A Dual-Fusion Cognitive Diagnosis Framework for Open Student Learning Environments

Yuanhao Liu*
51275901044@stu.ecnu.edu.cn
East China Normal University
Shanghai, China

Shuo Liu*
shuoliu@stu.ecnu.edu.cn
East China Normal University
Shanghai, China

Yimeng Liu
ymliu@cs.ecnu.edu.cn
East China Normal University
Shanghai, China

Chanjin Zheng
chjzheng@dep.ecnu.edu.cn
East China Normal University
Shanghai, China

Wei Zhang
zhangwei.thu2011@gmail.com
East China Normal University
Shanghai, China

Hong Qian†
hqian@cs.ecnu.edu.cn
East China Normal University
Shanghai, China

Abstract

Cognitive diagnosis model (CDM) is a fundamental component in intelligent education systems which aims to infer students' mastery levels based on historical response logs. However, existing CDMs usually follow the ID-based embedding paradigm, which could often diminish the effectiveness of CDMs in open student learning environments. This is mainly because most of them cannot directly infer new students' ability or utilize new exercises or knowledge concepts without retraining. Textual semantic information, due to its unified feature space and easy accessibility, can help alleviate this issue. However, directly incorporating textual semantic information may not benefit traditional CDMs due to the following challenges: the diversity and complexity of the original text corpus, lack of response-relevant features, and difficulty in integrating multi-source features. To this end, this paper proposes a Dual-Fusion Cognitive Diagnosis Framework (DFCD) to address the above challenges in open student learning environments. Specifically, to standardize the original text corpus and make it easier for CDMs to capture relevant textual semantic information, this paper first proposes the exercise-refiner and concept-refiner to make the exercises and knowledge concepts more coherent and reasonable in educational scenario via large language models. Then, DFCD encodes the refined features using text embedding models to obtain the textual semantic features. To construct response-relevant features, we propose a unified response-relevant feature construction to fully incorporate the information within the response logs. Finally, DFCD designs a dual-fusion module to merge the features from two sources, namely textual semantic features and response-relevant features. The ultimate representations possess the capability of inference in open student learning environments and can be also plugged in existing CDMs. Extensive experiments across three real-world datasets show that DFCD achieves superior performance and strong adaptability by improving the performance in three different scenarios of open student learning environments around 5% on average.

CCS Concepts

• **Applied computing** → Education; • **Computing methodologies** → Machine learning.

Keywords

Cognitive Diagnosis, Open Student Learning Environments, Inductive Learning, Intelligent Education

ACM Reference Format:
Yuanhao Liu, Shuo Liu, Yimeng Liu, Chanjin Zheng, Wei Zhang, and Hong Qian. 2025. A Dual-Fusion Cognitive Diagnosis Framework for Open Student Learning Environments. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '25), August 3–7, 2025, Toronto, ON, Canada.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3711896.3736820

1 Introduction

Nowadays, intelligent education is gaining increasing attention in the field of computer science [18, 21, 43]. Cognitive diagnosis (CD), which is a fundamental upstream task in intelligent education, acts as a pivotal role in current student learning environments [22]. It has a significant and primary impact on subsequent components such as computerized adaptive testing [49], course recommendations [44], and learning path recommendations [24]. As illustrated in the left part of Figure 1, its goal is to infer students' mastery level on each knowledge concept and other attributes, such as the difficulty and discrimination of exercises, through historical response logs and a Q-matrix (an exercise-concept correlation matrix labeled by educational experts).

Classical cognitive diagnosis models (CDMs), such as item response theory (IRT) [7] and the deterministic input, noisy and gate model (DINA) [5], either rely on hand-crafted interaction functions or stringent assumptions (e.g., students must master all concepts associated with an exercise to answer it correctly) or complex parameter estimation methods. These make them unsuitable for large-scale student learning environments. Consequently, neural-based CDMs have recently emerged rapidly. Most existing neural-based

*These authors contribute equally to this work.
†Hong Qian is the corresponding author.

KDD '25, Toronto, ON, Canada

ACM ISBN 979-8-4007-1454-2/2025/08
https://doi.org/10.1145/3711896.3736820

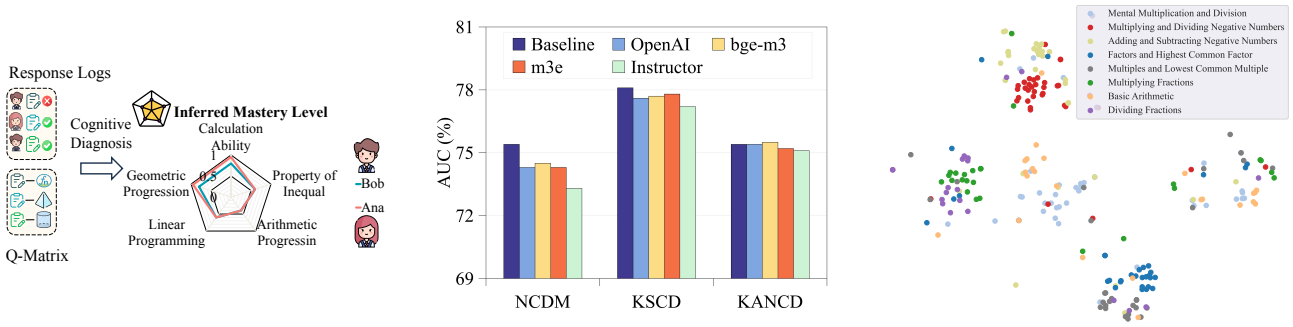


Figure 1: The left subfigure denotes the process of CD. The middle subfigure shows the results of the motivation study on NeurIPS2020 dataset. The right subfigure shows the t-SNE visualization of exercise textual semantic embeddings via text-embedding-ada-002 from the NeurIPS2020 dataset, with each exercise point colored according to its corresponding concept. Notably, we select the subfigures of certain datasets for brevity. Similar results for other datasets are presented in Appendix A.

CDMs [27, 36, 37] follow the traditional ID-based embedding paradigm, vectorizing students, exercises and concepts through embeddings and distinguishing them by IDs. They subsequently update the ID-embeddings by recovering historical response logs (i.e., predict student score on exercises) through binary cross entropy (BCE) loss. However, adhering to this paradigm can lead to failure in open student learning environments where the number or content of students, exercises and concepts are dynamically changing. Students today often complete tests on online education platforms such as IELTS, TOEFL, and GMAT. New students with a large number of their own response records can join at any time, and the assessment content may vary widely. And the online system must quickly diagnose the abilities of these new students and select subsequent test questions accordingly. Such a dynamic open student learning environment presents a significant drawback for the traditional ID-based CDM framework which relies on retraining to accommodate new students, exercises or concepts, because the extensive time required for retraining is often unacceptable given the low-latency demands of real-time testing. ***Therefore, the core idea is to design a framework that enables existing CDMs to be effective in open student learning environments without the need for retraining.***

Textual semantic features (e.g., exercise text and concept name) have demonstrated the ability of generalizing to various downstream tasks in natural language processing due to their unified nature, even in unseen domains [20, 29, 45]. It provides valuable insights into addressing the aforementioned issue in open student learning environments. However, directly incorporating textual semantic features in open student learning environments may not benefit CDMs and can even perform worse than the original CDMs. As shown in the middle part of Figure 1, we directly employ the original text corpus of exercises and concepts from NeurIPS2020 [39] dataset, leveraging cutting-edge text embedding models to transform the text corpus into textual semantic embeddings. These embeddings will then undergo a linear transformation and be used to replace the original ID-embeddings in CDMs. In this context, "Baseline" denotes the performance of the original CDM, while the other legends represent the names of different text embedding models. And we use three representative CDMs, namely NCDM [36], KSCD [27] and KaNCD [37], for exhibition. It is evident from the results that such an approach does not directly enhance the performance of the original CDMs. Therefore, we have analyzed the underlying reasons and summarized the following three key challenges of integrating textual semantic information into the CDMs: **i) Diversity and complexity of the original text corpus.** As shown in the right part of Figure 1, t-SNE [35] is used to visualize textual semantic embeddings of exercises transformed by text-embedding-ada-002 [8] from the NeurIPS2020 dataset. It can be observed that exercises with the same concept are not well-clustered together and are even quite dispersed. It indicates that due to the diversity and complexity of the original text corpus, most of the text lacks standardized expressions and template-based integration. This undoubtedly presents a significant challenge to the ability of CDMs to capture the relevant textual semantic information. **ii) Lack of response-relevant features.** As shown in Figure 8 of Appendix A, we visualize the textual semantic embeddings of exercises by shading the scatter plot according to the corresponding correct rates of exercise. And it can be observed that exercises with similar correctness rates are far apart. It indicates that textual semantic features are not able to capture response-relevant features which CDMs are better at processing. **iii) Difficulty in integrating multi-source features.** If we aim to incorporate both textual semantic information and response-relevant information in the representation of CDMs, a major challenge lies in effectively integrating these two sources due to the entirely different feature space.

To this end, this paper proposes a Dual-Fusion Cognitive Diagnosis Framework (DFCD) to address the above challenges and enables existing CDMs to be effective in open student learning environments without the need for retraining. Specifically, to standardize the original text corpus and make it easier for CDMs to capture relevant textual semantic information, this paper first proposes the exercise-refiner and concept-refiner to make the exercises and concepts more coherent and reasonable in educational scenario via large language models. Then, DFCD encodes the refined features using cutting-edge text embedding models to obtain the textual

semantic features. For response-relevant features, we propose a unified response-relevant feature construction to fully incorporate the information within the response logs and Q-Matrix, effectively balancing the size of feature spaces of students, exercises and concepts. Finally, DFCD designs a dual-fusion module to merge the features from two sources, namely textual semantic features and response-relevant features. The ultimate representations possess the capability of inference in open student learning environments and can be also plugged in existing CDMs. Extensive experiments across real-world datasets show that DFCD achieves superior performance and strong adaptability by integrating representations from different sources in open student learning environments.

The subsequent sections respectively recap the related work, present the preliminaries, introduce the proposed DFCD, show the empirical analysis and finally conclude the paper.

2 Related Work

2.1 Cognitive Diagnosis Models

ID-based Cognitive Diagnosis Models. Most existing CDMs adhere to the ID-based embedding paradigm, which involves vectorizing students, exercises, and concepts through embeddings and distinguish them by their IDs. They can be categorized by the dimension of mastery levels into two types: latent factor models (e.g., using a fixed length vector to represent students' latent mastery levels), such as multidimensional item response theory (MIRT)[32], and models based on patterns of concept mastery (i.e., the dimension of mastery level is the number of concepts), such as DINA [5]. These two methods either rely on hand-crafted interaction functions or impose stringent assumptions and complex parameter estimation methods, which may not be effective in today's large-scale student learning environments. NCDM [36] employs multi-layer perceptrons (MLPs) as interaction function and represents mastery patterns as continuous variables within the range of $[0, 1]$. Various approaches have been employed to capture fruitful information in the response logs, such as MLP-based [27, 37], graph neural network based [10, 28, 46], Bayesian network based [17]. However, this paradigm can fail in open student learning environments. Due to the limitations of IDs, for instance, ID-embedding methods require model retraining for new students, which is unacceptable in real online platforms where timely diagnostic results are expected.

Cognitive Diagnosis Models for Open Student Learning Environments. As online education platforms become increasingly popular, designing CDMs for open student learning environments is crucial. ICD [34] makes the first attempt to target streaming log data with the goal of updating students' mastery levels in real-time without the need for retraining. However, it may require substantial time when there are numerous records in a short period. The most related work are IDCD [16] and ICDM [25] which rely on simple interaction matrices or hand-crafted graph structures as the feature space, but they either demonstrate unpromising performance in open student learning environments or solely focus on a single scenario (e.g., new students). And it is worth noting that unlike the cold-start issues addressed by TechCD [12], ZeroCD [11] or BetaCD [1], open student learning environments focus on inferring the attributes for new students, new exercises and new concepts based on abundant response logs which is not used during training phase without retraining. That is to say, it is an inductive scenario but not like the cold-start scenario which focus on limited response logs.

2.2 Text-based Representation Learning in Intelligent Education Systems

Text-based representation learning in intelligent education systems has recently gained significant popularity. NCDM+ [36] utilizes exercise text via TextCNN [13] to complete the Q-Matrix in CD. EKT [23] enhances student performance prediction in knowledge tracing by utilizing exercise text descriptions. However, neither of them fuse the exercise text or concept name into representations in CD. ECD [48] fuses student context-aware features (e.g., parental education level, monthly study expenses) into representations of students in cognitive diagnosis. However, such features are often difficult to obtain in real-world scenarios due to the need to protect the privacy of students and teachers. TechCD [12] and ZeroCD [11] use BERT [6] for simply extracting exercise text feature as baseline but get limited performance improvement and they focus on cross domain CD which is different from open student learning environments.

3 Preliminaries

Let us consider open student learning environments with N students, M exercises and K knowledge concepts which contain three sets: $S = \{s_1, \ldots, s_N\}, E = \{e_1, \ldots, e_M\}$, and $C = \{c_1, \ldots, c_K\}$. The relationship between the exercises and the knowledge concepts can be recorded in a binary Q-matrix $\mathbf{Q} = \{Q_{ij}\}_{M \times K}$ where $Q_{ij} \in \{0, 1\}$ indicates whether the knowledge concept is involved in the exercise or not. In this paper, we consider three types of open learning environments: unseen students, unseen exercises, and unseen concepts. For instance, in the unseen students scenario, the number of exercises and concepts **remains unchanged**. Notably, this means we do not consider overlapping open scenarios, such as the simultaneous occurrence of a large number of new students and new exercises. This is because data from online learning platforms can always be divided into the aforementioned three types of open learning environments based on timestamps.

Problem Definition. Suppose that the open learning student environments has collected a large number of observed response logs, represented as triplets $T^O = \{(s, e, r) | s \in S^O, e \in E^O, r_{se} \in \{0, 1\}\}$. $r_{se} = 1$ represents correct and $r_{se} = 0$ represents wrong. $S^O \in S$ denotes the observed student set in T^O, and similarly, $E^O \in E$ and $C^O \in C$ represent the observed sets of exercises and concepts, respectively. CDMs have been trained on these observed sets. And assume that there are a certain number of unobserved upcoming response logs T^U involving $S^U \in S, E^U \in E$ and $C^U \in C$ which have not been used to train CDMs. The goal of CD in open student learning environments is to infer the $\mathbf{Mas} \in \mathbb{R}^{|S^U| \times |C^O \cup C^U|}$ which denotes the latent mastery level of students on each concept in unobserved sets without retraining.

4 Methodology: The Proposed DFCD

This section presents the textual feature constructor and the response feature constructor. Following that, we delve into the proposed dual-fusion module. The section is concluded by discussing

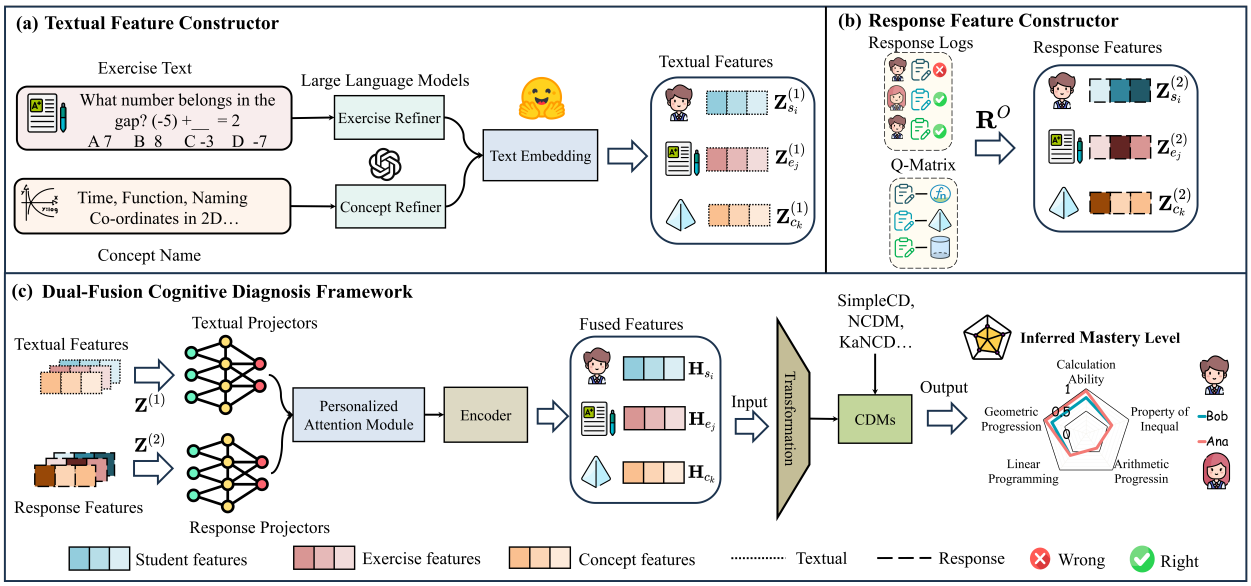


Figure 2: The overall framework of the proposed DFCD. (a) Textual feature constructor. Examples in it are all from real data. (b) Response feature constructor. (c) Detailed components of DFCD.

the model's training. Notably, the strength of DFCD lies in addressing CD in open learning environments. Hence, all its underlying notions are derived from this scenario. Nevertheless, DFCD is versatile enough to be applied in standard scenarios like previous works [36]. The framework of DFCD is shown in Figure 2.

4.1 Textual Feature Constructor

The diversity and complexity of the original text corpus pose challenges for CDMs in accurately capturing the relevant concepts. As shown in Figure 2(a), exercise text may relate to multiple concepts (e.g., trigonometric functions, calculate ability), but the annotated concept might be more specific (e.g., Square Roots), leading to ambiguity. Additionally, the same concept (e.g., time) may have different meanings across disciplines (e.g., physics vs. mathematics), making it difficult for CDMs to infer students' mastery levels on the intended concepts. Therefore, to bridge the gap between real text and its inherent concepts, inspired by the recent successes of large language models (LLMs) in reasoning, LLMs are utilized as exercise refiner and concept refiner. Specifically inspired by recent advancements [30, 42], we design the system prompt α_e, α_c to function as part of the input for LLMs. This prompt aims to explicitly outline the LLMs' role in creating precise summarizations for exercises or concepts by clearly defining the input-output content and the desired output format. By combining this system prompt with the exercise/concept summarization generation prompts β_e and β_c, we can effectively harness LLMs to create precise summarizations. The mathematical process is as Eq. (1).

$$\mathcal{S}_{e_j} = \mathbf{LLM}(\alpha_e, \beta_e, \gamma_{e_j}), \quad \mathcal{S}_{c_k} = \mathbf{LLM}(\alpha_c, \beta_c, \gamma_{c_k}), \tag{1}$$

where $\mathcal{S}_{e_j}$ denotes the summarization result of e_j, $\mathcal{S}_{c_k}$ denotes the summarization result of c_k. γ_{e_j} represents the related concept name of e_j, γ_{c_k} represents exercises which assess c_k. Finally, we can obtain the refined textual features of exercises and concepts using advanced text embedding models. These models effectively transform diverse text inputs into fixed-length vectors, preserving their inherent meaning and contextual information. It can be expressed as Eq. (2).

$$\mathbf{Z}_{e_j}^{(1)} = \mathbf{TEM}(\mathcal{S}_{e_j}), \quad \mathbf{Z}_{c_k}^{(1)} = \mathbf{TEM}(\mathcal{S}_{c_k}), \tag{2}$$

where $\mathbf{Z}_{e_j}^{(1)} \in \mathbb{R}^{1\times d_l}$ denotes the refined textual feature of exercise e_j, $\mathbf{Z}_{c_k}^{(1)} \in \mathbb{R}^{1\times d_l}$ denotes the refined textual feature of concept c_k. **TEM** denotes any text embedding modules (e.g., text-embedding-ada-002 [8]). d_l is the dimension of text embedding in **TEM**. Notably, since student textual profiles are difficult to obtain due to privacy and educational sensitivity, student textual features $\mathbf{Z}_{s_i}^{(1)}$ are derived as the pooled (e.g., mean) result of the exercises they have completed.

4.2 Response Feature Constructor

We contend that one of the main reasons why directly replacing the ID-embedding with text embedding fails is that textual descriptions do not accurately reflect the actual context of student responses. For example, a simple textual description may lead to an embedding representing lower difficulty, while the actual difficulty could be higher due to error-prone details in the question. Additionally, CDMs are better at processing response-relevant features since they are designed based on response logs. Therefore, integrating these features into the representations is crucial in open student learning environments.The previous work [4, 16], following the paradigm of recommendation systems [19], utilizes the historical interaction matrix as features for students or exercises to solve this problem in inductive scenario. However, this approach may lead to mismatch in the dimension of the student and exercise feature space,

causing it fail in certain open student learning environments due to the lack of unification. Additionally, it also fails to incorporate characteristics of the concepts, which have shown success in recent works [27, 37]. To this end, unified response-relevant features are constructed based on historical interaction matrix $\mathbf{I}$ and Q-matrix $\mathbf{Q}$. We can define the historical interaction matrix $\mathbf{I} = \{I_{ij}\}_{|S|\times|E|}$ as Eq. (3).

$$I_{ij} = \begin{cases} 1, & \text{if } r_{ij} = 1, \\ 0, & \text{if } (s_i, e_j, 0) \notin T \text{ and } (s_i, e_j, 1) \notin T, \\ -1, & \text{otherwise,} \end{cases} \tag{3}$$

where T can be either observed or unobserved response logs. Based on $\mathbf{I}$ and $\mathbf{Q}$, we can construct students' response-relevant features $\mathbf{Z}_S^{(2)} = [\mathbf{O}, \mathbf{I}, \mathbf{O}]$ which consist of their responses to exercises, exercises' response-relevant features $\mathbf{Z}_E^{(2)} = [\mathbf{I}^\top, \mathbf{O}, \mathbf{Q}]$ which consist of student responses and their related concepts, and concepts' response-relevant features $\mathbf{Z}_C^{(2)} = [\mathbf{O}, \mathbf{Q}^\top, \mathbf{O}]$ which consist of the exercises that include them. $[\cdot]$ means concatenation operation and $\mathbf{O}$ means zero matrix. Based on above construction, we unify these three kinds of response-relevant features $\mathbf{Z}_{s_i}^{(2)}$, $\mathbf{Z}_{e_j}^{(2)}$, $\mathbf{Z}_{c_k}^{(2)} \in \mathbb{R}^{1\times(|S|+|E|+|C|)}$. Such unified response-relevant features allow for the rapid establishment through any observed or unobserved response logs, without the need for retraining to update.

4.3 Dual Fusion Module

Projectors. After obtaining the textual semantic features and response-relevant features, the key challenge is how to fuse these two multi-source features, which have different feature space, in a personalized manner. Firstly, T-Projector and R-Projector are introduced to align features from two sources in the same dimension, facilitating subsequent processing. Concretely, in each projector, three different MLPs are utilized for students, exercises, and concepts. Here, we take student s_i as an example. It can be expressed as Eq. (4).

$$\tilde{\mathbf{Z}}_{s_i}^{(1)} = \text{MLP}_s^{(1)}(\mathbf{Z}_{s_i}^{(1)}), \quad \tilde{\mathbf{Z}}_{s_i}^{(2)} = \text{MLP}_s^{(2)}(\mathbf{Z}_{s_i}^{(2)}), \tag{4}$$

where $\tilde{\mathbf{Z}}_{s_i}^{(1)}, \tilde{\mathbf{Z}}_{s_i}^{(2)} \in \mathbb{R}^{1\times d}$ denotes the aligned student features from dual sources. $\text{MLP}_s^{(1)}$ and $\text{MLP}_s^{(2)}$ are trainable neural networks to change the dimension into d.

Personalized Attention Module. As our goal is to infer the mastery level of students, which is determined by the aforementioned two sources, each student should have different weights assigned to these sources. This reflects the personalized nature of student learning in reality. Therefore, inspired by [25, 38], we design a personalized attention module. The weight corresponding to the two sources can be computed as

$$w_{s_i}^{(1)} = \mathbf{a}_s \tanh\left(\tilde{\mathbf{Z}}_{s_i}^{(1)}\mathbf{W}_s + \mathbf{b}_s\right)^\top, \; w_{s_i}^{(2)} = \mathbf{a}_s \tanh\left(\tilde{\mathbf{Z}}_{s_i}^{(2)}\mathbf{W}_s + \mathbf{b}_s\right)^\top, \tag{5}$$

where $\mathbf{a}_s \in \mathbb{R}^{1\times d}$ denotes attention vector, $\mathbf{W}_s^g \in \mathbb{R}^{d\times d}$ and $\mathbf{b}_s^g \in \mathbb{R}^{1\times d}$ are trainable parameters in the students' features fusion phase. the ultimate representation of s_i can be derived by normalized weighted summed of $\tilde{\mathbf{Z}}_{s_i}^{(1)}$ and $\tilde{\mathbf{Z}}_{s_i}^{(2)}$ which can be expressed as

$$\mathbf{Z}_{s_i} = \tilde{w}_{s_i}^{(1)}\tilde{\mathbf{Z}}_{s_i}^{(1)} + \tilde{w}_{s_i}^{(2)}\tilde{\mathbf{Z}}_{s_i}^{(2)}, \tag{6}$$

where $\tilde{w}_{s_i}^{(1)} = (1 + e^{w_{s_i}^{(2)} - w_{s_i}^{(1)}})^{-1}$ and $\tilde{w}_{s_i}^{(2)} = (1 + e^{w_{s_i}^{(1)} - w_{s_i}^{(2)}})^{-1}$ denotes the normalized weights. $\mathbf{Z}_{s_i}$ represents the fused representation of student s_i. Similarly, one can obtain $\mathbf{Z}_{e_j}$ and $\mathbf{Z}_{c_k}$ through the same process.

Graph Encoder. Previous works [10, 25] have shown that extracting the relationships among students, exercises, and concepts is crucial, as it can enhance the model's generalization and interpretability performance. Therefore, graph encoder is utilized to obtain the final representation of s_i, which can be expressed as $\mathbf{H} = \textbf{Encoder}(\mathbf{Z}_s, \mathbf{Z}_e, \mathbf{Z}_c)$ where **Encoder** can be any graph encoder like graph attention network [2] or graph transformer [41]. And the classic student-exercise-concept graph is chosen to be the graph structure in our framework. Details can found in Appendix B.1.

4.4 Training of DFCD

Integrating Existing CDMs. To integrate DFCD with most existing CDMs, we need to modify the dimensions to align with the specific type of CDM being used. Since our goal is to infer the students' mastery levels in a fixed dimension, we assume that the total number of concepts is already known (i.e., $|C^O| + |C^U|$). For CDMs where the embedding size is a latent dimension (e.g., KaNCD), $\mathbf{H}_{s_i}, \mathbf{H}_{e_j}$ and $\mathbf{H}_{c_k}$ are directly employed as the input embedding for the integrated CDMs. Otherwise (e.g., NCDM), following [25], transformation layers are introduced. Here, we take student s_i as an example, which can be formulated as

$$\tilde{\mathbf{H}}_{s_i} = \mathbf{H}_{s_i}\mathbf{W}_{\text{t}}^{(s)} + \mathbf{b}_{\text{t}}^{(s)}, \tag{7}$$

where $\tilde{\mathbf{H}}_{s_i}$ will be employed as input embedding for incorporated CDMs and $\mathbf{W}_{\text{t}}^{(s)} \in \mathbb{R}^{d\times(|C^O|+|C^U|)}$, $\mathbf{b}_{\text{t}}^{(s)} \in \mathbb{R}^{1\times(|C^O|+|C^U|)}$ are trainable parameters. This significantly reduces the time complexity of graph convolution by encoder which will be further analyzed in the Appendix B.2. Therefore, DFCD can be trained with integrated CDMs in an end-to-end manner.

SimpleCD. Existing neural-based CDMs [10, 25, 37] except NCDM often have numerous parameters, which may not be effective in open learning environments because they tend to overfit the historical response logs [16]. Therefore, this paper proposes a CDM called "SimpleCD" which is **parameter-free** except for the interaction function. It can be expressed as

$$\hat{y}_{ij} = \mathcal{F}((\sigma(\mathbf{H}_{s_i}\mathbf{H}_c^\top) - \sigma(\mathbf{H}_{e_j}\mathbf{H}_c^\top)) \odot \mathbf{Q}_{e_j}), \tag{8}$$

where $\hat{y}_{ij} \in [0, 1]$ represents the prediction score of i-th student practice j-th exercise, $\mathcal{F}(\cdot)$ denotes the Positive MLP which is commonly utilized in CD and σ typically employs the Sigmoid. $\mathbf{Mas}_{s_i} = \sigma(\mathbf{H}_{s_i}\mathbf{H}_c^\top) \in \mathbb{R}^{1\times(|C^O|+|C^U|)}$ denotes the mastery level of student s_i. "$\odot$" represents the element-wise product. $\mathbf{Q}_{e_j} \in \mathbb{R}^{1\times(|C^O|+|C^U|)}$ signifies the concepts associated with the j-th exercise. We empirically find that it works well in open student learning environments which can be shown in Section 5.2.

Optimization. Given input features of students, exercises and concepts, existing CDMs can predict the score of students on certain exercises, which can be formulated as

$$\hat{y}_{ij} = \mathcal{M}_{\text{CD}}(\mathbf{H}_{s_i}, \mathbf{H}_{e_j}, \mathbf{H}_c), \tag{9}$$

Table 1: Statistics of real-world datasets for experiments.

Datasets	NeurIPS2020	XES3G5M	MOOCRadar
#Students	2,000	2,000	2,000
#Exercises	454	1,624	915
#Concepts	38	241	696
#Response Logs	258,233	207,204	385,323
Sparsity	0.284	0.063	0.210
Q Density	1.000	1.000	2.240

where $\mathcal{M}_{\mathrm{CD}}(\cdot)$ denotes the CDMs, and $\mathbf{H}$ represents the input features that contains the representation of the student, exercises and concepts. In the CD task, the main loss function involves computing the BCE loss between the actual response scores and the model's predicted outcomes in a mini-batch. This overall loss can be expressed as follows

$$\mathcal{L}_{\mathrm{BCE}} = -\frac{1}{|T^O|}\sum_{(s,e,r_{se})\in T^O}\left[r_{se}\log\hat{y}_{se} + (1-r_{se})\log(1-\hat{y}_{se})\right]. \quad (10)$$

Training Cost. Time complexity analysis and training speed comparison are conducted in Appendix B.2. Notably, after training, the mastery level of 126 newly arrived students with 1,024 response logs can be inferred in just 64 ms. For the cost of using large language models, the strategy for selecting large language models is discussed in Appendix C.2. We found that using cost-effective models like OpenAI's GPT-3.5-Turbo or Google's Gemini-pro achieves relatively satisfactory results.

5 Experiments

In this section, we first delineate three real-world datasets and evaluation metrics. Then through comprehensive experiments, we aim to manifest the preeminence of DFCD in both open student learning environment and standard scenario. *Experiments in the standard scenario are in Appendix C.3.* To ensure reproducibility and robustness, all experiments are conducted ten times. Our code is available at https://github.com/BW297/DFCD.

5.1 Experimental Settings

Datasets. The experiments are conducted on three real-world datasets, i.e., NeurIPS2020 [39], XES3G5M [26] and MOOCRadar [47]. These three datasets represents diverse educational contexts and subject, which are collected from a wide variety of courses includes the educational contexts and subjects from chinese, history, economics, math, physics and so on. We randomly selected 2,000 students in each dataset. This number is already a relatively large number for cognitive diagnosis tasks which can well support the training of the different cognitive diagnosis algorithms and evaluate their performance. At the same time, in order to ensure that each selected student has enough exercise data to support his or her cognitive diagnosis, we only select students who answered more than 50 questions. For more detailed statistics on these three datasets, please refer to Table 1. The details about datasets source are depicted in the Appendix C.1. Notably, "Sparsity" refers to the sparsity of the dataset, which is calculated as $\frac{|T|}{|S||E|}$ and "**Q** Density" indicates the average number of concepts per exercise.

Evaluation Metrics. To assess the efficacy of DFCD, we utilize both score prediction and interpretability metrics following the previous works [4, 36]. This approach offers a holistic evaluation from both the predictive accuracy and interpretability standpoints.

Generalization Metric: Evaluating the efficacy of CDMs poses difficulties owing to the absence of the true mastery level. A prevalent workaround is to appraise these models based on their capability to predict students' scores on exercises in the test data. The classic classification metrics such as area under the curve (AUC), accuracy (ACC) are used in our paper.

Interpretability Metric: Diagnostic results are highly interpretable hold significant importance in CD. In this regard, we employ the degree of agreement (DOA), which is consistent with the approach used in [17, 36]. The underlying intuition here is that, if s_a has a greater accuracy in answering exercises related to c_k than student s_b, then the probability of s_a mastering c_k should be greater than that of s_b. Namely, $\mathbf{Mas}_{s_a,c_k} > \mathbf{Mas}_{s_b,c_k}$. DOA is defined as (11)

$$\mathrm{DOA}_k = \frac{1}{Z}\sum_{a,b\in S}\delta\left(\mathbf{Mas}_{s_a,c_k},\mathbf{Mas}_{s_b,c_k}\right)\frac{\sum_{j=1}^{M}\mathbf{Q}_{jk}\wedge\varphi(j,a,b)\wedge\delta(r_{aj},r_{bj})}{\sum_{j=1}^{M}\mathbf{Q}_{jk}\wedge\varphi(j,a,b)\wedge I(r_{aj}\neq r_{bj})}, \quad (11)$$

where $Z = \sum_{a,b\in S}\delta(\mathbf{Mas}_{s_a,c_k},\mathbf{Mas}_{s_b,c_k})$, $\mathbf{Q}_{jk}$ indicates exercise e_j's relevance to concept c_k, $\varphi(j,a,b)$ checks if both students s_a and s_b answered e_j, r_{aj} represents the response of s_a to e_j, and $I(r_{aj} \neq r_{bj})$ verifies if their responses are different, $\delta(r_{aj}, r_{bj})$ is 1 for a right response by s_a and a wrong response by s_b, and 0 otherwise. We compute the top 10 concepts with the highest number of response logs in our experiment and refer to it as DOA@10.

Implementation Details. For parameter initialization, we employ the Xavier, and for optimization purposes, Adam [14] is adopted. All experiments are run on a Linux server with two 3.00GHz Intel Xeon Gold 6354 CPUs and one RTX3090 GPU. The batch size is set as 1024 for all datasets. The learning rate is fixed as $1e^{-4}$. We adjust the dimension d within the range $\{32, 64, 128, 256\}$. We choose text-embedding-ada-002 [8] as our text embedding model and Graph Transformer [41] as our graph encoder. We utilize four attention heads for attention-based encoders, with all other parameters set to the PyG [9] defaults. For all methods that involve using Positive MLP as the interaction function, we adopt the commonly used two-layer tower structure with hidden dimensions of 512 and 256. We employ grid search to find the best hyperparameters using the validation set. Selection related to LLMs is introduced in Appendix C.2.

5.2 Experimental Results in Open Student Learning Environment

Baselines. We compare DFCD against other methods and utilize the hyperparameter settings described in their respective original publications.

- KaNCD-Mean [37]: As the original KaNCD is designed solely for the standard scenario, we assign the embedding of unseen students or exercises to the average of the seen ones [25].
- KaNCD-Nearest [37]: For each unseen students, exercises or concepts in T^U, we assign their embedding based on the most similar one in T^O, who is selected based on the similarity of response logs. Here, we use cosine similarity as the similarity measure function [25].
- IDCD [16]: It propose an identifiable cognitive diagnosis framework based on a novel response-proficiency response paradigm and

Table 2: Overall performance in open student learning environments. In each column, an entry with the best mean value is marked in bold and underline for the runner-up. The standard deviation is not shown in the table since it is very small (less than 0.01). If the mean value of the best model significantly differs from the runner-up, passing a t-test with a significance level of 0.05, then we denote it with "*" at the corresponding position. "-" indicates that the model is not suitable of calculating this metric.

Dataset	NeurIPS2020			XES3G5M			MOOCRadar		
Metric	AUC	ACC	DOA@10	AUC	ACC	DOA@10	AUC	ACC	DOA@10
Unseen Student									
KaNCD-Mean	66.60	62.18	-	71.23	82.32	-	81.60	88.70	-
KaNCD-Nearest	74.59	68.00	71.15	71.55	81.97	60.27	89.37	90.34	77.98
IDCD	77.64	70.65	74.15	75.68	82.29	69.75	92.36	91.32	81.26
ICDM	67.67	62.99	62.53	70.34	81.53	61.82	86.94	89.23	71.10
DFCD	**78.19**	**71.39***	**74.33**	**77.79***	**83.05**	**71.99***	**92.91**	**91.68**	**82.15**
Unseen Exercise									
KaNCD-Mean	67.61	62.86	70.49	55.68	77.60	58.63	59.60	62.03	74.17
KaNCD-Nearest	69.58	69.12	70.01	55.34	74.12	58.58	65.14	69.85	75.59
IDCD	74.63	68.28	73.90	62.30	77.27	67.09	78.52	87.79	81.07
ICDM	69.49	64.17	64.80	61.10	79.03	63.18	79.79	87.06	73.71
DFCD	**77.76***	**71.29***	**74.17***	**76.15***	**82.61***	**71.82***	**91.98***	**91.61***	**81.93**
Unseen Concept									
KaNCD-Mean	67.91	65.61	68.21	63.01	71.57	58.89	82.30	85.58	76.48
KaNCD-Nearest	70.53	65.80	68.53	65.38	81.67	57.95	84.69	87.22	76.44
IDCD	73.55	66.36	68.04	72.50	82.04	69.51	91.12	91.01	81.27
ICDM	73.43	66.40	61.08	70.75	82.04	61.53	92.15	91.18	68.08
DFCD	**77.68***	**70.68***	**73.85***	**78.83***	**83.41***	**72.14***	**92.89***	**91.56***	**82.10**

Table 3: Overall prediction performance of ablation study for DFCD in open student learning environments. Details are as same as Table 2.

Dataset	NeurIPS2020			XES3G5M			MOOCRadar		
Metric	AUC	ACC	DOA@10	AUC	ACC	DOA@10	AUC	ACC	DOA@10
Unseen Student									
DFCD-w.o.TE	78.02	71.28	74.23	77.78	83.12	72.20	92.67	91.53	82.24
DFCD-w.o.RE	78.12	71.08	74.14	77.72	83.04	72.07	92.90	91.35	**82.64**
DFCD-w.o.attn	78.11	71.31	74.26	77.80	83.10	72.17	92.90	91.60	81.32
DFCD	**78.19**	**71.39**	**74.33**	**77.81**	**83.18**	**72.21**	**92.91**	**91.68**	82.15
Unseen Exercise									
DFCD-w.o.TE	77.72	71.14	74.13	75.90	82.41	72.06	91.97	91.52	82.02
DFCD-w.o.RE	74.59	68.38	74.11	68.21	81.06	71.71	85.94	89.16	**82.37**
DFCD-w.o.attn	77.74	71.27	74.10	76.10	82.56	71.91	91.92	91.51	81.96
DFCD	**77.76**	**71.31**	**74.17**	**76.11**	**82.62**	**72.29**	**91.98**	**91.61**	81.93
Unseen Concept									
DFCD-w.o.TE	77.67	70.80	74.07	78.82	83.38	72.03	92.55	91.33	82.34
DFCD-w.o.RE	76.80	69.72	74.13	76.83	82.45	72.03	91.84	90.76	**82.67**
DFCD-w.o.attn	77.63	70.63	73.85	78.46	83.30	72.02	92.88	91.50	80.81
DFCD	**77.68**	**70.83**	**74.14**	**78.83**	**83.41**	**72.14**	**92.89**	**91.56**	80.56

its diagnostic module leverages inductive learning representations which can be used in the open student learning environment.

• ICDM [25]: It utilizes a student-centered graph and inductive mastery levels as the aggregated outcomes of students' neighbors in student-centered graph which enables to infer the unseen students by finding the most suitable representations for different node types.

Details. To evaluate the effectiveness of our proposed DFCD in open student learning environments, we conduct experiments following [25] on datasets with unseen students, unseen exercises, and unseen concepts. For the unseen student scenario, we randomly select students who do not appear in the training data. So as unseen exercise and concept. The test size p_t is set to 0.2, following the previous researches [17, 36]. In order to prevent data leakage, we retain the test data intact and partition the training data by students, exercises, or concepts at a ratio of 0.2, with the validation ratio set at 0.1. In this approach, we can obtain two sets from training data: T^O and T^U. We train the DFCD using only the T^O. Then we use the T^U for inference. *Ultimately, the score prediction metrics is computed only by the prediction of students set S^U in T^U for exercises in the test data.* KaNCD-Mean which assigns the embedding of unseen students to the average of the seen ones during the training process has the same representation on every students in test set. So it is not suitable for calculating DOA. In Table 2, we use "-" to indicate this inapplicability.

Results. The comparison results are listed in Table 2. We have the following key observations:

• ID-Based CDMs with a simple postprocessing such as the strategy of mean or finding the nearest representation may solve the problem of open student learning environment to some extent. However, they still don't produce satisfactory results and fall significantly short compared to the outcomes of other models. For IDCD and ICDM, which is specifically designed for open student learning

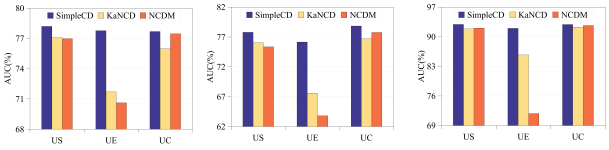
environment, they perform better than the standard CDMs in most of the cases,

• DFCD consistently outperforms the other models on all datasets and in three different scenarios. This demonstrates that DFCD is more effective in the open student learning environment scenario in CD. *And it is worth mentioning that DFCD has such a great performance gap between other models especially in the unseen exercise and knowledge scenario*, this may be because the CD designed for open student learning environment like IDCD and ICDM focus mainly on the unseen student. Due to the fusion of textual and response-relevant features, DFCD offers strong adaptability and interpretability in open student learning environments.

Ablation Study. To showcase the contributions of each component in DFCD, we conduct an ablation study on DFCD, which is divided into the following three versions: DFCD-w.o.TE means removing the textual semantic embeddings. DFCD-w.o.RE means removing the response-relevant embeddings. DFCD-w.o.attn means removing the attention module when fuse the textual semantic embeddings and response-relevant embeddings, the fusion ratio is simply set to 0.5 on both embeddings. As shown in Table 3, DFCD surpasses almost all the versions in both prediction and interpretability performance. This suggests that these components, when combined, enhance DFCD. When each component is removed individually, either the prediction performance decreases or the interpretability performance suffers, indicating that textual semantic features and response-relevant features are both important for the performance of the DFCD and the dual-fusion method of these two representations is also crucial. The DOA@10 on MOOCRadar is higher removing the response-relevant features. This may be due to the fact that the exercises in MOOCRadar include higher-level concepts, which tends to rely more on the textual semantic expression for interpretability. Therefore, when only textual semantic features are used, the interpretability may be higher.

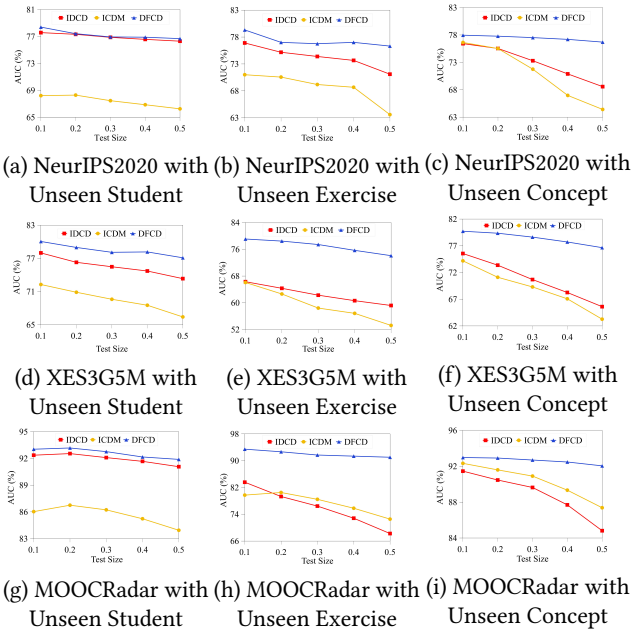
Versatility Analysis. To demonstrate the versatility of DFCD, we integrate the fused features it generates into commonly used CDMs. In this experiment, we compare our proposed SimpleCD with NCDM [36] and KaNCD [37]. We use the abbreviations US (unseen students), UE (unseen exercises), and UC (unseen concepts). As shown in Figure 3, SimpleCD outperforms the others in open student learning environments. This improvement is likely due to the overly simplistic interaction function in NCDM, which struggles to integrate the concepts information [37], leading to poor performance in data-scarce environments. On the other hand, KaNCD suffers from overfitting due to excessive parameters [16]. The fewer parameters and better information acquisition of SimpleCD result in its superior performance.

Generalization Analysis. To assess the efficacy of DFCD's generalization ability, we conduct experiments on three datasets with varying test size $p_t = \{0.1, 0.2, 0.3, 0.4, 0.5\}$. As p_t increases, the generalization ability of CDMs is tested more stringently. As depicted in Figure 4, with an increasing p_t, the number of response logs used for training decreases. However, DFCD consistently outperforms IDCD and ICDM in the open student learning environments, indicating that DFCD can provide more accurate diagnosis results with fewer response logs. Moreover, DFCD decrease more slightly with the increasing p_t than others. This is particularly suitable for current online education platform, where students often have



(a) NeurIPS2020 (b) XES3G5M (c) MOOCRadar

Figure 3: Comparison of DFCD with different integrated CDMs. US means the scenario of unseen student, UE means the scenario of unseen exercise, and UC means the scenario of unseen concept.



(a) NeurIPS2020 with Unseen Student (b) NeurIPS2020 with Unseen Exercise (c) NeurIPS2020 with Unseen Concept

(d) XES3G5M with Unseen Student (e) XES3G5M with Unseen Exercise (f) XES3G5M with Unseen Concept

(g) MOOCRadar with Unseen Student (h) MOOCRadar with Unseen Exercise (i) MOOCRadar with Unseen Concept

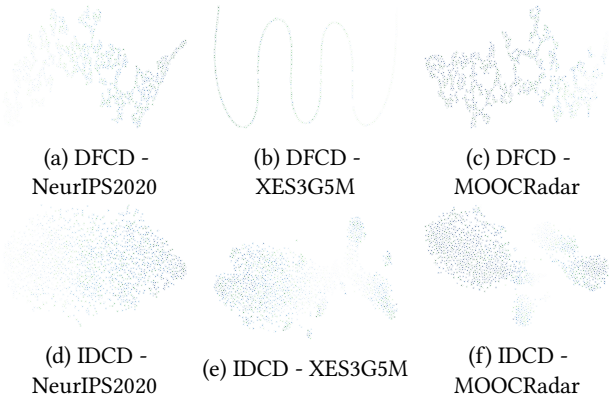
Figure 4: Comparison with other CDMs in different test sizes.

limited response logs. Additionally, we also conduct the experiment on cold-start scenario where response logs per new students are sparse. We compare our DFCD with the SOTA model BetaCD [1] and show a competitive result with it in Table 4.

Diagnosis Result Analysis. Students can naturally be grouped based on their scores, reflecting differences in their mastery levels. We employ t-SNE [35] to map the inferred **Mas** by CDMs onto a two-dimensional plane. By shading the scatter plot according to the corresponding correct rates, with deeper shades of color indicating higher correct rates, we achieve a visual representation of the students' **Mas** distribution. Notably, historical students are colored in blue, while newly arrived students are colored in green. As shown in Figure 5, DFCD exhibits a strip pattern, with colors shifting from lighter to darker shades, indicating it captures both historical and new students' **Mas** trends. In contrast, IDCD's color distribution is more scattered, suggesting it struggles to capture **Mas** information accurately. Additionally, DFCD provides more reliable mastery level inferences for new students, with those of similar correct rates clustering with historical students of comparable performance.

Table 4: The performance comparison with DFCD and BetaCD in cold-start scenario where new student response logs are sparse. Size means the size of response logs per new student.

Datasets		NeurIPS2020		XES3G5M	
Metric	Size	BetaCD	DFCD	BetaCD	DFCD
AUC	3	69.17	68.42	72.05	71.43
	5	69.71	68.81	72.64	72.01
	10	71.23	71.46	73.25	73.22
ACC	3	64.14	63.53	82.40	81.53
	5	64.56	64.53	82.47	81.54
	10	65.13	65.81	82.41	81.78
RMSE	3	46.95	47.21	36.47	37.16
	5	46.80	46.84	36.38	37.09
	10	46.36	46.26	36.28	36.85

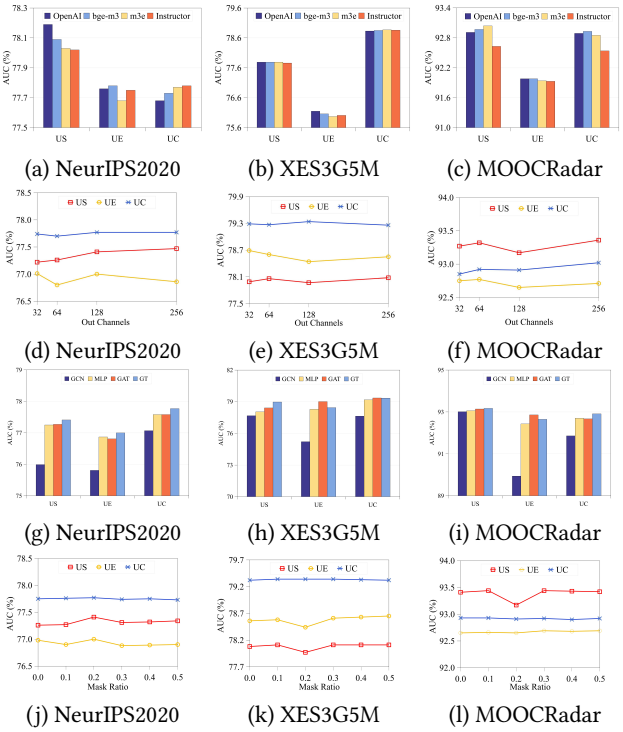


(a) DFCD - NeurIPS2020 (b) DFCD - XES3G5M (c) DFCD - MOOCRadar

(d) IDCD - NeurIPS2020 (e) IDCD - XES3G5M (f) IDCD - MOOCRadar

Figure 5: t-SNE visualization of students' mastery levels for DFCD and IDCD.

Hyperparameter Analysis. We study the impact of the hyperparameters on the text embedding model, embedding dimension, graph encoder and graph mask ratio in Figure 6. Among text embedding model, we select four competitive text embedding models currently available: "OpenAI" refers to text-embedding-ada-002 [8], "bge-m3" refers to BGE-M3 [3], "m3e" refers to M3E-base [40], and "Instructor" refers to Instructor-base [31]. It has been observed that text-embedding-ada-002 and BGE-M3 demonstrate superior performance, likely due to their extensive training data, which supports them to better captures semantic information. Their versatility across multiple languages and functions makes them effective for both English exercise text in the NeurIPS2020 dataset and Chinese exercise text in the XES3G5M and MOOCRadar datasets. We also evaluate the impact of different graph encoders. "GCN" refers to graph convolution network [15], "MLP" refers to normal multi-layer perceptrons, "GAT" refers tograph attention network [2], and "GT" refers to Graph Transformer [41]. Attention-based encoders (e.g., GAT, GT) outperform GCN. While MLP benefits from strong fused representations, incorporating graph structure better captures student–exercise–concept relations, leading to improved performance.



(a) NeurIPS2020 (b) XES3G5M (c) MOOCRadar

(d) NeurIPS2020 (e) XES3G5M (f) MOOCRadar

(g) NeurIPS2020 (h) XES3G5M (i) MOOCRadar

(j) NeurIPS2020 (k) XES3G5M (l) MOOCRadar

Figure 6: Hyperparameters analysis.

And it is recommended to set embedding dimension either 64 and graph mask ratio within the range of 0.2 to 0.3.

6 Conclusion

This paper proposes a Dual-Fusion Cognitive Diagnosis Framework (DFCD), where most existing CDMs can be integrated. For the first time, we identify that directly utilizing exercise text features may not benefit CDMs and can even degrade their performance in the open student learning environments. Therefore, we leverage LLMs as refiners to enhance the textual content. Via DFCD, we fuse the textual semantic features with response-relevant features and integrating them into existing CDMs. Our work enhances CDMs by leveraging LLMs' inference capabilities to better understand exercise semantics and combine textual and response-relevant data for a more comprehensive view of student performance. In the future, we will strengthen the theory of multi-source data fusion to ensure solid theoretical guarantees for fusion performance.

Acknowledgments

We would like to thank the anonymous reviewers for their constructive comments. The algorithms and datasets in the paper do not involve any ethical issue. This work is supported by the National Natural Science Foundation of China (No. 62476091), the Natural Science Foundation of Shanghai (No. 24ZR1418500) and the Education and Scientific Research Project of Shanghai (No. C2025017).

References

[1] Haoyang Bi, Enhong Chen, Weidong He, Han Wu, Weihao Zhao, Shijin Wang, and Jinze Wu. 2023. BETA-CD: A Bayesian meta-learned cognitive diagnosis framework for personalized learning. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence*, Vol. 37. Washington, DC, USA, 5018–5026.

[2] Shaked Brody, Uri Alon, and Eran Yahav. 2022. How Attentive are Graph Attention Networks?. In *Proceedings of the 10th International Conference on Learning Representations*. Virtual Event.

[3] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216* (2024).

[4] Xiangzhi Chen, Le Wu, Fei Liu, Lei Chen, Kun Zhang, Richang Hong, and Meng Wang. 2023. Disentangling Cognitive Diagnosis with Limited Exercise Labels. In *Advances in Neural Information Processing Systems 36*. Louisiana, NO, 18028–18045.

[5] Jimmy De La Torre. 2009. DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics* 34, 1 (2009), 115–130.

[6] Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[7] Susan E Embretson and Steven P Reise. 2013. *Item Response Theory*. Psychology Press.

[8] Tom B. Brown et al. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems 33*. Virtual Event.

[9] Matthias Fey and Jan E. Lenssen. 2019. Fast Graph Representation Learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*.

[10] Weibo Gao, Qi Liu, Zhenya Huang, Yu Yin, Haoyang Bi, Mu-Chun Wang, Jianhui Ma, Shijin Wang, and Yu Su. 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*. Virtual Event, 501–510.

[11] Weibo Gao, Qi Liu, Hao Wang, Linan Yue, Haoyang Bi, Yin Gu, Fangzhou Yao, Zheng Zhang, Xin Li, and Yuanjing He. 2024. Zero-1-to-3: Domain-Level Zero-Shot Cognitive Diagnosis via One Batch of Early-Bird Students towards Three Diagnostic Objectives. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, Vol. 38. Vancouver, Canada, 8417–8426.

[12] Weibo Gao, Hao Wang, Qi Liu, Fei Wang, Xin Lin, Linan Yue, Zheng Zhang, Rui Lv, and Shijin Wang. 2023. Leveraging transferable knowledge concept graph embedding for cold-start cognitive diagnosis. In *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval*. Taipei, Taiwan, 983–992.

[13] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. Doha, Qatar, 1746–1751.

[14] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference on Learning Representations*. San Diego, California.

[15] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *Proceedings of the 5th International Conference on Learning Representations*. Toulon, France.

[16] Jiatong Li, Qi Liu, Fei Wang, Jiayu Liu, Zhenya Huang, Fangzhou Yao, Linbo Zhu, and Yu Su. 2024. Towards the Identifiability and Explainability for Personalized Learner Modeling: An Inductive Paradigm. In *Proceedings of the ACM on Web Conference 2024*. Singapore, 3420–3431.

[17] Jiatong Li, Fei Wang, Qi Liu, Mengxiao Zhu, Wei Huang, Zhenya Huang, Enhong Chen, Yu Su, and Shijin Wang. 2022. HierCDF: A Bayesian Network-based Hierarchical Cognitive Diagnosis Framework. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. Virtual Event, 904–913.

[18] Mingjia Li, Hong Qian, Jinglan Lv, Mengliang He, Wei Zhang, and Aimin Zhou. 2025. Foundation model enhanced derivative-free cognitive diagnosis. *Frontiers of Computer Science* 19, 1 (2025), 191318.

[19] Dawen Liang, Rahul G. Krishnan, Matthew D. Hoffman, and Tony Jebara. 2018. Variational Autoencoders for Collaborative Filtering. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, Pierre-Antoine Champin, Fabien Gandon, Mounia Lalmas, and Panagiotis G. Ipeirotis (Eds.). Lyon, France, 689–698.

[20] Jiayu Liu, Zhenya Huang, Qi Liu, Zhiyuan Ma, Chengxiang Zhai, and Enhong Chen. 2025. Knowledge-Centered Dual-Process Reasoning for Math Word Problems with Large Language Models. *IEEE Transactions on Knowledge and Data Engineering* (2025).

[21] Jiayu Liu, Zhenya Huang, Tong Xiao, Jing Sha, Jinze Wu, Qi Liu, Shijin Wang, and Enhong Chen. 2024. SocraticLM: Exploring socratic personalized teaching with large language models. *Advances in Neural Information Processing Systems 38* (2024).

[22] Qi Liu. 2021. Towards a New Generation of Cognitive Diagnosis. In *Proceedings of 30th International Joint Conference on Artificial Intelligence*. Montreal, Canada, 4961–4964.

[23] Qi Liu, Zhenya Huang, Yu Yin, Enhong Chen, Hui Xiong, Yu Su, and Guoping Hu. 2021. EKT: Exercise-Aware Knowledge Tracing for Student Performance Prediction. *IEEE Transactions on Knowledge and Data Engineering* 33, 1 (2021), 100–115.

[24] Qi Liu, Shiwei Tong, Chuanren Liu, Hongke Zhao, Enhong Chen, Haiping Ma, and Shijin Wang. 2019. Exploiting Cognitive Structure for Adaptive Learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery*. Anchorage, AK, 627–635.

[25] Shuo Liu, Junhao Shen, Hong Qian, and Aimin Zhou. 2024. Inductive Cognitive Diagnosis for Fast Student Learning in Web-Based Intelligent Education Systems. In *Proceedings of the ACM on Web Conference 2024*. Singapore, 4260–4271.

[26] Zitao Liu, Qiongqiong Liu, Teng Guo, Jiahao Chen, Shuyan Huang, Xiangyu Zhao, Jiliang Tang, Weiqi Luo, and Jian Weng. 2024. A Knowledge Tracing Benchmark Dataset with Auxiliary Information. *Advances in Neural Information Processing Systems* 36 (2024).

[27] Haiping Ma, Manwei Li, Le Wu, Haifeng Zhang, Yunbo Cao, Xingyi Zhang, and Xuemin Zhao. 2022. Knowledge-Sensed Cognitive Diagnosis for Intelligent Education Platforms. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. Atlanta, GA, 1451–1460.

[28] Hong Qian, Shuo Liu, Mingjia Li, Bingdong Li, Zhi Liu, and Aimin Zhou. 2024. ORCDF: An Oversmoothing-Resistant Cognitive Diagnosis Framework for Student Learning in Online Education Systems. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. Barcelona, Spain.

[29] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. 9 pages.

[30] Xubin Ren, Wei Wei, Lianghao Xia, Lixin Su, Suqi Cheng, Junfeng Wang, Dawei Yin, and Chao Huang. 2024. Representation Learning with Large Language Models for Recommendation. In *Proceedings of the ACM on Web Conference 2024*. Singpore, 3464–3475.

[31] Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A Smith, Luke Zettlemoyer, and Tao Yu. 2022. One embedder, any task: Instruction-finetuned text embeddings. *arXiv preprint arXiv:2212.09741* (2022).

[32] James B Sympson. 1978. A model for testing with multidimensional items. In *Proceedings of the 1977 Computerized Adaptive Testing Conference*. Minneapolis, MN.

[33] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2023).

[34] Shiwei Tong, Jiayu Liu, Yuting Hong, Zhenya Huang, Le Wu, Qi Liu, Wei Huang, Enhong Chen, and Dan Zhang. 2022. Incremental Cognitive Diagnosis for Intelligent Education. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. Washington, DC, 1760–1770.

[35] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9, 11 (2008).

[36] Fei Wang, Qi Liu, Enhong Chen, Zhenya Huang, Yuying Chen, Yu Yin, Zai Huang, and Shijin Wang. 2020. Neural Cognitive Diagnosis for Intelligent Education Systems. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*. New York, NY, USA.

[37] Fei Wang, Qi Liu, Enhong Chen, Zhenya Huang, Yu Yin, Shijin Wang, and Yu Su. 2023. NeuralCD: A General Framework for Cognitive Diagnosis. *IEEE Transactions on Knowledge and Data Engineering* 35, 8 (2023).

[38] Xiao Wang, Nian Liu, Hui Han, and Chuan Shi. 2021. Self-supervised Heterogeneous Graph Neural Network with Co-contrastive Learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. Virtual Event, 1726–1736.

[39] Zichao Wang, Angus Lamb, Evgeny Saveliev, Pashmina Cameron, Yordan Zaykov, José Miguel Hernández-Lobato, Richard E Turner, Richard G Baraniuk, Craig Barton, Simon Peyton Jones, et al. 2020. Instructions and guide for diagnostic questions: The neurips 2020 education challenge. *arXiv preprint arXiv:2007.12061* (2020).

[40] He sicheng Wang Yuxin, Sun Qingxuan. 2023. *M3E: Moka Massive Mixed Embedding Model*.

[41] Zhanghao Wu, Paras Jain, Matthew A. Wright, Azalia Mirhoseini, Joseph E. Gonzalez, and Ion Stoica. 2021. Representing Long-Range Context for Graph Neural Networks with Global Attention. In *Advances in Neural Information Processing Systems 34*. Virtual Event, 13266–13279.

[42] Yunjia Xi, Weiwen Liu, Jianghao Lin, Jieming Zhu, Bo Chen, Ruiming Tang, Weinan Zhang, Rui Zhang, and Yong Yu. 2023. Towards Open-World Recommendation with Knowledge Augmentation from Large Language Models. *CoRR* abs/2306.10933 (2023).

[43] Hefei Xu, Min Hou, Le Wu, Fei Liu, Yonghui Yang, Haoyue Bai, Richang Hong, and Meng Wang. 2025. Fair Personalized Learner Modeling Without Sensitive Attributes. In *Proceedings of the ACM on Web Conference 2025*, Guodong Long, Michale Blumestein, Yi Chang, Liane Lewin-Eytan, Zi Helen Huang, and Elad Yom-Tov (Eds.). ACM, Sydney, Australia, 4612–4624.

[44] Wei Xu and Yuhan Zhou. 2020. Course video recommendation with multimodal information in online learning platforms: A deep learning framework. *British Journal of Educational Technology* 51, 5 (2020), 1734–1747.
[45] Shangzi Xue, Zhenya Huang, Jiayu Liu, Xin Lin, Yuting Ning, Binbin Jin, Xin Li, and Qi Liu. 2024. Decompose, analyze and rethink: Solving intricate problems with human-like reasoning cycle. *Advances in Neural Information Processing Systems 38* (2024).
[46] Shangshang Yang, Mingyang Chen, Ziwen Wang, Xiaoshan Yu, Panpan Zhang, Haiping Ma, and Xingyi Zhang. 2024. DisenGCD: A Meta Multigraph-assisted Disentangled Graph Learning Framework for Cognitive Diagnosis. In *Advances in Neural Information Processing Systems 38*, Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (Eds.). Vancouver, Canada.
[47] Jifan Yu, Mengying Lu, Qingyang Zhong, Zijun Yao, Shangqing= Tu, Zhengshan Liao, Xiaoya Li, Manli Li, Lei Hou, Haitao Zheng, Juanzi Li, and Jie Tang. 2023. MoocRadar: A Fine-grained and Multi-aspect Knowledge Repository for Improving Cognitive Student Modeling in MOOCs. (2023).
[48] Yuqiang Zhou, Qi Liu, Jinze Wu, Fei Wang, Zhenya Huang, Wei Tong, Hui Xiong, Enhong Chen, and Jianhui Ma. 2021. Modeling Context-aware Features for Cognitive Diagnosis in Student Learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. Virtual Event, 2420–2428.
[49] Yan Zhuang, Qi Liu, GuanHao Zhao, Zhenya Huang, Weizhe Huang, Zachary Pardos, Enhong Chen, Jinze Wu, and Xin Li. 2023. A Bounded Ability Estimation for Computerized Adaptive Testing. In *Advances in Neural Information Processing Systems 36*. New Orleans, LA.

Appendix

The appendix is organized as follows:

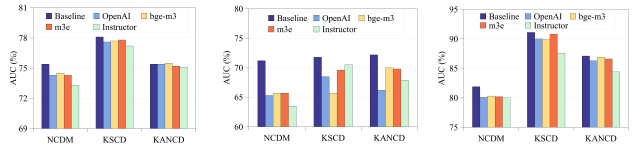
• Appendix A provide additional details about the motivation study in this paper.

• Appendix B provide additional technical details of DFCD in this paper.

• Appendix C provide additional experimental details in this paper.

A Details about Motivation Study

Details about middle subfigure in Figure 1. The complete experimental results are shown in Figure 7.



(a) NeurIPS2020 (b) XES3G5M (c) MOOCRadar

Figure 7: Motivation study: Comparison between original CDMs and directly utilizing textual information in CDMs.

The visualization of exercise text embeddings. The complete experimental results are shown in Figure 8. We employ t-SNE [35] to map the exercise text semantic embeddings onto a two-dimensional plane. By shading the scatter plot according to the corresponding correct rates of exercise, with deeper shades of color indicating higher correct rates, we achieve a visual representation of the exercise' text feature distribution. The exercise text embeddings are relatively loose in the distribution of accuracy, which cause exercises with high accuracy are not clustered together. This distribution will lead to difficulties in using the exercise textual semantic embeddings, which is the main reasons why we use response-relevant features.

B Technical Details about DFCD

B.1 Graph Structure

Since we use graph encoder to obtain the final representation, here we provide the details on the graph structure we use. Inspired by [10], we construct the student-exercise-concept heterogeneous graph, denoted as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Here, $\mathcal{V} = \{\mathcal{V}_s, \mathcal{V}_e, \mathcal{V}_c\}$ represents the set of three node types, with $\mathcal{V}_s \subseteq S$, $\mathcal{V}_e \subseteq E$, and $\mathcal{V}_c \subseteq C$, where S, E, and C denote the sets of students, exercises and concepts, respectively. The set of relationships among the nodes is denoted as $\mathcal{E} = \{\mathcal{E}_1, \mathcal{E}_2\}$, where $\mathcal{E}_1$ involves the interactions between students set S and exercises set E (i.e., student s_i complete the question e_j) and $\mathcal{E}_2$ involves the relationships between exercises set E and concepts set C (i.e., question e_j is related to the concept c_k).

B.2 Time Complexity of DFCD

In this subsection, we present a detailed time complexity analysis of our proposed DFCD. For brevity, we do not include the time complexity of the integrated CDMs, as it can easily add to the overall time complexity of DFCD. Suppose that we have obtain the refined textual feature of students, exercises and concepts. We set the default graph encoder as Graph Transformer. Firstly, we introduce some notions for clarity. d is the latent dimension transformed after projectors. L denotes the graph layers used in the graph encoder, d_l denotes the dimension of textual features, and F denotes the total number of students, exercises, and concepts. As the Textual-Projector and Response-Projector each have three MLPs, the total time complexity is $O(3d_l d + 3Fd)$. The time complexity of personalized attention module is $O(3Fd^2)$. The main time complexity of graph convolution is $O(LFd^2)$. So the ultimate time complexity of DFCD is $O(LFd^2 + 3d_l d + 3Fd + 3Fd^2)$. Therefore, the running speed of DFCD is related to the size of Fd^2, where F depends on the nature of the dataset, and d is a variable parameter. The smaller d is, the slower the speed. In fact, as shown in Figure 9, our proposed DFCD has a faster training speed than ICDM, though it is slightly slower than IDCD. However, it achieves a higher AUC compared to both. *Notably, after training, we can infer the mastery level of 126 newly arrived students with 1,024 response logs in just 64 ms.*

C Experimental Details

C.1 Datasets Sources

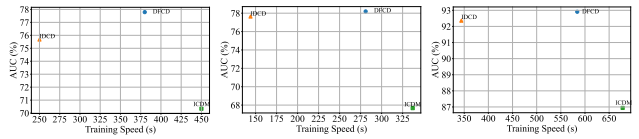
• NeurIPS2020 [39]: NeurIPS2020 comes from the public competition dataset of the NeurIPS 2020 Education Challenge. This competition mainly provides data on students' response logs to Eedi math problems in two school years (September 2018 to May 2020). Eedi provides diagnostic questions for students in elementary school

(a) NeurIPS2020 (b) XES3G5M (c) MOOCRadar

Figure 8: Motivation study: The visualization of exercise text features.

Table 5: Overall prediction performance in standard scenario. Details are the same as Table 2.

Datasets	NeurIPS2020			XES3G5M			MOOCRadar		
Metric	AUC	ACC	DOA@10	AUC	ACC	DOA@10	AUC	ACC	DOA@10
MIRT	77.79	70.72	-	**79.47**	83.45	-	92.52	91.23	-
NCDM	75.44	68.61	72.33	71.18	81.15	62.80	81.87	88.60	76.94
RCD	77.84	70.83	74.27	78.83	83.25	72.29	OOM	OOM	OOM
KSCD	78.07	**71.23**	58.53	71.80	81.75	57.92	91.05	87.92	49.79
KaNCD	75.74	68.85	71.25	72.16	82.17	58.35	87.13	89.22	73.58
DCD	75.93	69.71	73.09	52.66	81.75	55.02	63.90	88.97	55.22
IDCD	77.33	70.24	74.27	76.28	82.60	70.40	92.18	91.28	80.93
ICDM	77.16	70.33	64.29	74.49	82.07	63.64	92.96	91.36	73.14
DFCD	**78.11***	71.20	**74.37**	79.34	**83.48**	**72.53**	**92.97**	**91.61***	**81.01**



(a) NeurIPS2020 (b) XES3G5M (c) MOOCRadar

Figure 9: Training speed comparision with IDCD and ICDM.

through high school (approximately ages 7 to 18). Each diagnostic question is a multiple choice question with 4 possible answer choices, only one of which is correct. This competition mainly has 4 tasks. We choose the datasets of the 3rd and 4th tasks which include the English contextual information about the exercises and concepts, and the text information of the exercises does not exist in the datasets of tasks 1 and 2.

• XES3G5M [26]: XES3G5M is a large-scale knowledge tracing benchmark dataset which consists of student interaction logs collected from a K-12 online learning platform in China. It contains rich auxiliary information about questions and their associated knowledge components. It contains the rich Chinese contextual information including tree structured KC relations, question types, textual contents and analysis. It is worth noting that since XES3G5M is designed for knowledge tracing task, which assumes an evolving cognitive state, unlike our CD setting. We initially limited the dataset to a relatively short time windows to align with CD assumptions. And because the knowledge concepts of XES3G5M are displayed by tree structure, in order to avoid ambiguity, we only use the knowledge concepts of leaf nodes.

• MOOCRadar [47]: MOOCRadar is a dataset for supporting the developments of cognitive student modeling in MOOCs. It provides the relevant learning resources, structures, and contents about the students' exercise behaviors. It also contains the Chinese contextual information about the exercises and concepts.

C.2 Selection of LLMs

In exercise-refiner and concept-refiner, we use OpenAI's large language model GPT-3.5-Turbo. Although OpenAI's GPT-4 has superior performance in terms of text generation quality, it is relatively expensive to use. Since the task of this paper is not that complicated, using GPT-3.5-Turbo can also achieve a relatively satisfactory result. The overall inference cost of GPT-3.5-Turbo in the task is about 3-4 US dollars, which is very cost-effective. At the same time, we also try Google's Gemini Pro [33]. Although Gemini Pro is not as good as GPT-3.5-Turbo in terms of text generation quality, the performance on the task of this paper did not drop too much. And due to the free-use of Gemini Pro, it may also be a good choice.

C.3 Experiment for the Standard Scenario

Baselines. We conduct a comparison of DFCD against other baselines and utilize the hyperparameter settings described in their respective original publications. Among them, ICDM and IDCD can also be used in standard scenario, so we also add them in the baselines. As these two models has been introduced in the Section 5.2, introduction will not be given again. Due to the **Mas** inferred by MIRT being non-interpretable (i.e., the dimensions do not correspond to the number of concepts), we follow previous work [4] by presenting MIRT results but not comparing them.

• MIRT [32] is a representative model of latent factor CDMs, which uses multidimensional $\boldsymbol{\theta}$ to model the latent abilities. We set the latent dimension as 16 which is the same as [36].

• NCDM [36] is a deep learning based CDM which uses MLPs to replace the traditional interaction function (i.e., logistic function).

• KaNCD [37] improves NCDM by exploring the implicit association among knowledge concepts to address the problem of knowledge coverage.

• KSCD [27] explores the implicit association among knowledge concepts and leverages a knowledge-enhanced interaction function.

• RCD [10] leverages GNN to explore the relations among students, exercises and knowledge concepts. We utilize the student-exercise-concept component of RCD to construct the relation graph.

• DCD [4] utilize students' response records to model student proficiency, exercise difficulty and exercise label distribution concepts.

The implementation of MIRT, NCDM and KaNCD comes from the public repository https://github.com/bigdata-ustc/EduCDM. For RCD, DCD, IDCD, ICDM and KSCD, we adopt the implementation from the authors in https://github.com/bigdata-ustc/RCD ,https://github.com/CSLiJT/ID-CDF, https://github.com/kervias/DCD, https://github.com/ECNU-ILOG/ICDM and https://github.com/BIMK/Intelligent-Education/tree/main/KSCD_Code_F.

Details. In line with prior CDM studies [36], in the standard scenario, we partition the data into train and test data and assess our model's performance on the test data. The test size is also set to 0.2, following the setting of the open student learning environment scenario. To ensure fairness in comparison, we adhere to the hyperparameter settings as specified in their original publications. MIRT are non-interpretable models, namely latent factor CDMs, the **Mas** it learns cannot be correlated directly with specific knowledge concepts. Therefore, it is not suitable for calculating DOA. In Table 5, we use "-" to indicate this inapplicability. If CDMs signify out-of-memory on an NVIDIA 3090 GPU, we use the term "OOM" to denote this occurrence.

Results. The comparison results are listed in Table 5. As we can see, despite DFCD is primarily tailored for the open student learning environment scenario in CD, it performs competitively with or even outperforms most of the current state-of the-art CDMs in predictive performance. Moreover, DFCD demonstrates commendable interpretability performance across all three datasets.