Front. Digit. Educ., 2025, 2(2): 20
https://doi.org/10.1007/s44366-025-0057-8

RESEARCH ARTICLE

LLM-Driven Cognitive Diagnosis with SOLO Taxonomy: A Model-Agnostic Framework

Zhiang Dong, Jingyuan Chen, Fei Wu

College of Computer Science and Technology, Zhejiang University, Hangzhou 310058, China

Abstract With the development of the Internet and intelligent education systems, the significance of cognitive diagnosis has become increasingly acknowledged. Cognitive diagnosis models (CDMs) aim to characterize learners' cognitive states based on their responses to a series of exercises. However, conventional CDMs often struggle with less frequently observed learners and items, primarily due to limited prior knowledge. Recent advancements in large language models (LLMs) offer a promising avenue for infusing rich domain information into CDMs. However, integrating LLMs directly into CDMs poses significant challenges. While LLMs excel in semantic comprehension, they are less adept at capturing the fine-grained and interactive behaviours central to cognitive diagnosis. Moreover, the inherent difference between LLMs' semantic representations and CDMs' behavioural feature spaces hinders their seamless integration. To address these issues, this research proposes a model-agnostic framework to enhance the knowledge of CDMs through LLMs extensive knowledge. It enhances various CDM architectures by leveraging LLM-derived domain knowledge and the structure of observed learning outcomes taxonomy. It operates in two stages: first, LLM diagnosis, which simultaneously assesses learners via educational techniques to establish a richer and a more comprehensive knowledge representation; second, cognitive level alignment, which reconciles the LLM's semantic space with the CDM's behavioural domain through contrastive learning and mask-reconstruction learning. Empirical evaluations on multiple real-world datasets demonstrate that the proposed framework significantly improves diagnostic accuracy and underscoring the value of integrating LLM-driven semantic knowledge into traditional cognitive diagnosis paradigms.

Received January 5, 2025; revised February 20, 2025; accepted March 3, 2025

Jingyuan Chen (✉)
E-mail: jingyuanchen@zju.edu.cn
Fei Wu (✉)
E-mail: wufei@zju.edu.cn

Keywords large language models, cognitive diagnosis models, intelligent education system, SOLO taxonomy, knowledge representation

1 Introduction

The advancement of online intelligent education systems has placed growing emphasis on algorithms and models for automated evaluation. Cognitive diagnosis, as illustrated in Figure 1, assesses student's learning proficiency by examining their responses to a series of exercises, and thus serves as a foundational component of intelligent education platforms. The insights gained from cognitive diagnosis are central to a range of educational applications, including adaptive educational recommendation and computerized adaptive testing (Bi et al., 2020; Huang et al., 2019; Zhuang et al., 2022). As a result, ensuring accurate and reliable cognitive diagnosis is imperative for improving the overall effectiveness of these educational technologies.

Traditional cognitive diagnosis models (CDMs) primarily stem from psychometric theories and utilize manually crafted interaction functions informed by principles of psychometrics and educational theories. Recent advances in deep learning stimulate a new generation of CDMs that utilize neural networks to effectively obtain collaborative information, such as student–exercise interactions (Gao et al., 2021; Wang et al., 2020). Nevertheless, these existing CDMs struggle when diagnosing students and exercises with limited prior interactions, a challenge commonly termed the cold-start problem. This issue arises primarily from an absence of pre-existing knowledge within the models, which impedes their capacity to adapt to unfamiliar conditions. As illustrated in Figure 1, empirical results on the PTADisc dataset highlight the diminished performance of current CDMs in cold scenarios and ultimately undermine their overall diagnostic effectiveness (Hu et al., 2023).

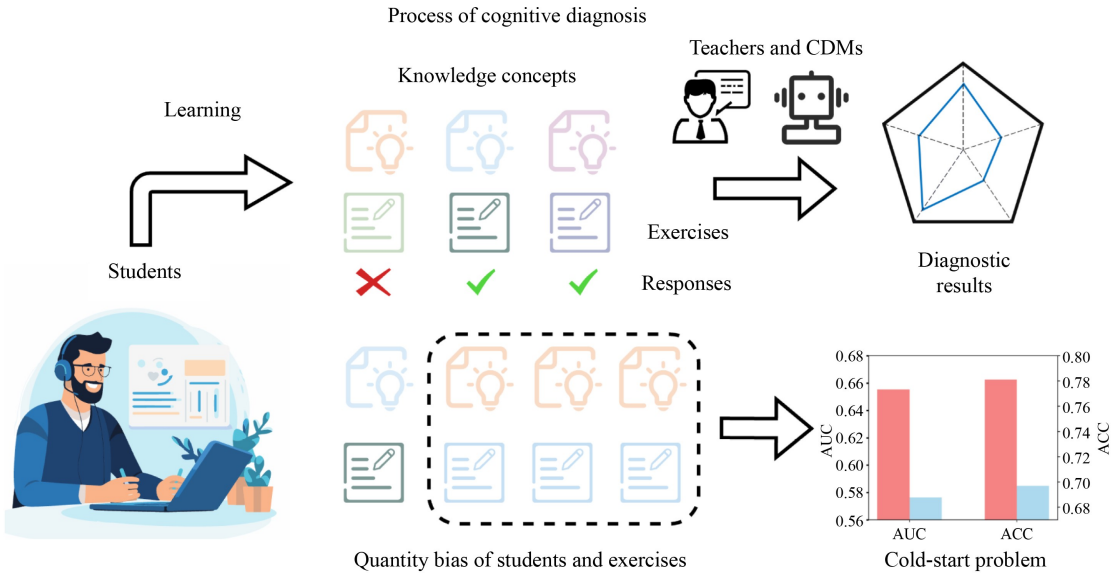

Figure 1 Cognitive diagnosis and the cold-start problem. CDMs: cognitive diagnosis models, AUC: area under curve, ACC: accuracy.

Recent advancements in large language models (LLMs) have demonstrated their exceptional capabilities in logical reasoning and text comprehension, as evidenced by their success across a wide range of applications (Abbasiantaeb et al., 2024; Dong et al., 2025; Wang et al., 2024a; Xu et al., 2024; Zhang et al., 2024a; Zhu et al., 2024). The extensive prior knowledge encoded within LLMs presents a promising avenue for addressing the limitations of current CDMs. In particular, LLMs capitalize on rich conceptual understanding and intricate interrelations among knowledge domains to gain nuanced insights into learning behaviours and exercise attributes. By drawing on these knowledge-rich representations, LLMs emulate the reasoning processes of human educators and offer more accurate diagnoses even in cold-start scenarios. Consequently, this research aims to integrate LLMs with CDMs effectively to enhance diagnostic performance.

However, integrating LLMs with CDMs is a non-trivial endeavour for two reasons. First, LLMs struggle to capture the fine-grained collaborative information essential for modelling student–exercise interactions. The inherent input length limitations of LLMs restrict the inclusion of detailed textual data on relationships among exercises, knowledge concepts, and students. Second, LLMs and CDMs operate in fundamentally different representation spaces. While LLMs process textual content within a semantic space, CDMs focus on behavioural features derived from student interaction patterns. Achieving effective integration therefore requires bridging the gap between these semantic and behavioural domains.

To address these challenges, this research presents a novel model-agnostic framework that integrates LLMs into conventional CDMs, thereby bridging the semantic space of LLMs with the behavioural space of CDMs. The framework comprises two key modules, LLM diagnosis and cognitive level alignment. In the LLM diagnosis module, the research harnesses LLMs to emulate human educators and diagnoses students' learning status, as well as the characterization of exercise attributes. This stage leverages LLMs to consolidate collaborative information about students and exercises, followed by an in-depth analysis of student response logs through the structure of observed learning outcomes (SOLO) taxonomy (Biggs & Collis, 1982). SOLO taxonomy is a framework for classifying learning outcomes based on their complexity and quality. It delineates progressive levels of understanding and guides educators in designing instructional activities and assessments to facilitate learners from surface-level understanding to a deeper, more integrative, and more creative comprehension. These analyses yield textual diagnoses that reflect learners' cognitive states and exercise attributes. Cognitive level alignment module aligns these LLM-generated textual diagnoses which are rooted in the semantic domain with the behavioural representations employed by CDMs. This alignment results in refined cognitive representations and ultimately enhances the diagnostic accuracy and adaptability of the integrated system.

Nevertheless, this enhancement inevitably incurs additional computational overhead. The alignment process introduces extra computational costs, maximizes mutual information from an information-theoretic perspective, and increases

computational complexity. Moreover, due to the characteristics of LLMs, real-time feedback in adaptive learning systems introduces biases, as this is not the model's strength. Experimental results show that the method performs well on large-scale datasets. It demonstrates that despite the additional overhead, it achieves significant advantages and makes the computational cost worthwhile. Furthermore, the context window limitations of LLMs may pose challenges for scaling the model to larger-scale datasets.

Four contributions of this work are summarized. First, the research introduces the framework in a model-agnostic approach that combines the complementary strengths of LLMs and CDMs to deliver more accurate and robust cognitive diagnoses. Second, the research presents the LLM diagnosis module, which integrates both collaborative information and student response logs via SOLO taxonomy to produce detailed textual diagnoses for students and exercises. Third, the research introduces the cognitive level alignment module, which aligns the textual diagnoses generated by LLMs with the behavioural representations utilized by CDMs through mixed contrastive learning and interactive reconstruction learning. Fourth, the research evaluates the proposed framework using multiple public datasets across various CDM architectures, and the results demonstrate its effectiveness in enhancing cognitive diagnosis performance.

2 Related Work

2.1 | Cognitive Diagnosis

Cognitive diagnosis, initially rooted in educational psychology, is a foundational task within intelligent education. Its primary objective is to model students' learning states and knowledge proficiency by assessing their feedback to a series of questions (Liu, 2021). Existing cognitive diagnosis methodologies can be broadly categorized into two groups, psychometric theory-based methods and neural network-based methods (Bi et al., 2023; de la Torre, 2009; Gao et al., 2021; Liu et al., 2024a; Lord, 1952; Reckase, 2009; Wang et al., 2020; Wang et al., 2023; Wang et al., 2024b; Zhang et al., 2024b). Psychometric theory-based methods, such as item response theory (IRT) (Lord, 1952), multidimensional IRT (MIRT) (Reckase, 2009), and the deterministic inputs, noisy, "and" gate (DINA) model (de la Torre, 2009), rely on established psychological theories to infer students' latent proficiency factors. In contrast, neural network-based methods leverage deep learning techniques to model students' learning states. Pioneering work, like neural cognitive diagnosis (NCD), introduced neural networks into cognitive diagnosis and enabled finer-grained modelling of student–exercise interactions (Wang et al., 2020). Building on this foundation, relation map-driven cognitive diagnosis (RCD) (Gao et al., 2021), and relation-guided dual-side graph transformer (RDGT) applied graph-based architectures to uncover intricate relationships among exercises (Yu et al., 2024), knowledge concepts, and students. More recently, Bayesian mETA-learned cognitive diagnosis framework employed meta-learning strategies to rapidly adapt diagnoses for new students (Bi et al., 2023), while affect-aware cognitive diagnosis (ACD) integrated affective states into the cognitive diagnosis process, thereby expanding the range of factors considered (Wang et al., 2024b). The path-specific causal reasoning for fairness-aware cognitive diagnosis employs causal reasoning to eliminate the negative impact of sensitive information in the process of cognitive diagnosis (Zhang et al., 2024a). Wang et al. (2024a) proposed a method to address uncertainty in cognitive diagnosis and utilized a unified uncertainty estimation approach that was applied to a wide range of cognitive diagnostic models.

However, most existing cognitive diagnosis models pay insufficient attention to incorporating prior knowledge, which impedes to delivery of accurate diagnoses, particularly in scenarios where student and exercise data is scarce.

2.2 | Large Language Models

With the emergence of the transformer architecture (Vaswani et al., 2017), LLMs with extensive parameters and large-scale training corpora are prominent. Typically, LLMs are trained through a pre-training and fine-tuning paradigm, which allows them to adapt to a wide range of downstream tasks. They have substantially advanced the state of the art in various natural language processing applications, such as text summarization (Laskar et al., 2022; Zhang et al., 2023b), sentiment analysis (Deng et al., 2023; Hoang et al., 2019), machine translation (Moslem et al., 2023; Zhang et al., 2023a), and multimodal understanding (Huang et al., 2024; Wu et al., 2024).

Given the strong comprehension and reasoning capabilities, as well as the vast knowledge repositories, LLMs show considerable potential in educational domains. In particular, LLMs offer researchers novel insights by emulating the roles of teachers or students (Li et al., 2023; Liu et al., 2024a; Liu et al., 2024b; Liu et al., 2024c; Wang et al., 2024a; Xu et al., 2024), and generating educational resources (Dai et al., 2024; Lin et al., 2024a; Lin et al., 2024b; Lin et al., 2024c). Nevertheless, the application of LLMs to cognitive diagnosis remains relatively underexplored. The established successes of LLMs in text summarization and other educational settings suggest that they are well-positioned to advance cognitive diagnosis tasks.

3 Methodology

This research first formalizes the task and outlines the overall framework, followed by a detailed discussion of the specific strategies employed within our proposed approach.

3.1 | Task Formulation

Formally, let $S = \{s_1, s_2, \ldots, s_{|S|}\}$, $\mathscr{E} = \{e_1, e_2, \ldots, e_{|\mathscr{E}|}\}$, and $\mathcal{K} = \{k_1, k_2, \ldots, k_{|\mathcal{K}|}\}$ denote the sets of students, exercises, and knowledge concepts, respectively. The response logs $\boldsymbol{R}$ are represented as a collection of triplets $(s_i, e_j, k_j, r_{ij}) \in \boldsymbol{R}$, where r_{ij} indicates whether student (s_i) correctly answered exercise (e_j), and $k_j \subseteq \mathcal{K}$ denotes the subset of knowledge concepts associated with e_j. i denotes the student's number and j means the number of the exercises or knowledge. In some datasets, each exercise e_j may also be accompanied by textual content that serves as an attribute. The central aim of cognitive diagnosis is to assess students' mastery of various knowledge concepts by leveraging their response logs $\boldsymbol{R}$ to predict their performance on exercises.

3.2 | Framework Overview

The proposed framework comprises two essential modules, including LLM diagnosis and cognitive level alignment, as illustrated in Figure 2. This framework incorporates collaborative information while capitalizing on the extensive prior knowledge embedded in LLMs. Through the cognitive level alignment module, it bridges the gap between the semantic space of LLMs and the behavioural space of CDMs. By integrating the complementary strengths of both models, it aims to achieve more accurate and robust cognitive diagnostic performance.

The LLM diagnosis module operates in two sequential phases, relevant collaborative information collection and diagnosis generation. During the first phase, relevant collaborative information is extracted from the response logs. In the second phase, both the newly gathered collaborative information and the original response logs are utilized to assess students' cognitive statuses and the attributes of exercises. In this phase, the research employs the SOLO taxonomy to diagnose students' learning processes accurately (Biggs et al., 1982).

The cognitive level alignment module introduces these LLM-generated diagnoses into conventional CDMs, thereby enhancing the cognitive-level representations of both students and exercises. To achieve this, the module employs two distinct alignment strategies, mixed contrastive alignment, and interactive reconstruction alignment. This module ensures that textual diagnoses from the LLM's semantic

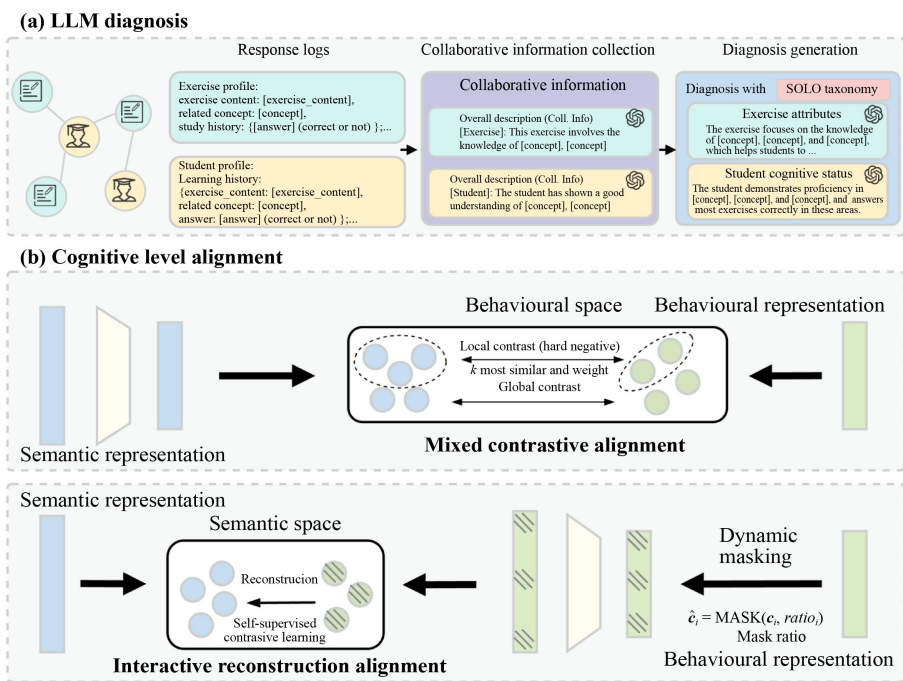

Figure 2 Framework overview: (a) LLM diagnosis module employs LLMs to generate textual diagnoses for both students and exercises, (b) cognitive level alignment module incorporates these LLM-derived diagnoses into conventional CDMs, facilitating a joint representation of students and exercises in both semantic and behavioural spaces. LLM: large language model, SOLO: structure of observed learning outcomes, CDMs: cognitive diagnosis models, Coll. Info: collaborative information.

domain are effectively integrated into the CDM's behavioural domain. Notably, the entire framework is model-agnostic, which allows for the selection of the most suitable CDM for any given educational setting and leads to more accurate diagnostic outcomes.

3.3 | LLM Diagnosis Module

LLMs can be more effectively guided through formulated natural language instructions, thereby producing higher-quality outputs. Two categories of input instructions have been employed, system prompt (M), and input prompt (P). The system prompt (M) outlines the tasks to be performed by the LLM, which details both the required inputs and the expected outputs. In contrast, the input prompt (P) provides specific input data, such as students' response logs.

3.3.1 Collaborative Information Collection

Inspired by the strategies employed by experienced teachers, who refine their evaluations by incorporating insights from both students and exercises, a collaborative information collection stage is utilised. This stage is designed to capture student-level collaborative information by examining each student's performance across all completed exercises, as well as exercise-level collaborative information by considering the aggregated performance of all students who have engaged with a particular exercise.

Specifically, for the student-focused diagnosis, the system prompts (M_s) define the format of the input prompts (P_s) and guide the LLM in generating textual collaborative information. The input prompts (P_s) include the exercise content (t), associated knowledge concepts (n), and the student's response (r) for all exercises (e) the student has completed.

Similarly, for exercise-focused diagnosis, the system prompt (M_e) defines the format of the input prompt (P_e). Following a similar pattern, P_e includes the exercise content (t), relevant knowledge concepts (n), and the response (r) from all students (S) who have attempted that exercise.

3.3.2 Diagnosis Generation

After collecting the collaborative information (I) for both students and exercises, the next phase involves generating diagnoses of students' cognitive statuses and exercise attributes. To achieve this, this research first merges the previously gathered collaborative information (I) for each student and exercise with the corresponding response logs, thereby forming a new input prompt (P') enriched with more comprehensive details.

Subsequently, this research adjusts the system prompt (M') to specify the revised format of P' and guides the LLMs to produce diagnoses for students' cognitive states and the attributes of exercises. Specifically, SOLO taxonomy is introduced to assess the learning process (Biggs et al., 1982). SOLO taxonomy provides a systematic framework for classifying the quality of learners' understanding. It is designed to guide educators in developing instructional activities and assessments that facilitate a progression from superficial knowledge toward deeper and more integrative comprehension. As shown in Figure 3, the taxonomy comprises five core levels of cognitive engagement: First, at the prestructural level, learners exhibit minimal relevant understanding, often characterized by confusion or reliance on irrelevant information. Second, at the unistructural level, learners grasp one or a limited number of aspects of a concept but lack a broader perspective or connection to other elements. Third, at the multistructural level, learners understand several important components independently, but they have yet to connect them into a coherent whole. Fourth, at the relational level, learners integrate multiple elements into a structured and meaningful pattern, which demonstrates how they

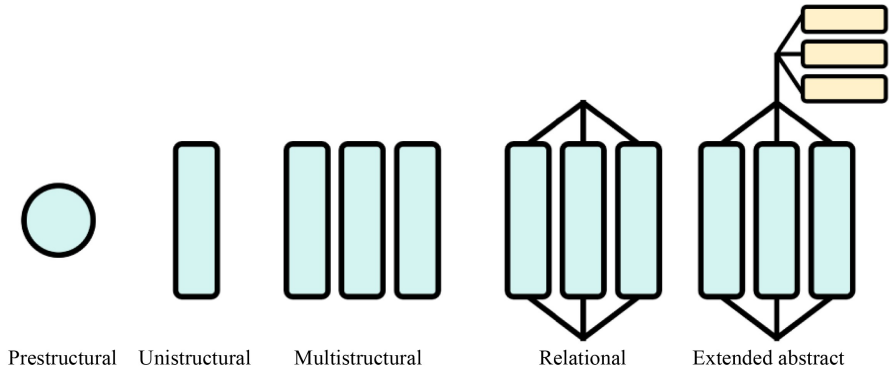

Figure 3 Illustration of the SOLO taxonomy. SOLO: structure of observed learning outcomes.

relate to one another. Fifth, at the extended abstract level, learners generalize and transfer their understandings beyond the immediate context. They theorize, apply concepts in new domains, and propose novel solutions, which demonstrate a capacity for abstraction and creative synthesis.

At the same time, it allows for a more comprehensive analysis of the questions themselves. Formally, the diagnoses T are obtained via T = LLMs(M', P'). The example of input prompts is provided in Figure 4.

With the assistance of SOLO taxonomy, LLMs attain a more fine-grained understanding of the problem, enabling a better analysis of the student's learning process. Firstly, when processing student-submitted records, LLMs extract key concepts and their interrelationships from the text through their knowledge analyses and comprehension capabilities. Secondly, by employing the SOLO framework, LLMs can systematically analyze the cognitive origins of erroneous responses. In data-sparse scenarios such as new questions and new students, traditional CDMs are limited due to the lack of interaction history. SOLO taxonomy provides a channel for the injection of prior

Prompt of students collaborative information collection

System prompt:
You will serve as an experienced teacher to help me determine the students' learning status.
I will provide you with information about exercises that the student has finished, described as study history, as well as his or her answer of those exercises.
Here are the instructions:
1. Each finished exercise contained in the Study history will be described in JSON format, with the following attributes:
{
'content': 'the content of the exercise',
'concept': 'the concept related to this exercise',
'answer': 'whether the student answer this exercise correctly or not'
}
2. The information I will give you:
Study history: a list of JSON strings describing the exercises that the students has finished in the format mentioned above.
Requirements:
1. Please provide your answer in JSON format and in English, following this structure:
{
"summarization": "A summarization of the students' learning status, including what types of exercises this user is good at or not good at, which aspect of knowledge needs to be strengthened and other necessary information for diagnosing students' learning status",
"reasoning": "briefly explain your reasoning for the summarization"
}
2. Please ensure that the "summarization" and "reasoning" is no longer than 200 words.
3. Do not provide any other text outside the JSON string.

Input prompt:
Study history: [{'content': '下列字符中，ASCII码最小的是 ()…', 'concept': '文本信息的表示; ', 'answer': '回答正确', …

Prompt of students diagnosis generation

System prompt:
You will serve as an experienced teacher to help me determine the students' learning status.
I will provide you with information about exercises that the student has finshed and their feedback as well as detailed description of the exercise. Based on these information, please diagnose the student's learning status from the teacher's perspective following SOLO taxonomy.
SOLO taxnomy:
SOLO taxonomy is a framework for classifying learning outcomes based on their complexity and quality. It delineates progressive levels of understanding, guiding educators in designing instructional activities and assessments that move learners from surface-level to deeper, more integrative and creative comprehension.
SOLO taxonomy comprises five core levels of understanding:
1. Prestructural: The learner displays little to minimal relevant understanding, often characterized by confusion or reliance on irrelevant information.
2. Unistructural: The learner grasps one or a limited number of aspects of a concept, but lacks a broader perspective or connection to other elements.
3. Multistructural: The learner understands several important elements independently but has yet to connect them into a coherent whole.
4. Relational: The learner integrates multiple elements into a structured and meaningful pattern, which demonstrates how they relate to one another.
5. Extended Abstract: The learner goes beyond what is given, generalizing ideas to new domains, theorizing,or creating novel solutions.
Here are the instructions:
1. Each finished exercise will be described in JSON format, with the following attributes:
{
'content': 'the content of the exercise',
'concept': 'the concept related to this exercise',
'answer': 'whether the student answer this exercise correctly or not',
'exercise_profile': 'the overall description of the exercise'
}
2. The information I will give you:
Basic information: a JSON string describing the basic information about the student.
Study history: a list of JSON strings describing the exercises that the students has finished.
Requirements:
1. Please provide your answer in JSON format and in English, following this structure:
{
"summarization": "A summarization of the students' learning status, including what types of the student answer this exercises this user is good at or not good at, which aspect of knowledge needs to be strengthened and other necessary information for diagnosing students' learning status",
"reasoning": "briefly explain your reasoning for the summarization"
}
2. Please ensure that the "summarization" and "reasoning" is no longer than 1,000 words and in string format.
3. Do not provide any other text outside the JSON string.

Input prompt:
Basic information:
The student has shown good understanding in topics related to text information representation, cloud computing basic concepts...
Study history:
[{'content': '下列字符中, ASCII码最小的是 ()…', 'concept': '文本信息的表示;', 'answer': '回答正确'
…

Prompt of exercises collaborative information collection

System prompt:
You will serve as an experienced teacher to help me summarize the characteristics of the exercise and what kind of students might correctly answer this exercise.
I will provide you with the Basic information including exercise content and related concept) of the exercise and also study history of students for it.
Here are the instructions:
1. The basic information will be described in JSON format, with the following attributes:
{
'exercise_content': 'the content of the exercise',
'related_concept': 'the concept related to this exercise'
}
Feedback from users will be managed in the following List format:
'history': [{'answer': 'whether the student correctly answer the exercise'...]
2. The information I will give you:
Basic information: a JSON string describing the basic information about the exercise.
Study history: a List object containing some feedbacks from students who have answered the exercise.
Requirements:
1. Please provide your answer in JSON format and in English, following this structure:
{
"summarization": "A summarization of the detailed characteristics of the exercise and what kind of students might correctly answer this exercise",
"reasoning": "briefly explain your reasoning for the summarization"
}
2. Please ensure that the "summarization" and "reasoning" is no longer than 200 words.
3. Do not provide any other text outside the JSON string.

Input prompt:
Basic information: 'exercise_content': '下列字符中, ASCII码最小的是 ()…', 'related_concept': '文本信息的表示;'} Study history: [{'answer': '回答正确',...

Prompt of exercises diagnosis generation

System prompt:
You will serve as an experienced teacher to help me summarize the characteristics of the exercise and what kind of students might correctly answer this exercise following SOLO taxonomy.

I will provide you with the Basic information, such as exercise content and related concept, of that exercise and also Study history of students for it. The students' feadback contains the students' overall knowledge status and whether they answer correctly.
SOLO taxnomy:
SOLO taxonomy is a framework for classifying learning outcomes based on their complexity and quality. It delineates progressive levels of understanding, guiding educators in designing instructional activities, and assessments that move learners from surface-level to deeper more integrative and creative comprehension.
SOLO taxonomy comprises five core levels of understanding:
1. Prestructural: The learner displays little to minimal relevant understanding, often characterized by confusion or reliance on irrelevant information.
2. Unistructural: The learner grasps one or a limited number of aspects of a concept, but lacks a broader perspective or connection to other elements.
3. Multistructural: The learner understands several important elements independently but has yet to connect them into a coherent whole.
4. Relational: The learner integrates multiple elements into a structured and meaningful pattern, which demonstrates how they relate to one another.
5. Extended abstract: The learner goes beyond what is given, generalizing ideas to new domains, theorizing, or creating novel solutions.

Here are the instructions:
1. The basic information will be described in JSON format, with the following attributes:
{
'exercise_profile': 'the overall description of the exercise',
'exercise_content': 'the content of the exercise',
'related_concept': 'the concept related to this exercise'
}

2. Feedback from users will be managed in the following List format:
'history': [
{'answer': 'whether the student correctly answer the exercise'
'student_profile': 'the learning status of the student'},...
]
3. The information I will give you:
Basic information: a JSON string describing the basic information about the exercise
Study history: a List object containing some feedbacks from students who have answered the exercise
Requirements:
1. Please provide your answer in JSON format and in English, following this structure:
{
"summarization": "A summarization of the detailed characteristics of the exercise and what kind of students might correctly answer this exercise",
"reasoning": "briefly explain your reasoning for the summarization"
}
2. Please ensure that the "summarization" and "reasoning" is no longer than 1,000 words.
3. Do not provide any other text outside the JSON string.

Input prompt:
Basic information:
{'exercise profile': 'This exercise involves identifying the character...', 'exercise content': '下列字符中, ASCII码最小的是 ()…', 'related_concept': '文本信息的表示；'}
Study history:
[{'answer': '回答正确', 'student profile': 'The student has shown good understanding in topics related to text information representation...

Figure 4 Example of used prompts in LLMs diagnosis. The Chinese in the input prompt is the exercise content and related knowledge concepts of the PTADisc dataset. LLMs: large language models, JSON: JavaScript objective notation, SOLO: structure of observed learning outcomes, ASCII: American Standard Code for Information Interchange.

knowledge into LLMs by defining universal stages of cognitive development. Thirdly, the hierarchical nature of SOLO taxonomy offers a structured expression framework for the diagnostic reports generated by LLMs, which enables the production of more specific and precise student diagnoses and item attribute analyses. By leveraging this taxonomy, this research presents that the derived cognitive statuses and attributes are both reliable and insightful.

3.4 | Cognitive Level Alignment Module

By integrating LLMs, this research obtains textual diagnoses of students' cognitive statuses and exercise attributes. However, LLMs alone cannot fully interpret response logs due to input length constraints, which limit their ability to incorporate detailed student-exercise interactions. To address this issue, it is essential to align the LLM-generated diagnoses with CDM-derived diagnoses at a cognitive level. Given that LLMs operate in a semantic space, while CDMs analyze behaviours in a distinct behavioural space, both representations should be projected onto a shared space to facilitate effective alignment. Before applying the alignment method, this research first obtains the semantic representations of the LLM-generated textual diagnoses. Then, a high-performance text embedding model is employed to encode the diagnoses, yielding $\boldsymbol{L} = \boldsymbol{E}(T)$, where $\boldsymbol{E}(\cdot)$ denotes the embedding function and each $\boldsymbol{l} \in \boldsymbol{L}$ represents the semantic embedding of a student or exercise derived from the LLM outputs (Su et al., 2023). Concurrently, corresponding representations of students and exercises from the CDMs have been extracted, as illustrated in Figure 2.

Using the alignment method effectively enhances the robustness and precision of LLMs and CDMs in modelling both students and problems. This is particularly crucial in scenarios where data is sparse or noisy, as the aligned representations provide a more stable foundation for cognitive diagnosis. For instance, in cold-start scenarios where students or exercises have limited prior interactions, the enriched semantic information from LLMs can compensate for the lack of behavioural data, thereby improving the overall diagnostic performance. Moreover, the precision of the models is significantly improved through this alignment. The fine-grained semantic understanding provided by LLMs, when aligned with the behavioural features of CDMs, enables a more accurate characterization of students' learning states and exercise attributes. This is achieved by leveraging the SOLO taxonomy. In addition to enhancing robustness and precision, the alignment method also plays a crucial role in reducing bias and noise in the embeddings of CDMs. CDMs often rely on interaction functions, structure of knowledge concepts, and limited prior knowledge, which can introduce biases and noise into the model. By incorporating the rich semantic information from LLMs, the embeddings generated by CDMs become more reliable and less susceptible to such biases. In this research, two alignment methods have been introduced, mixed contrastive alignment and interactive reconstruction alignment.

3.4.1 Mixed Contrastive Alignment

Mixed contrastive alignment seeks to project the LLM-derived student and exercise representations into the CDM's behavioural space. To achieve this, this research employs contrastive learning, a commonly used technique for bidirectional alignment across distinct representation views (Cui et al., 2024; Khosla et al., 2020). The rationale for adopting contrastive learning is that each pair $(\boldsymbol{c}_i, \boldsymbol{l}_i)$ should be highly similar since they both encode the same information about the student or exercise. To perform the mapping, a multi-layer perceptron (MLP) network is utilized to transform the semantic embeddings $\boldsymbol{l}$ generated by LLMs into the behavioural space of CDMs. This transformation is expressed as $\boldsymbol{l}' = \mathrm{MLP}(\boldsymbol{l})$

Inspired by hard negative mining techniques (Xia et al., 2022), this research collects alternative similar example sets for contrastive learning. The contrastive learning is conducted in both simple perspective, denoted as global contrast, and hard negative perspective, denoted as local contrast. For the global contrast, the entire set $\boldsymbol{L}$ has been employed to capture broad and overarch characteristics of the data. The local contrast, regarded as a hard negative mining process, focuses on a subset $\boldsymbol{L}' \subset \boldsymbol{L}$ composed of the most similar students and exercises for each student or exercise. This subset is derived by computing pairwise cosine similarities and selecting the top k most similar examples ($k = 20$ in the experiments). According to the hard negative mining techniques, these similar negative samples can improve the effectiveness of training and enhance the discriminative ability of the model. Meanwhile, the research assigns different weights based on the similarity between these similar samples and the anchor point, the higher the similarity, the higher the weight; the lower the similarity, the lower the weight. While the global contrast highlights general features shared across the entire dataset, the local contrast utilizes similar negative samples.

During training, this research uses the information noise-contrastive estimation (InfoNCE) loss to optimize both global and local contrast, to maximize the mutual information between $\boldsymbol{c}$ and $\boldsymbol{l}$ in the behavioural space (van den Oord et al., 2018). The loss is defined as:

$$\begin{cases} f = -\frac{1}{N}\sum \log \frac{\exp\left(\frac{x_i \cdot y_{i^+}}{T}\right)}{\sum_{j=1}^{N} \exp\left(\frac{x_i \cdot y_{j^-}}{T}\right)}, \\ g = -\frac{1}{N}\sum \log \frac{\exp\left(\frac{x_i \cdot y_{i^+}}{T}\right)}{\sum_{j=1}^{N} w_j \exp\left(\frac{x_i \cdot y_{j^-}}{T}\right)}, \end{cases} \tag{1}$$

where x_i and y_{i^+} form a positive pair, y_{j^-} indicates a negative sample, T is a temperature parameter, w is the weight of the hard sample, and N denotes the sample size. For CDMs, an additional loss, $\mathcal{L}_{\text{cdm}}$ is introduced. For instance, in the NCD model, the cross-entropy loss between the predicted probability y and the ground-truth response r has been adopted:

$$\mathcal{L}_{\text{cdm}} = -\sum_i [r_i \log y_i + (1 - r_i) \log(1 - y_i)]. \tag{2}$$

The complete loss function is formulated as:

$$\begin{cases} \mathcal{L}_{\text{global}} = f(\boldsymbol{c}_i, \boldsymbol{l}'_i, \boldsymbol{L}'), \\ \mathcal{L}_{\text{local}} = g(\boldsymbol{c}_i, \boldsymbol{l}'_i, \boldsymbol{L}'_k), \\ \mathcal{L} = \mathcal{L}_{\text{cdm}} + \alpha \mathcal{L}_{\text{global}} + \beta \mathcal{L}_{\text{local}}, \end{cases} \tag{3}$$

where $f(x_i, x_j, X_k)$ represents the InfoNCE loss function, x_i and x_j are positive samples, and X_k denotes the set of negative samples. The loss $\mathcal{L}_{\text{cdm}}$ corresponds to the CDM-specific objective function, while $\mathcal{L}_{\text{global}}$ and $\mathcal{L}_{\text{local}}$ represent the global and local contrastive loss terms for mixed contrastive alignment, respectively. The parameters α and β are tunable hyper-parameters that balance these terms.

By optimizing the overall loss $\mathcal{L}$, CDMs integrate the modelling insights derived from LLMs for both students and exercises, which ensures that their representations are well-aligned within the CDMs' behavioural space.

3.4.2 Interactive Reconstruction Alignment

Beyond aligning within the CDM behavioural space, this research seeks alignment within the LLMs' semantic space. Drawing inspiration from masked autoencoders (He et al., 2022), a mask-reconstruction strategy is adopted to align the two models from different spaces. An MLP is utilized to map $\boldsymbol{c}$ from the CDM's behavioural space into the LLM's semantic space, resulting in $\boldsymbol{c}'$ = MLP($\boldsymbol{c}$). Specifically, inspired by interactive masked representation learning, an interactive and adaptive mask-reconstruction approach is introduced. Firstly, a dynamic masking approach is implemented, where the mask ratio is adjusted based on the frequency of each student and exercise. Conversely, for less frequently observed instances, the mask ratio is decreased to mitigate noise. Formally, this process can be expressed as:

$$\hat{\boldsymbol{c}}_i = \text{MASK}(\boldsymbol{c}_i, ratio_i), \tag{4}$$

where $\hat{\boldsymbol{c}}_i$ denotes the masked embedding for student i, and $ratio_i$ represents the corresponding mask ratio applied.

Secondly, self-supervised contrastive learning is employed to compute a reconstruction loss $\mathcal{L}_{\text{recon}}$ encouraging the preservation of mutual information between $\boldsymbol{c}$ and $\boldsymbol{l}$ and assisting in accurately reconstructing $\boldsymbol{c}$. The reconstruction loss $\mathcal{L}_{\text{recon}}$ can be expressed as:

$$\mathcal{L}_{\text{recon}} = -\frac{1}{N}\sum \log \frac{\exp\left(\frac{\sin(\hat{\boldsymbol{c}}'_i, \boldsymbol{l}_i)}{T}\right)}{\exp\left(\frac{\sin(\hat{\boldsymbol{c}}'_i, \boldsymbol{l}_i)}{T}\right) + \sum_{i \neq j} \exp\left(\frac{\sin(\hat{\boldsymbol{c}}'_i, \boldsymbol{l}_j)}{T}\right)}. \tag{5}$$

The entire loss function can be formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{cdm}} + \lambda \mathcal{L}_{\text{recon}}, \tag{6}$$

where λ is a hyperparameter controlling the relative importance of the reconstruction loss.

By optimizing this combined loss $\mathcal{L}$, the model benefits from a more faithful reconstruction of masked inputs and achieves better alignment between the representations from CDMs and LLMs within the semantic space. The improved alignment enables CDMs to fully leverage the rich semantic information encoded by LLMs, ultimately enhancing diagnostic accuracy and robustness.

4 Experiments

4.1 | Experimental Settings

4.1.1 Datasets

During the experiments, the statistics of four courses are employed, including Python programming, Linux, database technology and application, and literature and history from the PTADisc dataset (Hu et al. 2023). The PTADisc dataset, sourced from actual student responses on the PTA educational platform, includes both textual descriptions of exercises and associated knowledge concepts. Table 1 presents the statistics for each dataset. The data has been split into training, validation, and testing sets following an 8:1:1 ratio.

4.1.2 Evaluation Metrics

Following established practice in cognitive diagnosis, students' latent cognitive states are assessed to predict

Table 1 Statistics of datasets

Dataset	Number of students	Number of exercises	Number of concepts	Number of response logs	Logs per student	Sparsity (%)
Python	22,953	11,807	713	170,844	7.44	0.06
Linux	1,253	1,335	221	34,758	27.74	2.08
Database	11,891	3,106	313	122,642	10.38	0.33
Literature	2,264	885	31	30,273	13.37	1.51

Note. Sparsity here refers to the matrix numbers including many values without any significant impacts on the computing process.

their future performance on the testing set, given that cognitive proficiency cannot be measured directly. To verify the effectiveness of the CDMs, the widely used evaluation metrics are employed, including area under curve (AUC), accuracy (ACC), and root mean square error (RMSE).

4.1.3 Baseline Methods

To assess the generalizability and effectiveness of the proposed approach, experiments in the research utilize a diverse set of representative CDMs. IRT is a basic CDM that represents students and exercises as unidimensional traits and captures their interactions utilizing a linear function (Lord, 1952). Moreover, MIRT expands the traits of students and exercises in IRT to multiple dimensions (Reckase, 2009). DINA models the factors of students and exercises as binary vectors, including guessing and slipping parameters to evaluate performance (de la Torre, 2009). NCD utilizes deep neural networks to model the interactions between students and exercises (Wang et al., 2020). RCD captures the relationships between students, exercises, and knowledge concepts using graph convolutional networks (Gao et al., 2021). The slowly changing dimensions (SCD) utilize self-supervised learning to enhance the modelling of students and exercises in graph-based cognitive diagnosis, which alleviates the long-tail distribution problem (Wang et al., 2023). ACD takes into account the emotional state of students while answering questions, designs an emotional cognition module, and combines it with traditional cognitive models, achieving good results (Wang et al., 2024b). Furthermore, RDGT employs relation-guided graph transformers to model the associations between students and exercises (Yu et al., 2024).

4.1.4 Implementation Details

Both the baseline methods and the proposed framework are implemented by using PyTorch. For the baseline models, the default hyperparameters are adopted. For the proposed framework, the consistent hyperparameter settings are employed. ChatGPT, specifically the GPT-4o-mini model, is employed to represent the LLMs and use text-embedding-ada-002 as the text embedding model. The mixed contrastive alignment approach is referred to as '-Con', while the interactive reconstruction alignment approach is referred to as '-Rec'.

4.2 | Performance Comparison

To demonstrate the effectiveness of the proposed method in enhancing cognitive diagnosis, the framework of 7 different cognitive diagnosis models is employed in the research. The resulting performance measures are presented in Table 2.

Moreover, the performance of NCD and MIRT under both warm and cold scenarios is employed, as depicted in Figure 5. Specifically, the terms, cold scenario and warm scenario, are defined in this research. A cold scenario is one where an exercise has fewer than 3 interactions in the training set, and a warm scenario is one where an exercise has more than 10 interactions. Based on these definitions, this research partitions the testing set into corresponding cold and warm subsets.

The observations have been made from the results. Both mixed contrastive alignment and interactive reconstruction alignment yield notable improvements over the baseline CDMs, indicating that the proposed framework is widely applicable across various CDM architectures. Moreover, the observed enhancements suggest that both alignment strategies effectively bridge the gap between the semantic space of LLMs and the behavioural space of CDMs. In most models, the mixed contrastive alignment outperforms the interactive reconstruction alignment, which suggests that aligning representations within the CDMs' native behavioural space more effectively integrates the knowledge from LLMs. The behavioural space provides a more intuitive and interpretable framework for modelling student–exercise interactions. By aligning within this space, the semantic insights from LLMs are seamlessly integrated with the behavioural patterns.

Compared to the baseline CDMs, both mixed contrastive alignment and interactive reconstruction alignment demonstrate enhanced performance in both cold and warm scenarios, with particularly pronounced improvements in the cold scenario. This indicates that the proposed framework not only addresses the cold-

Table 2 Performance comparison with baseline methods

Baseline method	Python			Linux			Database			Literature		
	AUC	ACC	RMSE	AUC	ACC	RMSE	AUC	ACC	RMSE	AUC	ACC	RMSE
IRT	0.6338	0.7749	0.4031	0.8146	0.7874	0.3943	0.7312	0.7989	0.3948	0.8086	0.7818	0.3866
IRT-Con	0.6573	0.7974	0.3907	0.8342	0.8056	0.3772	0.7536	0.8037	0.3739	0.8245	0.8013	0.3644
IRT-Rec	0.6639	0.7946	0.3865	0.8291	0.7992	0.3795	0.7441	0.8110	0.3812	0.8210	0.7942	0.3678
MIRT	0.6434	0.7693	0.4415	0.8183	0.7899	0.4056	0.7220	0.7973	0.4217	0.8289	0.8002	0.4083
MIRT-Con	0.6882	0.8051	0.4063	0.8348	0.8069	0.3864	0.7623	0.8192	0.4068	0.8596	0.8237	0.3664
MIRT-Rec	0.6654	0.7883	0.4170	0.8324	0.8016	0.3882	0.7450	0.8109	0.4052	0.8473	0.8192	0.3817
DINA	0.6001	0.5521	0.4962	0.6791	0.5469	0.4964	0.6581	0.5981	0.4716	0.7021	0.6162	0.4735
DINA-Con	0.6482	0.6253	0.4347	0.7252	0.6102	0.4446	0.6941	0.6726	0.4284	0.7452	0.7006	0.4349
DINA-Rec	0.6359	0.6097	0.4462	0.7049	0.6178	0.4581	0.6804	0.6589	0.4361	0.7279	0.6711	0.4562
NCD	0.6522	0.7758	0.4027	0.8256	0.7759	0.3926	0.7375	0.7932	0.3953	0.8449	0.7805	0.3896
NCD-Con	0.6831	0.8014	0.3857	0.8473	0.7935	0.3768	0.7563	0.8213	0.3722	0.8697	0.8202	0.3655
NCD-Rec	0.6695	0.7961	0.3886	0.8490	0.7982	0.3739	0.7524	0.8165	0.3760	0.8631	0.8089	0.3681
RCD	0.6781	0.7767	0.3901	0.8557	0.8086	0.3865	0.7583	0.7948	0.3897	0.8494	0.7879	0.3809
RCD-Con	0.6987	0.7952	0.3762	0.8744	0.8302	0.3614	0.7885	0.8210	0.3719	0.8652	0.8164	0.3659
RCD-Rec	0.6916	0.7916	0.3795	0.8732	0.8284	0.3628	0.7863	0.8159	0.3725	0.8603	0.8153	0.3662
SCD	0.6815	0.7792	0.3882	0.8594	0.8113	0.3806	0.7598	0.7973	0.3824	0.8537	0.7902	0.3781
SCD-Con	0.7041	0.7988	0.3742	0.8756	0.8327	0.3578	0.7952	0.8241	0.3674	0.8701	0.8234	0.3601
SCD-Rec	0.6963	0.7959	0.3763	0.8733	0.8305	0.3597	0.7904	0.8226	0.3682	0.8696	0.8213	0.3602
ACD	0.6738	0.7932	0.4007	0.8374	0.7573	0.4079	0.7578	0.8137	0.3786	0.8517	0.7924	0.3765
ACD-Con	0.7072	0.8064	0.3832	0.8564	0.8017	0.3719	0.7742	0.8337	0.3611	0.8733	0.8134	0.3594
ACD-Rec	0.7046	0.8013	0.3775	0.8532	0.7749	0.3825	0.7786	0.8265	0.3667	0.8690	0.8104	0.3621
RDGT	0.6831	0.7765	0.3862	0.8572	0.8087	0.3855	0.7614	0.7992	0.3815	0.8512	0.7933	0.3754
RDGT-Con	0.7029	0.7932	0.3745	0.8734	0.8302	0.3621	0.7933	0.8231	0.3643	0.8724	0.8245	0.3582
RDGT-Rec	0.6983	0.7894	0.3758	0.8712	0.8269	0.3667	0.7917	0.8206	0.3667	0.8689	0.8227	0.3605

Notes. AUC: area under curve, ACC: accuracy, RMSE: root mean square error, IRT: item response theory, MIRT: multidimensional item response theory, Con: mixed contrastive alignment approach, Rec: interactive reconstruction alignment, DINA: deterministic inputs, noisy, "and" gate, NCD: neural cognitive diagnosis, RCD: relation map driven cognitive diagnosis, SCD: slowly changing dimensions, ACD: affect-aware cognitive diagnosis, RDGT: relation-guided dual-side graph transformer.

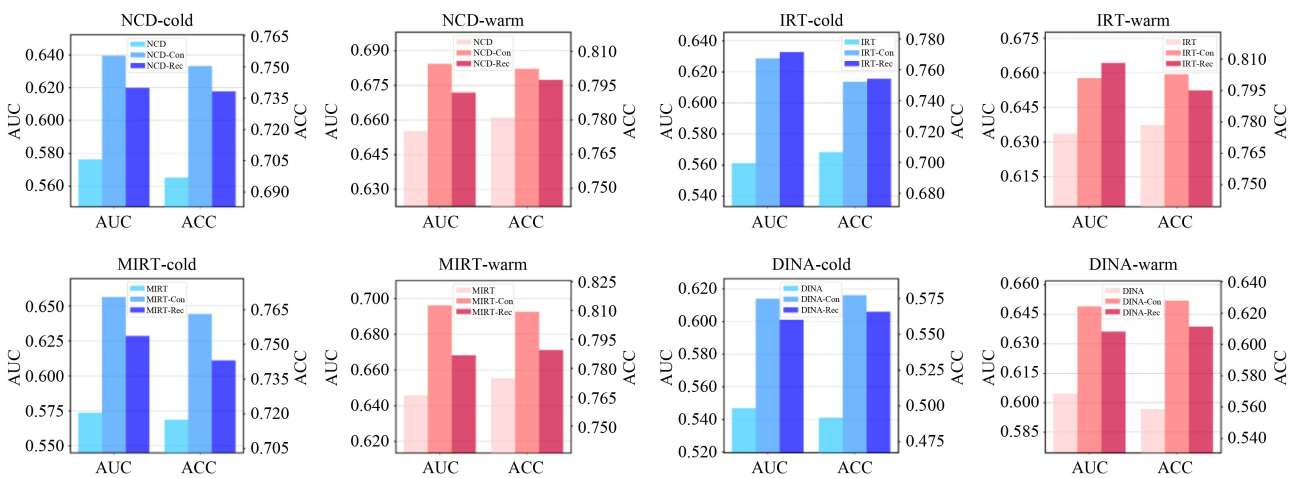

Figure 5 Performance comparison in the cold (blue) and warm (red) scenarios on the Python dataset. NCD: neural cognitive diagnosis, IRT: item response theory, MIRT: multidimensional item response theory, DINA: deterministic inputs, noisy, "and" gate, Con: mixed contrastive alignment approach, Rec: interactive reconstruction alignment, AUC: area under curve, ACC: accuracy.

start problem but also improves overall diagnostic accuracy by enriching the models with additional semantic knowledge. The combination of behavioural and semantic information allows the models to capture both the detailed interaction patterns and the broader conceptual understanding, which leads to more reliable and accurate cognitive diagnoses.

4.3 | Ablation Study

To evaluate the contribution of individual components within the proposed approach, the ablation study has been conducted to assess the impact of several elements used in the LLM diagnosis and cognitive level alignment stages. In Table 3, the effectiveness of incorporating collaborative information denoted as "Coll. Info", both local and global contrastive learning strategies denoted as "Local Con." and "Global Con.", and the dynamic masking mechanism denoted as "Dym. Mask" are implemented. Table 3 presents the experimental results on the Python dataset, which compares the performance of the full model against variations in which specific components are removed as indicated by "w/o." For instance, "w/o Coll. Info" refers to a setting where no collaborative information is included during diagnosis generation, and "w/o Dym. Mask" replaces the dynamic masking strategy with a fixed mask ratio.

Table 3 Ablation study on Python dataset

Condition	Method	AUC	ACC	RMSE
No conditions	NCD	0.6522	0.7758	0.4027
No conditions	NCD-Con	0.6831	0.8014	0.3857
No conditions	NCD-Rec	0.6695	0.7961	0.3886
w/o Coll. Info	NCD-Con	0.6752	0.7958	0.3902
w/o Coll. Info	NCD-Rec	0.6645	0.7879	0.3907
w/o Local Con.	NCD-Con	0.6739	0.7881	0.3916
w/o Global Con.	NCD-Con	0.6733	0.7874	0.3944
w/o Dym. Mask	NCD-Rec	0.6627	0.7882	0.3928

Notes. w/o: without, Coll. Info: collaborative information, Local Con.: local contrast, Global Con.: global contrast. Dym. Mask: dynamic masking strategy, NCD: neural cognitive diagnosis, Con: mixed contrastive alignment approach, Rec: interactive reconstruction alignment, AUC: area under curve, ACC: accuracy, RMSE: root mean square error.

The findings reveal that excluding any of these components leads to a decrease in overall model performance, which underscores the significance of each element in enhancing the final diagnostic accuracy and robustness.

4.4 | Performance on Cold-Start Scenarios

To further assess the robustness of the approach under varying data sparsity conditions, additional experiments on sub-datasets is conducted, which is created by applying random dropout to the training sets of the Python and Linux datasets at ratios of 10%, 20%, 30%, 40%, and 50%. As shown in Figure 6, both AUC and ACC values decline as the dropout ratio increases. This arises from the training data becoming progressively sparse, effectively approximating a cold-start scenario. Notably, for the more sparse Python dataset, the AUC exhibits a more pronounced decrease compared to the Linux dataset.

Despite these challenges, the proposed approach consistently improves the performance of CDMs across different dropout ratios. Moreover, examining the differing outcomes of NCD-Con and NCD-Rec on the Linux and Python datasets highlights that selecting the appropriate alignment method based on the dataset characteristics can yield more favourable diagnostic results.

4.5 | Visualization of Semantic and Behavioural Embeddings

To further demonstrate the effectiveness of the alignment strategies, *t*-distributed stochastic neighbour embedding (*t*-SNE) is employed to visualize the distribution of features in the LLMs' semantic space and the CDMs' behavioural space (van der Maaten & Hinton, 2008). About 200 students are selected randomly, and their corresponding behavioural and semantic embeddings into a 2-dimensional space are mapped.

As shown in Figure 7, both NCD-Con and NCD-Rec yield significantly closer distributions of semantic and behavioural embeddings compared to the unaligned CDMs. For NCD, the semantic and behavioural embeddings are separated, which indicates a significant gap between the two spaces. This visual confirmation is crucial because it provides empirical evidence that the proposed alignment strategies are not only theoretically sound but also practically effective. It confirms that the proposed alignment methods bridge the gap between the semantic and behavioural spaces effectively, which enhances the integration of LLM-derived and CDM-derived representations.

4.6 | Comparision of Different LLMs

The impact of using different LLMs on the LLM diagnosis process is investigated to replace GPT-4o-mini with alternative models. Specifically, GLM-4 and Qwen, two popular and powerful LLMs are considered (Hui et al., 2024; Zeng et al., 2024).

As shown in Table 4, the use of GPT-4o-mini and GLM-4 leads to marginal improvements. This can be attributed to several factors. Firstly, GPT-4o-mini is

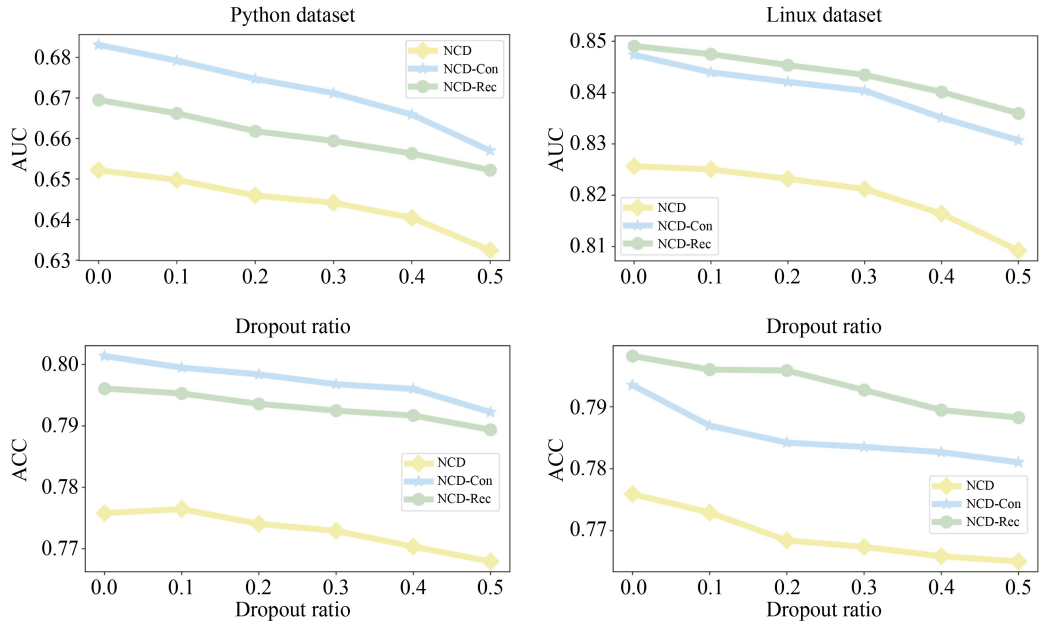

Figure 6 Dropout ratios of Python and Linux databases. NCD: neural cognitive diagnosis, Con: mixed contrastive alignment approach, Rec: interactive reconstruction alignment, AUC: area under curve, ACC: accuracy.

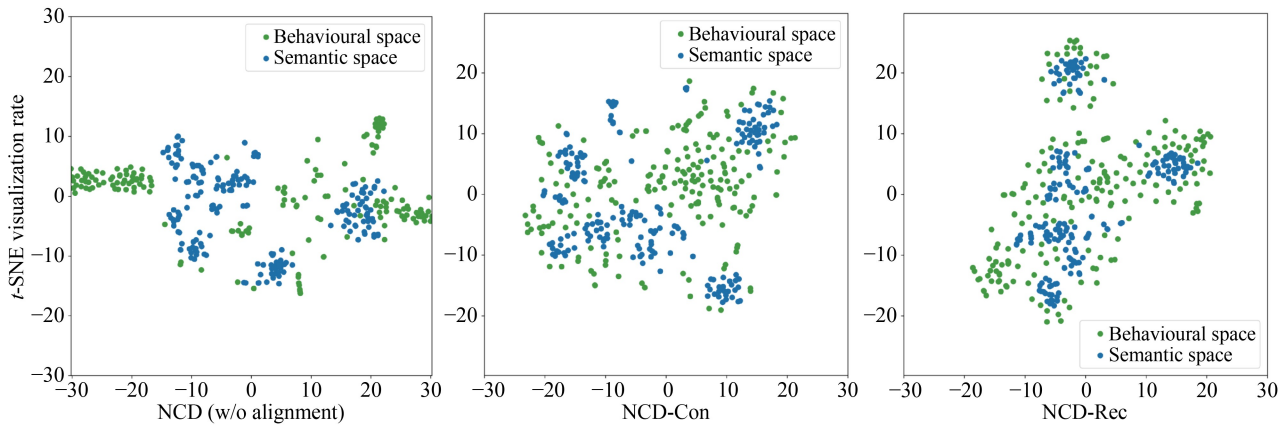

Figure 7 *t*-SNE visualization of student embeddings on literature dataset. *t*-SNE: *t*-distributed stochastic neighbour embedding, NCD: neural cognitive diagnosis, w/o: without, Con: mixed contrastive alignment approach, Rec: interactive reconstruction alignment, AUC: area under curve, ACC: accuracy.

known for its enhanced reasoning capabilities, which enable it to effectively interpret and analyze complex questions. Secondly, GLM-4 excels in its strong comprehension of the Chinese language, which is particularly beneficial given the linguistic nuances present in our dataset. These strengths allow both models to better understand the questions and generate more accurate diagnoses.

However, since these models all possess substantial prior knowledge and reasoning proficiency, their overall diagnostic performance remains relatively similar. Consequently, the final diagnostic outcomes vary only slightly among different LLMs. It is worth mentioning that the smallest LLM, Qwen2.5-14B, exhibited some limitations in our experiments. Specifically, it occasionally generated outputs that did not conform to the required format. This can be attributed to its relatively weaker ability to follow instructions, which is crucial for producing structured and relevant diagnostic results.

To illustrate the differences in diagnoses generated by distinct LLMs including GPT-4o-mini, GLM-4, Qwen2.5-72B, and Qwen2.5-14B more correctively, the diagnostic results for particular students are presented. Each model offers insights into the student's cognitive state from unique perspectives and assesses mastery of specific knowledge concepts accurately, such as network management and file and directory management. Notably, these LLMs not only list the relevant knowledge points but also leverage their

Table 4 Different LLMs on the Linux dataset

LLM	Method	AUC	ACC	RMSE
w/o LLM	NCD	0.8256	0.7759	0.3926
GPT-4o-mini	NCD-Con	0.8473	0.7935	0.3768
GPT-4o-mini	NCD-Rec	0.8490	0.7982	0.3739
GLM-4	NCD-Con	0.8452	0.7955	0.3758
GLM-4	NCD-Rec	0.8483	0.7986	0.3749
Qwen2.5-72B	NCD-Con	0.8413	0.7919	0.3780
Qwen2.5-72B	NCD-Rec	0.8429	0.7947	0.3765
Qwen2.5-14B	NCD-Con	0.8374	0.7891	0.3807
Qwen2.5-14B	NCD-Rec	0.8369	0.7912	0.3782

Notes. w/o: without, LLM: large language model, AUC: area under curve, ACC: accuracy, RMSE: root mean square error, Con: mixed contrastive alignment approach, Rec: interactive reconstruction alignment, NCD: neural cognitive diagnosis.

reasoning capabilities to provide more nuanced interpretations and summaries. Moreover, by drawing on the exercise content, the LLMs supplement the set of identified knowledge points which yields a more detailed and fine-grained representation of the student's cognitive status.

For example, the GPT-4o-mini model generates the diagnostic results of a particular student as follows. The student demonstrates a solid understanding of various Linux concepts, particularly in file and directory management, software package management, and general Linux characteristics, as evidenced by their high number of correct responses. However, the student struggles with some concepts, such as network management, shell command usage, and understanding specifics about links and connections, as indicated by multiple incorrect answers in these areas. The student seems to have a clear grasp of theoretical aspects and practical commands but needs to focus on refining their understanding of the main configurations and functionalities associated with shell commands and network management. It is important for the student to strengthen their knowledge in these weaker areas to achieve a more comprehensive understanding of Linux.

The diagnostic results of the GLM-4 model are presented as follows. The student shows a strong understanding of Linux concepts, such as file and directory management, packaging and compression, and system management. They perform well on exercises related to these concepts. However, some areas need improvements, such as network management and SOLO taxonomy, where the student's understanding is less comprehensive. Moreover, the student seems to struggle with exercises that require a deeper, more integrative understanding, as evidenced by incorrect answers to questions involving complex concepts or those that require theory application.

The diagnostic results of the Qwen2.5-72B model are presented as follows. The student demonstrates a strong grasp of basic Linux concepts and commands, particularly in areas such as file and directory management, software package management, and system information queries. However, there are notable weaknesses in network management, shell and command line usage, and some aspects of file and directory operations. The student often struggles with more complex or nuanced questions, especially those involving multiple steps and a deeper understanding of system internals. Specific areas needing improvement include understanding telnet, terminal usage, and advanced file operations like symbolic links. The student also shows inconsistency in answering questions about the Linux kernel and system features.

The diagnostic results of the Qwen2.5-14B model are presented as follows. The student performs well in Linux basics, especially in Linux common directory structure, file and directory management, package management, network management, and permission management. However, the student has some deficiencies in system information query, shell and command line introduction, and Linux kernel. Moreover, although some answers are correct, they are not completely accurate, indicating that the student does not have a deep understanding of some concepts.

4.7 | Comparison of Diagnostic Results

Table 5 shows the diagnostic result of collaborative information and diagnosis generation about a student and exercise, where the main difference lies in whether collaborative information is included. From the results, it is evident that obtaining collaborative information allows for a more comprehensive diagnosis of the student's cognitive status and the attributes of the exercises.

Compared to the student, the exercise benefits more from collaborative information. This is because, the exercise, having information on all participating students allows for a more comprehensive assessment of the exercise's attributes. For the student, the information included in their response logs about the exercises they completed is usually sufficient to diagnose their cognitive status, so the addition of collaborative information does not provide a substantial improvement. As illustrated in the figure, knowledge-enriched LLMs are capable of diagnosing both students and exercises, thus producing comprehensive and explainable diagnostic results. Such diagnostic results can be better integrated with CDMs to achieve more accurate and comprehensive cognitive diagnoses.

4.8 | Case Study

To further illustrate the enhancements introduced by

the proposed methods, a specific student case from the Linux dataset is examined and the prediction results of NCD are compared to those of NCD-Con. As depicted in Figure 8, a student is selected randomly and the predicted mastery of various knowledge concepts is examined as estimated by both NCD and NCD-Con.

This student demonstrated accurate responses to exercises related to system calls, common commands, and account permission, which suggests a solid grasp of these concepts. Conversely, the student's incorrect responses to other exercises indicate unfamiliarity with the corresponding knowledge concepts. The LLM-based diagnostic results provide insight into similar question-answer patterns observed in the training set, which allows for more informed and nuanced inferences. The LLM diagnosis notes the student's strong understanding of basic Linux concepts but also identifies areas for improvement, particularly

Table 5 Diagnostic results of collaborative information and diagnosis generation about a student and exercise

Diagnostic object	w/o collaborative information and SOLO taxonomy	w/ collaborative information and SOLO taxonomy
Diagnostic results of student	Based on the study history provided, the student answers the exercises correctly. The student is good at concepts related to function definitions, operators and expressions, strings, file operations, and basic data types, such as lists, tuples, and dictionaries. However, the student needs to strengthen their understanding of concepts related to classes and objects, exception handling, and some specific topics like recursion and library management. It is recommended to provide additional exercises and examples in these areas to enhance the student's learning.	The student has demonstrated a strong understanding of many core Python concepts, successfully answering 93 out of 100 exercises correctly. They exhibit proficiency in topics such as data types, functions, and Python's syntax and semantics. However, mistakes in specific areas indicate a need for improvement in understanding operation precedence and advanced function behaviour. Emphasizing deeper engagement with multi-step reasoning and application of knowledge might enhance their learning outcomes further. Potential focus areas include operator precedence and distinctions between similar concepts, especially in function definitions and calling behaviours.
Diagnostic results of exercise	The exercise focuses on understanding Linux file permissions, specifically the structure of the 10-character permission string divided into segments. Successful students are likely to demonstrate a solid grasp of Linux basic knowledge, particularly the role of each segment in the permission string. Generally, learners who can correctly answer this question function at the relational level of the knowledge, as they exhibit an integrated understanding of the components involved in Linux permissions.	The exercise centers on understanding the Linux file permission system, specifically the importance of the third segment in the permission string. This assessment aims for students to demonstrate an intermediate grasp of Linux system administration concepts, notably file permission management. Students who succeed in this exercise likely possess practical experience with Unix and Linux systems, have studied permission management deeply, and can differentiate segments within the permission string. The feedback indicates a mix of proficient students and those needing improvement, emphasizing that while a foundational understanding is common, some struggle with specific aspects of file permissions.

Notes. w/o: without, w/: with, SOLO: structure of observed learning outcomes.

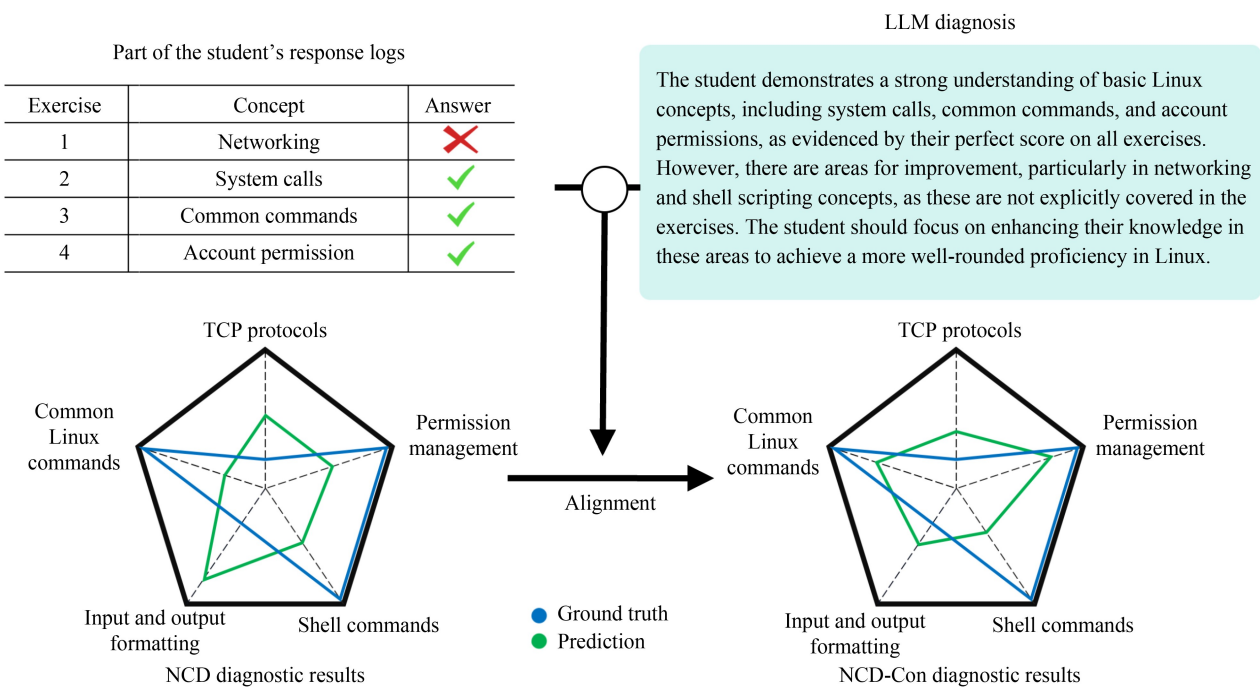

Figure 8 Case study of a student with multiple knowledge concepts on Linux dataset. NCD: neural cognitive diagnosis, Con: mixed contrastive alignment approach, TCP: transmission control protocol.

in networking and shell scripting concepts. This integration of LLM-derived knowledge is instrumental in enabling NCD-Con to more accurately predict the student's mastery levels than the baseline NCD model.

The radar charts in Figure 8 visually compare the diagnostic results of NCD and NCD-Con. The NCD-Con diagnostic results chart shows a closer alignment between the predicted mastery levels in green lines and the ground truth in blue lines, particularly in the areas of networking, TCP protocols, and shell commands. This closer alignment indicates that the integration of LLM-derived knowledge through mixed contrastive alignment has improved the accuracy of the diagnostic predictions.

5 Conclusions

In this research, a model-agnostic framework is introduced and designed to integrate LLMs into CDM effectively. By incorporating an LLM diagnosis module and a cognitive-level alignment module, it harnesses the extensive prior knowledge inherent in LLMs with SOLO taxonomy and bridges the gap between the semantic space of LLMs and the behavioural space of CDMs. This integration yields improved diagnostic accuracy and robustness.

The extensive experiments conducted on four real-world cognitive diagnosis datasets demonstrate that the proposed framework consistently outperforms all baseline CDMs and underscores the efficacy of incorporating LLM-derived insights. In future research, alternative integration strategies will be investigated to achieve more accurate and explainable cognitive diagnoses. This will facilitate downstream applications and enhance the overall utility of cognitive diagnosis models.

Acknowledgments This research was partially supported by the National Natural Science Foundation of China (Grant Nos. 62037001 and 62307032), and the Zhejiang Province Leading Geese Plan (Grant No. 2025C02022).

Conflict of Interest Fei Wu is a member of the Editorial Board and Jingyuan Chen is a Senior Editor of *Frontiers of Digital Education*, who were excluded from the peer-review process and all editorial decisions related to the acceptance and publication of this article. Peer-review was handled independently by the other editors to minimise bias.

Ethics Statements The authors declare that their Institutional Ethics Committee confirmed that no ethical review was required for this study. Written informed consent for participation was not required because all participants' data was anonymized before the statistical analyses were conducted.

Data Availability Statements The authors confirm that all data generated or analysed during this study are included in this published article.

Authors Contributions Zhiang Dong made substantial contributions to the conception of the work, the acquisition, analysis, and interpretation of data, and drafted the work. Jingyuan Chen made substantial contributions to the conception of the work and revised it critically for important intellectual content. Fei Wu made substantial contributions to the acquisition of data and revised it critically for important intellectual content. All authors approved the version to be published and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References

Abbasiantaeb, Z., Yuan, Y. F., Kanoulas, E., & Aliannejadi, M. (2024). Let the LLMs talk: Simulating human-to-human conversational QA via zero-shot LLM-to-LLM interactions. In: *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*. Merida: ACM, 8–17.

Bi, H. Y., Chen, E. H., He, W. D., Wu, H., Zhao, W. H., Wang, S. J., & Wu, J. Z. (2023). BETA-CD: A Bayesian meta-learned cognitive diagnosis framework for personalized learning. In: *Proceedings of the 37th AAAI Conference on Artificial Intelligence*. Washington: AAAI Press, 5018–5026.

Bi, H. Y., Ma, H. P., Huang, Z. Y., Yin, Y., Liu, Q., Chen, E. H., Su, Y., & Wang, S. J. (2020). Quality meets diversity: A model-agnostic framework for computerized adaptive testing. In: *Proceedings of 2020 IEEE International Conference on Data Mining*. Sorrento: IEEE, 42–51.

Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy* (*structure of the observed learning outcome*). New York: Academic Press.

Cui, J. Q., Zhong, Z. S., Tian, Z. T., Liu, S., Yu, B., & Jia, J. Y. (2024). Generalized parametric contrastive learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *46*(12), 7463–7474.

Dai, Z. L., Yao, C., Han, W. K., Yuan, Y., Gao, Z. P., & Chen, J. Y. (2024). MPCoder: Multi-user personalized code generator with explicit and implicit style representation learning. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*. Bangkok: ACL, 3765–3780.

de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, *34*(1), 115–130.

Deng, X., Bashlovkina, V., Han, F., Baumgartner, S., & Bendersky, M. (2023). LLMs to the moon? Reddit market sentiment analysis with large language models. In: *Proceedings of the ACM Web Conference 2023*. New York: ACM, 1014–1019.

Dong, Z., Chen, J. Y., & Wu, F. (2025). Knowledge is power: Harnessing large language models for enhanced cognitive

diagnosis. *arXiv Preprint*, arXiv:2502.05556.

Gao, W. B., Liu, Q., Huang, Z. Y., Yin, Y., Bi, H. Y., Wang, M. C., Ma, J. H., Wang, S. J., & Su, Y. (2021). RCD: Relation map driven cognitive diagnosis for intelligent education systems. In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York: ACM, 501–510.

He, K. M., Chen, X. L., Xie, S. N., Li, Y. H., Dollár, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. In: *Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. New Orleans: IEEE, 16000–16009.

Hoang, M., Bihorac, O. A., & Rouces, J. (2019). Aspect-based sentiment analysis using BERT. In: *Proceedings of the 22nd Nordic Conference on Computational Linguistics*. Turku: ACL, 187–196.

Hu, L. Y., Dong, Z. A., Chen, J. Y., Wang, G. F., Wang, Z. H., Zhao, Z., & Wu, F. (2023). PTADisc: A cross-course dataset supporting personalized learning in cold-start scenarios. In: *Proceedings of the 37th International Conference on Neural Information Processing Systems*. New Orleans: Curran Associates Inc., 44976–44996.

Huang, Z. C., Jin, X. J., Lu, C. Z., Hou, Q. B., Cheng, M. M., Fu, D. M., Shen, X. H., & Feng, J. S. (2024). Contrastive masked autoencoders are stronger vision learners. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *46*(4), 2506–2517.

Huang, Z. Y., Liu, Q., Zhai, C. X., Yin, Y., Chen, E. H., Gao, W. B., & Hu, G. P. (2019). Exploring multi-objective exercise recommendations in online education systems. In: *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. New York: ACM, 1261–1270.

Hui, B. Y., Yang, J., Cui, Z. Y., Yang, J. X., Liu, D. Y. H., Zhang, L., Liu, T. Y., Zhang, J. J., Yu, B. W., Lu, K. M., & et al. (2024). Qwen2.5-coder technical report. *arXiv Preprint*, arXiv:2409.12186.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y. L., Isola, P., Maschinot, A., Liu, C., & Krishnan, D. (2020). Supervised contrastive learning. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. Vancouver: Curran Associates Inc., 18661–18673.

Laskar, T. R., Hoque, E., & Huang, J. X. (2022). Domain adaptation with pre-trained transformers for query-focused abstractive text summarization. *Computational Linguistics*, *48*(2), 279–320.

Li, Q. Y., Fu, L. Y., Zhang, W. M., Chen, X. Y., Yu, J. W., Xia, W., Zhang, W. N., Tang, R. M., & Yu, Y. (2023). Adapting large language models for education: Foundational capabilities, potentials, and challenges. *arXiv Preprint*, arXiv:2401.08664.

Lin, W., Chen, J. Y., Shi, J. X., Guo, Z. R., Zhu, Y. C., Wang, Z. H., Jin, T., Zhao, Z., Wu, F., Yan, S. C., & Zhang, H. W. (2024a). Action imitation in common action space for customized action image synthesis. In: *Proceedings of the 38th Annual Conference on Neural Information Processing Systems*. Vancouver.

Lin, W., Chen, J. Y., Shi, J. X., Zhu, Y. C., Liang, C., Miao, J. Z., Jin, T., Zhao, Z., Wu, F., Yan, S. C., & Zhang, H. W. (2024b). Non-confusing generation of customized concepts in diffusion models. In: *Proceedings of the 41st International Conference on Machine Learning*. Vienna: JMLR, 1206.

Lin, W., Feng, Y. Y., Han, W. K., Jin, T., Zhao, Z., Wu, F., Yao, C., & Chen, J. Y. (2024c). E^3: Exploring embodied emotion through a large-scale egocentric video dataset. In: *Proceedings of the 38th Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. Vancouver.

Liu, Q. (2021). Towards a new generation of cognitive diagnosis. In: *Proceedings of the 30th International Joint Conference on Artificial Intelligence*. Montreal: ijcai, 4961–4964.

Liu, J. Y., Huang, Z. Y., Xiao, T., Sha, J., Wu, J. Z., Liu, Q., Wang, S. J., & Chen, E. H. (2024a). SocraticLM: Exploring Socratic personalized teaching with large language models. In: *Proceedings of the 38th Annual Conference on Neural Information Processing Systems*. Vancouver.

Liu, S., Shen, J. H., Qian, H., & Zhou, A. M. (2024b). Inductive cognitive diagnosis for fast student learning in web-based intelligent education systems. In: *Proceedings of the ACM Web Conference 2024*. New York: ACM, 4260–4271.

Liu, Z. Y., Yin, S. X., Lin, G. Y., & Chen, N. F. (2024c). Personality-aware student simulation for conversational intelligent tutoring systems. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Miami: ACL.

Lord, F. (1952). *A theory of test scores*. Psychometric Monographs No. 7. Richmond: Psychometric Corporation.

Moslem, Y., Haque, R., Kelleher, J. D., & Way, A. (2023). Adaptive machine translation with large language models. In: *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*. Tampere: European Association for Machine Translation, 227–237.

Reckase, M. D. (2009). Multidimensional item response theory models. In: Reckase, M. D., ed. *Multidimensional item response theory*. New York: Springer, 79–112.

Su, H. J., Shi, W. J., Kasai, J., Wang, Y. Z., Hu, Y. S., Ostendorf, M., Yih, W. T., Smith, N. A., Zettlemoyer, L., & Yu, T. (2023). One embedder, any task: Instruction-finetuned text embeddings. In: *Proceedings of the Findings of the Association for Computational Linguistics*. Toronto: ACL, 1102–1121.

van den Oord, A., Li, Y. Z., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv Preprint*, arXiv:1807.03748.

van der Maaten, L., & Hinton, G. (2008). Visualizing data using *t*-SNE. *Journal of Machine Learning Research*, *9*(86), 2579–2605.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Long Beach: Curran Associates Inc., 6000–6010.

Wang, F., Liu, Q., Chen, E. H., Huang, Z. Y., Chen, Y. Y., Yin, Y., Huang, Z., & Wang, S. J. (2020). Neural cognitive diagnosis for intelligent education systems. In: *Proceedings of the 34th*

AAAI Conference on Artificial Intelligence. New York: AAAI Press, 6153–6161.

Wang, F., Liu, Q., Chen, E. H., Liu, C. R., Huang, Z. Y., Wu, J. Z., & Wang, S. J. (2024a). Unified uncertainty estimation for cognitive diagnosis models. In: *Proceedings of the ACM Web Conference 2024*. New York: ACM, 3545–3554.

Wang, S. S., Zeng, Z., Yang, X., Xu, K., & Zhang, X. Y. (2024b). Boosting neural cognitive diagnosis with student's affective state modeling. In: *Proceedings of the 38th AAAI Conference on Artificial Intelligence*. Vancouver: AAAI Press, 620–627.

Wang, S. S., Zeng, Z., Yang, X., & Zhang, X. Y. (2023). Self-supervised graph learning for long-tailed cognitive diagnosis. In: *Proceedings of the 37th AAAI Conference on Artificial Intelligence*. Washington: AAAI Press, 110–118.

Wu, T., Li, M. Z., Chen, J. Y., Ji, W., Lin, W., Gao, J. Y., Kuang, K., Zhao, Z., & Wu, F. (2024). Semantic alignment for multimodal large language models. In: *Proceedings of the 32nd ACM International Conference on Multimedia*. New York: ACM, 3489–3498.

Xia, J., Wu, L. R., Wang, G., Chen, J. T., & Li, S. Z. (2022). ProGCL: Rethinking hard negative mining in graph contrastive learning. In: *Proceedings of the 39th International Conference on Machine Learning*. Baltimore: PMLR, 24332–24346.

Xu, S. L., Zhang, X. Y., & Qin, L. H. (2024). EduAgent: Generative student agents in learning. *arXiv Preprint*, arXiv:2404.07963.

Yu, X. S., Qin, C., Shen, D. Z., Ma, H. P., Zhang, L., Zhang, X. Y., Zhu, H. S., & Xiong, H. (2024). RDGT: Enhancing group cognitive diagnosis with relation-guided dual-side graph transformer. *IEEE Transactions on Knowledge and Data Engineering*, *36*(7), 3429–3442.

Zeng, A. H., Xu, B., Wang, B. W., Zhang, C. H., Yin, D., Zhang, D., Rojas, D., Feng, G. Y., Zhao, H. L., Lai, H. Y., & et al. (2024). ChatGLM: A family of large language models from GLM-130B to GLM-4 all tools. *arXiv Preprint*, arXiv:2406.12793.

Zhang, B., Haddow, B., & Birch, A. (2023a). Prompting large language model for machine translation: A case study. In: *Proceedings of the 40th International Conference on Machine Learning*. Honolulu: JMLR, 41092–41110.

Zhang, D. C., Zhang, K., Wu, L., Tian, M., Hong, R. C., & Wang, M. (2024b). Path-specific causal reasoning for fairness-aware cognitive diagnosis. In: *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. New York: ACM, 4143–4154.

Zhang, H. P., Liu, X., & Zhang, J. W. (2023b). SummIt: Iterative text summarization via ChatGPT. In: *Proceedings of the Findings of the Association for Computational Linguistics*. Singapore: ACL, 10644–10657.

Zhang, J. J., Hou, Y. P., Xie, R. B., Sun, W. Q., McAuley, J., Zhao, W. X., Lin, L. Y., & Wen, J. R. (2024a). AgentCF: Collaborative learning with autonomous language agents for recommender systems. In: *Proceedings of the ACM Web Conference 2024*. New York: ACM, 3679–3689.

Zhu, L. X., Huang, X. W., & Sang, J. T. (2024). How reliable is your simulator? Analysis on the limitations of current LLM-based user simulators for conversational recommendation. In: *Proceedings of the ACM Web Conference 2024*. New York: ACM, 1726–1732.

Zhuang, Y., Liu, Q., Huang, Z. Y., Li, Z., Jin, B. B., Bi, H. Y., Chen, E. H., & Wang, S. J. (2022). A robust computerized adaptive testing approach in educational question retrieval. In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York: ACM, 416–426.