IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 22 February 2025, accepted 4 March 2025, date of publication 7 March 2025, date of current version 21 March 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3549309

RESEARCH ARTICLE

LLM-CDM: A Large Language Model Enhanced Cognitive Diagnosis for Intelligent Education

XIN CHEN[1], JIN ZHANG[1], TONG ZHOU[2], AND FENG ZHANG[1]
[1]College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao 266590, China
[2]College of Civil Engineering and Architecture, Shandong University of Science and Technology, Qingdao 266590, China

Corresponding author: Xin Chen (xxwar@163.com)

This work was supported in part by the Distinguished Teachers Training Plan Program of Shandong University of Science and Technology under Grant MS20211105, in part by the National Higher Education Research Project of the Coal Industry of China under Grant 2021MXJG105, in part by the Education Ministry Humanities and Social Science Research Planning Fund Project of China "Personalized Learning Path Recommendation Driven by Multi-Source Educational Data and Its Quantitative Evaluation" under Grant 23YJAZH192, and in part by the General Project of the 14th Five-Year Plan for Educational Science of Shandong Province "Research on the Generation Method of Test Resource Based on Large Language Models" under Grant 2023YB162.

ABSTRACT Cognitive diagnosis is a key component of intelligent education to assess students' comprehension of specific knowledge concepts. Current methodologies predominantly rely on students' historical performance records and manually annotated knowledge concepts for analysis. However, the extensive semantic information embedded in exercises, including latent knowledge concepts, has not been fully utilized. This paper presents a novel cognitive diagnosis model based on the LLAMA3-70B framework (referred to as LLM-CDM), which integrates prompt engineering with the rich semantic information inherent in exercise texts to uncover latent knowledge concepts and improve diagnostic accuracy. Specifically, this study first inputs exercise texts into a large language model and develops an innovative prompting method to facilitate deep mining of implicit knowledge concepts within these texts by the model. Following the integration of these newly extracted knowledge concepts into the existing Q matrix, this paper employs a neural network to diagnose students' understanding of knowledge concepts while applying the monotonicity assumption to ensure the interpretability of model factors. Experimental results from an examination data set for course completion assessments demonstrate that LLM-CDM exhibits superior performance in both accuracy and explainability.

INDEX TERMS Cognitive diagnosis, large language models, exercise texts, higher education and intelligent education.

I. INTRODUCTION

With the rapid advancement of educational informatization in China, smart education has increasingly emerged as a dominant trend in higher education. In this context, personalized teaching, one of the fundamental objectives of smart education, seeks to tailor instructional plans according to the needs of individual students, thus enhancing learning outcomes [1]. However, traditional assessment methods (such as examination scores) rely predominantly on macro-level evaluations and do not capture the internal cognitive states and learning processes of students. This limitation underscores the need for scientifically rigorous and accurate assessments of student learning statuses as a critical prerequisite for realizing personalized teaching [2].

The primary objective of higher education is to cultivate students' abilities to adapt to societal demands [3], encompassing both undergraduate and postgraduate programs. In comparison with secondary education, interactions between teachers and students in higher education are relatively constrained, manifesting primarily in two ways: first, direct communication is diminished due to large-class teaching formats that hinder teachers from gaining an in-depth understanding of each student's learning capabilities

The associate editor coordinating the review of this manuscript and approving it for publication was Gustavo Olague.

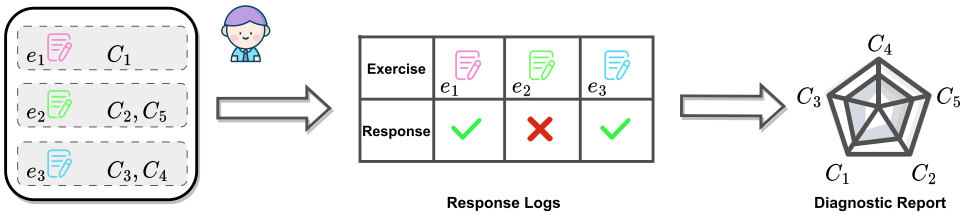


FIGURE 1. A case study of cognitive diagnosis for student performance prediction.

through classroom engagement; second, conventional assessment methods (such as examinations) often inadequately reflect students' comprehension of knowledge. The sole reliance on exam scores complicates the identification of specific areas where students may struggle, thus limiting the effectiveness of personalized instruction.

To address these challenges, Cognitive Diagnosis (CD) has emerged as a significant educational assessment tool that is gradually gaining attention. Fig. 1 illustrates a typical example of cognitive diagnosis. In this process, students first complete a set of exercises (e.g., e_1, e_2, e_3) and submit their answers (correct or incorrect). Based on the students' responses, the model can infer their level of mastery of the relevant knowledge points. For example, if the student answers exercise e_1 correctly, it indicates a good understanding of knowledge point C_1; whereas if the student answers exercise e_2 incorrectly, it suggests a weaker grasp of knowledge points C_2 and C_5. This diagnostic mechanism not only helps teachers or educational systems accurately assess students' knowledge levels but also provides a scientific basis for personalized learning resource recommendations and the optimization of teaching strategies, thereby improving teaching effectiveness and learning efficiency.

In cognitive diagnosis, the Q matrix serves as a fundamental tool for delineating the mapping relationship between practice questions and knowledge points. The rows of the Q matrix represent practice questions, while the columns correspond to knowledge points. If a specific practice question assesses a particular knowledge point, the corresponding element in the matrix is assigned a value of 1; otherwise, it is designated as 0. As illustrated in Figure 2, when the exercise e_i refers to a certain knowledge point c_j, the associated element $Q_{ij} = 1$ in the Q matrix is marked as 1; if not, it is set to 0 [4].

Despite its rich semantic content, experts typically construct the Q matrix based on manual annotations. This process can introduce subjective bias and often results in incomplete annotations. For instance, a question may be annotated solely as assessing one specific knowledge point; however, its content might implicitly encompass additional relevant knowledge points [5], [6]. Such incomplete annotation can lead to inaccuracies within the Q matrix regarding its representation of true relationships between practice questions and knowledge points, ultimately impacting the precision of cognitive diagnosis models.

Furthermore, existing cognitive diagnostic models—such as the DINA model [7], Item Response Theory (IRT) [8], Multidimensional IRT (MIRT) [9]and NeuralCDM [10] primarily concentrate on the interaction between students and practice questions, while overlooking the intricate relationship between practice questions and knowledge points. This oversight not only limits the diagnostic capabilities of these models but also impedes the further advancement of cognitive diagnostic technology.

Another pressing concern is the scarcity of cognitive diagnostic data within higher education. Currently, most cognitive diagnostic research relies on datasets from secondary education (such as ASSIST [11]and MATH [12]), whereas datasets specifically tailored for higher education are relatively limited. This deficiency in data hampers cognitive diagnostic research in higher education by lacking adequate support and constrains educators' ability to accurately evaluate students' knowledge acquisition.

$$Q = \begin{array}{c} \begin{matrix} c_1 & c_2 & c_3 & c_4 & c_5 \end{matrix} \\ \begin{bmatrix} 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \end{bmatrix} \end{array} \begin{array}{l} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{array} \quad \begin{array}{l} c_1 : \textit{Sequential List} \\ c_2 : \textit{Queue} \\ c_3 : \textit{Linked List} \\ c_4 : \textit{Stack} \\ c_5 : \textit{Complete Binary Tree} \end{array}$$

FIGURE 2. An example of Q-matrix.

In recent years, Large Language Models (LLMs) have exhibited remarkable capabilities in the domain of natural language processing. LLMs are capable of generating human-like text responses without being specifically trained for particular tasks [13], which enables them to excel in activities such as sentiment analysis [14]named entity recognition [15]. For instance, they have demonstrated exceptional performance in content generation [16], text mining [17], [18], and knowledge extraction [19]. These attributes render LLMs promising candidates for applications in education, particularly in areas such as automated question generation, assessment of students' knowledge mastery, and the design of personalized learning pathways.

However, large language models also encounter several challenges when applied practically; one notable issue

is the ''hallucination'' problem [20], [21], [22], [23], [24]. This phenomenon occurs when models generate content that appears plausible but is actually inaccurate. Such inaccuracies may undermine their reliability within educational contexts. Nevertheless, the potential benefits offered by large language models cannot be overlooked. For example, LLAMA3-70B serves as an illustrative case: as an open-source large language model, it demonstrates performance comparable to ChatGPT [25] while requiring fewer computational resources. This characteristic makes it more suitable for researchers seeking to deploy and utilize these models within local environments.

In response to the aforementioned challenges, this paper proposes a series of innovative solutions. Firstly, to address the inadequacy of cognitive diagnostic data in higher education, we have constructed a comprehensive dataset derived from the final exam results of the university's data structure course. This dataset provides reliable support for cognitive diagnostic research at the higher education level. Secondly, to tackle the issue of incomplete Q-matrix annotation, we have integrated LLAMA3-70B (a large language model) into our cognitive diagnostic framework. By leveraging its advanced text comprehension capabilities, we can automatically extract implicit knowledge points from practice questions, thereby reducing reliance on manual annotation and enhancing both the comprehensiveness and accuracy of the Q-matrix. Finally, to mitigate potential ''hallucination'' issues—where large language models generate seemingly plausible yet inaccurate content—we have employed prompt engineering techniques. This approach guides the model in accurately extracting knowledge points through meticulously designed prompts, ensuring the reliability of our results.

In summary, this paper makes several key contributions:

1) To address subjective bias and incomplete annotations in Q-matrix labeling, this study introduces large language models (LLMs) to automatically extract implicit knowledge points from exercise texts, enhancing the Q-matrix. Using prompt engineering, we guide the model to mitigate ''hallucination'' issues, reducing subjectivity and inaccuracies in manual annotations. This approach improves the precision of cognitive diagnosis models by addressing incomplete labeling challenges.
2) To address the scarcity of cognitive diagnosis data in higher education, this study collects final exam data from university data structure courses to build a cognitive diagnosis dataset, including exercise texts. This dataset provides new resources for research and practice, enabling more accurate assessment of students' abilities and optimization of teaching strategies, thereby improving teaching effectiveness.
3) Experimental results derived from extensive real-world datasets demonstrate that our model exhibits significant advantages in terms of both accuracy and interpretability.

The remainder of this paper is structured as follows. Section II provides a review of pertinent research on cognitive diagnosis and large language models, along with a detailed explanation of the literature review methodology employed. Section III outlines the concepts, hypotheses, and data preparation involved in the study. Section IV elaborates on the selection process for large language models, offers a comprehensive introduction to the LLM-CDM framework, and presents an in-depth analysis of the design and functionalities of each module. Section V details the experimental setup and analyzes the results obtained. Section VI discusses the implications of these findings within the context of existing research, addresses limitations encountered during the study, and suggests directions for future research endeavors. Finally, Section VII concludes by summarizing key points and insights from this paper while emphasizing its main contributions and conclusions.

II. RELATED WORK

A. COGNITIVE DIAGNOSIS

Currently, the majority of cognitive diagnosis models concentrate on modeling the interactions between learners and exercises, often overlooking the integrity of knowledge concepts within these exercises. However, as research progresses, an increasing number of scholars have begun to acknowledge the critical importance of knowledge concepts in cognitive diagnosis. For example, the ICD [26] model diagnoses learners' cognitive states by fitting the quantitative relationships between exercises and concepts as well as examining interactions among those concepts. DeepCDM [27] employs neural networks and attention mechanisms to learn both the interactions among concepts and the relationships between exercises and concepts. RCD [28] represents learners, exercises, and concepts as nodes within three local relationship graphs while constructing a multi-layer attention network to aggregate node-level and graph-level relationships. CDGK [29] captures interactions among exercise features, learner scores, and learners' mastery of concepts through neural networks. Most of these models depend on manually labeled Q matrices that emphasize existing knowledge concept relationships but fail to thoroughly investigate implicit knowledge points; thus they are unable to eliminate subjectivity and limitations associated with Q matrices.

For this reason, Liu et al. [10] employed a pre-trained convolutional neural network (CNN) to predict the knowledge points associated with practice texts, integrating these predictions into the cognitive diagnosis framework. Cheng et al. [30] introduced an enhanced item response theory (DIRT), which incorporates deep learning techniques such as deep neural networks (DNN) and long short-term memory networks (LSTM) to analyze the semantics of exercise texts and elucidate relationships between exercises and knowledge points, to diagnose students' latent traits. The QRCDM method proposed by Yang et al. [31] quantitatively extracts implicit knowledge points through relational

TABLE 1. Example exercises.

ID	Exercise
Example 1	Given that a complete binary tree has 8 leaves at layer 6 (with the root designated as layer 1), the maximum number of nodes in this complete binary tree is: A. 39 B. 52 C. 111 D. 119
Example 2	If the primary operations performed on a table involve inserting a node after the last node or deleting the last node, then the most efficient storage method for minimizing computational time is: A. Singly linked list B. Doubly linked list C. Circular singly linked list D. Circular doubly linked list with leading nodes.
Example 3	To address the issue of speed mismatch between the computer host and the printer, a print data buffer is typically established. The host sequentially writes data to this buffer, while the printer retrieves data from it in a similar manner. The logical structure of the buffer should be: A. Stack B. Queue C. Tree D. Diagram.

analysis, aiming to identify unmarked knowledge points from the existing Q matrix. Although these text-based cognitive diagnosis models have improved diagnostic effectiveness to some extent, they often overlook variations in exercise types and fail to fully exploit the knowledge points embedded within different categories of exercises. Some approaches merely enhance the existing Q matrix without expanding it or uncovering new knowledge points beyond its current scope.

B. LARGE LANGUAGE MODELS

In the realm of higher education, the utilization of large language models (LLMs) is progressively on the rise, showcasing considerable potential in both teaching and learning processes [32], [33], [34], [35]. LLMs possess the capability to process and generate natural language, rendering them invaluable in various applications such as automatic scoring [36], personalized learning path recommendations [37], and data mining [38]. For example, LLMs can analyze students' learning habits, offer tailored educational resources, or assist educators in course preparation and content development [37]. Furthermore, LLMs are capable of providing immediate feedback and tutoring to students through natural language interactions [39], thereby enriching the overall learning experience.

Although LLMs have broad application prospects in education, they also face the "hallucination" problem [20], [21], [22], [23], [24], in which the model may generate information that sounds reasonable but is inaccurate or fabricated. This issue is particularly critical in cognitive diagnosis because inaccurate knowledge point extraction can lead to an incomplete Q-matrix, affecting the evaluation of students' mastery. To address this problem, prompt engineering [40], [41] has become a key research direction. Using prompts, LLMs can be guided to extract knowledge points from exercises more accurately, optimize the Q-matrix, and reduce the occurrence of hallucinations. This improves the accuracy and reliability of cognitive diagnosis. This method provides strong technical support for cognitive diagnosis, allowing LLMs to be more effectively applied in education, helping with personalized learning [42] and precise student ability assessment.

C. LITERATURE REVIEW METHOD

In this study, a systematic literature search and screening method was employed to ensure the comprehensiveness and scientific rigor of the literature review. Core literature in the fields of generative AI, large language models (LLM), and cognitive diagnosis was retrieved from databases such as Google Scholar, Web of Science, and Science Direct using keywords including "Generative AI," "Large Language Models," and "Cognitive Diagnosis." The aim was to encompass the latest developments in these related fields as thoroughly as possible.

To guarantee both quality and relevance, only peer-reviewed articles published within the past five years were selected. Priority was given to those with high citation counts from leading journals and conferences. Non-peer-reviewed articles and technical reports were excluded from consideration. Furthermore, all included literature had to be directly pertinent to generative AI, LLMs, or cognitive diagnosis, with a particular emphasis on practical aspects concerning model training, application, or educational contexts.

The screening process comprised three steps: initial search, detailed evaluation, and quality review. Initially, relevant literature was identified through keyword searches; subsequently, studies that did not meet predefined criteria were excluded. Finally, an analysis of the remaining literature's quality and academic value was conducted. This analysis focused on evaluating research methods used in each study along with their results, and innovation contributions—thereby ensuring both reliability and scientific integrity in the selected works.

Through this methodological approach, this study established a comprehensive theoretical foundation while also ensuring rigor and transparency throughout the literature review process.

III. PROBLEM DEFINITION

Cognitive diagnosis in wisdom education consists of three components: student $S = \{s_1, s_2, \ldots, s_N\}$, exercise $E = \{e_1, e_2, \ldots, e_M\}$, and knowledge concept $C = \{c_1, c_2, \ldots, c_K\}$, where N, M, and K denote the number of students, the number of exercises, and the number of

knowledge concepts, respectively. Each student selects a certain number of exercises for practice; their response log R is represented as a set of triples (s, e, r), where $s \in S, e \in E$, and r are the scores obtained by student s on exercise e. Furthermore, this paper defines an optimized Q-matrix $\tilde{Q} = \left\{\tilde{Q}_{ij}\right\}_{M\times K}$ such that if exercise e_i is related to knowledge concept k_j it is denoted as $\tilde{Q}_{ij} = 1$; otherwise, it is denoted as $\tilde{Q}_{ij} = 0$. The objective of the cognitive diagnosis task presented in this paper is to assess students' proficiency in knowledge concepts by predicting student performance based on R the $\tilde{Q}$ and matrices derived from response records and exercise texts.

A. CLASSIFICATION OF EXERCISES

This paper begins by analyzing the expression and evaluation methods of exercises, categorizing them into three types: explicit exercises, semi-explicit exercises, and implicit exercises. As illustrated in Table 1, explicit exercises directly assess specific knowledge points; for instance, the exercise presented in Example 1 pertains to 'complete binary trees' and 'number of nodes.' Semi-explicit exercises require some analysis to identify the relevant knowledge points under examination. For example, although Example 2 includes the keyword 'insert delete operation,' it fundamentally assesses the storage methods used by different linked lists and their operational time efficiency. In contrast, implicit exercises lack overt knowledge points or keywords; thus, understanding these underlying concepts requires a thorough text analysis of the exercise stem. For instance, in Example 3, the exercise stem primarily addresses the 'first-in-first-out problem associated with buffers,' specifically focusing on queue characteristics. All exercises in this study were sourced from an online examination platform specifically designed for university education. Students completed and submitted their answers through the platform, which monitored the entire process. This setup effectively minimized the influence of external interference and mitigated risks of dishonest behavior, thereby ensuring the reliability of the research outcomes.

B. THE MONOTONICITY ASSUMPTION

The monotonicity assumption [10] is the foundation of cognitive diagnostic models. Qualitatively describe the relationship between a student's cognitive state on a knowledge concept and the probability of answering a question correctly, thus ensuring the interpretability of the model's results. Specifically, the monotonicity assumption is defined as follows:

Monotonicity assumption: The probability of correct response to the exercise is monotonically increasing at any dimension of the student's knowledge proficiency

In simple terms, if a student s_1 has a better understanding of a concept (e.g. queues) than another student s_2, the student s_1 is more likely to correctly answer questions related to that concept than the other student s_2. This assumption should be converted as a property of the interaction function. Intuitively, we assume that the student s answers exercise e correctly. During training, the optimization algorithm should increase the student's proficiency if the model produces a wrong prediction (that is, a value below 0.5).

IV. A LARGE LANGUAGE MODEL ENHANCED COGNITIVE DIAGNOSIS MODEL

A. CHOICE OF LLM

Language models based on the Transformer architecture, especially pre-trained large language models (LLMs), have made notable advancements in various natural language processing (NLP) tasks [16], [17], [18], [19]. These models demonstrate excellent understanding and reasoning abilities through large-scale pre-training and cross-disciplinary text corpora. For example, GPT-4 achieved a performance level exceeding 20 points on the USMLE (United States Medical Licensing Examination), even without domain-specific fine-tuning or prompt engineering [43]. This efficiency across a wide range of text domains indicates that LLMs can perform effectively in various NLP tasks when applied to education. However, choosing the right LLM is very important to ensure accurate and reliable results. Studies have shown that using the wrong model can cause errors and affect later evaluations and decisions [44].

To implement large language models (LLMs) in the field of education, we have selected the Meta open-source LLAMA3-70B model. This model represents the third iteration of the LLAMA series and emphasizes privacy protection as well as fine-tuning control over generated content, thereby addressing the need for customized enterprise solutions and data security [25]. The LLAMA3-70B has demonstrated exceptional performance across various natural language processing (NLP) tasks, particularly in knowledge extraction and text analysis within the domains of medicine and chemistry. For example, Cui et al. [15] investigated the application of LLAMA3-70B for text mining, showcasing its robust capabilities in knowledge extraction pertinent to medical contexts. Similarly, Ofir Ben Shoham et al. [45] employed LLAMA3-70B to elucidate medical codes and differentiate between medical concepts; their findings indicated that this model outperformed existing state-of-the-art clinical language models specifically tailored for the medical domain. Furthermore, Zhang et al. [46] utilized LLAMA3-70B to extract knowledge from intricate chemical texts, underscoring its strengths in managing complex text processing tasks. These research outcomes suggest that LLAMA3-70B possesses a profound understanding of complex texts and exhibits cross-domain adaptability, making it an optimal choice for extracting key information from exercises.

B. MODEL OVERVIEW

To better uncover the implicit knowledge concepts within exercise texts, we propose a Cognitive Diagnosis Model (CDM) enhanced by large language models (LLM), referred

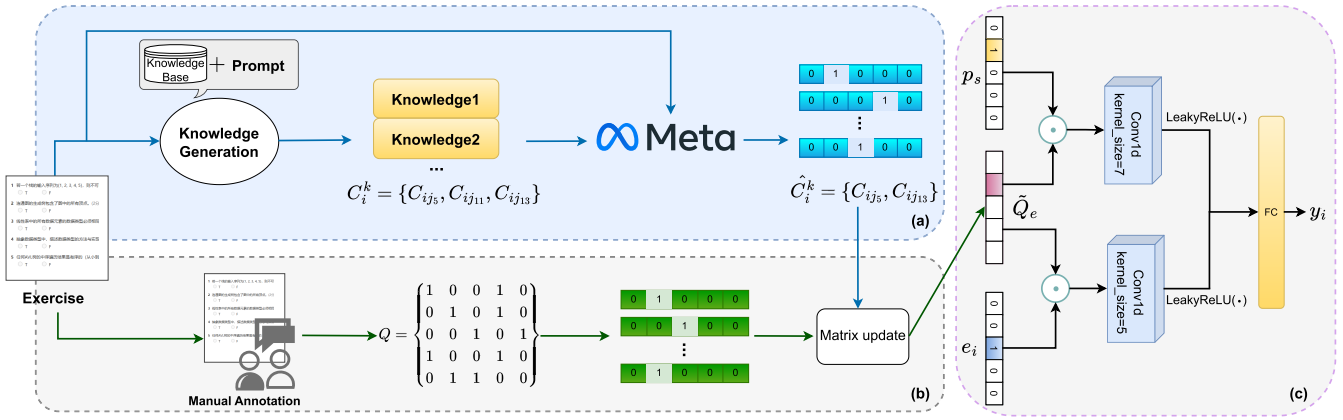


FIGURE 3. The Framework of Large Language Model-enhanced Cognitive Diagnosis Model (LLM-CDM) and Its Core Modules:(a) Knowledge Generation Module; (b) Q-Matrix Optimization Module; (c) Diagnostic Module.

to as LLM-CDM. As shown in Fig. 3, LLM-CDM consists of three core modules: the Knowledge Generation Module, the Q-Matrix Optimization Module, and the Diagnostic Module. In the Knowledge Generation Module (Fig. 3(a)), we utilize large language models (LLMs) and prompt engineering techniques to extract implicit knowledge concepts from the exercise texts. These knowledge concepts are generated through the deep semantic understanding capabilities of the LLM, enabling it to comprehensively capture the underlying knowledge points in the exercises.In the Q-Matrix Optimization Module (Fig. 3(b)), we integrate the generated knowledge concepts to refine the existing Q-matrix. This step enhances the accuracy and completeness of the Q-matrix, allowing it to more precisely reflect the relationships between exercises and knowledge points.In the Diagnostic Module (Fig. 3(c)), we map the students' cognitive states and exercise features into latent factors. The optimized Q-matrix is then used as input features, which, together with the latent factors, are fed into the diagnostic model. Based on these inputs, the model can accurately predict the students' mastery of relevant knowledge points, providing a scientifically grounded and reliable foundation for personalized learning.

C. KNOWLEDGE GENERATION MODULE

The primary objective of the knowledge generation module is to leverage prompt words to guide large language models (LLMs) in extracting implicit knowledge concepts from exercises. This paper introduces a novel method for utilizing prompt words to steer the generation process of LLMs, as illustrated in Fig. 4. The prompt words encompass both instructions and fixed task examples. These examples are crafted by humans, with each example comprising an exercise text alongside its corresponding knowledge concept. Furthermore, to mitigate the potential uncertainty associated with LLMs when generating knowledge concepts, this study constructs a comprehensive knowledge base that encompasses all key knowledge points relevant to the data structure course. This resource aims to assist large language models in achieving more precise knowledge matching within this established framework. Specifically, for each given exercise, this paper concatenates the prompt word P with the exercise text E and subsequently inputs this combined text into a large model to facilitate the search for related concepts within an existing knowledge base, thereby generating a corresponding set of knowledge concepts $C_i^k = \{C_{ij_5}, C_{ij_{11}}, C_{ij_{13}}\}$. Here, k denotes the number of knowledge concepts while C_i^k represents the set of knowledge concepts contained in exercise e_i; additionally, C_{ij_5} indicates that exercise e_i assesses the knowledge concept labeled as 5.

Then, the predicted knowledge concepts are input into the large model again as prompt words to verify the similarity between these knowledge concepts and the exercises. If the verification is successful, these knowledge concepts are added to the knowledge concept set associated with the exercise, and the final knowledge concept set $\hat{C}_i^k = \{C_{ij_5}, C_{ij_{13}}\}$ is obtained. In this paper, the correlation between exercise E and knowledge concept C is expressed as $f(p_{e_i}, p_{c_j})$, and cosine similarity is employed to compute it. The calculation formula is as follows:

$$f(p_{e_i}, p_{c_j}) = \frac{p_{e_i} \cdot p_{c_j}}{\|p_{e_i}\| \, \|p_{c_j}\|} \tag{1}$$

Here, vector p_{e_i} denotes the exercise E, while vector p_{c_j} represents the newly predicted knowledge concept C. $p_{e_i} \cdot p_{c_j}$ signifies the dot product of the two vectors, and $\|p_{e_i}\|$ and $\|p_{c_j}\|$ denote the magnitudes of these vectors, respectively. Subsequently, this paper filters the generated knowledge concepts based on the following formula:

$$f(p_{e_i}, p_{c_j}) \geq \theta \tag{2}$$

where θ represents the predefined threshold. If $f(p_{e_i}, p_{c_j})$ is greater than or equal to this threshold, the corresponding knowledge concept is incorporated into the final knowledge concept set $\hat{C}_i^k$. In this module, LLAMA3-70B is utilized as the generation model. In the experiment, no fine-tuning of the model is performed; instead, it is utilized directly for knowledge generation.

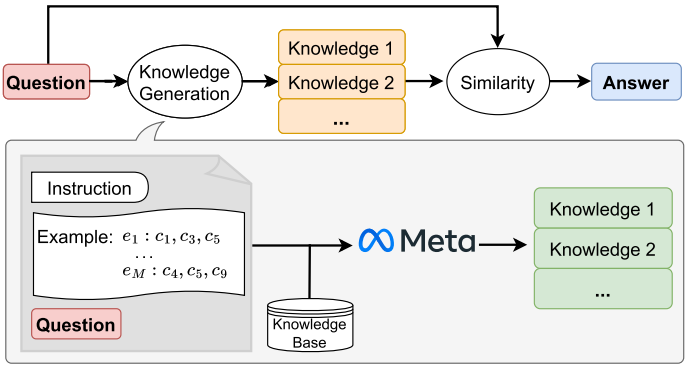


FIGURE 4. Knowledge Generation Module: Utilizes large language models (LLMs) and prompt engineering techniques to extract implicit knowledge points from exercise texts, providing a foundation for Q-matrix optimization.

D. Q MATRIX OPTIMIZATION MODULE

The Q matrix optimization module aims to enhance the accuracy of the cognitive diagnosis model. It does this by integrating knowledge concepts extracted by large language models (LLMs) and refining the existing Q matrix. As illustrated in Fig. 3(b), this module employs two branches to extract knowledge concepts embedded within the exercises. One branch utilizes a large language model and prompt words to uncover implicit knowledge concepts, while the other branch derives the Q matrix through manual labeling conducted by human experts. Despite certain limitations associated with manually labeled data, its reliability remains high; thus, the relevance of knowledge concepts derived from Q matrix labeling is greater than those from $\left\{c_j | c_j \in \hat{C}_i^k \text{ and } Q_{ij} = 0\right\}$. To achieve this integration, this paper retains existing knowledge concepts in the Q matrix and directly incorporates newly predicted knowledge concepts into it. To achieve this integration, the Q matrix is updated according to the following rule:

$$\tilde{Q}_{ij} = \begin{cases} 1, \textit{if} \quad Q_{ij} = 1 \quad \textit{or} \quad f(p_{e_i}, p_{c_j}) \geq \theta \\ 0, \textit{otherwise} \end{cases} \tag{3}$$

where: Q_{ij} denotes the Q-matrix prior to optimization. This method enables the Q-matrix to be dynamically updated, thereby enhancing the performance of cognitive diagnosis.

E. DIAGNOSIS MODULE

The diagnostic module integrates neural networks to model and explain the complex interactions among factor vectors, enabling precise diagnosis of students' mastery of knowledge concepts. In this module, the original Q-matrix is replaced with an optimized version, and the feature extraction module is incorporated to capture broader local interaction information. Finally, the output vectors of the module undergo non-linear transformation through a fully connected layer to generate the final assessment results. As illustrated in Fig. 3(c), this module consists of the following components:

1) STUDENT FACTOR

In LLM-CDM, p^s is used in this paper to represent each student's mastery of knowledge, which is derived by multiplying the student's unique heat representation vector x^s with a trainable matrix A. Specifically:

$$p^s = sigmoid\left(x^s \times A\right) \tag{4}$$

2) EXERCISE FACTOR

Each exercise factor is denoted by $\tilde{Q}_e$, $\tilde{Q}$ is the optimized Q-matrix.

$$\tilde{Q}_e = x^e \times \tilde{Q} \tag{5}$$

where: $\tilde{Q}_e \in \{0, 1\}^{1\times K}$, $x^e \in \{0, 1\}^{1\times M}$ denotes the unique heat vector representation of the exercise. Simultaneously, $p^{diff} \in (0, 1)^{1\times K}$ and $p^{disc} \in (0, 1)$ are utilized to represent the difficulty and differentiation of the exercises, respectively.

3) FEATURE EXTRACTION MODULE

To more accurately capture the characteristics of students and exercises, we designed a feature extraction module based on a one-dimensional convolutional neural network. This module processes the one-hot encoded vectors of students and exercises separately, interacts with the Q-matrix, and ultimately outputs the latent feature representations of students and exercises.

First, for a student p^s, we obtain the interaction vector x^p through the following operation:

$$x^p = \tilde{Q}_e \circ p^s \times p^{disc} \tag{6}$$

where $\circ$ denotes element-wise multiplication.

To model the complex dependencies between students and knowledge points, we employ a one-dimensional convolutional layer for feature extraction. The kernel size is set to 7×7 to capture broader local interaction information. Simultaneously, the LeakyReLU activation function is used to avoid the dying neuron problem and ensure that sparse feature information is preserved. Ultimately, we obtain the

latent feature representation P^s of the student p^s:

$$P^s = LeakyReLU\left(Conv1d\left(x^p, Kernelsize = 7\right)\right) \quad (7)$$

Similarly, to extract the local features of an exercise, we input the one-hot encoded vector of the exercise into a one-dimensional convolutional layer with a kernel size of 5×5 and use the LeakyReLU activation function to obtain the latent feature representation E^i of the item e^i:

$$e^i = \tilde{Q}_e \circ p^{diff} \times p^{disc} \quad (8)$$

$$E^i = LeakyReLU\left(Conv1d\left(e^i, Kernelsize = 5\right)\right) \quad (9)$$

Finally, we input the latent feature representation of the student $P^s$and the latent feature representation of the exercise E^iinto a fully connected layer, and obtain the final predicted score y through a nonlinear transformation:

$$y = Z_3 \times \phi\left(Z_2 \times \phi\left(Z_1 \times \left(P^s - E^i\right)^T + b_1\right) + b_2\right) + b_3 \quad (10)$$

where ϕ denotes the activation function and *Sigmoid* is employed. To satisfy the monotonicity hypothesis, a straightforward strategy is used: each element of Z_1, Z_2, Z_3 is constrained to be a positive number. Consequently, for each entry p_i^s in p^s, $\frac{\partial y}{\partial p_i^s}$ remains positive. Thus, the monotonicity assumption is consistently upheld during training.

In this way, the feature extraction module can effectively capture the local features of students and test items, integrate the global information from the Q-matrix, and significantly improve the predictive performance of the model.

4) LOSS FUNCTION

In cognitive diagnosis tasks, the primary loss function employed is the binary cross-entropy loss. This function quantifies the discrepancy between the model's prediction y and the true label r. The overall loss function can be expressed as follows:

$$Loss = -\sum_i \left(r_i logy_i + (1 - r_i)\, log\,(1 - y_i)\right) \quad (11)$$

Upon completion of training, the value of p^s represents the diagnostic outcome obtained in this study, reflecting students' knowledge proficiency.

V. EXPERIMENT

In this section, we first outline our self-constructed exercise text dataset and the evaluation metrics employed. Subsequently, the effectiveness of the model is validated through comprehensive experiments. Specifically, the experiment encompasses the following aspects: (1) Results of predicting student performance; (2) Student performance prediction under different proportional training sets; (3)Performance Comparison and Analysis of Large Language Models; (4)Ablation Study; (5) Comparative Analysis of Q Matrix Optimization Before and After; (6) Quantitative analysis of the Q matrix.

A. EXPERIMENTAL SETUP

1) DATASET DESCRIPTION

To address the limitations of existing open datasets regarding the quantity and diversity of test questions, this study developed a novel dataset encompassing all final exam exercises from a university's data structure course, with each exercise annotated with relevant knowledge concepts. This dataset comprises 14,878 interaction records, covering exercises of varying levels and difficulties across knowledge domains such as data structures, algorithms, and programming. The exercises were manually annotated by educational experts to ensure the accuracy of the Q-matrix and to provide a robust foundation for cognitive diagnosis models. The dataset, including exercise texts, score matrices, and the Q-matrix, is publicly available at https://pan.baidu.com/s/1YoQh6a8BoUbECuqZ7JLWJg?pwd=8qk6, providing valuable resources for research in educational cognitive diagnosis.

In the process of constructing the dataset, we employed a large language model (LLM), LLAMA3-70B, to extract knowledge concepts that had not been explicitly annotated by experts. This approach aimed to expand the Q-matrix and identify more granular knowledge components. The extraction process was designed not only to uncover implicit knowledge points but also to offer a more detailed representation of the underlying knowledge structure associated with the questions. To validate the accuracy of the knowledge concepts extracted by the LLM, we conducted expert validation by comparing these concepts with those annotated by experts and making necessary adjustments through discussions among specialists.

The dataset offers rich data for model training along with a clear framework for evaluating cognitive diagnosis models. Relevant statistics are presented in Table 2.

TABLE 2. Summary of the dataset for cognitive diagnosis research.

Attribute	Value
Students	565
Exercises	177
Knowledge Concepts	45
Response Logs	14838

B. EVALUATION INDICATORS

To comprehensively evaluate the performance of large language models (LLMs) in the task of knowledge point extraction and their contributions to various cognitive diagnostic models, this paper adopts different quantitative evaluation metrics tailored to the knowledge point extraction task and the performance of cognitive diagnostic models, respectively. Below is a detailed introduction to these metrics and their applications.

1) EVALUATION OF KNOWLEDGE POINT EXTRACTION

In the knowledge point extraction task, traditional metrics such as precision, recall, and F1 score are typically used to measure the alignment between model predictions and expert annotations (Ground Truth). However, the scenario addressed in this paper requires LLMs not only to extract knowledge points already identified by experts but also to identify potential knowledge points that are plausible but unannotated. To address this, we introduce a set of expert-corrected knowledge points as a reference standard and conduct a comprehensive assessment based on the following metrics:

Precision: This metric measures the proportion of correctly extracted knowledge points among those identified by the model. It is calculated as:

$$Precision = \frac{TP}{TP + FP}$$

Here, True Positives (TP) denote the number of knowledge points accurately extracted by the model and subsequently validated by experts, while False Positives (FP) refer to the number of knowledge points identified by the model but not acknowledged by experts.

Recall: This metric quantifies the proportion of true knowledge points that the model successfully extracts from the test questions. It is calculated as:

$$Recall = \frac{TP}{TP + FN}$$

Here, False Negatives (FN) denote the count of knowledge points that the model fails to extract, despite being recognized by experts.

F1 Score: This metric integrates precision and recall to provide a balanced assessment of the model's overall performance. It is calculated as:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{12}$$

These metrics collectively provide a comprehensive evaluation of the LLMs' ability to extract both explicit and implicit knowledge points, ensuring a thorough understanding of their contributions to the task.

2) EVALUATION OF COGNITIVE DIAGNOSTIC MODELS

Evaluating the performance of cognitive diagnostic models presents unique challenges, as the true level of students' knowledge mastery is often inaccessible. In most studies, the effectiveness of diagnostic models is assessed indirectly through their ability to predict student performance. This paper adopts a similar methodology, evaluating the diagnostic model's effectiveness by predicting student exercise scores. To ensure a comprehensive evaluation, we employ widely used metrics, including accuracy (ACC), area under the curve (AUC) [47], and root mean square error (RMSE) [48]. These metrics collectively capture the model's classification and regression capabilities, providing a holistic assessment of its performance.

By combining these metrics, we ensure a comprehensive evaluation of the cognitive diagnostic models, capturing both their classification and regression capabilities. This evaluation approach not only validates the effectiveness of LLMs in knowledge point extraction but also provides a robust assessment of the cognitive diagnostic models' performance in predicting student outcomes.

C. BASELINE

To validate the effectiveness of the proposed LLM-CDM model, we compared it with several widely used cognitive diagnostic models. Additionally, to select the most suitable large language model (LLM) for extracting hidden knowledge points, we evaluated multiple pre-trained language models based on evaluation metrics. Below is a detailed description of these baseline models.

1) COGNITIVE DIAGNOSTIC MODELS

The following cognitive diagnostic models were selected as baselines to evaluate the performance of LLM-CDM in cognitive diagnosis tasks:

1) DINA [7]: A cognitive diagnosis method that models candidates' mastery of knowledge concepts through a Q matrix while accounting for clerical errors and guessing parameters.
2) IRT [8]: As a widely adopted cognitive diagnosis approach, IRT employs linear functions to model one-dimensional student and exercise features.
3) MIRT [9]: MIRT is a multidimensional extension of IRT, capable of modeling multiple levels of knowledge proficiency among students and exercises.
4) NeuralCDM [10]: NeuralCDM is a deep learning-based cognitive diagnostic model that captures higher-order and complex interactions between students and exercises through neural networks.

It is important to emphasize that methods such as ICD [26], DeepCDM [27], RCD [28], and CDGK [29], which are based on the Q-matrix, primarily focus on the interrelationships among knowledge points. This focus differs from that of our study, which seeks to extract new knowledge points from educational texts without modeling the relationships between them. Consequently, due to differences in datasets and research objectives, these Q-matrix-based methods do not meet the criteria for direct comparison with our approach and were therefore excluded as baseline methods.

2) LARGE LANGUAGE MODEL

To select the most suitable large language model (LLM) for extracting hidden knowledge points, we evaluated the following five representative pre-trained language models based on evaluation metrics such as Precision, Recall, and F1 score:

1) LLAMA3-70B [25]: Developed by Meta, this open-source large model comprises 70 billion parameters and is based on the Transformer architecture. It demonstrates exceptional capabilities in natural

TABLE 3. Results of predicting student performance.

Model	Accuracy	RMSE	AUC
DINA	0.531	0.515	0.611
IRT	0.770	0.406	0.748
MIRT	0.693	0.521	0.647
NeuralCDM	0.876	0.313	0.923
LLM-CDM	**0.942**	**0.211**	**0.960**

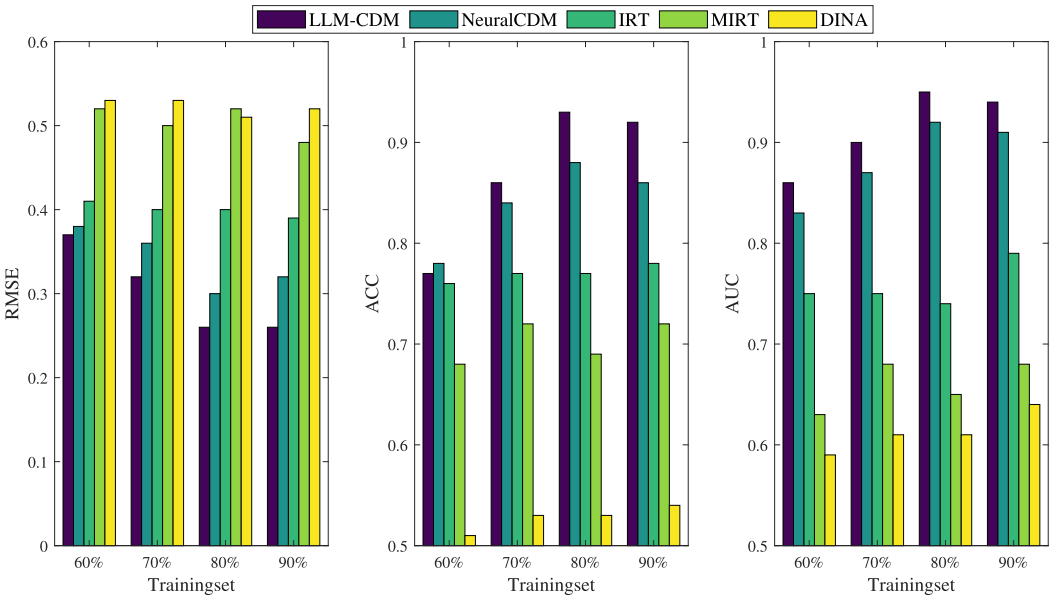


FIGURE 5. Student performance prediction under different proportional training sets.

language understanding and generation tasks, particularly excelling at managing complex semantics and long-range text dependencies.

2) ChatGPT [49]: Developed by OpenAI, this general-purpose dialogue model is based on the GPT architecture and is known for its powerful context understanding and generation capabilities. Its performance is particularly outstanding in multi-round dialogues and complex contexts.
3) ChatGLM [50]: Developed by Tsinghua University, this Chinese dialogue model is grounded in the GLM architecture and has been specifically optimized to cater to the unique characteristics of the Chinese language. It demonstrates exceptional proficiency in semantic understanding and generation tasks within the context of the Chinese language.
4) Mistral [51]: This efficient and lightweight open-source large model is distinguished by its low resource consumption coupled with high performance. The primary design objective is to achieve superior task performance while simultaneously minimizing computational and storage costs.
5) T5 [52]: Developed by Google, this text-to-text transfer model employs a unified framework to convert various natural language processing tasks into text generation tasks. The flexibility and robust generative capabilities of T5 render it an exemplary candidate for cognitive diagnosis applications.

By comparing the LLM-CDM model with other cognitive diagnostic models, we were able to comprehensively evaluate its effectiveness in cognitive diagnosis tasks. At the same time, through the evaluation of multiple pre-trained language models, we selected the most suitable model for extracting hidden knowledge points, thereby providing a solid foundation for the knowledge point extraction module of LLM-CDM.

D. EXPERIMENTAL RESULT

1) RESULTS OF PREDICTING STUDENT PERFORMANCE

This paper conducts an experiment on student achievement prediction based on the baseline model and LLM-CDM. A random segmentation method is employed to extract 80% of each student's response records from the dataset, which is a standard practice in student achievement prediction tasks.

The experimental results are presented in Table 3, where LLM-CDM outperforms all baseline models, particularly when compared to IRT and MIRT. Specifically, LLM-CDM demonstrates significant improvements over NeuralCDM across all metrics: a 6.6% increase in accuracy (ACC), a 3.7% increase in area under the curve (AUC), and a 10.2% reduction in root-mean-square error (RMSE). The values highlighted in bold within the table indicate the best results. These findings suggest that extracting implicit knowledge concepts from exercises by integrating large language models with prompt words is effective for student achievement prediction, significantly enhancing predictive accuracy. Furthermore, this approach underscores the importance of rich semantic information present within exercise texts for cognitive diagnosis.

2) STUDENT PERFORMANCE PREDICTION UNDER DIFFERENT PROPORTIONAL TRAINING SETS

To comprehensively evaluate the model's performance, this paper divides the dataset into various proportions and compares the effect of each proportion on predicting student performance. For each student's practice record, we used 60%, 70%, 80%, and 90% of the exercises as the training set, with the remaining exercises constituting the test set. The average results were reported using multiple evaluation metrics. The experimental findings are illustrated in Fig. 5. The results indicate that LLM-CDM exhibits superior classification performance compared to other models when trained with 60% of the data, underscoring its effectiveness in handling small sample sizes. As the proportion of training data increases to 70%, there is a further enhancement in model performance, suggesting a positive correlation between increased data availability and improved learning outcomes. As the training data increases to 80%, model performance continues to improve, but at a diminishing rate. At 90%, performance stabilizes, showing no significant difference from the 80% training data. These findings suggest that LLM-CDM models possess distinct advantages when dealing with small samples and can effectively learn from limited datasets.

Among all baseline models, LLM-CDM consistently outperforms others across all indices, providing further evidence for its efficacy in extracting knowledge concepts from exercise text by leveraging large-scale language models.

3) PERFORMANCE COMPARISON AND ANALYSIS OF LARGE LANGUAGE MODELS

To comprehensively evaluate the performance of various large language models in the task of knowledge point extraction, we conducted a systematic comparative experiment involving LLAMA3-70B, ChatGPT, ChatGLM, Mistral, and T5. The evaluation metrics employed in this experiment included the adjusted F1 score, recall rate (Recall), and precision rate (Precision). As illustrated in Table 4,

The experimental results indicate that LLAMA3-70B significantly outperforms the other comparison models concerning both the adjusted F1 score (0.871) and precision rate (0.853). Although model ChatGPT exhibits a slightly higher recall rate (0.904) compared to LLAMA3-70B's recall rate of 0.890, the performance gap between these two models is relatively minor.

TABLE 4. Performance comparison of different large language models in knowledge point extraction tasks.

Model	Precision	Recall	F1
T5	0.734	0.784	0.758
Mistral	0.825	0.803	0.814
Chatglm	0.794	0.871	0.830
ChatGPT	0.831	**0.904**	0.865
LLAMA3-70B	**0.853**	0.890	**0.871**

More importantly, LLAMA3-70B demonstrates distinct advantages across multiple dimensions that make it particularly suitable for knowledge point extraction tasks within educational contexts. Firstly, its open-source nature confers substantial benefits regarding model transparency, customizability, and scalability, which allow for targeted optimization tailored to specific pedagogical needs. Secondly, LLAMA3-70B displays enhanced stability when processing long texts and managing cross-paragraph semantic associations—especially when extracting implicit knowledge points and complex conceptual relationships—where its performance is comparable to that of model ChatGPT and even superior in certain intricate scenarios.

Furthermore, the open-source ecosystem surrounding LLAMA3-70B, coupled with robust community support, provides an abundance of tools and resources that significantly lower deployment costs while expediting practical applications within educational settings. Based on a thorough consideration of performance metrics alongside model characteristics and actual application requirements, we have selected LLAMA3-70B as our final choice for hidden knowledge point extraction.

To further understand the limitations of the model, we conducted a systematic analysis of false positives (FPs) and false negatives (FNs) in the knowledge point extraction process. For example, in the input text "The definition of data structures includes arrays, linked lists, and stacks," the model incorrectly labeled "the definition of data structures" as a knowledge point, whereas it is merely a descriptive statement. In another input text, "The applications of recursive algorithms include tree traversal and dynamic programming," the model failed to recognize "dynamic programming" as a knowledge point.

Through the analysis of FPs and FNs, we identified the following main issues: semantic similarity interference leading to the model misjudging certain terms or phrases related to knowledge points; insufficient contextual understanding causing the model to mistake non-critical information for knowledge points when processing complex

sentences; and the omission of implicit knowledge points resulting in the model failing to extract knowledge points that were not explicitly mentioned. Based on these findings, we propose several potential directions for improvement, including optimizing prompt engineering to guide the model more precisely, introducing context-aware mechanisms to enhance the understanding of complex contexts, and adding post-processing rules to filter out obvious false positives. These improvement directions provide valuable insights for future research and are expected to further enhance the model's performance.

4) ABLATION STUDY

To thoroughly assess the necessity and contribution of each core component within the LLM-CDM architecture, we designed and conducted ablation experiments. We used accuracy (ACC), root mean square error (RMSE), and area under the curve (AUC) as performance evaluation metrics. The experiments consisted of four model configurations: 1) Base: the baseline model without any additional modules; 2) Base + KGM: the baseline model with the addition of the Knowledge Generation Module (KGM); 3) Base + FEM: the baseline model with the integration of the Feature Extraction Module (FEM); 4) Base + KGM + FEM: the full LLM-CDM architecture, incorporating both the Knowledge Generation Module and Feature Extraction Module.

TABLE 5. Performance analysis of ablation experiments for the LLM-CDM architecture.

Model Setup	ACC	RMSE	AUC
BASE	0.865	0.301	0.901
BASE+KGM	0.929	0.261	0.947
BASE+FEM	0.927	0.232	0.949
BASE+KGM+FEM	**0.942**	**0.211**	**0.960**

The experimental results are presented in Table 5. It is evident from the findings that incorporating either the Knowledge Generation Module (Base + KGM) or the Feature Extraction Module (Base + FEM) alone leads to a significant enhancement in model performance. Specifically, Base + KGM results in increases of 6.4% in accuracy (ACC) and 4.6% in area under the curve (AUC), while reducing root mean square error (RMSE) by 4%. In contrast, Base + FEM leads to improvements of 6.2% in ACC and 4.8% in AUC, along with an 6.9% reduction in RMSE. These findings highlight the critical roles played by both modules in cognitive diagnosis tasks.

When both the Knowledge Generation Module and the Feature Extraction Module are integrated simultaneously (Base + KGM + FEM), optimal model performance is achieved. Compared to the baseline model, there is a notable increase of 7.7% in ACC, a rise of 5.9% in AUC, and a decrease of 9% in RMSE. This outcome not only validates the effectiveness of both modules but also suggests their synergistic effect within the LLM-CDM architecture, collectively enhancing cognitive diagnosis accuracy.

The observed improvements in ACC and AUC clearly indicate that the LLM-CDM architecture excels in classification tasks. Furthermore, the substantial reduction in RMSE reinforces its superiority in predictive accuracy. These results comprehensively demonstrate the necessity and contribution of incorporating the Knowledge Generation and Feature Extraction Modules into the LLM-CDM framework. In summary, the findings from the ablation experiments robustly validate the effectiveness of these modules across multiple performance metrics(ACC, RMSE, and AUC),highlighting their significant role in enhancing overall model performance and providing stronger technical support for cognitive diagnosis.

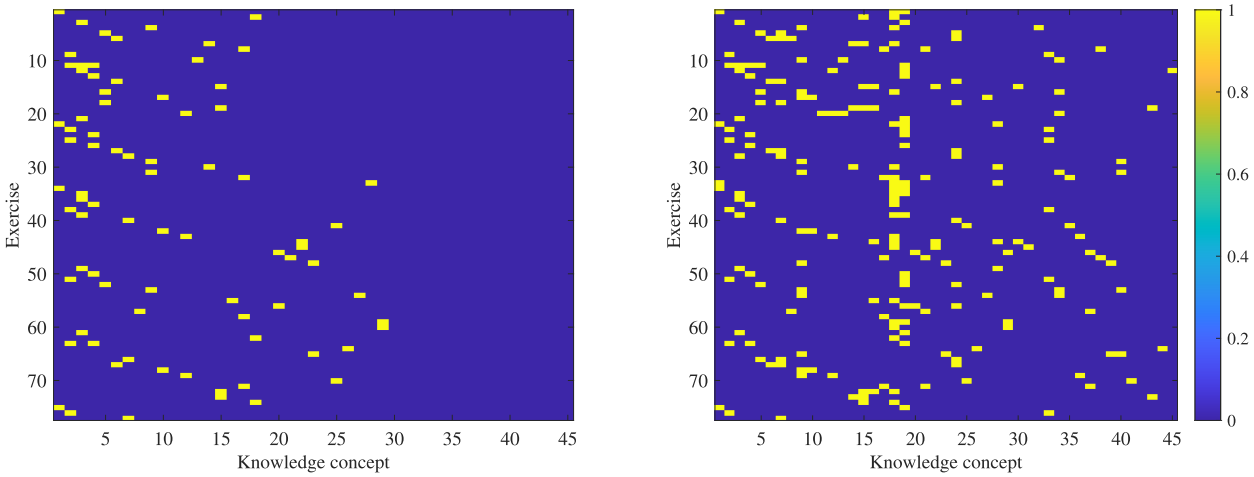
5) COMPARATIVE ANALYSIS OF Q MATRIX OPTIMIZATION BEFORE AND AFTER

To validate the capacity of the large model in the extraction of knowledge concepts, the Q matrix was optimized, enabling a visual comparison of its state before and after optimization. As illustrated in Fig. 6, the optimized Q matrix shows significant improvements in both the quantity and the coverage of knowledge concepts. This enhancement enables a more comprehensive evaluation of students' understanding across various knowledge concepts, thereby increasing the accuracy and robustness of the cognitive diagnostic model. It is important to note that this study encompasses a total of 177 exercises; however, only 80 have been selected for visual presentation to clearly illustrate the effects of optimization. The experimental results further indicate that the large-scale model exhibits strong generalization capabilities during knowledge concept extraction, effectively addressing diverse types of exercises while uncovering deeper knowledge structures. This ability is crucial for developing more precise cognitive diagnostic models and lays a solid foundation for future educational applications.

In general, the optimized Q matrix not only validates the knowledge extraction capabilities of the large model but also provides new insights and methods to improve cognitive diagnosis models.

6) QUANTITATIVE ANALYSIS OF THE Q-MATRIX

The quality of the Q matrix directly impacts the accurate prediction of students' mastery of knowledge points. To evaluate the optimized Q matrix, this study conducts a quantitative analysis of knowledge points that were manually annotated and those extracted by LLAMA3-70B, focusing on two key aspects: coverage of knowledge points and their distribution within exercises, as shown in Table 6. Compared to the total number of manually annotated knowledge points, which is 183, LLAMA3-70B extracted 442 knowledge points. This demonstrates that LLAMA3-70B can identify a broader range of implicit or hard-to-annotate knowledge areas, thereby significantly enhancing the Q matrix's coverage and addressing potential



(a) $\mathbf{Q}$-matrix (b) $\tilde{\mathbf{Q}}$-matrix

FIGURE 6. Comparative analysis of Q matrix optimization before and after.

omissions in manual annotations. Additionally, while the average number of manually annotated knowledge points per exercise is 1.03, LLAMA3-70B extracted an average of 2.49 knowledge points per exercise. The analysis further highlights that manually annotated knowledge points are unevenly distributed across exercises, whereas the knowledge points extracted by LLAMA3-70B exhibit a more balanced distribution, reflecting the underlying structure of exercises more accurately. In conclusion, LLAMA3-70B substantially contributes to optimizing the Q matrix by improving both coverage and the balance of knowledge point distribution, thereby enhancing the overall quality of the cognitive diagnosis model.

TABLE 6. Quantitative analysis of the Q matrix.

Annotation method	Knowledge Points	AVG
Manual annotation	183	1.03
Large Language Model	442	2.49

VI. DISCUSSION

This study proposes a method for extracting knowledge points and enhancing the Q matrix based on the LLAMA3-70B generative AI model, aimed at improving the accuracy of cognitive diagnostic models. Experimental results indicate that the generated knowledge points can effectively optimize the Q matrix, thereby increasing the model's assessment accuracy regarding students' mastery of knowledge. However, several challenges persist. The application of generative AI models in education inevitably encounters the "hallucination" problem; specifically, generated knowledge points may be inaccurate or misaligned with educational objectives [20], [21], [22], [23], [24]. Although this study mitigates the hallucination issue to some extent by designing targeted prompts to guide large language models in knowledge generation, these measures are still insufficient to eliminate potential risks. Future research should further investigate the underlying causes of hallucinations, such as the model's reasoning pathways and information integration processes within task contexts, while also optimizing prompt design and generation mechanisms. By incorporating feedback from domain experts alongside internal diagnostic tools within the model, comprehensive strategies can be developed to enhance both the accuracy and educational relevance of generated knowledge points.

Knowledge point extraction and Q-matrix optimization are critical components of cognitive diagnostic models. While this study has enhanced the Q-matrix to improve the accuracy of assessing students' knowledge mastery, there remain instances where relevant knowledge points for certain exercises are either incompletely or inaccurately extracted. This underscores the necessity for further optimization of the algorithm, as well as the development of extraction methods that cater to a diverse array of exercise types. Future research should prioritize enhancing these processes to ensure comprehensive and precise knowledge point extraction across a broader spectrum of educational materials.

Although the methods employed in this study primarily focus on exercises and knowledge points within a specific domain, their potential extends well beyond these confines. Future research may explore how to adapt model architectures and prompt designs to accommodate the unique characteristics of various disciplines (such as operating systems, software engineering, etc.), thereby broadening their applicability across multiple fields. Concurrently, integrating multimodal data (including student behavior records and learning videos) with other educational technologies (such as intelligent recommendation systems) is expected to further enhance the model's capacity for comprehensive analysis of student learning behaviors and improve personalized learning outcomes.

As generative AI models gain traction in education, concerns regarding student data privacy protection and algorithmic fairness have increasingly emerged. Moving forward, it is imperative to implement stringent data protection measures that ensure the security and anonymity of student information during its collection, processing, and storage phases. Moreover, an interpretable algorithmic framework should be developed to elucidate the rationale behind model decisions while mitigating potential biases and inequities. To ensure the responsible application of these technologies, it is also essential to evaluate educational scenarios involving generative AI from an ethical perspective and establish clear technical guidelines that promote a harmonious integration of technological advancements with educational practices.

From the perspective of practical management significance (PMS), the findings of this study provide substantial value to educational management practices. By enhancing the Q-matrix to improve the accuracy of cognitive diagnostic models, educators, and administrators can gain deeper insights into students' knowledge gaps, thereby optimizing teaching resource allocation. For instance, based on precise cognitive diagnostic results, teachers can design targeted personalized learning plans and offer more effective tutoring support to students while also increasing the relevance of classroom instruction. Furthermore, university administrators can modify course structures and assessment systems to scientifically enhance students' learning experiences and overall teaching effectiveness. In higher education settings, such improvements enable instructors to identify student needs more efficiently, promote individualized instruction, and achieve better educational outcomes with limited teaching resources.

To validate the practical applicability of this method, this study designed a scenario evaluation example. In a data structure course at a university, instructors utilized knowledge points generated by LLAMA3-70B to refine the Q-matrix and applied cognitive diagnostic models to assess students' mastery of knowledge. The results indicated that instructors could accurately identify students' weaknesses and significantly improve their final exam performance through personalized review materials—further substantiating the effectiveness of this approach in educational practice.

In conclusion, while this study has initially demonstrated the potential of LLAMA3-70B in optimizing the Q-matrix and enhancing the performance of cognitive diagnostic models, numerous challenges and unanswered questions remain. Future research should prioritize improving the accuracy of knowledge point extraction, expanding the application of this method across interdisciplinary and multimodal data, and addressing critical issues such as data privacy protection and algorithmic ethics. Through further exploration and optimization in these areas, generative AI models are anticipated to yield significant breakthroughs and practical value for the advancement of educational technology.

VII. CONCLUSION

This paper introduces an innovative cognitive diagnostic model (LLM-CDM) designed to enhance the precise assessment of students' knowledge mastery. By integrating large language models (LLMs) with prompt engineering, this model can automatically extract knowledge concepts from exercise texts and effectively incorporate them into the existing Q matrix, thereby significantly improving the accuracy of cognitive diagnosis. Experimental results indicate that comparative experiments conducted on real-world datasets demonstrate substantial advantages in predictive performance for this model, validating the effectiveness of employing prompts to guide large language models in extracting and optimizing the Q matrix to enhance diagnostic accuracy. While this research acknowledges certain limitations, such as challenges in accurately interpreting highly complex or ambiguous exercise texts, it offers novel insights and methodological advancements in cognitive diagnosis for educational assessment. This study highlights the promising application prospects of large language models in the realm of educational intelligence.

REFERENCES

[1] S. S. Mohamed, N. B. Al Barghuthi, and H. Said, "An analytical study towards the UAE universities smart education innovated approaches," in *Proc. IEEE 19th Int. Conf. High Perform. Comput. Commun., IEEE 15th Int. Conf. Smart City, IEEE 3rd Int. Conf. Data Sci. Syst. (HPCC/SmartCity/DSS)*, Dec. 2017, pp. 200–205.

[2] C. Li, "Development and application of assessment tools based on cognitive diagnosis," in *Proc. 3rd Int. Conf. Inf. Sci. Educ. (ICISE-IE)*, Nov. 2022, pp. 98–102.

[3] H. L. A. Gazca, S. S. I. Parra, H. G. L. Sánchez, Z. A. Omar, and G. D. I. Gaona, "Cross-sectional study of digital competences in the school trajectory higher education students (e-skills)," in *Proc. IEEE Int. Conf. Eng. Veracruz (ICEV)*, vol. 1, Oct. 2019, pp. 1–6.

[4] W. Wang, H. Ma, Y. Zhao, Z. Li, and X. He, "Tracking knowledge proficiency of students with calibrated Q-matrix," *Exp. Syst. Appl.*, vol. 192, Apr. 2022, Art. no. 116454. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417421017383

[5] L. DiBello, L. Roussos, and W. Stout, "31A review of cognitively diagnostic assessment and a summary of psychometric models," *Handbook Statist.*, vol. 26, pp. 979–1030, Dec. 2006.

[6] J. Liu, G. Xu, and Z. Ying, "Data-driven learning of Q-matrix," *Appl. Psychol. Meas.*, vol. 36, no. 7, pp. 548–564, Oct. 2012.

[7] J. de la Torre, "DINA model and parameter estimation: A didactic," *J. Educ. Behav. Statist.*, vol. 34, no. 1, pp. 115–130, Mar. 2009.

[8] S. E. Embretson and S. P. Reise. (2000). *Item Response Theory for Psychologists*. [Online]. Available: https://api.semanticscholar.org/CorpusID

[9] M. D. Reckase. (2009). *Multidimensional Item Response Theory Models*. [Online]. Available: https://api.semanticscholar.org/CorpusID

[10] F. Wang, Q. Liu, E. Chen, Z. Huang, Y. Yin, S. Wang, and Y. Su, "NeuralCD: A general framework for cognitive diagnosis," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 8, pp. 8312–8327, Aug. 2023.

[11] M. Feng, N. Heffernan, and K. Koedinger, "Addressing the assessment challenge with an online system that tutors as it assesses," *User Model. User-Adapted Interact.*, vol. 19, no. 3, pp. 243–266, Aug. 2009.

[12] Q. Liu, R. Wu, E. Chen, G. Xu, Y. Su, Z. Chen, and G. Hu, "Fuzzy cognitive diagnosis for modelling examinee performance," *ACM Trans. Intell. Syst. Technol.*, vol. 9, no. 4, pp. 1–26, Jul. 2018.

[13] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutiérrez, T. F. Tan, and D. S. W. Ting, "Large language models in medicine," *Nature Med.*, vol. 29, no. 8, pp. 1930–1940, Jul. 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID

[14] I.-C. Chiu and M.-W. Hung, "Finance-specific large language models: Advancing sentiment analysis and return prediction with LLaMA 2," *Pacific-Basin Finance J.*, vol. 90, Apr. 2025, Art. no. 102632. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0927538X24003846

[15] S. Cui, S. Yu, H.-Y. Huang, Y.-C.-D. Lin, Y. Huang, B. Zhang, J. Xiao, H. Zuo, J. Wang, Z. Li, G. Li, J. Ma, B. Chen, H. Zhang, J. Fu, L. Wang, and H.-D. Huang, "MiRTarBase 2025: Updates to the collection of experimentally validated microRNA–target interactions," *Nucleic Acids Res.*, vol. 53, no. 1, pp. 147–156, Jan. 2025, doi: 10.1093/nar/gkae1072.

[16] S. Al Faraby, A. Romadhony, and Adiwijaya, "Analysis of LLMs for educational question classification and generation," *Comput. Educ., Artif. Intell.*, vol. 7, Dec. 2024, Art. no. 100298. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666920X24001012

[17] Y. Zhang, M. Zhong, S. Ouyang, Y. Jiao, S. Zhou, L. Ding, and J. Han, "Automated mining of structured knowledge from text in the era of large language models," in *Proc. 30th ACM SIGKDD Conf. Knowl. Discovery Data Mining*. New York, NY, USA: Association for Computing Machinery, Aug. 2024, pp. 6644–6654, doi: 10.1145/3637528.3671469.

[18] M. Wan, T. Safavi, S. K. Jauhar, Y. Kim, S. Counts, J. Neville, S. Suri, C. Shah, R. W. White, L. Yang, R. Andersen, G. Buscher, D. Joshi, and N. Rangan, "TnT-LLM: Text mining at scale with large language models," in *Proc. 30th ACM SIGKDD Conf. Knowl. Discovery Data Mining*. New York, NY, USA: Association for Computing Machinery, Aug. 2024, pp. 5836–5847, doi: 10.1145/3637528.3671647.

[19] H. C. Jami, P. R. Singh, A. Kumar, B. R. Bakshi, M. Ramteke, and H. Kodamana, "CCU-llama: A knowledge extraction LLM for carbon capture and utilization by mining scientific literature data," *Ind. Eng. Chem. Res.*, vol. 63, no. 41, pp. 17585–17598, Oct. 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID

[20] Y. Xiao and W. Y. Wang, "On hallucination and predictive uncertainty in conditional language generation," in *Proc. 16th Conf. Eur. Chapter Assoc. Comput. Linguistics, Main Volume*, P. Merlo, J. Tiedemann, and R. Tsarfaty, Eds., 2021, pp. 2734–2744. [Online]. Available: https://aclanthology.org/2021.eacl-main.236

[21] A. Rohrbach, L. A. Hendricks, K. Burns, T. Darrell, and K. Saenko, "Object hallucination in image captioning," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, Eds., Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 4035–4045. [Online]. Available: https://aclanthology.org/D18-1437

[22] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," *ACM Comput. Surv.*, vol. 55, no. 12, pp. 1–38, Mar. 2023, doi: 10.1145/3571730.

[23] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald, "On faithfulness and factuality in abstractive summarization," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds., 2020, pp. 1906–1919. [Online]. Available: https://aclanthology.org/2020.acl-main.173

[24] K. Filippova, "Controlled hallucinations: Learning to generate faithfully from noisy data," in *Proc. Findings Assoc. Comput. Linguistics*, T. Cohn, Y. He, and Y. Liu, Eds., 2020, pp. 864–870. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.76

[25] AI@Meta. (2024). *Llama 3 Model Card*. [Online]. Available: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md

[26] T. Qi, M. Ren, L. Guo, X. Li, J. Li, and L. Zhang, "ICD: A new interpretable cognitive diagnosis model for intelligent tutor systems," *Exp. Syst. Appl.*, vol. 215, Apr. 2023, Art. no. 119309. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417422023272

[27] L. Gao, Z. Zhao, C. Li, J. Zhao, and Q. Zeng, "Deep cognitive diagnosis model for predicting students' performance," *Future Gener. Comput. Syst.*, vol. 126, pp. 252–262, Jan. 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167739X21003277

[28] W. Gao, Q. Liu, Z. Huang, Y. Yin, H. Bi, M.-C. Wang, J. Ma, S. Wang, and Y. Su, "RCD: Relation map driven cognitive diagnosis for intelligent education systems," in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.* New York, NY, USA: Association for Computing Machinery, Jul. 2021, pp. 501–510, doi: 10.1145/3404835.3462932.

[29] X. Wang, C. Huang, J. Cai, and L. Chen, "Using knowledge concept aggregation towards accurate cognitive diagnosis," in *Proc. 30th ACM Int. Conf. Inf. Knowl. Manage.* New York, NY, USA: Association for Computing Machinery, Oct. 2021, pp. 2010–2019, doi: 10.1145/3459637.3482311.

[30] S. Cheng, Q. Liu, E. Chen, Z. Huang, Z. Huang, Y. Chen, H. Ma, and G. Hu, "DIRT: Deep learning enhanced item response theory for cognitive diagnosis," in *Proc. 28th ACM Int. Conf. Inf. Knowl. Manage.* New York, NY, USA: Association for Computing Machinery, Nov. 2019, pp. 2397–2400, doi: 10.1145/3357384.3358070.

[31] H. Yang, T. Qi, J. Li, L. Guo, M. Ren, L. Zhang, and X. Wang, "A novel quantitative relationship neural network for explainable cognitive diagnosis model," *Knowl.-Based Syst.*, vol. 250, Aug. 2022, Art. no. 109156. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705122005755

[32] J. Ling and M. Afzaal, "Automatic question-answer pairs generation using pre-trained large language models in higher education," *Comput. Educ., Artif. Intell.*, vol. 6, Jun. 2024, Art. no. 100252. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666920X24000559

[33] K. Mohamed, M. Yousef, W. Medhat, E. H. Mohamed, G. Khoriba, and T. Arafa, "Hands-on analysis of using large language models for the auto evaluation of programming assignments," *Inf. Syst.*, vol. 128, Feb. 2025, Art. no. 102473. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306437924001315

[34] A. W. Ou, C. Stöhr, and H. Malmström, "Academic communication with AI-powered language tools in higher education: From a post-humanist perspective," *System*, vol. 121, Apr. 2024, Art. no. 103225. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0346251X24000071

[35] T. Alqahtani, H. A. Badreldin, M. Alrashed, A. I. Alshaya, S. S. Alghamdi, K. Bin Saleh, S. A. Alowais, O. A. Alshaya, I. Rahman, M. S. Al Yami, and A. M. Albekairy, "The emergent role of artificial intelligence, natural learning processing, and large language models in higher education and research," *Res. Social Administ. Pharmacy*, vol. 19, no. 8, pp. 1236–1242, Aug. 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1551741123002802

[36] G.-G. Lee, E. Latif, X. Wu, N. Liu, and X. Zhai, "Applying large language models and chain-of-thought for automatic scoring," *Comput. Educ., Artif. Intell.*, vol. 6, Jun. 2024, Art. no. 100213. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666920X24000146

[37] C. Ng and Y. Fung, "Educational personalized learning path planning with large language models," 2024, *arXiv:2407.11773*.

[38] Y. Zhang, Y. Pan, T. Zhong, P. Dong, K. Xie, Y. Liu, H. Jiang, Z. Wu, Z. Liu, W. Zhao, W. Zhang, S. Zhao, T. Zhang, X. Jiang, D. Shen, T. Liu, and X. Zhang, "Potential of multimodal large language models for data mining of medical images and free-text reports," *Meta-Radiol.*, vol. 2, no. 4, Dec. 2024, Art. no. 100103. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2950162824000572

[39] S. Maity and A. Deroy, "Generative AI and its impact on personalized intelligent tutoring systems," Oct. 2024, *arXiv:2410.10650*.

[40] H. Liu, H. Yin, Z. Luo, and X. Wang, "Integrating chemistry knowledge in large language models via prompt engineering," *Synth. Syst. Biotechnol.*, vol. 10, no. 1, pp. 23–38, Jul. 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2405805X24001029

[41] S. Zhao and X. Sun, "Enabling controllable table-to-text generation via prompting large language models with guided planning," *Knowl.-Based Syst.*, vol. 304, Nov. 2024, Art. no. 112571. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S095070512401205X

[42] D. Yigci, M. Eryilmaz, A. K. Yetisen, S. Tasoglu, and A. Ozcan, "Large language model-based chatbots in higher education," *Adv. Intell. Syst.*, Aug. 2024, doi: 10.1002/aisy.202400429.

[43] H. Nori, N. King, S. M. McKinney, D. Carignan, and E. Horvitz, "Capabilities of gpt-4 on medical challenge problems," 2023, *ArXiv:2303.13375*. [Online]. Available: https://api.semanticscholar.org/CorpusID:257687695

[44] Y. Xia, F. Kong, T. Yu, L. Guo, R. A. Rossi, S. Kim, and S. Li, "Which LLM to play? Convergence-aware online model selection with time-increasing bandits," in *Proc. ACM Web Conf.* New York, NY, USA: Association for Computing Machinery, May 2024, pp. 4059–4070, doi: 10.1145/3589334.3645420.

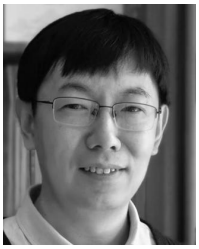
[45] O. B. Shoham and N. Rappoport, "MedConceptsQA: Open source medical concepts QA benchmark," *Comput. Biol. Med.*, vol. 182, Nov. 2024, Art. no. 109089. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482524011740

[46] W. Zhang, Q. Wang, X. Kong, J. Xiong, S. Ni, D. Cao, B. Niu, M. Chen, Y. Li, R. Zhang, Y. Wang, L. Zhang, X. Li, Z. Xiong, Q. Shi, Z. Huang, Z. Fu, and M. Zheng, "Fine-tuning large language models for chemical text mining," *Chem. Sci.*, vol. 15, no. 27, pp. 10600–10611, 2024.

[47] H. Pei, B. Yang, J. Liu, and L. Dong, "Group sparse Bayesian learning for active surveillance on epidemic dynamics," in *Proc. AAAI Conf. Artif. Intell.*, 2018, vol. 32, no. 1, pp. 800–807.
[48] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognit.*, vol. 30, no. 7, pp. 1145–1159, Jul. 1997. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320396001422
[49] A. Bahrini, M. Khamoshifar, H. Abbasimehr, R. J. Riggs, M. Esmaeili, R. M. Majdabadkohne, and M. Pasehvar, "ChatGPT: Applications, opportunities, and threats," 2023, *arXiv:2304.09103*.
[50] T. Glm et al., "ChatGLM: A family of large language models from GLM-130B to GLM-4 all tools," 2024, *arXiv:2406.12793*.
[51] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. Le Scao, T. Lavril, T. Wang, T. Lacroix, and W. El Sayed, "Mistral 7B," 2023, *arXiv:2310.06825*.
[52] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," 2019, *arXiv:1910.10683*.

XIN CHEN received the B.S. and M.S. degrees in computer science and technology from the College of Computer Science and Technology, Shandong University of Science and Technology, Qingdao, China, in 1998 and 2006, respectively, and the Ph.D. degree in computer software and theory from the College of Information Science and Engineering, Shandong University of Science and Technology, in 2017. She is currently an Associate Professor with the College of Computer Science and Engineering, Shandong University of Science and Technology. Her research interests include deep learning and the application of big data technologies.

JIN ZHANG received the bachelor's degree from Yanshan College, Shandong University of Finance and Economics, Jinan, China, in 2022. She is currently pursuing the master's degree with the College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao, China. Her current research interests include deep learning and cognitive diagnosis.



TONG ZHOU received the B.S. degree in architecture from Shandong Jianzhu University, Jinan, China, in 1999, and the M.S. and Ph.D. degrees in architectural design and theory from the College, Tianjin University, Tianjin, China, in 2006 and 2014, respectively. He is currently an Associate Professor with the College of Civil Engineering and Architecture, Shandong University of Science and Technology. His research interests include big data analysis in urban renewal and green building evaluation.

FENG ZHANG received the B.S. and M.S. degrees in computer science and technology from the College of Computer Science and Technology, Shandong University, Jinan, China, in 2002 and 2005, respectively, and the Ph.D. degree in computer software and theory from the College of Information Science and Engineering, Shandong University of Science and Technology, Qingdao, China, in 2014. He is currently an Associate Professor with the College of Computer Science and Engineering, Shandong University of Science and Technology. His research interests include machine learning and business process management.