Language Representation Favored Zero-Shot Cross-Domain Cognitive Diagnosis

Shuo Liu*
shuoliu@stu.ecnu.edu.cn
Shanghai Institute of AI Education, and School of Computer Science and Technology
East China Normal University
Shanghai, China

Zihan Zhou*
zhzhou@stu.ecnu.edu.cn
Shanghai Institute of AI Education, and School of Computer Science and Technology
East China Normal University
Shanghai, China

Yuanhao Liu
51275901044@stu.ecnu.edu.cn
Shanghai Institute of AI Education, and School of Computer Science and Technology
East China Normal University
Shanghai, China

Jing Zhang
jzhang@ed.ecnu.edu.cn
Department of Educational Psychology, Faculty of Education
East China Normal University
Shanghai, China

Hong Qian†
hqian@cs.ecnu.edu.cn
Shanghai Institute of AI Education, and School of Computer Science and Technology
East China Normal University
Shanghai, China

Abstract

Cognitive diagnosis aims to infer students' mastery levels based on their historical response logs. However, existing cognitive diagnosis models (CDMs), which rely on ID embeddings, often have to train specific models on specific domains. This limitation may hinder their directly practical application in various target domains, such as different subjects (e.g., Math, English and Physics) or different education platforms (e.g., ASSISTments, Junyi Academy and Khan Academy). To address this issue, this paper proposes the language representation favored zero-shot cross-domain cognitive diagnosis (LRCD). Specifically, LRCD first analyzes the behavior patterns of students, exercises and concepts in different domains, and then describes the profiles of students, exercises and concepts using textual descriptions. Via recent advanced text-embedding modules, these profiles can be transformed to vectors in the unified language space. Moreover, to address the discrepancy between the language space and the cognitive diagnosis space, we propose language-cognitive mappers in LRCD to learn the mapping from the former to the latter. Then, these profiles can be easily and efficiently integrated and trained with existing CDMs. Extensive experiments show that training LRCD on real-world datasets can achieve commendable zero-shot performance across different target domains, and in some cases, it can even achieve competitive performance with some classic CDMs trained on the full response data on target domains. Notably, we surprisingly find that LRCD can also provide interesting insights into the differences between various subjects (such as humanities and sciences) and sources (such as primary and secondary education).

CCS Concepts

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

Keywords

Cognitive diagnosis, Zero-shot, Cross domain, Student Score Prediction, Intelligent education systems

ACM Reference Format:
Shuo Liu, Zihan Zhou, Yuanhao Liu, Jing Zhang, and Hong Qian. 2025. Language Representation Favored Zero-Shot Cross-Domain Cognitive Diagnosis. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1 (KDD '25), August 3–7, 2025, Toronto, ON, Canada.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3690624.3709281

1 Introduction

Online intelligent education platforms (OIDP) (e.g., ASSISTments, Junyi Academy and Khan Academy) serve as tools for proactive learning [15, 29], providing personalized practice opportunities that enable students to rapidly enhance their mastery of specific concepts. As shown in Figure 1, cognitive diagnosis (CD) [15, 20], as a crucial component of these platforms, aims to uncover students' mastery levels (a.k.a., diagnosis results) of specific concepts and the characteristics of exercises based on their historical response logs. The results can support further customized applications, such as exercise recommendation [8, 37, 38] and computerized adaptive testing [40, 41].

Over recent years, a diverse array of cognitive diagnosis models (CDMs) has been developed, notably including Item Response Theory (IRT) and the Neural Cognitive Diagnosis Model (NCDM). IRT [32] employs latent factors to represent mastery levels and

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

KDD '25, August 3–7, 2025, Toronto, ON, Canada

ACM ISBN 979-8-4007-1245-6/25/08
https://doi.org/10.1145/3690624.3709281

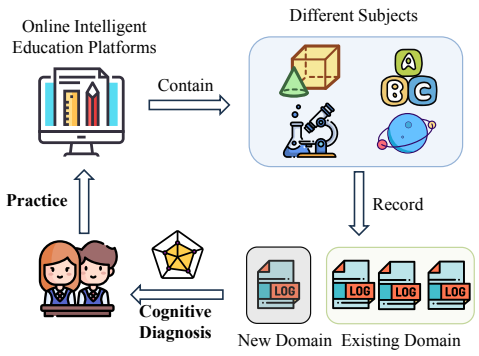


Figure 1: An example of zero-shot cross-domain cognitive diagnosis.

utilizes the logistic function as the interaction function (IF) to predict student performance on exercises. In contrast, NCDM [35], a pioneering neural-based CDM, replaces the traditional manual IF with multi-layer perceptrons (MLP) and has achieved success in large-scale OIDP. Consequently, neural-based CDMs [1, 5, 23, 36] have rapidly gained prominence. Most existing of them continue the ID-based embedding paradigm, vectorizing students, exercises, and concepts through embeddings and distinguishing them by IDs.

However, as shown in Figure 1, with the increasing diversification in education, students' demands for a variety of subjects are also rising. OIDP now encompasses an increasing number of subjects, ranging from sciences like mathematics and physics to humanities like English and political science. It also caters to a wide range of students, from primary school to university level. However, existing ID-based embedding paradigm forces teachers or researchers to train specific CDMs on specific response logs, causing practical difficulties and inconveniences in applying CDMs across different domains, such as varying subjects and age groups of students. Although some studies [6, 7] have made significant efforts to tackle this task, they still follow the ID-based paradigm or rely on strong assumptions to achieve zero-shot cognitive diagnosis. Nonetheless, these assumptions could be difficult to meet in real educational settings. For instance, TechCD [7] requires that the source domain and the target domain have anchor students, which means that there are some students common to both domains. Similarly, Zero-1-3 [6] requires early-bird students in the target domain to learn the shared cognitive signals, which can then be transferred to the target domain.

Motivation. Generally, OIDP already possesses response logs from various domains. When new domain's response logs emerge, it often needs to leverage prior knowledge to quickly and accurately provide diagnostic results for students without retraining models. Therefore, in this paper, we concentrate on a more generalized scenario where one only possesses the response logs from the source domain and lacks any information about the target domain. We designate this critical task as zero-shot cross-domain cognitive diagnosis (ZSCD).

Contribution. To this end, this paper proposes the language representation favored zero-shot cross-domain cognitive diagnosis (LRCD). ***Our core idea is to articulate the behavior patterns of students, exercises and concepts in response logs from different domains through textual descriptions, referred to as textual cognitive profiles.*** Specifically, LRCD first analyzes the behavior patterns of students, exercises, and concepts in various response data, and then describes the profiles of students, exercises and concepts using textual descriptions. Utilizing recent advanced text-embedding modules (e.g., Llama or OpenAI-3-large), these profiles derived from various domains can be transformed into vectors in the unified language space. Nonetheless, this approach introduces a new challenge, specifically, the inherent discrepancy between the language space and the space of CD. Therefore, we propose language-cognitive mappers in LRCD to learn the mapping from the language space to the cognitive space. Consequently, these profiles can be seamlessly and efficiently integrated and trained with existing CDMs. Notably, we observe that even a simple MLP with two linear layers and ReLU can achieve commendable performance. Extensive experiments demonstrate that training LRCD on most real-world datasets can achieve commendable zero-shot performance on totally different target domains, and in some cases it can even match the performance of classic CDMs trained on the complete target response data. Interestingly, through LRCD's performance across different source domains, we find that data from science subjects tend to be more transferable, demonstrating better performance across various target domains. Additionally, LRCD trained on data from students at higher educational levels exhibit greater transferability when applied to students at lower educational levels in different domains.

The following sections respectively review the related work, outline the preliminaries, introduce the proposed LRCD, present the empirical analysis, and ultimately conclude the paper. More details of LRCD and experimental setup are provided in the Appendix.

2 Related Work

2.1 Traditional Cognitive Diagnosis

Cognitive diagnosis has been extensively researched for decades in educational measurement [32] or intelligent education [15], aiming to infer students' mastery levels on concepts based on their historical response logs. Item response theory (IRT) and multidimensional IRT (MIRT) [32] employ latent factors to represent students' mastery levels and use logistic functions as item functions (IF) to predict students' performance on exercises. NCDM [35], as a pioneer of neural-based CDM, directly models students' mastery levels on specific concepts with ID-embeddings and employs MLP as IF, which is successful and remarkable. SCD [26] introduces the symbolic tree to explicably represent IF to further improve the interpretability. Subsequently, following the ID-embedding paradigm, MLP-based (e.g., CDMFKC [14], KSCD [23], KaNCD [36]), Bayesian network-based [12], and GNN-based approaches (e.g., RCD [5], ORCDF [24]) have swiftly achieved even greater success. Recently, a large language model based method called FineCD [13] has been proposed to further enhance CD. FineCD incorporates side information such as question statements to realize fine-grained CD. However, due to the absence or scarcity of training data in target domains, CDMs typically focus on training specific models for specific domains, which may hinder their direct application in entirely different target domains (e.g., different subjects or platforms [9]).

2.2 Cross-Domain Cognitive Diagnosis

Recently, cross-domain cognitive diagnosis has garnered increasing attention, as the proliferation of subjects on OIDP continues, making it challenging to obtain or access abundant domain-specific data during training. TechCD [7] initially proposed the knowledge concept graph to link concepts across different domains, thus effectively transferring student cognitive signals from source domains to target domains. However, the validity of the hand-crafted knowledge concept graph and the feasibility of directly connecting concepts (e.g., concepts from different subjects) across different domains may significantly influence its cross-domain cognitive diagnosis performance. Zero-1-3 [6] capitalizes on early bird students in the target domains to learn shared cognitive signals, which can be transferred to the target domain, thereby enriching the cognitive priors for the new domain. However, the necessity of having early bird students in the target domain may not always be feasible in real educational scenarios. More importantly, both TechCD and Zero-1-3 require an overlap of students between the source domain and the target domain, which does not constitute a completely zero-shot cross-domain cognitive diagnosis. In this paper, we focus on a more generalized and challenging task: zero-shot cross-domain cognitive diagnosis, where there is no overlap of information between the source domain and the target domain. We will elaborate on this task in the following sections.

3 Preliminaries

Let us consider an intelligent education platform with numerous response logs across M different domains, which can be formulated as $R = \{R_1, R_2, \cdots, R_M\}$. In a specific domain, the response logs consist of a vast number of quadrants, which can be represented as $R_m = \{(s, e, \{c \mid \mathbf{Q}_{e,c} = 1\}, y_{se}) \mid s \in S_m, e \in E_m, c \in C_m, y_{se} \in \{0, 1\}\}$. S_m, E_m, and C_m denote the sets of students, exercises, and concepts inherent in the domain m, respectively. $\mathbf{Q}$ represents the relationship between exercises and concepts, where $\mathbf{Q}_{e,c} = 1$ denotes that the exercise e is related to the concept c. Next, we will present the formal definition of Zero-Shot Cross-Domain Cognitive Diagnosis (ZSCD).

Problem Definition of ZSCD. Suppose there are M_o source domains' response logs, namely, $R_o = \{R_1, R_2, \cdots, R_{M_o}\}$. The goal is to train CDMs on R_o and infer the mastery level of students in the target domains in a zero-shot manner. Namely, there are no overlapping students, exercises and concepts between the source domain and the target domain.

4 Methodology: The Proposed LRCD

This section introduces the proposed LRCD. It begins by introducing the textual cognitive profiles, specifically how to describe students, exercises, and concepts in different domains. Moreover, leveraging recent advances in text-embedding modules, we can transform these descriptions into vectors within a unified language space. Next, we explore the language-cognitive mapper, a technique designed to map vectors from the language space to the cognitive space. Following this, we explain how these vectors are trained alongside existing diagnostic models. Finally, we discuss how to perform zero-shot inference in a completely new target domain and analyze model complexity. An overview of LRCD is provided in Figure 2.

4.1 Textual Cognitive Profiles

Due to the sensitivity and sparsity of educational data, the input of CD is very simple, consisting only of response logs, where each log contains students' IDs, exercises' IDs, and concept IDs. Recently, researchers have followed the mature paradigm in user modeling by utilizing ID embeddings to represent the features of students, exercises, and concepts [35, 36]. Initially, these ID embeddings are randomly initialized. Subsequently, through supervised training, each embedding is updated via back propagation. However, this paradigm results in embeddings trained in different domains being on completely different scales and in different spaces, making them unsuitable for direct application in a different target domain.

To tackle this issue, we need to consider what data can be unified in response logs across different domains. Our core idea is to analyze the behavior patterns of students, exercises, and concepts in response logs and utilize textual descriptions, which we refer to as the ***textual cognitive profiles*** of response logs. We begin by elucidating the profiles of the concepts.

Concepts' Profiles. As students use OIDP to assess their mastery of concepts across different domains, it is crucial to effectively articulate the profiles of concepts. Different from TechCD [7] who introduces the hand-crafted knowledge concept graph, leveraging concepts as a pivotal bridge to interconnect various domains. Here, we employ the concepts' names as their profiles, as they often reflect their intrinsic interconnections, such as calculus and multivariable calculus. Moreover, these names can be accessed in nearly all OIDPs (e.g., ASSISTments). It can be expressed as

$$\mathcal{P}_{c_k} = \mathbf{Name}(c_k), \tag{1}$$

where c_k denotes the k-th concept and $\mathcal{P}_{c_k}$ is the processed textual profile of c_k. **Name** represents the concept name.

Exercises' Profiles. Exercises serve as the intermediaries to measure students' mastery of concepts in response logs. By analyzing the response logs, we can infer that it is crucial not only to differentiate exercises based on distinct concepts but also to categorize exercises of the same concept according to their difficulty levels. However, expert labeling of exercise difficulty is a time-consuming and laborious process that often lacks precision. For instance, if a supposedly challenging exercise is administered to a group of high-achieving students, the results may inaccurately suggest that the exercise is not difficult. Consequently, expert labels may fail to accurately reflect the true difficulty of exercises within a specific domain. To address this issue, we utilize the concept names and the average accuracy rates (ACR) of exercises as their profiles. It can formulated as

$$\mathrm{ACR}_{e_j} = \frac{1}{Z_j} \sum_i y_{ij}\,,\ \mathcal{P}_{e_j} = [\{c_k \mid \mathbf{Q}_{j,k} = 1\}, \mathrm{ACR}_{e_j}]\,, \tag{2}$$

where e_j denotes the j-th exercise. ACR_{e_j} denotes the average correct rate of e_j. Z_j denotes the number of students who have practiced e_j. c_k is the related concepts labeled by experts. $\mathcal{P}_{e_j}$ is the processed textual profile of e_j. Notably, when calculating ACR_{e_j}, we exclusively utilize the available training data to mitigate the risk of information leakage.

Students' Profiles. Students' profiles are pivotal to the success of CDMs, as highlighted by recent researchers [11, 17]. We assert that excellent students' profiles should not only consist of their

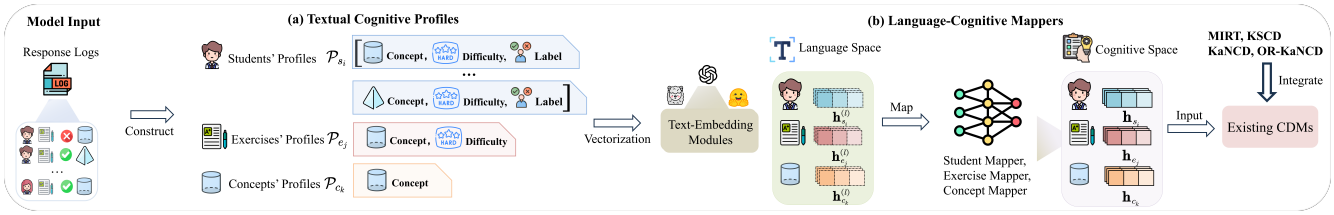


Figure 2: An overview of our proposed LRCD framework. The left side (i.e., sub-figure (a)) provides an overview of the proposed Textual Cognitive Profiles. The right side (i.e., sub-figure (b)) shows the proposed Language-Cognitive Mappers.

interaction records with exercises but also reflect their mastery of different concepts. However, given that the goal of CD is to infer students' mastery levels across various concepts, this creates a classic "the chicken or the egg" dilemma.

Consequently, we integrate the concept name, exercises' average correct rate, and the student score as proxies. Thus, each student-exercise interaction can be considered an integral part of the students' profiles. It can be mathematically expressed as

$$\mathcal{I}_{ij} = [\{c_k \mid \mathbf{Q}_{j,k} = 1\}, \mathrm{ACR}_{e_j}, y_{ij}], \ \mathcal{P}_{s_i} = \{\mathcal{I}_{i0}, \ldots, \mathcal{I}_{ij}, \ldots\}, \quad (3)$$

where $\mathcal{I}_{ij}$ denotes the textual profile of interaction between student s_i and exercise e_j, and y_{ij} represents the i-th student's score on exercise e_j. $\mathcal{P}_{s_i}$ is the processed textual profile of s_i. Notably, unlike other profiles, $\mathcal{P}_{s_i}$ is a set comprising the processed textual profiles of interactions by student s_i.

4.2 Language-Cognitive Mappers

Vectorization. Obviously, through the aforementioned methods, we can analyze each response log and convert them into textual cognitive profiles. Although these textual descriptions cannot be directly used in existing CDMs, we can leverage advanced text-embedding modules (e.g., OpenAI-3-large) to transform these profiles into vectors in the language space which can be formulated as

$$\mathbf{h}_{s_i}^{(l)} = \sum_j \frac{\mathbf{TEM}(\mathcal{I}_{ij})}{|\mathcal{P}_{s_i}|}, \ \mathbf{h}_{e_j}^{(l)} = \mathbf{TEM}(\mathcal{P}_{e_j}), \ \mathbf{h}_{c_k}^{(l)} = \mathbf{TEM}(\mathcal{P}_{c_k}), \quad (4)$$

where $\mathbf{h}_{s_i}^{(l)}, \mathbf{h}_{e_j}^{(l)}, \mathbf{h}_{c_k}^{(l)} \in \mathbb{R}^{1\times d_l}$ represent the vectors of student s_i, exercise e_j and concept c_k in the language space, respectively. d_l is the dimension of the text embedding (e.g., 3072 in OpenAI-3-large). **TEM** denotes the text-embedding module (e.g., OpenAI-3-large) which can be seen as hyperparameters in LRCD, we give further experiments in Section. The representation of s_i can be considered as the mean pooling of the text embeddings of each of his or her interactions. Notably, we exclusively utilize the available training interactions to mitigate the risk of information leakage. Detailed information can be found in Section 5.

Mappers. Leveraging the aforementioned text cognitive profiles and advanced text-embedding modules, we can harmonize data from diverse domains into a unified language space. This enables our model to diagnose students' abilities without the necessity of training on target-domain data. However, the language space is substantially disparate from the space of CD, as the former is trained on extensive corpora (e.g., Common Crawl, Wikipedia, and BooksCorpus) that are entirely unrelated to education data. Therefore, we propose language-cognitive mappers to learn the projection of the language space to the cognitive space which can be formulated as

$$\mathbf{h}_{s_i} = \mathcal{F}_s(\mathbf{h}_{s_i}^{(l)}; \boldsymbol{\theta}_s), \ \mathbf{h}_{e_j} = \mathcal{F}_e(\mathbf{h}_{e_j}^{(l)}; \boldsymbol{\theta}_e), \ \mathbf{h}_{c_k} = \mathcal{F}_c(\mathbf{h}_{c_k}^{(l)}; \boldsymbol{\theta}_c), \quad (5)$$

where $\mathbf{h}_{s_i}, \mathbf{h}_{e_j}, \mathbf{h}_{c_k} \in \mathbb{R}^{1\times d}$ represent the vectors of student s_i, exercise e_j and concept c_k in the space of CD. d is the dimension of space of CD (e.g. 32 in KaNCD [36]). $\mathcal{F}_s$, $\mathcal{F}_e$ and $\mathcal{F}_c$ indicate the student mapper, exercise mapper and concept mapper where $\boldsymbol{\theta}_s, \boldsymbol{\theta}_e, \boldsymbol{\theta}_c$ are model parameters. Concretely, each mapper follows the identical model architecture. Details can be found in Section 5.

4.3 Training and Inference

Model-Agnostic. Given that the primary focus of LRCD is to handle ZSCD tasks, we do not design an additional CDM. Instead, we incorporated recent advanced CDMs into LRCD to illustrate its efficacy and adaptability. It can be expressed as

$$\hat{y}_{ij} = \mathcal{M}_{\mathrm{CD}}(\mathbf{h}_{s_i}, \mathbf{h}_{e_j}, \mathbf{h}_c; \boldsymbol{\theta}_{\mathrm{CD}}), \quad (6)$$

where $\hat{y}_{ij}$ is the score prediction of student s_i practice exercise e_j. $\mathcal{M}_{\mathrm{CD}}$ denotes the integrated CDM (e.g., KSCD [23], KaNCD [36], ORCDF [24]). $\mathbf{h}_{s_i}, \mathbf{h}_{e_j}, \mathbf{h}_c$ indicate the representations of s_i, e_j, and concepts, which are the outputs of LRCD. $\boldsymbol{\theta}_{\mathrm{CD}}$ represents the parameters of the integrated CDM.

Multi-Domain Training. Consistent with previous papers [35], we employ supervised learning to recover the response logs while simultaneously estimating the whole model parameters (i.e., $\boldsymbol{\theta}_s, \boldsymbol{\theta}_e, \boldsymbol{\theta}_c, \boldsymbol{\theta}_{\mathrm{CD}}$). Specifically, we compute the loss between the model's predictions and the actual response scores within a mini-batch and utilize binary cross-entropy (BCE) as the loss function. Notably, LRCD can be trained on multi-domain data, suppose that there are M_o source domains as introduced in Section 3, it can be formulated as follows

$$\begin{aligned} \mathcal{L}_{R_m} &= -\sum_{(s,e,c,y_{se})\in R_m} [y_{se}\log\hat{y}_{se} + (1-y_{se})\log(1-\hat{y}_{se})], \\ \mathcal{L} &= \sum_m^{|R_o|} w_m \mathcal{L}_{R_m}. \end{aligned} \quad (7)$$

$\mathcal{L}_{R_m}$ denotes the BCE loss of domain R_m. w_m is the weight of loss in domain m. In implementation, we directly choose w as a constant value, namely $\frac{1}{|R_o|}$. For more sophisticated selections, we defer to future work.

Zero-Shot Inference in Target Domains. Once the training of LRCD reaches convergence, we can undertake zero-shot inference in any target domain. We first **frozen the parameters in LRCD**. Then, we tackle the response logs in the train data of target domains

into textual cognitive profiles and vectorized them with identical TEM. Benefiting from the proposed mappers, we can rapidly acquire the representations of students, exercises, and concepts in a completely distinct target domain. Ultimately, through the integrated CDM, we can predict the student score on certain exercises or infer students' mastery levels and exercises' difficulty levels.

4.4 Discussion

In this section, we engage in some pivotal discussions about the proposed LRCD, namely time complexity, scalability, and distinction from previous CDMs.

Time Complexity Analysis. Analyzing the time complexity of the CDM is crucial, as students often wish to quickly receive their diagnostic results after completing exercises. Since we do not design a tailored CDM and instead integrate existing CDMs, we will focus our discussion on the time complexity of obtaining text embeddings and the proposed language-cognitive mappers. Favored by the fast API provided by OpenAI [25], we can obtain all representations of a normal domain within 10 minutes. ***More importantly, we only need to run this process once and store the results locally. Therefore, the runtime for this aspect is practically negligible.*** The time complexity of the proposed language mappers is approximately $O(d_l d)$ which depends on the chosen text-embedding module and integrated CDMs. We will provide further details to illustrate the actual training time and inference time for LRCD in Appendix A.1 and Appendix A.2. Notably, LRCD can directly infer students' mastery levels in a new domain with 1,500 students and 50,000 response logs in just ***0.1 seconds***, making it ***406 times*** faster than model retraining.

Scalability. As we do not design computationally intensive modules (e.g., no need for language model inference [2]), LRCD is both efficient and easy to implement, allowing for the seamless integration of all existing CDMs. ***Thus, LRCD can handle thousands of students, exercises and concepts with millions of response logs***, and we will verify this capability in experiments.

Distinction with Zero-1-3 [6]. Herein, we primarily focus on discussing the differences between LRCD and Zero-1-3 at the model level, while the differences in tasks are mentioned in Section 2.2. Compared with Zero-1-3, which uses exercise content [16, 21] as exercise features and employs pre-trained CDMs to initialize student representations, we argue that exercise descriptions are highly variable, as questions on the same concept can take numerous forms. Therefore, as proposed in our textual cognitive profiles, we describe students, exercises, and concepts through the analysis of their behavior patterns in the response logs. Therefore, we do not need the pre-trained CDMs, and we can place all features into a unified language space using text embedding, thereby successfully achieving ZSCD.

5 Experiments

This section first introduces three real-world datasets and evaluation metrics. Subsequently, through comprehensive experiments, we endeavor to substantiate the superiority of LRCD's zero-shot performance across various target domains (e.g., different subjects and different platforms). To ensure the reliability and reproducibility of our experiments, they are independently repeated ten times with different seeds, and our code is available at https://github.com/ECNU-ILOG/LRCD.

5.1 Experimental Settings

Datasets Description. We conduct our experiments on three real-world datasets: SLP [22], EDM [4], and MOOC [39]. To ensure that each student has a sufficient number of diagnostic records and meets the requirements of the baseline as well as equipment needs, we exclude students with fewer than 5, 20 and 50 responses, respectively. Moreover, following the paradigm established by [12, 35], we restrict to the first attempt for each exercise to ensure the stability of students' mastery levels. This approach ensures that the data reflect the students' initial understanding, thereby maintaining the static nature of their mastery levels. Importantly, we have uploaded all the processed data to the aforementioned repository. Table 1 provides detailed statistics of those datasets, where "Average Correct Rate" indicates the average accuracy of students on exercises, and "Q Density" represents the average number of concepts per exercise. Moreover, the entry for SLP represents the aggregate values across all subjects (e.g., Math, Physics). Due to space constraints, we provide the detailed values for each subject in Table 8.

Table 1: Statistics of the three real-world datasets.

Datasets	SLP	MOOC	EDM
#Students	7,663	6,077	6,061
#Exercises	4,873	2,366	1,534
#Concepts	179	2,611	323
#Response Logs	299,391	769,541	198,058
Average Correct Rate	0.484	0.829	0.601
Q Density	1.000	2.233	1.000

Evaluation and Settings. To evaluate the efficacy of LRCD, we adopt the methodology of previous research by assessing the predictive accuracy of student performance, given that the true mastery levels of students are inherently unobservable in real-world contexts. Consistent with previous studies [35, 36], we validate the accuracy of the diagnostic outcomes produced by CDMs by predicting students' performance on assessments. We employ both performance prediction and interpretability metrics [30, 35] to measure effectiveness. Concretely, for the score prediction metric, given that the task is a binary classification problem [35], we employ the Area Under the Curve (AUC) as our evaluation metric. For the interpretability metric, in alignment with previous methodologies [12], we utilize the degree of agreement (DOA) to assess the interpretability of the inferred mastery levels. For a more detailed explanation of the DOA, please refer to Appendix B.2.

For evaluation, we follow the methodology of previous work [17, 36], dividing the response logs of each domain similarly. Specifically, the response logs of each student are partitioned into three parts: 70% for training, 20% for validation, and 10% for testing. During the training phase, we utilize the training data from the source domains to train the model and determine the best hyperparameters using the validation data. During the inference phase, we use the training data in the target domain to infer the students' mastery levels and

evaluate the results on the test data. Notably, the training data in the target domain are not available during the training phase, thus preventing information leakage.

Then, to evaluate the performance of LRCD in zero-shot cross-domain cognitive diagnosis, we conduct comparisons under the following four experimental settings. We solely compare Zero-1-3 in the third setting as it requires overlapping students between the source domain and the target domain.

• **Subject-Level Zero-Shot Student Score Prediction.** In this setting, the source domain and the target domain encompass distinct academic subjects, yet both are derived from the same platform, SLP. For instance, the source domains include Math and Physics, while the target domain is English.

• **Platform-Level Zero-Shot Student Score Prediction.** In this setting, the source domain and the target domain encompass distinct intelligent education platforms, yet both include the same subject, Math. For instance, the source domain is SLP, while the target domain is EDM.

• **Student Score Prediction with Overlap Students.** In this setting, the source domain and the target domain encompass distinct academic subjects, yet both are derived from the same platform, SLP. Different from the first setting, here we consider scenarios where there is an overlap of students between the source domain and the target domain.

• **Standard Setting.** In this setting, the source domain and the target domain are identical. For brevity, we only consider three domains (i.e., SLP-Math, MOOC, EDM) within the same subject, Math. Due to space limitations, the results are presented in Appendix B.4.

Baseline Description. Here, we will introduce our baselines in detail as follows:

• Oracle: It is trained using training data from the target domains, employing integrated traditional CDMs which stands for the upper bound of the performance.

• NCDM [35]: This is a recent classic neural-based CDM. In this study, we train it using the training data from the target domains to compare it with LRCD's zero-shot performance.

• Random: The embeddings of integrated traditional CDMs in the target domain are randomly initialized with values between 0 and 1 which stands for the lower bound of the performance.

• TechCD [7, 42]: It uses a pedagogical knowledge concept graph as a mediator to connect students in the source domain with those in the target domain, thus effectively transferring student cognitive signals from source domains to target domains. Following [6], we utilize the statistical method proposed in [5] to construct the graph.

• GCN-based [7, 42]: Derived from TechCD, the original TechCD obtains more transferable representations by pruning the outputs of some lower GCN layers. The GCN-based method is an ablation of TechCD that directly uses only the embeddings from the final layer of nodes.

• NLP-based [7, 42]: Derived from TechCD, it utilizes learnable embeddings for students and encodes the texts of exercises with OpenAI-3-large [25] to represent them.

• Zero-1-3 [6]: It employs dual regularizers to decouple student embeddings into domain-shared and domain-specific parts. Additionally, it devises a strategy to generate simulated practice logs for new students in the target domain by analyzing the behavioral patterns of early-bird students in the same domain.

Implementation Details. All parameters are initialized using Xavier initialization and optimized with Adam [10] optimizer. We set d as 64 which is the dimension of vectors utilized in integrated CDMs, and set the batch size to 256. For a fair comparison between LRCD and all baselines, we utilize OpenAI-3-large [25] as the default text-embedding module and OR-KaNCD, proposed in ORCDF [36], as the default integrated CDM. The dimensions of the MLP for all methods are consistent, 512 and 256, respectively. We provide a detailed hyperparameter analysis regarding the selection of text-embedding modules and integrated CDMs in Section 5.7. The learning rate is chosen from $\{1e^{-5}, 5e^{-5}, 1e^{-4}, 2.5e^{-4}, 5e^{-4}, 2e^{-3}\}$. All experiments are run on a Linux server with two 3.00GHz Intel Xeon Gold 6354 CPUs and two RTX3090 GPUs. All models are implemented by PyTorch. Details about baselines can be found in Appendix B.3.

5.2 Subject Level Zero-Shot Student Score Prediction

In this subsection, we will compare the performance of our proposed LRCD with other baselines for Student Score Prediction at the subject level. Specifically, the source domains and target domains are derived from datasets pertaining to different academic subjects (e.g., Math, English). Here, for brevity, we denote each subject by its initial letter, for instance, M for Math. Consequently, PB-M signifies that the source domains are Physics and Biology, while the target domain is Math. Notably, we categorize different subjects based on their attributes into Humanities (History, Geography, English, Chinese) and Sciences (Physics, Biology) to analyze the influence between subjects.

Results. As shown in Table 2, we can obtain two key observations. LRCD significantly outperforms other baselines, ***achieving at least 97.30% of the oracle performance and demonstrating competitive results with NCDM in certain scenarios***. This indicates that LRCD is highly effective in subject-level zero-shot cross-domain cognitive diagnosis. TechCD generally performs better than other baselines but still exhibits suboptimal performance. This may be due to the hand-crafted knowledge concept graph, which links completely unrelated concepts across different subjects, potentially hindering its effectiveness in cross-domain scenarios. Moreover, we can find some interesting discoveries: the cross-domain performance is better when the source domain is a science subject compared to when it is a humanities subject, regardless of whether the target domain is a humanities or science subject. This may be due to the greater inherent differences in concepts within humanities subjects, whereas science subjects tend to have relatively smaller disparities. Science data may be more universally applicable in ZSCD. Of course, we cannot rule out the influence of the data size within each subject. Nonetheless, we believe that this warrants further investigation in future work.

5.3 Platform Level Zero-Shot Student Score Prediction

In this subsection, we will compare the performance of our proposed LRCD with other baselines for Student Score Prediction at the platform level. Specifically, the source domains and target domains are derived from datasets pertaining to different education

Table 2: Overall student score prediction performance in subject-level zero-shot cross-domain cognitive diagnosis. Within each method, the highest mean value is highlighted in bold, and the runner-up is underlined. The value following "±" represents the standard deviation of the model's performance. If the mean value significantly differs from the runner-up, passing a t-test with a significance level of 0.01, it is marked with "*". We use the first letter of each subject to represent it. For example, PB-M signifies that the source domains are Physics and Biology, while the target domain is Math.

Datasets	Metrics	Random	GCN	NLP	TechCD	**LRCD**	NCDM	Oracle
PB-M	AUC (%)	50.03±0.75	51.57±0.03	48.86±0.03	52.66±0.03	**80.23*±0.15**	81.18±0.09	81.74±0.10
	DOA (%)	51.32±0.00	50.83±2.11	49.41±1.03	50.59±2.33	**77.08*±0.06**	81.68±0.02	81.31±0.08
HGE-M	AUC (%)	49.24±2.31	50.82±0.01	50.00±0.00	52.53±0.04	**79.54*±0.16**	81.18±0.09	81.74±0.10
	DOA (%)	50.66±0.81	50.58±3.78	50.48±1.35	53.82±1.53	**76.70*±0.09**	81.68±0.02	81.31±0.08
PB-C	AUC (%)	50.41±1.72	49.89±0.02	49.96±0.01	51.68±0.02	**83.32*±0.23**	83.47±0.07	84.30±0.04
	DOA (%)	49.51±0.43	45.09±2.78	50.14±1.65	52.36±4.38	**75.53*±0.07**	81.97±0.06	80.39±1.17
HGE-C	AUC (%)	50.12±1.79	50.46±0.02	50.84±0.02	50.38±0.01	**83.10*±0.22**	83.47±0.07	84.30±0.04
	DOA (%)	49.64±0.66	44.97±2.96	50.44±0.87	48.61±4.81	**75.12*±0.18**	81.97±0.06	80.39±1.17

Table 3: Overall student score prediction performance in platform-level zero-shot cross-domain cognitive diagnosis. Details are the same as Table 2.

Datasets	Metrics	Random	GCN	NLP	TechCD	**LRCD**	NCDM	Oracle
EDM-MATH	AUC (%)	49.09±0.65	51.44±0.03	51.30±0.03	51.70±0.03	**79.76*±0.33**	81.18±0.09	81.74±0.10
	DOA (%)	50.61±1.51	54.90±6.09	50.03±0.68	52.88±2.66	**76.92*±0.33**	81.68±0.02	81.31±0.08
MATH-EDM	AUC (%)	49.37±0.94	51.09±0.01	50.35±0.03	50.17±0.03	**79.41*±0.15**	80.38±0.07	83.48±0.04
	DOA (%)	49.65±0.27	50.53±0.89	49.59±0.89	50.57±3.15	**77.17*±0.14**	85.42±0.02	79.64±0.88
MOOC-MATH	AUC (%)	51.39±3.67	50.17±0.03	48.30±0.03	49.99±0.00	**77.06*±0.34**	81.19±0.09	81.74±0.10
	DOA (%)	50.05±0.94	47.25±0.00	50.21±1.24	50.82±1.52	**75.25*±0.46**	81.68±0.02	81.31±0.08
MATH-MOOC	AUC (%)	48.92±8.06	50.13±0.02	51.26±0.01	50.15±0.01	**81.57*±0.97**	81.97±0.06	89.63±0.05
	DOA (%)	50.14±0.60	50.37±0.06	49.93±0.22	50.13±3.37	**75.87*±0.85**	83.52±0.05	80.05±0.31

Table 4: Ablation Study of LRCD. Details are the same as Table 2.

Datasets	Metrics	LRCD-w/o-TCP	LRCD-w/o-LCM	**LRCD**
PB-M	AUC (%)	75.74	61.32	**80.30***
	DOA (%)	71.44	71.20	**76.75***
PB-C	AUC (%)	80.67	79.10	**83.78***
	DOA (%)	72.94	72.81	**76.25***
EHG-C	AUC (%)	80.38	77.09	**83.17***
	DOA (%)	70.86	65.68	**75.72***

platforms (e.g., MOOC, SLP). For brevity, we use MATH to represent Math from SLP.

Results. As shown in Table 3, LRCD significantly outperforms other baselines, ***achieving at least 94.91% of the oracle performance and demonstrating competitive results with NCDM in certain scenarios***. This indicates that LRCD is highly effective in platform-level zero-shot cross-domain cognitive diagnosis. At the same time, we can discover from the experimental results that when the source domain or target domain is EDM, the model performs better. This may be due to the higher similarity between the exercise on the MATH and EDM platforms. The exercises on these two platforms are entirely within the field of mathematics and are similar in difficulty, leaning towards simpler elementary or middle school math problems. This makes students' problem-solving patterns more similar, which allows the model to perform better on EDM using the data patterns extracted from MATH and vise versa. On the other hand, the MOOC includes not only math problems, but also issues from other fields such as physics and economics. Moreover, the MOOC platform is aimed at older students, typically college students, which means the exercise tends to be more challenging and involve more complex mathematical problems. The significant difference in difficulty between the math problems in MATH and MOOC makes it harder for the model to learn mutually between these two platforms.

5.4 Student Score Prediction with Overlap Students

This subsection compares the performance of the proposed LRCD with other baselines for Student Score Prediction on the same platform SLP. Unlike previous experiments, this time we have constructed a new dataset from SLP, referred to as SLP†, where the source domain and the target domain have overlapping students, consistent with the requirements of Zero-1-3 and TechCD. It comprises 107 students, 2,239 exercises, 93 concepts, and 17,955 response logs. The source domains encompass Biology and Physics, whereas the target domain is Math. Detailed implementations of the baselines are provided in Appendix B.3.

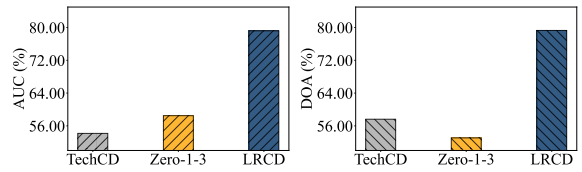


Figure 3: Overall student score prediction performance in overlap students setting on SLP†.

Results. As shown in Figure 3, both TechCD and Zero-1-3 benefit from the overlapping student setting, consistent with their original claims. Their performance has significantly improved compared with previous results. Moreover, LRCD significantly outperforms TechCD and Zero-1-3, achieving an improvement of nearly **35%**. This further validates the effectiveness of LRCD in settings with overlapping students.

5.5 Ablation Study

To validate the efficacy of the Text Cognitive Profiles (TCP) and Language-Cognitive Mappers (LCM) in LRCD, we conduct an ablation study. We present two ablated versions of LRCD: LRCD-w/o-TCP and LRCD-w/o-LCM. Specifically, the former replaces the TCP with a random vector sampled from a standard normal distribution. The latter omits the use of Mappers and directly employs the obtained text embeddings in the language space. Since the dimension of OpenAI-3-large is 3072, which is excessively large, omitting the proposed LCM would result in high GPU demands. Therefore, we use Bert embeddings as a substitute. Unfortunately, LRCD-w/o-LCM still exceeds memory limits in some cases, which also underscores the importance of the proposed LCM. Therefore, we will only report the results for the successful instances.

Results. As illustrated in Table 4, it is evident that LRCD markedly outperforms both ablated versions, substantiating the synergistic effect of integrating both modules. Furthermore, each ablated version of LRCD surpasses the best-performing baseline in Table 1, corroborating the efficacy of each individual module. Surprisingly, we find that LRCD-w/o-TCP also achieves commendable performance, demonstrating the effectiveness of the LRCD framework. This underscores the importance of representing students, exercises, and concepts within the unified space for the ZSCD task. The subpar performance of LRCD-w/o-LCM underscores the considerable disparity between the language space and the cognitive space. This highlights the inadequacy of directly utilizing representations from the language space and emphasizes the critical importance of establishing a mapping between the two spaces.

5.6 Scaling Up in Datasets

This subsection aims to investigate whether expanding the scope of the source domain can enhance the performance of LRCD in the target domain. In Table 5, when our target domain, Math, is included in the training set, we can see that as we continue to add questions from different related domains to the dataset, the model's performance improves steadily. This suggests that adding more related domain questions to the training set can enhance the predictive performance of the target domain in this situation. In Table 6, our target domain, Chinese, is not included in the training set. When we add questions from related domains to the training set, the model's performance shows some fluctuation. This may be because the model is relatively sensitive to the types of source domain in the training set. Adding data that are not closely related to the target domain might affect performance. However, such additions can also enhance the model's generalization ability to some extent, improving the model's robustness and leading to an overall upward trend in performance.

Table 5: Scaling up in training datasets to predict SLP-Math.

Integrated CDM	Metrics	M-M	P-M	PB-M
KaNCD	AUC (%)	81.34	81.39	**81.49**
	DOA (%)	**79.59**	78.92	78.62

Table 6: Scaling up in training datasets to predict SLP-CHI.

Integrated CDM	Metrics	H-C	HE-C	HEG-C
KaNCD	AUC (%)	81.11	80.68	**81.49**
	DOA (%)	75.38	**75.43**	75.20

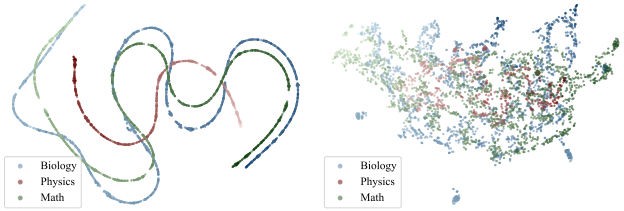
Table 7: Comparison of LRCD with different TEM. Details are the same as Table 2.

Datasets	Metrics	Bert	LLama	ada	3-large
PB-M	AUC (%)	80.30	79.50	80.09	**80.32**
	DOA (%)	76.75	76.16	76.92	**77.21**
EHG-M	AUC (%)	79.72	79.09	**80.17**	79.98
	DOA (%)	76.71	75.93	76.56	**76.75**
PB-C	AUC (%)	**83.78**	83.16	83.71	83.55
	DOA (%)	**76.25**	75.33	75.93	75.48
EHG-C	AUC (%)	83.17	81.37	**83.41**	83.30
	DOA (%)	**75.72**	74.17	75.44	75.18
EDM-MATH	AUC (%)	**80.06**	79.05	80.15	79.99
	DOA (%)	**77.38**	76.10	76.81	76.93
MATH-EDM	AUC (%)	78.52	77.55	78.96	**79.46**
	DOA (%)	76.77	76.45	76.81	**77.14**
MOOC-MAT	AUC (%)	75.39	77.28	**78.97**	77.23
	DOA (%)	73.93	74.15	75.56	**75.76**
MATH-MOOC	AUC (%)	82.79	81.26	**84.17**	82.28
	DOA (%)	**79.22**	77.81	78.43	75.58

5.7 Hyperparameter Analysis

The Effect of TEM. After obtaining the TCP, we utilize advanced TEM to generate representations in the language space. Here, we employ four renowned TEMs (i.e., Bert [3], OpenAI-ada, OpenAI-3-large [25]) for experiments following [18, 19, 27, 28, 31, 33]. As shown in Table 7, within LRCD, using either OpenAI-ada or OpenAI-3-large results in strong performance. We recommend these two options. Notably, Bert also achieves commendable results, making it suitable for resource-constrained scenarios.

The Effect of the Integrated CDM. As LRCD is model-agnostic, we can integrate any existing CDMs. In this study, we employ



(a) Integrated CDM: KaNCD (b) Integrated CDM: OR-KaNCD

Figure 4: t-SNE visualization of students' mastery levels in different domains.

Figure 5: Case study of student profile editing.

four renowned CDMs (i.e., MIRT [32], KSCD [23], KaNCD [36], OR-KaNCD [24]) for our experiments. As shown in Figure 7 in Appendix, OR-KaNCD generally outperforms both KaNCD and KSCD. This shows that OR-KaNCD is not only effective in standard settings but also across various ZSCD settings. Therefore, we recommend OR-KaNCD as the default CDM.

5.8 Diagnosis Report Analysis

Visualization of the Inferred Mastery Levels. This subsection aims to further illustrate the implications of the students' mastery levels inferred by LRCD. We consider Biology and Physics as the source domains and Math as the target domain. Consequently, the inferred mastery levels of students in the source domains are obtained through supervised training, while those in the target domain are inferred without training. Following previous work [36], we employ t-SNE [34], a renowned dimensionality reduction method, to map the mastery levels onto a two-dimensional plane. Firstly, we use different colors to represent different subjects. Then, we use shading to indicate the students' average correct rates, with darker shades representing higher correct rate. Finally, we visualize this using a scatter plot, as shown in Figure 4. We can observe that students with similar average correct rates tend to cluster together. Moreover, the mastery levels of students from different domains are well separated. This indicates that, although LRCD does not explicitly use methods to distinguish different domains and instead trains them together, it can still accurately differentiate students from various domains due to the distinct behavior patterns of each domain.

Student Profile Editing. An additional advantage of LRCD is its capability to directly adjust students' mastery levels by editing their profiles, without necessitating data alteration and model retraining. Here, we select a student s_i as an example who has not done any related exercises on the concept of "Angle" before editing. We give a detailed ID within the corresponding dataset in Appendix B.6. After LRCD training achieves convergence, we fix the parameters and obtain its current representation $\mathbf{h}_{s_i}^{(l)^-}$. We then acquire the student's new interactions in "Angle" and, using the method described in Section 4.1, derive the new representation $\mathbf{h}_{s_i}^{(l)^+}$. Finally, we combine the two representations in a 7:3 ratio and pass them through the student mapper to infer the student's new mastery level. To evaluate the accuracy of the diagnostic results, we assess whether the student's performance on Angle-related exercises has improved. As shown in Figure 5, the third row (i.e., Response Logs) indicates that this student answers all exercises correctly except for e_2. In the second row (i.e., Before Editing), we see the predictions given by LRCD before editing. It is evident that, although the model has not been trained on this student's interactions related to the "Angle", it still provides reasonable predictions. In the first row (i.e., After Editing), after editing, our predictions have all improved compared to the original ones. Notably, on exercise e_2, where the student initially got wrong, our prediction has become more accurate. This shows that LRCD can effectively adjust its assessment of a student's abilities through profile editing. This editing capability is highly beneficial in real educational scenarios, allowing students to preview their potential improvements in advance. By doing so, they can select appropriate exercises to focus on, thereby alleviating their academic burden.

6 Conclusion

This paper proposes a language representation favored zero-shot cross-domain cognitive diagnosis framework (LRCD) to address the limitations of existing CDMs, which often require specific models trained for specific domains. By leveraging textual descriptions to profile students, exercises, and concepts, LRCD transforms these profiles into vectors within a unified language space using advanced text-embedding modules. To bridge the gap between language space and cognitive diagnosis space, we introduce language-cognitive mappers in LRCD, enabling efficient integration and training with existing CDMs. Extensive experiments validate that LRCD achieves commendable zero-shot performance across different target domains and, in some instances, competes with classic CDMs trained on full response data. Notably, LRCD provides intriguing insight into the distinctions between various subjects and educational sources. However, while LRCD shows significant efficacy, further development of more interpretable methods is needed to fully elucidate the mapping process and enhance the framework's applicability in online intelligent education systems.

Acknowledgments

We would like to thank the anonymous reviewers for their constructive comments. We also would like to thank Xinyue Ma for the reliable help. 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, No. 62106076).

References

[1] 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*. New Orleans, LA.

[2] Yuxin Chen, Junfei Tan, An Zhang, Zhengyi Yang, Leheng Sheng, Enzhi Zhang, Xiang Wang, and Tat-Seng Chua. 2024. On Softmax Direct Preference Optimization for Recommendation. *CoRR* abs/2406.09215 (2024).

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Minneapolis, MN, 4171–4186.

[4] NHeffernan Ethan Prihar. 2023. EDM Cup 2023. https://kaggle.com/competitions/edm-cup-2023

[5] 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.

[6] 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*. Vancouver,Canada, 8417–8426.

[7] 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, China, 983–992.

[8] Hengnian Gu, Zhiyi Duan, Pan Xie, and Dongdai Zhou. 2024. Modeling Balanced Explicit and Implicit Relations with Contrastive Learning for Knowledge Concept Recommendation in MOOCs. In *Proceedings of the ACM on Web Conference 2024*. Singapore, Singapore, 3712–3722.

[9] H Tolga Kahraman, Seref Sagiroglu, and Ilhami Colak. 2010. Development of adaptive and intelligent web-based educational systems. In *Proceedings of the 4th international conference on application of information and communication technologies*. Halifax Nova Scotia, Canada, 1–5.

[10] 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.

[11] 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, Singapore, 3420–3431.

[12] 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.

[13] 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.

[14] Sheng Li, Quanlong Guan, Liangda Fang, Fang Xiao, Zhenyu He, Yizhou He, and Weiqi Luo. 2022. Cognitive Diagnosis Focusing on Knowledge Concepts. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. Atlanta, GA, 3272–3281.

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

[16] 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.

[17] 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, Singapore, 4260–4271.

[18] Shuo Liu, An Zhang, Guoqing Hu, Hong Qian, and Tat-seng Chua. 2024. Preference Diffusion for Recommendation. *arXiv preprint arXiv:2410.13117* (2024).

[19] Xiaohao Liu, Zhulin Tao, Jiahong Shao, Lifang Yang, and Xianglin Huang. 2022. Elimrec: Eliminating single-modal bias in multimedia recommendation. In *Proceedings of the 30th ACM International Conference on Multimedia*. Lisboa, Portugal, 687–695.

[20] Yingjie Liu, Tiancheng Zhang, Xuecen Wang, Ge Yu, and Tao Li. 2023. New Development of Cognitive Diagnosis Models. *Frontiers of Computer Science* 17, 1 (2023), 171604.

[21] Zitao Liu, Qiongqiong Liu, Jiahao Chen, Shuyan Huang, and Weiqi Luo. 2023. simpleKT: A Simple But Tough-to-Beat Baseline for Knowledge Tracing. In *Proceedings of the 11th International Conference on Learning Representations*. Kigali, Rwanda.

[22] Yu Lu, Yang Pian, Ziding Shen, Penghe Chen, and Xiaoqing Li. 2021. SLP: A Multi-Dimensional and Consecutive Dataset from K-12 Education. In *Proceedings of the 29th International Conference on Computers in Education*. 261–266.

[23] 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.

[24] 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, 2455–2466.

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

[26] Junhao Shen, Hong Qian, Wei Zhang, and Aimin Zhou. 2024. Symbolic Cognitive Diagnosis via Hybrid Optimization for Intelligent Education Systems. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*. Vancouver, Canada, 14928–14936.

[27] Xiaoteng Shen, Rui Zhang, Xiaoyan Zhao, Jieming Zhu, and Xi Xiao. 2024. PMG: Personalized Multimodal Generation with Large Language Models. In *Proceedings of the ACM on Web Conference 2024*. Singapore, Singapore, 3833–3843.

[28] Leheng Sheng, An Zhang, Yi Zhang, Yuxin Chen, Xiang Wang, and Tat-Seng Chua. 2024. Language Models Encode Collaborative Signals in Recommendation. *arXiv preprint arXiv:2407.05441* (2024).

[29] Jianwen Sun, Fenghua Yu, Sannyuya Liu, Yawei Luo, Ruxia Liang, and Xiaoxuan Shen. 2023. Adversarial Bootstrapped Question Representation Learning for Knowledge Tracing. In *Proceedings of the 31st ACM International Conference on Multimedia*. Ottawa, Canada, 8016–8025.

[30] Jianwen Sun, Fenghua Yu, Qian Wan, Qing Li, Sannyuya Liu, and Xiaoxuan Shen. 2024. Interpretable Knowledge Tracing with Multiscale State Representation. In *Proceedings of the ACM on Web Conference 2024*. 3265–3276.

[31] Yang Sun, Fajie Yuan, Min Yang, Alexandros Karatzoglou, Li Shen, and Xiaoyan Zhao. 2022. Enhancing Top-N Item Recommendations by Peer Collaboration. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Madrid, Spain, 1895–1900.

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

[33] Zhulin Tao, Xiaohao Liu, Yewei Xia, Xiang Wang, Lifang Yang, Xianglin Huang, and Tat-Seng Chua. 2022. Self-supervised learning for multimedia recommendation. *IEEE Transactions on Multimedia* (2022).

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

[35] 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.

[36] 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).

[37] 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.

[38] Shangshang Yang, Xiaoshan Yu, Ye Tian, Xueming Yan, Haiping Ma, and Xingyi Zhang. 2023. Evolutionary Neural Architecture Search for Transformer in Knowledge Tracing. In *Advances in Neural Information Processing Systems 36*. New Orleans, LA.

[39] 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).

[40] Yan Zhuang, Qi Liu, Zhenya Huang, Zhi Li, Shuanghong Shen, and Haiping Ma. 2022. Fully Adaptive Framework: Neural Computerized Adaptive Testing for Online Education. In *Proceeddings of the 36th AAAI Conference on Artificial Intelligence*. Virtual Event, 4734–4742.

[41] 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.

[42] Jianhuan Zhuo, Jianxun Lian, Lanling Xu, Ming Gong, Linjun Shou, Daxin Jiang, Xing Xie, and Yinliang Yue. 2022. Tiger: Transferable Interest Graph Embedding for Domain-Level Zero-Shot Recommendation. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. Atlanta, GA, 2806–2816.

Appendix

The appendix is organized as follows:

• Appendix A presents a detailed comparison of training time and inference time with other baselines.

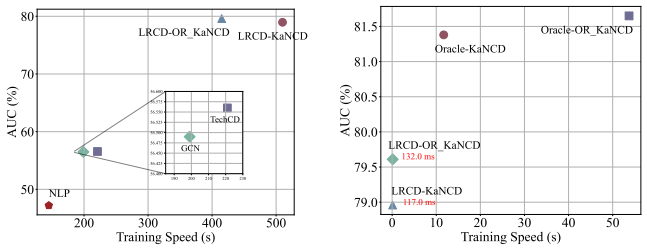
• Appendix B provides additional details on the experiments carried out in this paper, including the details about datasets, the explanation of DOA, implementation of baselines, experimental results in the standard setting, as well as details on hyperparameter analysis and case studies.

A Time Comparison

Here, we use Physics and Biology as the source domains, and Math as the target domain, as an example to explore the comparison of different models in terms of training time and inference time.

A.1 Training Time

As discussed in Section 4.4, by employing a space-for-time strategy, we can store the obtained text embeddings locally, thereby minimizing additional time requirements. Here, we provide the actual running time of LRCD compared with the baselines in ZSCD. Note that the training time here is the time it takes for the models to reach the optimal AUC. As shown in Figure 6(a), we select KaNCD and OR-KaNCD for our model. It can be seen that although our model takes longer time, it has fine predictive performance compared to other baselines.



(a) (b)

Figure 6: (a) Training time comparison with baselines. (b) Inference time comparison with the model retraining

A.2 Inference Time

As we claim in the introduction, when new domain's response logs emerges, OIDP often need to leverage prior knowledge to quickly and accurately provide diagnostic results for students without retraining models. Therefore, inference time, which equates to the system's response time to the user, is crucial. Here, we compare the inference time with the model retraining time.

As shown in Figure 6(b), we selected KaNCD and OR-KaNCD as diagnostic methods. For KaNCD, the diagnostic time of our model is 99 times longer than that of the oracle, while OR-KaNCD is 406 times longer. This indicates that our model has good time utility while maintaining great diagnostic performance, and can quickly diagnose "zero sample" students.

B Experiments

B.1 Details about the Datasets

Here, we provide detailed information regarding the SLP dataset, including the number of students and exercises for each subject in Table 8. The specific details align with those presented in Table 1.

• **SLP** [22]: The SLP dataset for K-12 education encompasses five dimensions: student demographics, psychometric intelligence, academic performance, family, and school information. It automatically records students' academic performance in eight subjects (Math, Physics, Chemistry, Biology, History, Chinese, Geography, English) over three years (7th to 9th grade).

• **MOOC** [39]: The MOOC dataset supports cognitive student modeling in Massive Open Online Courses by providing learning resources, structures, and content related to students' exercise behaviors. It also includes Chinese contextual information for exercises and concepts.

• **EDM** [4]: Originating from the EDM Cup 2023 competition, the EDM dataset focuses on predicting students' end-of-unit assignment scores using their click-stream data from previous in-unit assignments on the ASSISTments platform. It contains millions of student actions along with detailed information on the curricula, assignments, problems, and the tutoring provided.

Table 8: Details about different subjects in SLP.

Datasets	Math	Chinese	Physics	Biology	English	History	Geography
#Students	1,475	623	639	1,940	306	1,603	1,077
#Exercises	615	510	1,441	773	355	752	427
#Concepts	33	17	50	16	18	20	25
#Response logs	55,332	29,202	37,405	81,838	3,872	63,207	28,535
Average Correct Rate	0.550	0.588	0.611	0.503	0.366	0.418	0.371
Q Density	1,000	1.000	1.000	1.000	1.000	1.000	1.002
Category	Science	Humanity	Science	Science	Humanity	Humanity	Humanity

B.2 Degree of Agreement (DOA)

Here, we provide a further explanation regarding the degree of agreement. Suppose that the inferred students' mastery levels are represented by $\mathbf{Mas} \in \mathbb{R}^{N\times K}$, where N denotes the number of students and K signifies the number of concepts. The underlying intuition here is that if the student s_a demonstrates higher accuracy in answering exercises related to the concept c_k compared to the student s_b, then the probability of s_a mastering c_k should be greater than that of s_b. In other words, $\mathbf{Mas}_{s_a,c_k} > \mathbf{Mas}_{s_b,c_k}$. The Degree of Agreement (DOA) is defined as in Eq. (8)

$$\mathrm{DOA}_k = \frac{1}{Z} \sum_{a,b\in S} \delta\left(\mathbf{Mas}_{s_a,c_k}, \mathbf{Mas}_{s_b,c_k}\right) \frac{\sum_{j=1}^{M} \mathbf{Q}_{jk} \wedge \varphi(j,a,b) \wedge \delta(r_{aj}, r_{bj})}{\sum_{j=1}^{M} \mathbf{Q}_{jk} \wedge \varphi(j,a,b) \wedge I(r_{aj} \neq r_{bj})}, \quad (8)$$

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, $\varphi(j,a,b)$ verifies if both students s_a and s_b answered e_j, r_{aj} represents the response of s_a to e_j, and $I(r_{aj} \neq r_{bj})$ checks if their responses are different, $\delta(r_{aj}, r_{bj})$ is 1 for a correct response by s_a and an incorrect response by s_b, and 0 otherwise.

B.3 Details about Baselines

In the following, we elaborate on some details regarding the utilization of the compared methods.

• KSCD [23] explores the implicit association among concepts and leverages a knowledge-enhanced interaction function.

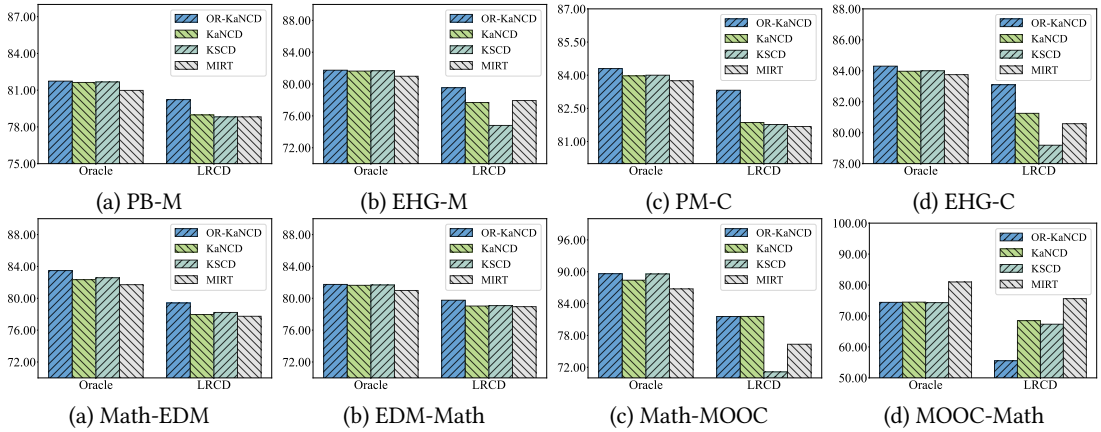


Figure 7: Comparison of LRCD with different integrated CDM.

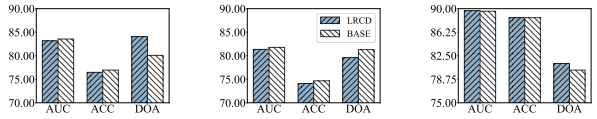


Figure 8: Comparison with OR-KaNCD in the standard setting.

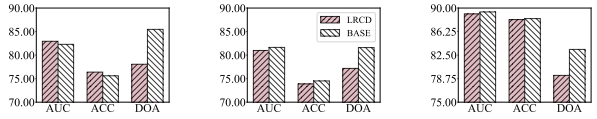


Figure 9: Comparison with KaNCD in the standard setting.

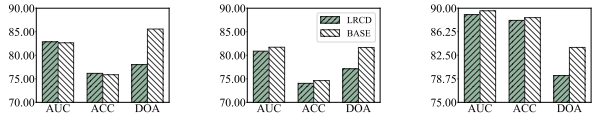


Figure 10: Comparison with KSCD in the standard setting.

• KaNCD [36] enhances NCDM by investigating the implicit associations among concepts to address the issue of knowledge coverage. Here, we adopt the default parameters in the paper.

• TechCD [7] uses a pedagogical knowledge concept graph as a mediator to connect students in the source domain with those in the target domain, thus effectively transferring student cognitive signals from source domains to target domains. Following [6], we utilize the statistical method proposed in [5] to construct the graph.

• Zero-1-3 [6] uses dual regularizers to split student embeddings into domain-shared and domain-specific components. It generates simulated practice logs for new target domain students by analyzing early-bird behaviors. Since exercise content is unavailable, we create exercise texts as described in Section 4.1 and embed them using BERT [3]. For early bird students, we randomly select 10% of the target domain due to the limited number of students.

The implementation of MIRT, KaNCD comes from the public repository https://github.com/bigdata-ustc/EduCDM. For KSCD, we adopt the implementation from the authors in https://github.com/BIMK/Intelligent-Education/tree/main/KSCD_Code_F. For ORKaNCD, we adopt the implementation from the authors in https://github.com/ECNU-ILOG/ORCDF. For TechCD, we adopt the implementation from the authors in https://github.com/bigdata-ustc/TechCD. For Zero-1-3, since there was no publicly available code when we submitted this paper, we have implemented it ourselves.

B.4 Standard Setting

In this subsection, we will compare the performance of our proposed LRCD with other baselines for Student Score Prediction in the standard setting. Specifically, the source domains and target domains originate from the same subject and platform. As shown in Figure 8, Figure 9 and Figure 10, even in the standard setting where the source domain and target domain are identical, our method, despite not being explicitly tailored for this scenario, still demonstrates commendable performance.

B.5 Hyperparameter Analysis

The Effect of the Integrated CDMs. Here, we provide the results in Figure 7. Detailed analysis can be found in Section 5.7.

B.6 Case Study

The student selected for the case study is from SLP-Math, identified as "00ad006d0bcf06158f49fb0580Abd957". We believe that student profile editing is highly beneficial in real educational scenarios. It allows students to preview their potential improvements in advance, allowing them to choose appropriate exercises to focus on, thereby reducing their academic burden. Detailed analysis can be found in the Student Profile Editing subsection in Section 5.8.