang, Li, & Li, 2022), so it attains a superior understanding of the Chinese context and the capacity to generate more authentic and accurate Chinese text. For the above reasons, GLM-4 is chosen as the LLM in the framework,

which was launched by Zhipu AI in January 2024, as the foundation model, inputs the prompt, generates a heuristic question sequence, and engages in continuous interaction with students to guide them to solve their doubts step by step.

4. Experiments

4.1 Problem generation performance

In order to measure the language quality of generated questions, two indexes, Perplexity and Diversity, are selected in this paper. A low Perplexity indicates better coherence, with values ranging from 0 to +∞. For Diversity, we report Distinct-2 (Li et al., 2016), which represents the lexical diversity of the generated questions by calculating the average of different 2-gram words in the generated problems. Distinct-2 is calculated as shown in formula (8).

$$Distinct\text{-}2 = \frac{Count(unique(2\text{-}gram))}{Count(2\text{-}gram)} \tag{8}$$

Where, "Count(unique(2-gram))" represents the number of unrepeated 2-gram words in the generated text, and "Count(2-gram)" represents the total number of 2-gram words in the text.

Beyond that, we argue that the security of question generation is significant for educational applications. Therefore, the Perspective API (Lees et al., 2022) is used in this paper to report the Toxicity of the generated questions, ranging from 0 to 1. The higher the Toxicity, the stronger the probability that the text contains an inappropriate expression.

To fully verify the importance and specific contribution of each part of the prompt construction algorithm adopted in this paper, based on the above metrics, this paper reviews the following methods: (1) the proposed prompt, (2) a prompt that does not retrieve personalized knowledge path (Prompt w/o. RAG), (3) prompt that does not contain examples, (4) prompt that do not include either (prompt w/o. RAG & Example). As for test data, 40 questions including "What do I need to pay attention to when building a class diagram?" raised by students in the learning of software engineering are selected, with the basic statistical information presented in Table 2. The questions are input into our algorithm to generate corresponding heuristic question sequences. Ultimately, a total of 327 sub-questions including "Summarize what basic elements a class consists of" were obtained. The number of sub-questions is determined by the knowledge paths obtained from the RAG to ensure the variables are controlled for fairness in the experimental comparisons. The evaluation results on all sub-problems are shown in Table 3. The prompt design studied in this paper has a low degree of confusion, and the information retrieval and demonstration examples only lose about 3.1% of the diversity level, indicating that the system has a rich vocabulary expression while ensuring that the generated problems are more easily understood by humans.

4.2 User study

To verify the performance of the heuristic problem sequence generation algorithm in real scenarios and explore whether it can help students solve their doubts through independent learning, we invited 20 students who took the course "Software Engineering" to carry out a preliminary user study, aiming to provide a reference for the optimization direction of the subsequent upgrade of this work. oncerning the manual evaluation conducted by (Wang et al., 2022) and the user survey designed by (Abu-Rasheed et al, 2024), we design the following statement list combined with this special scenario to make an adaptive supplement.

SQ1: I'm satisfied with the design of this chatbot.

SQ2: I'm satisfied with the quality of the generated questions.

SQ3: The generated question sequence keeps me going.

SQ4: I'm satisfied with the responsiveness of the algorithm.

SQ5: The generated questions are basically answerable.

SQ6: The generated questions can inspire me to think for myself.

SQ7: The chatbot asks questions with correct syntax.

SQ8: The first sub-question the chatbot asks me is based on the knowledge I'm familiar with.

SQ9: Through Q&A interaction, the chatbot can help me solve my original questions.

SQ10: The question sequence that the chatbot asks me is logically progressive.

SQ11: Based on my answers to the sub-questions, the chatbot can make correct judgments.

SQ12: After trying it out, I would like to continue using the system as a learning aid.

On this basis, Likert scales (Joshi et al., 2015) were used to record the user's feedback on each statement, that is, each statement have five options: strongly disagree, disagree, uncertain, agree and strongly agree, corresponding to the score of 1, 2, 3, 4 and 5 given by the user. Before students can ask the chatbot questions, they are first asked to complete chapter quizzes that allow the algorithm to assess students' knowledge level. Research result is shown in Fig. 6.

Fig. 6. User study result of our chatbot.

There are four statements with an average score of more than 4.5 points, indicating that there are few grammatical errors in the generated questions (SQ7). The vast majority of students are satisfied with the chatbot (SQ1) and believe that the system can effectively inspire their self-thinking (SQ6). Also, they are willing to continue to use this system to assist their daily learning (SQ12). Moreover, the average score of six statements is between 4 and 4.5, indicating that students were generally able to approve of the quality (SQ2) and answerability (SQ5) of the generated questions. The generated

Table 2. Ablation experiment of prompt construction.

Prompt Construction Method	Perplexity↓	Diversity↑	Toxicity↓
Prompt w/o. RAG & Example	46.952	0.924	0.126
Prompt w/o. RAG	37.854	0.917	0.112
Prompt w/o. Example	35.226	0.892	0.139
Prompt	27.581	0.896	0.093

Table 3. Statistics of the 40 questions proposed by students.

Average tokens	Number of tokens for the shortest sentence	Number of tokens for the longest sentence
16.43	12	37

questions are in line with the student's knowledge level (SQ8) and cognitive load, which enables students to consistently answer the questions according to the chatbot's inspiration (SQ3). At the same time, in response to the doubts originally raised by students, the system is capable of generating a series of progressive questions (SQ10) to inspire students to solve their doubts step by step (SQ9).

Because the response time to generate questions for students depends on many factors. Especially, the retrieval of personalized knowledge paths and the response of LLM's API are time-consuming. Therefore, it is reasonable that our chatbot has poor satisfaction with the response speed (SQ4), and we will continue optimizing the inference speed of the algorithm in the future. Furthermore, the outcomes of users' feedback on SQ11 demonstrate that the simultaneous input of questions, answers generated by the LLM, and students' responses into the LLM to make correct or incorrect judgments did not yield the desired results. Consequently, our subsequent research should investigate whether the injection of more detailed domain knowledge into the model can effectively address the issue.

5. Conclusion and future work

5.1 Conclusion

In this paper, a heuristic question sequence generation algorithm based on retrieval augmentation is proposed and developed into a chatbot. In response to students' questions about the course, the chatbot can synthesize students' knowledge mastery and course organization structure to build scaffolding, generate question sequences that inspire students to think independently and guide students to solve their doubts by gradually answering sub-questions. The works done in this paper are summarized as follows:

(1) Automatic planning of personalized knowledge path

Taking "Software Engineering" as an example, this paper constructs the course knowledge graph based on the syllabus and designs an automatic assessment algorithm for students' knowledge level. Aiming at the target questions raised by students, this paper utilized a question-knowledge matching model trained through three stages to identify the target knowledge point in the knowledge graph. Moreover, the most relevant knowledge point mastered by the students is searched as the starting point, and a personalized knowledge path that can reach the target knowledge point will be planned according to Dijkstra.

(2)The combination and improvement of prompting strategies

In this paper, the Least-to-Most are combined with the retrieval enhancement of personalized knowledge paths, and the information such as task description, reasoning example, knowledge path, and Bloom cognitive level are combined to form the context, prompting the LLM to generate question sequence. This allows the heuristic process to be explained and tailored to different students' learning states. Concurrently, an interactive Q&A algorithm is designed, so that the chatbot can output each sub-question in order, and students answer it in turn, building a multi-round dialogue learning mode of collaborative exploration.

(3) Development and verification of intelligent tutoring chatbot

The question sequence generation algorithm is embodied in a chatbot, and the effectiveness of the algorithm is verified from the two perspectives: experimental comparison and user study. The experimental results demonstrate that retrieval augmentation and demonstration examples can reduce the confusion of the LLM's generated questions. Furthermore, the user feedback indicates that our chatbot can inspire students to solve their doubts and generate question sequences with progressive relations.

5.2 Future work

The heuristic question sequence generation algorithm proposed in this paper is capable of fulfilling the functions of chapter tests, knowledge level assessment, question generation, etc. It is expected to play a more prominent role in the future educational practices. However, there are still numerous details to be addressed such as feature selection, response time, multidisciplinary integration, evaluation mechanism, and so on. For instance, in future work, students' learning behaviors such as watching videos and language feature such

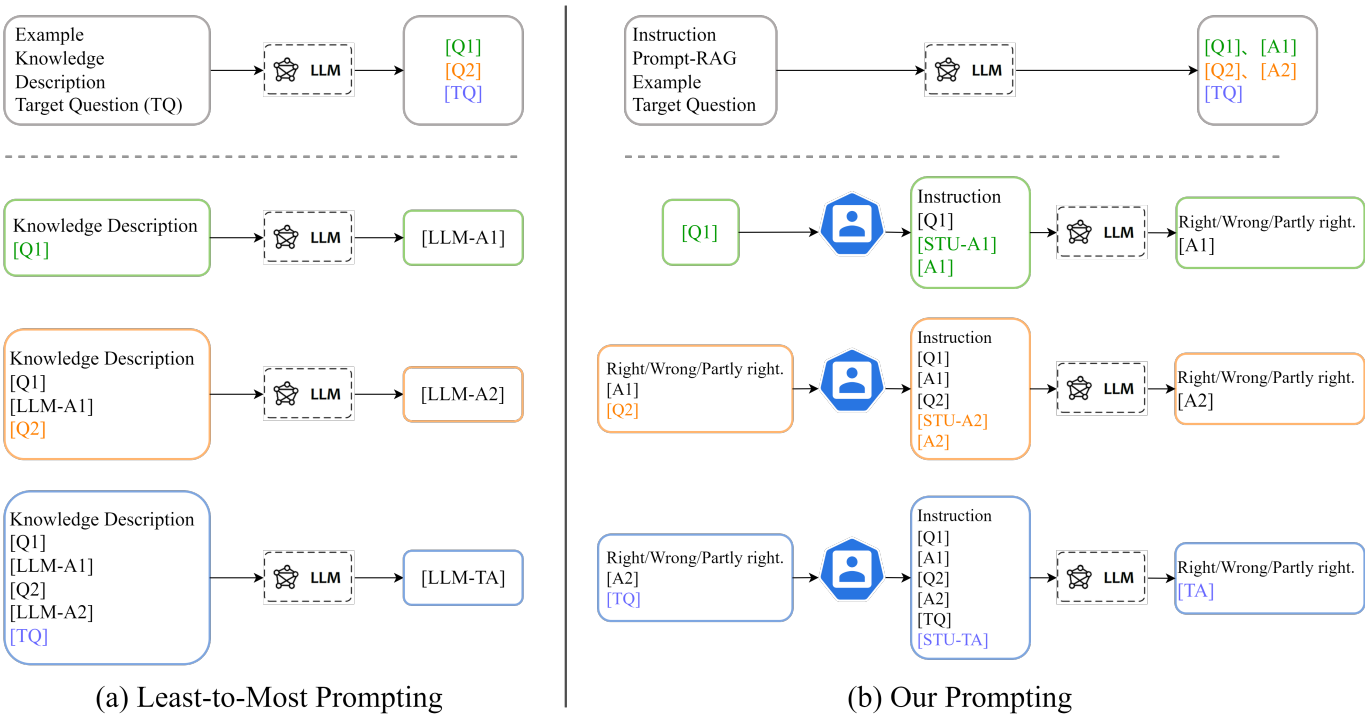

Fig. 5. Comparison diagram between Least-to-Most and ours.

as the context of students' questions can also be utilized as characteristics to assess students' mastery of knowledge points, thereby providing a more diverse and accurate means of evaluating knowledge levels and generating questions. By analyzing students' emotions, motivations, and learning styles, the algorithm can generate question sequences that are more consistent with the principles of educational psychology, thus promoting students' learning more effectively. In addition, we intend to design a long-term, multi-dimensional assessment framework, including regular learning outcome tests, follow-up surveys of students' emotions and attitudes, and data collection through case studies in real teaching scenarios.

To sum up, this paper is just the beginning, and the heuristic question sequence generation algorithm has a large optimization space in multiple directions, which is an exciting research direction. Through iterative optimization, our work is expected to help learners improve higher-order thinking and promote high-quality education.

Acknowledgements

This paper is supported by the Humanities and Social Sciences Research Planning Fund Project of the Ministry of Education: "Research on Metacognitive Diagnosis Theory and Technology Driven by Multimodal Learning Data" (23YJA880091).

Conflict of interest

The authors declare no competing interest.

References

Abu-Rasheed, H., Abdulsalam, M. H., Weber, C., et al. (2024). Supporting Student Decisions on Learning Recommendations: An LLM-Based Chatbot with Knowledge Graph Contextualization for Conversational Explainability and Mentoring. LAK Workshops.

Alhuzali, H., & Ananiadou, S. (2021). SpanEmo: Casting Multi-label Emotion Classification as Span-prediction. Conference of the European Chapter of the Association for Computational Linguistics.

Alkhatlan, A., & Kalita, J. (2019). Intelligent Tutoring Systems: A Comprehensive Historical Survey with Recent Developments. International Journal of Computer Applications, 181(43), 1-20.

Bulathwela, S., Muse, H., & Yilmaz, E. (2023, May). Scalable Educational Question Generation with Pre-trained Language Models. International Conference on Artificial Intelligence in Education.

Chen, G., Yang, J., Hauff, C., et al. (2018, June). LearningQ: A Large-scale Dataset for Educational Question Generation. International Conference on Web and Social Media.

Chen, T., Kornblith, S., Norouzi, M., et al. (2020, February). A simple framework for contrastive learning of visual representations. International Conference on Machine Learning.

Chowdhery, A., Narang, S., Devlin, J., et al. (2023). Palm: Scaling language modeling with pathways. Journal of Machine Learning Research, 24(240), 1-113.

Conklin, J. (2005). A Taxonomy for Learning, Teaching, and

Assessing: A Revision of Bloom's Taxonomy of Educational Objectives [Complete Edition, Lorin W. Anderson, David Krathwohl, Peter Airasian, Kathleen A. Cruikshank, Richard E. Mayer, Paul Pintrich, James Raths, Merlin C. Wittrock]. Educational Horizons, 83(3), 154-159.

Connor-Greene, P. A. (2000). Assessing and promoting student learning: Blurring the line between teaching and testing. Teaching of Psychology, 27(2), 84-88.

De La Torre, J. (2009). DINA model and parameter estimation: A didactic. Journal of Educational and Behavioral Statistics, 34(1), 115-130.

Devlin, J., Chang, M.-W., Lee, K., et al. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. North American Chapter of the Association for Computational Linguistics.

Gao, Y., Xiong, Y., Gao, X., et al. (2023). Retrieval-augmented generation for large language models: A survey. ArXiv, abs/2312.10997.

GLM, T., Zeng, A., Xu, B., Wang, B., Zhang, C., Yin, D., . . . Lai, H. (2024). ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools. arXiv:2406.12793.

Hu, H., Richardson, K., Xu, L., et al. (2020). OCNLI: Original chinese natural language inference. Findings of the Association for Computational Linguistics: EMNLP 2020, 3512.

Joshi, A., Kale, S., Chandel, S., et al. (2015). Likert scale: Explored and explained. British Journal of Applied Science & Technology, 7(4), 396-403.

Kang, M., Kwak, J. M., Baek, J., et al. (2023). Knowledge Graph-Augmented Language Models for Knowledge-Grounded Dialogue Generation. ArXiv, abs/2305.18846.

Khandelwal, U., Levy, O., Jurafsky, D., et al. (2019, November). Generalization through Memorization: Nearest Neighbor Language Models. International Conference on Learning Representations.

Kulshreshtha, D., Belfer, R., Serban, I. V., et al. (2021, April). Back-Training excels Self-Training at Unsupervised Domain Adaptation of Question Generation and Passage Retrieval. Conference on Empirical Methods in Natural Language Processing.

Kulshreshtha, D., Shayan, M., Belfer, R., et al. (2022, June). Few-Shot Question Generation for Personalized Feedback in Intelligent Tutoring Systems. Paper presented at the 11th Conference on Prestigious Applications of Artificial Intelligence, PAIS 2022, co-located with the 31st International Joint Conference on Artificial Intelligence and the 25th European Conference on Artificial Intelligence, IJCAI-ECAI 2022.

Lees, A., Tran, V. Q., Tay, Y., et al. (2022, February). A New Generation of Perspective API: Efficient Multilingual Character-level Transformers. Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington DC, USA.

Li, J., Galley, M., Brockett, C., et al. (2016, June). A Diversity-Promoting Objective Function for Neural Conversation Models. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

Li, T., Ma, X., Zhuang, A., et al. (2023). Few-shot In-context Learning on Knowledge Base Question Answering. Annual Meeting of the Association for Computational Linguistics.

Liang, Y., Wang, J., Zhu, H., et al. (2023, October). Prompting Large Language Models with Chain-of-Thought for Few-Shot Knowledge Base Question Generation. Paper presented at the The 2023 Conference on Empirical Methods in Natural Language Processing.

Lo, K., Wang, L. L., Neumann, M., et al. (2020, July). S2ORC: The Semantic Scholar Open Research Corpus. Paper presented at the Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.

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.

Nishikawa, S., Ri, R., Yamada, I.,et al. (2022, May). EASE: Entity-Aware Contrastive Learning of Sentence Embedding. Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

Raffel, C., Shazeer, N.M., Roberts, A., et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140), 1-67.

Rajpurkar, P., Zhang, J., Lopyrev, K., et al. (2016, June). SQuAD: 100,000+ Questions for Machine Comprehension of Text. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing.

Ram, O., Levine, Y., Dalmedigos, I., et al. (2023). In-context retrieval-augmented language models. Transactions of the Association for Computational Linguistics, 11, 1316-1331.

Reimers, N., & Gurevych, I. (2019, August). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP).

Sequeda, J., Allemang, D., & Jacob, B. (2023, November). A Benchmark to Understand the Role of Knowledge Graphs on Large Language Model's Accuracy for Question Answering on Enterprise SQL Databases. Proceedings of the 7th Joint Workshop on Graph Data Management Experiences & Systems (GRADES) and Network Data Analytics (NDA).

Srivastava, M., & Goodman, N.D (2021, June). Question Generation for Adaptive Education, Annual Meeting of the Association for Computational Linguistics.

St-Hilaire, F., Vu, D.D., Frau, A., et al. (2022, March). A New Era: Intelligent Tutoring Systems Will Transform Online Learning for Millions. ArXiv, abs/2203.03724.

Wang, Z., Valdez, J., Basu Mallick, D.,et al. (2022, July). Towards human-like educational question generation with large language models. International Conference on Ar-

tificial Intelligence in Education (pp. 153–166).

Wei, J., Wang, X., Schuurmans, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35, 24824-24837.

Welbl, J., Liu, N. F., & Gardner, M. (2017, July). Crowdsourcing Multiple Choice Science Questions. Proceedings of the 3rd Workshop on Noisy User-generated Text.

Yang, A., Li, Z., & Li, J. (2024). Advancing GenAI Assisted Programming–A Comparative Study on Prompt Efficiency and Code Quality Between GPT-4 and GLM-4. arXiv:2402.12782

Yang, Z., Gan, Z., Wang, J., et al. (2022, September). An empirical study of gpt-3 for few-shot knowledge-based vqa. Proceedings of the AAAI Conference on Artificial Intelligence.

Zhou, D., Scharli, N., Hou, L., et al.(2022, May). Least-to-Most Prompting Enables Complex Reasoning in Large Language Models. Paper presented at the The Eleventh International Conference on Learning Representations.