Embedding Enhancement via Fine-Tuned Language Models for Learner-Item Cognitive Modeling

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

Zihan Zhou*
East China Normal University
Shanghai, China
zhzhou@stu.ecnu.edu.cn

Kaiying Wu*
East China Normal University
Shanghai, China
10235102479@stu.ecnu.edu.cn

Shuo Liu
Tencent Inc
Shenzhen, China
seokliu@tencent.com

Yiyang Huang
East China Normal University
Shanghai, China
10235102470@stu.ecnu.edu.cn

Jiajun Guo
East China Normal University
Shanghai, China
jjguo@psy.ecnu.edu.cn

Aimin Zhou
East China Normal University
Shanghai, China
Shanghai Innovation Institute
Shanghai, China
amzhou@cs.ecnu.edu.cn

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

Abstract

Learner-item cognitive modeling plays a central role in the web-based online intelligent education system by enabling cognitive diagnosis (CD), the upstream and crucial component of the system, across increasingly diverse online educational scenarios. Although ID embedding remains the mainstream approach in cognitive modeling due to its effectiveness and flexibility, recent advances in language models (LMs) have introduced new possibilities for incorporating rich semantic representations to enhance CD performance. However, current studies often focus on a specific task, such as zero-shot CD, limiting the broader application of LMs in this field. This highlights the need for a comprehensive analysis of how LMs enhance embeddings through semantic integration across mainstream CD tasks. This paper identifies two key challenges in fully leveraging LMs in existing work: Misalignment between the training objectives of LMs and CD models creates a distribution gap in feature spaces, hindering the potential of LMs for embedding enhancement; A unified framework is essential for integrating textual embeddings across varied CD tasks while preserving the strengths of existing cognitive modeling paradigms, such as ID embeddings, to ensure the robustness of embedding enhancement. To address these challenges, this paper introduces EduEmbed, a unified embedding enhancement framework that leverages fine-tuned LMs to enrich learner-item cognitive modeling across diverse CD tasks. EduEmbed operates in two stages. In the first stage called role-aware interactive fine-tuning, we fine-tune LMs based on role-specific representations and an interaction diagnoser to bridge the semantic gap of CD models. In the second stage called adapter-aware representation integration, we employ a textual adapter to extract task-relevant semantics and integrate them with existing modeling paradigms to improve generalization across diverse CD tasks. We evaluate the proposed framework on four CD tasks and computerized adaptive testing (CAT) task, achieving robust performance. Further analysis reveals the impact of semantic information across diverse tasks, offering key insights for future research on the application of LMs in CD for online intelligent education systems.

CCS Concepts

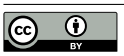
• **Computing methodologies** → Machine learning; • **Applied computing** → Education.

Keywords

Learner-Item Cognitive Modeling, Cognitive Diagnosis, Computerized Adaptive Testing, Embedding Enhancement, Web-based Intelligent Education Systems

ACM Reference Format:
Yuanhao Liu, Zihan Zhou, Kaiying Wu, Shuo Liu, Yiyang Huang, Jiajun Guo, Aimin Zhou, and Hong Qian. 2026. Embedding Enhancement via Fine-Tuned Language Models for Learner-Item Cognitive Modeling. In *Proceedings of the ACM Web Conference 2026 (WWW '26), April 13–17, 2026, Dubai, United Arab Emirates.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3774904.3792542

*These authors contribute equally to this research.
†Corresponding author: Hong Qian (hqian@cs.ecnu.edu.cn).



WWW '26, Dubai, United Arab Emirates.

ACM ISBN 979-8-4007-2307-0/2026/04
https://doi.org/10.1145/3774904.3792542

Resource Availability:
The source code of this paper has been made publicly available at https://doi.org/10.5281/zenodo.18301397 and https://github.com/BW297/EduEmbed.

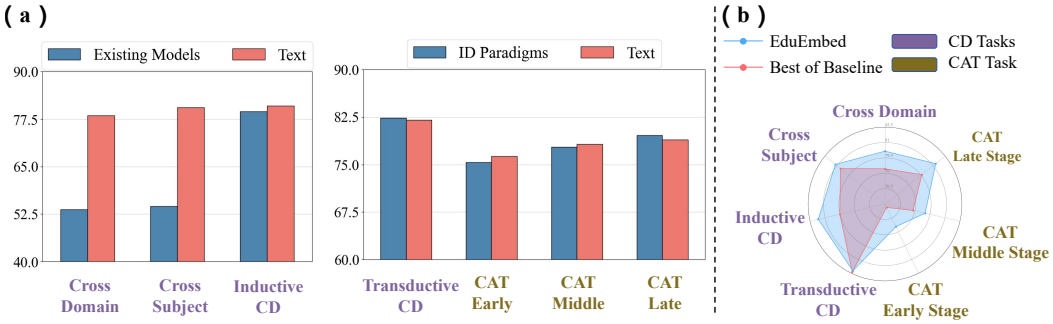


Figure 1: (a) Motivation study. (b) The comparison of our proposed EduEmbed with best-performing baseline methods on SLP.

1 Introduction

With the growing demands of personalized learning, web-based online intelligent education systems [17] have emerged as a critical development direction. Cognitive Diagnosis (CD) [16, 18, 22, 26, 32], as a crucial upstream component of the system, aims to infer students' mastery level of specific concepts by analyzing their past interaction records. The diagnosis results can also support further customized applications, such as Computerized Adaptive Testing (CAT) [39, 40]. Currently, these technologies have been widely applied in modern web-based online education platforms [5], and single-task scenario settings are no longer sufficient to meet real-world demands. For example, in the field of CD, a variety of scenarios have been proposed and actively studied, including traditional transductive CD [28, 33, 34] for daily practice tests, inductive CD [14, 19, 21] for large-scale, dynamic open student learning environments, zero-shot CD [7, 9, 20] for interdisciplinary and cross-domain settings and CAT [1, 39, 40], as a downstream application of CD, for online standardized testing scenarios.

As the foundational module for CD, learner-item cognitive modeling [6, 15, 25] learns latent representations of learners (e.g., students) and items (e.g., exercises, concepts) via embedding construction, and its quality directly affects aforementioned task performance. ID embedding, which maps entity IDs to latent vectors, has long been the dominant paradigm due to its effectiveness and flexibility, but it struggles to generalize across increasingly diverse CD tasks. Recently, the advancements in language models (LMs) [3, 29, 30] offer new possibilities. Natural language offers a unified interface for modeling diverse CD tasks and pretraining, particularly in large language models, captures rich open-world knowledge, enabling more informative semantic representations. However, most LM-based CD works remain limited to single tasks such as zero-shot CD [7, 20]. ***Therefore, there is a lack of a comprehensive analysis on the effectiveness of textual semantic embedding generated by LMs across mainstream CD tasks.***

As shown in Figure 1 (a), we compare the pure textual embeddings generated by the original LMs without any additional training against the best-performing models in each task that do not use textual embeddings, across multiple CD scenarios and different stages of CAT. Detailed experimental settings are provided in Appendix A. The results show that the embedding enhancement brought by textual semantic information varies across different CD tasks. Therefore, understanding the enhancement these embeddings bring to each task, as well as the potential improvement space introduced by incorporating textual semantic information in different CD tasks, is essential for assessing the value of textual semantic embedding enhancement and guiding future applications of LMs in CD. In investigating this, we identify two widespread challenges for applying LMs to CD in current research: **(1) Training objectives misalignment:** A key challenge lies in the misalignment between the training objectives of general LMs and learner-item cognitive modeling in CD models. This often leads to a distribution gap between LM-generated embeddings and the feature space of mainstream CD frameworks, limiting the potential of LMs for embedding enhancement. Aligning LMs semantic pattern with CD models representation may be crucial to unlock full potential of LMs in embedding enhancement. **(2) Lack of a unified integration framework:** Given the diversity of CD tasks, there is currently no unified integration paradigm that allows textual embeddings to be seamlessly incorporated across varied scenarios while preserving the strengths of existing learning paradigms, such as ID embeddings. This lack of generalizability makes it difficult to ensure a performance lower bound across tasks, thereby limiting the robustness of embedding enhancement.

To address these challenges, this paper proposes EduEmbed, a unified embedding enhancement framework that leverages fine-tuned LMs to enrich learner-item cognitive modeling across diverse CD tasks. The framework consists of two stages. In the first stage, it is assumed that LMs have acquired extensive external knowledge during pretraining. Therefore, we aim to activate their capacity for learner-item cognitive modeling through fine-tuning, which facilitates their adaptation to CD models by aligning the training objectives of LMs with those of CD models to a certain extent. We propose role-aware interactive fine-tuning, where we produce textual embeddings aligned with CD models feature spaces, thereby unlocking the full potential of embedding enhancement. In the second stage, adapter-aware representation integration, we propose a unified paradigm to integrate mainstream ID embeddings and textual embeddings. By preserving the strengths of ID embeddings, this paradigm enhances the generalization and robustness of embedding enhancement across diverse CD tasks. Benefiting from this two-stage design, EduEmbed consistently achieves robust performance on four representative CD tasks and a downstream CAT task. Moreover, the analysis of the impact of semantic information under

diverse CD tasks offers valuable insights for future research about LMs application in CD for online intelligent education systems.

2 Related Work

2.1 Learner-Item Cognitive Modeling in Cognitive Diagnosis

CD is a vital field in educational psychology, which is used to infer students' mastery levels for each concept by their response logs. Since responses are noisy indicators influenced by guessing and item properties, a student's mastery level is considered as latent, determining response correctness together with these related properties. Learner-item cognitive modeling serves as the representation learning module in CD, aiming to construct latent representations of learners (e.g., students) and items (e.g., exercises, concepts) via embedding. Most existing methods follow the ID-based embedding paradigm. They can be divided by mastery dimension into two types: latent factor models (e.g., MIRT [28]) that represent students' mastery as fixed-length vectors, and concept-based models (e.g., DINA [2]) that use concept-specific mastery patterns. With deep learning advancements, more flexible models have emerged. For example, NCDM[33] uses MLPs as interaction functions and models mastery as continuous variables in $[0, 1]$. Recent learner-item cognitive modeling methods include MLP-based[34], graph-based [6, 25], and Bayesian network-based methods [15].

However, with the increasing diversity of CD task scenarios, the ID-based paradigm is no longer sufficient to support all applications. In inductive CD, IDCD [14] replaces ID embeddings with interaction matrices to model the cognitive states of entities. In zero-shot CD, TechCD [9] leverages transferable hand-crafted knowledge graph structures to overcome the limitations of ID embeddings across domains. Meanwhile, models like ZeroCD [7] and LRCD [20] introduce textual semantic representation learning to replace ID embeddings, significantly enhancing generalization in zero-shot CD tasks. It is evident that LMs have begun to emerge in learner-item cognitive modeling, but their use in CD remains limited. Given the strong generalization ability of natural language, its potential across diverse CD scenarios deserves deeper exploration.

2.2 The Appilication of Language Models in Intelligent Education

Among the major application scenarios for LMs in education, two related scenarios are introduced as follows. First, LMs are employed as agents to simulate learner behavior. For example, EduAgent [36] leverages LLM-based agents to mimic learners' engagement with PowerPoint presentations and videos. Agent4Edu [8] uses LLM as response generators to simulate learner response data, thereby supporting the training and evaluation of downstream educational tasks. Second, LMs have been used as embedders to encode textual information into vector representations, which is the focus of our work. For instance, NCDM+ [33] utilizes exercise text via TextCNN [13] to complete the Q-Matrix in CD. ECD [38], which fuses student context-aware features (e.g., parental education level, monthly study expenses) into representations of students in cognitive diagnosis. ZeroCD [7] use exercise contents [27] as textual features to serve as a mediator between the students in source and target domains. LRCD [20] further analyzes the behavior patterns among students, exercises, and knowledge concepts to construct unified textual cognitive representations, supporting zero-shot CD. Depite these efforts, current applications of LMs in CD are still simplistic, lacking in-depth adaptation, which may limit their effectiveness. Moreover, most existing methods rely heavily on rich textual data, failing to fully leverage the broad knowledge coverage of LMs and thus, limiting the effectiveness of these methods in real-world educational scenarios.

Although these embedding-based approaches have shown improvements in educational tasks, most of them still rely on LLMs. The lack of deep adaptation to educational datasets often results in suboptimal embeddings, limiting the effectiveness of these methods in real-world educational scenarios.

3 Preliminaries

Consider an educational scenario of a web-based online intelligent education system, which involves M students $S = \{s_1, s_2, ..., s_M\}$, N exercises $E = \{e_1, e_2, ..., e_N\}$, and K concepts $C = \{c_1, c_2, ..., c_K\}$. The corresponding response logs $R = \{(s_i, e_j, r_{ij}) | s_i \in S, e_j \in E, r_{ij} \in \{0, 1\}\}$ consist of a set of triplets (s_i, e_j, r_{ij}), where r_{ij} represents the score obtained by student s_i on exercise e_j. $r_{ij} = 1$ indicates that the student answered the question correctly and $r_{ij} = 0$ indicates otherwise. Additionally, $\boldsymbol{Q} = \left\{q_{j,k}\right\}_{N \times K}$ is a binary matrix representing the relationship between exercises and concepts, where $q_{j,k} = 1$ indicates that exercise e_j relates to concept c_k and $q_{j,k} = 0$ indicates otherwise.

Cognitive Diagnosis Basis. Given the student's response log R and the matrix $\boldsymbol{Q}$, the goal of the CD task is to infer the student's mastery $\boldsymbol{Mas} \in \mathbb{R}^{M \times K}$ on knowledge concepts. Building on this, we will introduce the following four specific educational scenarios and provide detailed explanations of their application in experiments.

• Transductive Cognitive Diagnosis. In this scenario, we assume the set of students and exercises is known and fixed. The CD model uses the known student-exercise score matrix $\boldsymbol{A} \in \mathbb{R}^{M \times N}$ and the exercise-concept relationship matrix $\boldsymbol{Q} \in \mathbb{R}^{N \times K}$ to infer the latent knowledge mastery $\boldsymbol{Mas} \in \mathbb{R}^{M \times K}$ of all students. The goal of this method is to infer students' mastery based on the existing response data.

• Inductive Cognitive Diagnosis. This scenario takes into account the addition of new students and requires the model to evaluate the knowledge mastery of new students without retraining. Given that the set of existing students S_o and the set of new students S_u do not overlap, i.e., $S_o \cap S_u = \emptyset$, the goal is to predict the knowledge mastery of new students $\boldsymbol{Mas}_u \in \mathbb{R}^{|S_u| \times K}$ based on the response data of the existing students, thus enabling inductive reasoning of the model.

• Domain-Level Zero-Shot Cognitive Diagnosis. In this scenario, we assume we have response logs from H source domains $R_s = \{R_1, R_2, ..., R_H\}$. The goal is to train a CD model on the source domains and then infer in the target domain T, where the target domain has no overlap with the source domain in terms of exercises and concepts, i.e., $E_s \cap E_t = \emptyset, C_s \cap C_t = \emptyset$. In this case, the CD models adapts to the students S_t in the target domain and predict their knowledge mastery levels $\boldsymbol{Mas}_t \in \mathbb{R}^{M \times K}$.

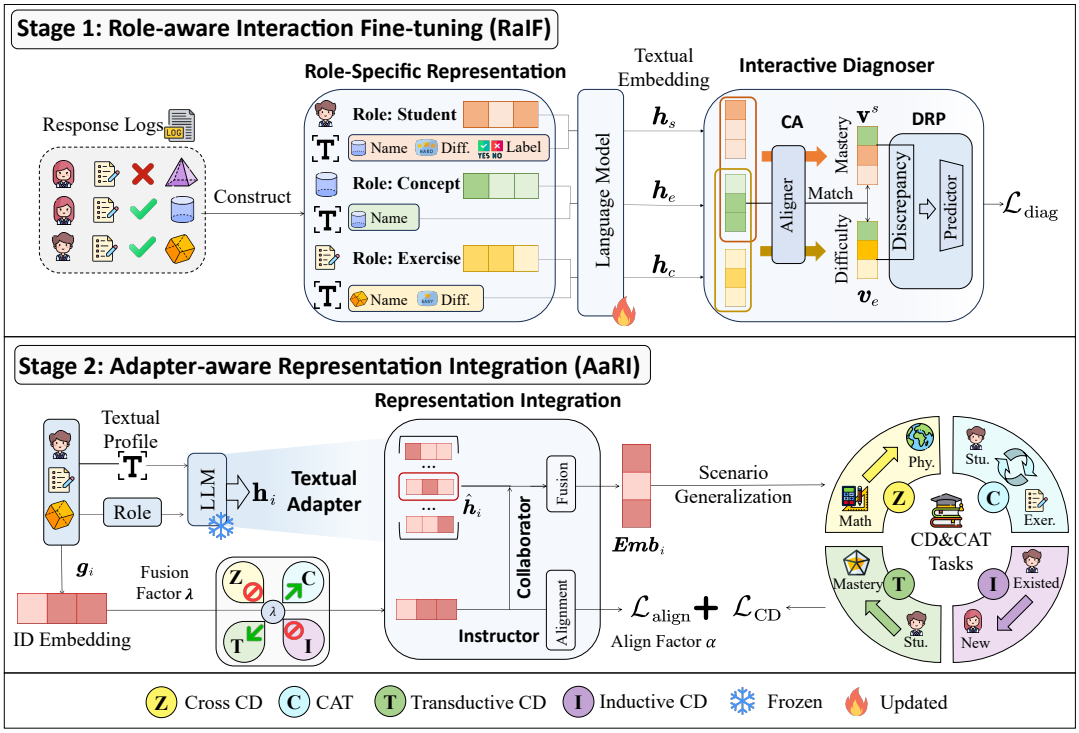


Figure 2: The overall framework of the proposed EduEmbed. Stage 1: Role-aware Interaction Fine-tuning (RaIF). Stage 2: Adapter-aware Representation Integration (AaRI).

• Computerized Adaptive Testing (CAT). In this scenario, the CD model alternates with the selection strategy to form a feedback loop. At each time step $t \in [1, T]$, a student i will update their mastery level based on the answered questions $R_{t-1,i} = \{(e_1, r_1), (e_2, r_2), ..., (e_{t-1}, r_{t-1})\}$. The CD models will estimate the student's mastery at time t as $\hat{\boldsymbol{Mas}}_i^t = \boldsymbol{Mas}(R_{t-1,i})$, i.e., the model infers the current mastery level based on previous performance. Then, based on the item selection strategy π, the systems will choose a new question e_t for the student to answer. The student's feedback will update the mastery level. This process will continue for T steps, with the ultimate goal being for the student's final mastery estimate $\hat{\boldsymbol{Mas}}_i^T$ to be as close as possible to the true ability $\boldsymbol{Mas}_i^*$ at the end of the test.

Learner-Item Cognitive Modeling. Given the response logs R and the Q matrix $\boldsymbol{Q}$, the objective of learner-item cognitive modeling is to learn latent representations of learner (e.g., students) and items (e.g., exercises and concepts). These representations for task t are denoted as $\boldsymbol{Emb}_s^t \in \mathbb{R}^{M \times d_t}$, $\boldsymbol{Emb}_e^t \in \mathbb{R}^{N \times d_t}$, and $\boldsymbol{Emb}_c^t \in \mathbb{R}^{K \times d_t}$, respectively, where d_t is the embedding dimension of task t. These embeddings serve as foundational representations to support various CD tasks.

4 Methodology: The proposed EduEmbed

In this section, we provide a detailed introduction to EduEmbed which consists of two main stages: Role-aware Interaction Fine-tuning and Adapter-aware Representation Enhancement. The overall framework of EduEmbed is illustrated in Figure 2.

4.1 Role-aware Interaction Fine-tuning (RaIF)

This subsection first describes how to design personalized descriptions for three educational roles, students, exercises, and concepts, combined with corresponding encodings to obtain role-specific representations. Then, the constructed textual inputs are fed into the LMs, followed by an explanation of how the model is fine-tuned using an interaction diagnoser to generate textual embeddings that align with CD models.

4.1.1 *Role-specific Representation.*

Inspired by [20], we design personalized descriptions for students, exercises, and concepts to capture their behavior patterns in the dataset. Specifically, the textual description for each educational role is constructed based on its corresponding attributes A, with the attribute description following a standardized format of $< name\ is\ value >$. Specifically, for concept c_k, the attribute is the concept name; for exercise e_j, the attributes include the concepts involved and the average accuracy rate $\text{ACR}_{e_j} = \frac{1}{z_j} \sum_i r_{ij}$, where z_j denotes the set of students s who have completed exercise e_j and r_{ij} denotes the response of student s_i to exercise e_j; for student s_i, the attributes are based on the exercises completed and the corresponding responses. The formal description of attribute A of the three roles is given below:

$$\begin{cases} \text{A}_{c_k} = \text{Name}_{c_k} \\ \text{A}_{e_j} = \left[\left\{ \text{A}_{c_k} \mid \boldsymbol{Q}_{j,k} = 1 \right\}, \text{ACR}_{e_j} \right] \\ \text{A}_{s_i} = \left\{ \left[\text{A}_{e_j}, r_{ij} \right] \mid (s_i, e_j, r_{ij}) \in R \right\} \end{cases}. \tag{1}$$

These attributes have minimal dataset demands, making them effective even when textual data is limited. This addresses a key challenge in current educational datasets and enhances real-world applicability. Further analysis on richer textual inputs such as exercise contents is provided in Section 5.2.4. However, relying solely on descriptions is often insufficient to effectively distinguish educational roles. For example, the textual descriptions of students and exercises may be highly similar, with the only difference being whether there is a response. Such semantic similarity may lead to ambiguity in role alignment within the LMs. Thus, we introduce a token-level learnable role embedding $\boldsymbol{p}_{\text{role}} \in \mathbb{R}^{1\times d_{\text{LM}}}$ with role $\in$ {Student, Exercise, Concept}, which distinguishes three entity types independent of the text descriptions. We define the token combination as follows:

$$\boldsymbol{p} = \boldsymbol{p}_{\text{base}} + \boldsymbol{p}_{\text{role}}, \tag{2}$$

where $\boldsymbol{p}_{\text{base}} \in \mathbb{R}^{1\times d_{\text{LM}}}$ is the base word token, $\boldsymbol{p} \in \mathbb{R}^{1\times d_{\text{LM}}}$ denotes the final token. Then we feed $\boldsymbol{p}$ into the LMs to obtain the sentence-level textual representation $\boldsymbol{h} \in \mathbb{R}^{1\times d}$, where d is the dimension produced by a classification head applied on the LMs' hidden state of the final layer. Notably, as the student s_i may have multiple responses, we apply average pooling to aggregate all corresponding embeddings to obtain the final textual representation h_{s_i}.

4.1.2 *Interactive Diagnoser.*

We introduce the interactive diagnoser to fine-tune LMs, thereby aligning the training objectives between LMs and CD models. Through this design, the textual embeddings generated by the LMs can mitigate the distribution gap in the feature space of CD models to some extent.

Concept Aligner. To enhance the educational interpretability of both students and exercises in the semantic space, we propose a Concept Aligner that projects the textual embeddings of both students and exercises into the concept space. Formally, given the personalized textual embedding of a student s_i as $\boldsymbol{h}_{s_i} \in \mathbb{R}^{1\times d}$ and that of an exercise e_j as $\boldsymbol{h}_{e_j} \in \mathbb{R}^{1\times d}$, we align both to the concept embedding matrix $\boldsymbol{H}_c \in \mathbb{R}^{K\times d}$, where K is the number of concepts. We get $\boldsymbol{v}_{s_i} = \boldsymbol{h}_{s_i} \cdot \boldsymbol{H}_c^\top \in \mathbb{R}^{1\times K}$ as the mastery level of student s_i on each concept c_k and $\boldsymbol{v}_{e_j} = \boldsymbol{h}_{e_j} \cdot \boldsymbol{H}_c^\top \in \mathbb{R}^{1\times K}$ as the difficulty level of exercise e_j on each concept c_k.

Discrepancy-based Response Predictor. Furthermore, we propose a Discrepancy-based Response Predictor (DRP) to model the interaction function between students and exercises. As mentioned in Section 2.1, MIRT [28] is a representative latent factor model that encodes students' mastery using fixed-dimensional vectors and has been widely used in prior CD studies, where it has consistently shown near-SOTA performance in transductive CD tasks. In this paper, we adopt MIRT as our interaction function to avoid introducing additional learnable parameters during the modeling of student-exercise interactions, which would otherwise require optimizing both the embeddings and the interaction process during fine-tuning, where the predicted score of student s_i on exercise e_j can be formulated as:

$$\hat{r}_{ij} = \sigma(\boldsymbol{q}_j^\top(\boldsymbol{v}_{s_i} - \boldsymbol{v}_{e_j})), \tag{3}$$

where $\sigma(\cdot)$ is the sigmoid function and $\boldsymbol{q}_j$ denotes the row in the Q matrix $\boldsymbol{Q}$ corresponding to exercise e_j, indicating the concepts included in exercise e_j. Building on this, we apply the BCE loss as the fine-tuning loss for task-specific supervision for interaction modeling. It can be formulated as:

$$\mathcal{L}_{\text{diag}} = -\frac{1}{|R|}\sum_{(s_i,e_j,r_{ij})\in R}\left[r_{ij}\log\hat{r}_{ij} + \left(1-r_{ij}\right)\log\left(1-\hat{r}_{ij}\right)\right], \tag{4}$$

where $r_{ij} \in \{0,1\}$ represents the actual response of student s_i to exercise e_j (correct or incorrect) in response logs R, and $\hat{r}_{ij}$ is the predicted score.

4.2 Adapter-aware Representation Integration (AaRI)

This subsection first introduces how to leverage the textual embeddings generated by fine-tuned LMs in Section 4.1 by employing a textual adapter to extract task-relevant semantics. Subsequently, we explain how the ID embeddings are utilized to assist in representation integration of the textual embeddings, ultimately producing high-quality embeddings that can be applied to diverse CD tasks.

4.2.1 *Textual Adapter.*

We believe that the textual embeddings generated through **RaIF** in Section 4.1 effectively capture general cognitive traits of educational roles. To preserve these general traits, we freeze the fine-tuned LM parameters to ensure consistency across CD tasks. However, since the educational domain involves multiple tasks, each with different demands for these traits, we introduce a textual adapter to extract task-specific semantics. It helps CD models focus on the core traits relevant to the task, thereby significantly enhancing the performance without additional training burdens. The adaptation process can be formulated as:

$$\hat{\boldsymbol{h}}_{s_i}^t = \mathcal{A}_s^t(\boldsymbol{h}_{s_i};\boldsymbol{\theta}_s^t),\ \hat{\boldsymbol{h}}_{e_j}^t = \mathcal{A}_e^t(\boldsymbol{h}_{e_j};\boldsymbol{\theta}_e^t),\ \hat{\boldsymbol{h}}_{c_k}^t = \mathcal{A}_c^t(\boldsymbol{h}_{c_k};\boldsymbol{\theta}_c^t), \tag{5}$$

where $\hat{\boldsymbol{h}}_{s_i}^t, \hat{\boldsymbol{h}}_{e_j}^t, \hat{\boldsymbol{h}}_{c_k}^t \in \mathbb{R}^{1\times d_t}$ are the task t-relevant embeddings corresponding to student s_i, exercise e_j, and concept c_k, and d_t is the latent dimension in task t. $\mathcal{A}_s^t$, $\mathcal{A}_e^t$, and $\mathcal{A}_c^t$ denote the adapters of students, exercises, and concepts for task t respectively, where $\boldsymbol{\theta}_s^t$, $\boldsymbol{\theta}_e^t$, and $\boldsymbol{\theta}_c^t$ are the parameters. In this paper, we represent the adapter as MLPs.

4.2.2 *Representation Integration.*

In this subsection, we propose a unified paradigm for integrating textual and ID embeddings, since ID embeddings serve as a mainstream and effective approach in most CD tasks, particularly in transductive CD [25, 28, 34] and CAT [39, 40] task. Specifically, ID embeddings act as both an instructor and a collaborator to guide the alignment and fusion process, aiming to preserve their strengths while ensuring a performance lower bound across various CD tasks.

ID Embedding-as-Collaborator. To ensure that the final entity embeddings retain rich semantic information while incorporating personalized traits, we introduce the ID embedding $\boldsymbol{g}^t$ as a collaborator to the textual embedding $\hat{\boldsymbol{h}}^t$ in task t. These two representations are jointly fused to produce the latent embedding $\boldsymbol{Emb}^t \in \mathbb{R}^{1\times d_t}$, which can be formally expressed as follows:

$$\boldsymbol{Emb}^t = \lambda\cdot\hat{\boldsymbol{h}}^t + (1-\lambda)\cdot\boldsymbol{g}^t, \tag{6}$$

where $\lambda \in [0, 1]$ is the fusion factor that controls the weight of the textual embedding in the fusion of representation. Finally, the learned latent representations are applied to various CD tasks.

ID Embedding-as-Instructor. Since the current textual embeddings are solely derived from learning the behavioral patterns of entities, they may struggle to effectively distinguish between individuals and tend to be sensitive to noisy data. In contrast, ID embeddings often possess stronger discriminative power. Therefore, we introduce ID embeddings as an instructor to align the textual embeddings accordingly, thereby alleviating these limitations. We define our alignment loss based on InfoNCE [24] and take students as an example. We set textual-ID pairs from same students as positive and pairs with other IDs as negative. Specifically,

$$\mathcal{L}^t_{\text{align},s} = -\frac{1}{|S|}\sum_{s_i \in S} \log\left(\frac{\exp(\hat{\boldsymbol{h}}^t_{s_i} \cdot \boldsymbol{g}^{t\top}_{s_i}/\tau)}{\sum_{j\neq i}\exp(\hat{\boldsymbol{h}}^t_{s_i} \cdot \boldsymbol{g}^{t\top}_{s_j}/\tau)}\right), \tag{7}$$

where S is the set of students, $\boldsymbol{g}^t_{s_i} \in \mathbb{R}^{1\times d_t}$ denotes the ID embeddings for the student s_i, and τ is the temperature hyperparameter. The computation of the alignment loss is similar for exercises and concepts. We obtain the final alignment loss, formulated as $\mathcal{L}^t_{\text{align}} = \mathcal{L}^t_{\text{align},s} + \mathcal{L}^t_{\text{align},e} + \mathcal{L}^t_{\text{align},c}$ for original CD task t. Let $\mathcal{L}^t_{\text{CD}}$ denote the loss of task t, which is formulated as:

$$\mathcal{L}^t = \mathcal{L}^t_{\text{CD}} + \alpha \cdot \mathcal{L}^t_{\text{align}}, \tag{8}$$

where α is the align factor used to balance the weight of alignment loss $\mathcal{L}^t_{\text{align}}$.

5 Experiments

We conduct experiments on real-world datasets to answer the following key research questions.

- **RQ1:** How effective is the textual embedding enhancement in EduEmbed across various CD tasks?
- **RQ2:** How does each component contribute to the performance of EduEmbed across various CD tasks?
- **RQ3:** How do the types and scale of LMs impact the performance of EduEmbed?
- **RQ4:** How does the textual attribute selection influence the performance of EduEmbed?
- **RQ5:** How do hyperparameters influence EduEmbed?

5.1 Experimental Settings

Datasets Description. We conduct experiments on four real-world datasets collected from different web-based online intelligent education systems: SLP [23], NeurIPS20 [35], EDM [4], and MOOC [37]. Table 1 provides detailed statistics of those datasets. Here, "Average Correct Rate" refers to the mean accuracy of students on exercises, and "Q Density" refers to the average number of concepts associated with each exercise. Specifically, we implement our Stage 1 RaIF on the SLP-Math dataset, using NeurIPS20 as the in-domain dataset, since both SLP-Math and NeurIPS20 cover junior and senior-level math, and EDM as the out-domain dataset, which focuses on elementary-level math. This setup allows us to evaluate the generalization performance of EduEmbed across different educational levels. Due to the rich exercise context, MOOC is employed to explore how different attribute selections for textual profiling affect the performance of EduEmbed in RQ5. All datasets largely satisfy normality due to scale and random splits. The detailed introduction of these datasets is summarized in Appendix B.1.

Table 1: Statistics of the real-world datasets.

Datasets	SLP-Math	SLP-Chi	NeurIPS20	EDM	MOOC
# Students	1080	562	4918	2699	3000
# Exercises	609	510	948	1479	1967
# Knowledge Concepts	32	17	86	319	2278
# Response Logs	52100	28686	1382727	116156	333602
Average Correct Rate	0.506	0.623	0.545	0.628	0.812
Q Density	1.000	1.000	4.017	1.000	2.284

Evaluation Metrics. Since students' true mastery levels are unobservable, we follow prior research [33] to evaluate the performance of EduEmbed by predicting the performance of students on CD tasks. We employ score-prediction metrics and interpretability metrics to assess its effectiveness. Specifically, for score prediction metrics, given that the CD task is a binary classification problem, we use the Area Under the Curve (AUC) and Accuracy (ACC) as evaluation metrics. For interpretability, following previous works [33], we employ the Degree of Agreement (DOA) to assess the interpretability of the mastery levels of students. For a more detailed explanation of DOA, please refer to Appendix B.2.

Compared Methods. The following provides a brief description of the baselines used in four representative CD tasks and a downstream CAT task.

- **Transductive CD.** As the most traditional task setting, Transductive CD has been extensively studied, with most methods adopting the ID embedding paradigm, which fits well within our framework. We select three representative models as both compared methods and integrated CD models in EduEmbed: the classic MIRT [28], the widely used KaNCD [34], and the recent SOTA model ORCDF [25].
- **Inductive CD.** In inductive CD, traditional ID embedding paradigm is no longer applicable. Therefore, EduEmbed relies solely on textual semantic features in this setting. We compare our approach with two recent models, IDCD [14] and ICDM [19].
- **Zero-shot CD.** Zero-shot CD can be further divided into two categories. The first is cross-subject CD, which focuses on transfer across different academic subjects, and the second is cross-CD, which addresses transfer across different datasets. In both tasks, the dominant paradigm is textual semantic embeddings. Accordingly, EduEmbed adopts pure textual semantic features in this setting. We compare our approach with three representative methods: TechCD [9], ZeroCD [7], and LRCD [20].
- **Computerized Adaptive Testing (CAT).** CAT is a downstream task of CD. It consists of two main components: a selection strategy and a CD model. We select NCD [33] and IRT [11] as the CD models and five selection strategies: RANDOM, MAAT [1], BOBCAT [10], NCAT [39] and BECAT [40]. Since CAT follows the ID embedding paradigm, we also integrate ID embeddings into our EduEmbed.

Implementation Details. For stage 1, we use Qwen2.5-3B [29] as the default LM. Large LMs are fine-tuned with LoRA [12], whereas smaller models undergo full fine-tuning. For stage 2, we set d_t to 64, which is the dimension of the learned latent representations in all tasks. The batch size is set to 256 for all CD tasks, and for CAT task,

the batch size is chosen from the set $\{32, 64, 128, 256\}$. The learning rate is chosen from $\{1e^{-4}, 5e^{-4}, 1e^{-3}, 5e^{-3}, 1e^{-2}\}$. All experiments are conducted on two A6000 GPUs. We employ a grid search on the validation set to obtain the best hyperparameters and the detailed hyperparameter analysis is provided in Appendix B.7.

5.2 Experimental Results

5.2.1 Effectiveness Analysis of Embedding Enhancement (To RQ1). As shown in Table 2 and 3, we conduct a detailed analysis of the effectiveness of textual embedding enhancement across different CD tasks. For CAT, the experimental results on SLP-Math dataset in Table 4 are shown as an instance. For zero-shot CD, we adopt both cross-subject and cross-domain settings. In the cross-subject CD, we illustrate a representative case where the source domain is the Chinese literature subject (SLP-Chi) and the target domain is the mathematics subject (SLP-Math) within the diverse SLP dataset. For cross-domain CD, for SLP-Math, we use EDM as the source domain and SLP-Math itself as the target domain. Additionally, for in-domain and out-of-domain datasets, we treat each dataset itself as the source domain, with the other dataset serving as the target domain. The complete analysis are provided in Appendix B.3.

Significant Enhancement in Cold Start and High Generalization Scenarios. Textual embedding shows clear performance enhancement in scenarios requiring strong generalization or having severe cold-start issues, such as inductive CD, zero-shot CD and the early stages of CAT.

Limited Enhancement in Low Generalization Requirements Tasks. In tasks with low generalization demands, such as transductive CD, textual semantic embedding offers limited enhancement. Therefore, EduEmbed effectively integrates the ID paradigm, ensuring the performance lower bound and maintaining competitive results.

Interpretability Analysis. For models relying entirely on textual semantic features like LRCD, the fine-tuned EduEmbed offers better interpretability. However, for pattern-driven models like IDCD, which use sparse handcrafted interaction features, these features often show clearer structure and thus outperform dense textual embeddings.

Domain-Sensitive Enhancement. The enhancement provided by fine-tuned LMs is sensitive to their training datasets. As our LM is fine-tuned on SLP-Math, it shows strong performance in in-domain datasets like NeurIPS20, but their generalization to out-domain datasets like EDM remains limited and requires further exploration.

Limited cases like low generalization and out-of-domain applications are discussed in Appendix C.

5.2.2 Ablation Study (To RQ2). To validate the efficacy of each module in EduEmbed, we conduct an ablation study. Five ablated versions of EduEmbed are presented. *EduEmbed-w/o-RaIF* omits all the fine-tuning designs, using the textual embeddings generated directly by LMs; *EduEmbed-w/o-RsR* removes the role embedding $\boldsymbol{r}_{\text{role}}$ from fine-tuning process; *EduEmbed-w/o-TA* skips the Textual Adapter which is MLPs in this paper; *EduEmbed-w/o-IDI* does not utilize the alignment loss in AaRI; In *EduEmbed-w/o-IDC*, ID embeddings are not integrated with textual embeddings. Specially,

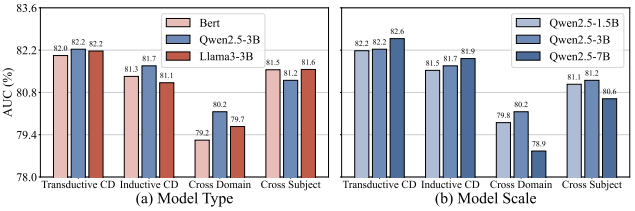


Figure 3: The performance of EduEmbed under varying LMs types and scales on SLP-Math.

EduEmbed-w/o-TA replaces MLPs with a simple linear layer in inductive CD and CAT. Also, some ablation experiments cannot be conducted in certain scenarios due to limitations. Corresponding explanations would be given in Appendix B.4.

Experimental Results. As shown in Table 5, our proposed EduEmbed outperforms most of its ablated versions, confirming the effectiveness of each module. However, we also observe that certain ablated versions exhibit superior performance in specific scenarios. In transductive CD, due to the relatively low requirement for generalization, the performance gains brought by fine-tuning are limited. In inductive CD, using a simple linear layer as the adapter in *EduEmbed-w/o-TA* helps mitigate potential overfitting and achieve strong predictive performance. In zero-shot CD, where a greater generalization of semantics is required, the lack of explicit semantic information in role embeddings limits the interpretability of EduEmbed compared with *EduEmbed-w/o-RsR*. For more results and further analysis, please refer to Appendix B.4.

5.2.3 Comparison of Types and Scales of the LMs (To RQ3). Here, we investigate the impact of LMs scales and types on the performance of EduEmbed. We conduct experiments on four CD tasks, and the corresponding results based on AUC are shown in Figure 3. For more detailed evaluations, please refer to Figure 5 and 6 in Appendix B.5.

Model Types. We fine-tune Qwen2.5-3B [29], Llama3.2-3B [30] and Bert-Base-Cased [3], respectively. As shown in Figure 3 (a), Qwen2.5-3B delivers optimal performance in most CD scenarios, likely due to its advanced text comprehension and generation capabilities. However, its performance in cross-subject CD is less satisfactory, possibly because it tends to memorize subject-specific patterns from the training data, leading to a limited capacity to generalize to unseen subjects.

Model Scales. We fine-tune the Qwen2.5-series [29] LMs with 1.5B, 3B, and 7B parameters, respectively.

As shown in the results of Figure 3 (b), we observe that in transductive CD and inductive CD, model performance improves as the parameter size increases. This is likely due to the similar distribution between training and testing data, which allows larger models to more effectively capture complex cognitive patterns during fine-tuning. However, in cross-domain and cross-subject CD, the performance initially improves but then declines as the model size increases. This trend may be attributed to domain bias in the training data. Larger models tend to overfit fine-grained, domain-specific features, improving in-domain learning but impairing generalization to new domains.

Table 2: The overall performance of EduEmbed compared with the baseline methods in four CD tasks. Within each method, the highest mean performance is highlighted in bold. The value following "±" denotes the standard deviation of the model's performance. If a mean value is significantly higher than the second-best result according to a t-test with a significance level of 0.05, it is marked with "*".

Datasets		SLP-Math			NeurIPS20			EDM		
Scenarios	Method	AUC	ACC	DOA	AUC	ACC	DOA	AUC	ACC	DOA
Transductive CD	MIRT	82.03±0.01	**74.81**±0.09	–	78.68±0.01	71.77±0.02	–	78.98±0.03	74.36±0.04	–
	KaNCD	82.12±0.13	74.67±0.11	77.81±0.13	78.57±0.03	71.73±0.04	66.61±1.92	79.92±0.13	74.40±0.23	**78.78***±0.12
	ORCDF	**82.37**±0.01	74.48±0.13	**78.24***±0.08	**78.70**±0.03	**71.79**±0.03	73.58±0.04	**82.63**±0.07	**76.88**±0.03	77.84±0.16
	EduEmbed	82.23±0.05	74.45±0.11	77.85±0.09	78.55±0.01	71.75±0.02	**73.60**±0.01	82.59±0.05	76.75±0.02	77.65±0.11
Inductive CD	ICDM	74.54±0.03	68.83±0.01	60.49±0.02	71.72±0.00	65.63±0.01	59.00±0.00	74.18±0.01	70.54±0.01	65.38±0.01
	IDCD	79.52±0.06	72.59±0.12	**80.96***±0.04	75.91±0.23	69.84±0.20	**73.16***±0.38	79.67±0.07	**75.41**±0.13	**79.93***±0.49
	EduEmbed	**81.68***±0.04	**73.78***±0.11	78.61±0.05	**76.59***±0.07	**70.01**±0.17	72.78±0.32	**80.66***±0.04	75.35±0.44	76.53±0.03
Cross-Domain CD	TechCD	52.52±0.14	53.27±0.41	54.03±1.16	52.05±0.08	53.65±0.27	52.89±0.71	54.05±0.21	63.67±0.83	58.71±0.43
	LRCD	79.67±0.69	72.11±0.33	76.15±0.42	76.05±0.31	68.47±1.03	73.00±0.03	**79.19***±0.21	73.02±1.77	76.91±0.10
	EduEmbed	**80.06***±0.38	**72.61***±0.23	**78.61***±0.14	**76.31**±0.16	**69.41***±0.43	**73.02**±0.03	78.28±1.13	**74.68***±0.00	**76.95**±0.00

Table 3: The performance of cross-subject CD on SLP. Other details are as same as Table 2.

Method	AUC	ACC	DOA
LRCD	80.56±0.12	72.59±0.32	76.87±0.04
EduEmbed	**81.20***±0.21	**73.69***±0.42	**77.11***±0.08

Table 4: The overall performance of EduEmbed with five CAT selection strategies on SLP-Math. "OL" stands for the original method under ID embedding paradigm.

Dataset		SLP-Math			
Metric		AUC / ACC (%)			
		IRT		NCD	
Strategy	step	OL	EduEmbed	OL	EduEmbed
RANDOM	5	74.61 / 68.03	**75.23*** / **69.42***	73.38 / 67.38	**74.01*** / **68.02***
	10	77.15 / 70.16	**78.56*** / **71.48***	76.47 / 69.59	**78.20*** / **71.22***
	15	78.44 / 71.34	**80.24*** / **72.02***	78.33 / 70.78	**79.28*** / **72.09***
MAAT	5	74.18 / 67.35	**76.66*** / **69.85***	73.66 / 60.07	**74.02*** / **60.82***
	10	76.26 / 68.35	**78.96*** / **71.17***	76.29 / 60.77	**77.32*** / **61.23***
	15	77.32 / 69.30	**79.42*** / **71.55***	77.88 / 63.65	**77.92*** / **64.21***
BOBCAT	5	75.67 / 68.75	**78.95*** / **71.91***	73.74 / 66.39	**74.52*** / **68.35***
	10	77.75 / 70.75	**80.44*** / **72.27***	75.69 / 69.05	**76.27*** / **70.14***
	15	78.89 / 71.65	**81.07*** / **73.54***	77.43 / 70.57	**77.44*** / **71.05***
NCAT	5	73.94 / 67.35	**77.63*** / **70.30***	**73.32** / 62.78	73.19 / **67.08***
	10	75.89 / 68.86	**80.14*** / **72.54***	76.30 / 68.71	**76.59*** / **70.03***
	15	77.45 / 70.21	**80.43*** / **72.57***	77.43 / 70.67	**79.41*** / **72.09***
BECAT	5	75.37 / 68.76	**77.45*** / **70.40***	71.85 / 64.70	**72.36*** / **65.74***
	10	77.81 / 70.95	**79.02*** / **71.48***	75.16 / 66.26	**77.26*** / **69.57***
	15	79.60 / 72.70	**81.33*** / **73.38***	77.21 / 69.73	**78.40*** / **70.20***

5.2.4 The Effect of Text Selection (To RQ4). Previous research [27] has shown that the textual content of exercises can serve as valuable attributes for learner-item cognitive modeling. However, many existing datasets lack such content, limiting the broader application of text-based features in CD. To assess the impact of this limitation, we conduct experiments on MOOC dataset which includes exercise content, under both inductive CD and transductive CD. Corresponding details are presented in Appendix B.6.

Table 5: Ablation study in four CD tasks on SLP-Math.

CD Scenario	Metric	EduEmbed w/o RaIF	EduEmbed w/o RsR	EduEmbed w/o TA	EduEmbed
Transductive CD	AUC	**82.27**	82.24	82.06	82.23
	ACC	74.40	74.40	74.38	**74.45**
	DOA	77.75	77.44	76.78	**77.85**
Inductive CD	AUC	81.04	81.59	81.62	**81.68**
	ACC	73.75	73.63	**73.97**	73.78
	DOA	78.60	77.33	**78.79**	78.61
Cross Domain CD	AUC	78.49	79.87	77.45	**80.06**
	ACC	71.24	71.12	64.05	**72.61**
	DOA	76.87	**78.91**	76.22	78.61
Cross Subject CD	AUC	80.41	81.14	78.01	**81.20**
	ACC	72.87	73.64	63.51	**73.69**
	DOA	77.01	**77.19**	76.12	77.11

5.2.5 Hyperparameter Analysis (To RQ5). We investigate the impact of two key hyperparameters on the performance of EduEmbed. For detailed results, please refer to Appendix B.7.

6 Conclusion and Discussion

In this paper, we systematically evaluate and reveal the task-based potential of LM-based textual embeddings across mainstream CD tasks for web-based online intelligent education systems. We introduce EduEmbed, a unified enhancement framework that leverages fine-tuned LMs to improve learner-item cognitive modeling. Comprehensive experiments verify the varying enhancement brought by semantic information, offering insights for future research. Limitations and future directions including performance robustness in low-generalization scenarios, further unified integration and computational cost are discussed in Appendix C.

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), and Tencent Inc Research Program.

References

[1] Haoyang Bi, Haiping Ma, Zhenya Huang, Yu Yin, Qi Liu, Enhong Chen, Yu Su, and Shijin Wang. 2020. Quality Meets Diversity: A Model-Agnostic Framework for Computerized Adaptive Testing. In *Proceedings of the 20th IEEE International Conference on Data Mining*. Sorrento, Italy, 42–51.

[2] Jimmy De La Torre. 2009. DINA Model and Parameter Estimation: A Didactic. *Journal of Educational and Behavioral Statistics* 34, 1 (2009), 115–130.

[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, volume 1 (long and short papers)*. Minneapolis, Minnesota, 4171–4186.

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

[5] Mingyu Feng, Neil T. Heffernan, and Kenneth R. Koedinger. 2009. Addressing the Assessment Challenge with an Online System That Tutors as it Assesses. *User Modeling and User-Adapted Interaction* 19, 3 (2009), 243–266.

[6] 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.

[7] 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*, Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan (Eds.). Vancouver, Canada, 8417–8426.

[8] Weibo Gao, Qi Liu, Linan Yue, Fangzhou Yao, Rui Lv, Zheng Zhang, Hao Wang, and Zhenya Huang. 2025. Agent4Edu: Generating Learner Response Data by Generative Agents for Intelligent Education Systems. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence*, Toby Walsh, Julie Shah, and Zico Kolter (Eds.). Philadelphia, PA, 23923–23932.

[9] 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*, Hsin-Hsi Chen, Wei-Jou (Edward) Duh, Hen-Hsen Huang, Makoto P. Kato, Josiane Mothe, and Barbara Poblete (Eds.). Taiwan, China, 983–992.

[10] Aritra Ghosh and Andrew S. Lan. 2021. BOBCAT: Bilevel Optimization-Based Computerized Adaptive Testing. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence*. Virtual Event, 2410–2417.

[11] Shelby J Haberman. 2005. Identifiability of Parameters in Item Response Models with Unconstrained Ability Distributions. *ETS Research Report Series* 2005, 2 (2005), i–22.

[12] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *Proceedings of the 10th International Conference on Learning Representations*. Virtual Event.

[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] 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.

[15] 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.

[16] Mingjia Li, Hong Qian, Jinglan Lv, Mengliang He, Wei Zhang, and Aimin Zhou. 2024. Foundation Model Enhanced Derivative-Free Cognitive Diagnosis. *Frontiers of Computer Science* (2024).

[17] Mingjia Li, Junkai Tong, Yiyang Huang, Yifei Ding, Hong Qian, and Aimin Zhou. 2025. Paper-Level Computerized Adaptive Testing for High-Stakes Examination via Multi-Objective Optimization. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. Toronto, Canada, 1435–1446.

[18] Shuo Liu, Hong Qian, Mingjia Li, and Aimin Zhou. 2023. QCCDM: A Q-Augmented Causal Cognitive Diagnosis Model for Student Learning. In *Proceedings of the 26th European Conference on Artificial Intelligence*. Kraków, Poland, 1536–1543.

[19] 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.

[20] 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*, Yizhou Sun, Flavio Chierichetti, Hady W. Lauw, Claudia Perlich, Wee Hyong Tok, and Andrew Tomkins (Eds.). Toronto, Canada, 836–847.

[21] 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*. Toronto, Canada, 1915–1926.

[22] 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.

[23] 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*. Virtual Event, 261–266.

[24] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation Learning with Contrastive Predictive Coding. *arXiv preprint arXiv:1807.03748* (2018).

[25] 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.

[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 AAAI conference on artificial intelligence*. Vancouver, Canada, 14928–14936.

[27] Yu Su, Qingwen Liu, Qi Liu, Zhenya Huang, Yu Yin, Enhong Chen, Chris H. Q. Ding, Si Wei, and Guoping Hu. 2018. Exercise-Enhanced Sequential Modeling for Student Performance Prediction. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, Sheila A. McIlraith and Kilian Q. Weinberger (Eds.). New Orleans, LA, 2435–2443.

[28] James B Sympson. 1978. A Model for Testing with Multidimensional Items. In *Proceedings of the 1977 Computerized Adaptive Testing Conference*. Minneapolis, MN.

[29] Qwen Team. 2024. Qwen2.5: A Party of Foundation Models. https://qwenlm.github.io/blog/qwen2.5/

[30] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *CoRR* abs/2302.13971 (2023).

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

[32] Fei Wang, Weibo Gao, Qi Liu, Jiatong Li, Guanhao Zhao, Zheng Zhang, Zhenya Huang, Mengxiao Zhu, Shijin Wang, Wei Tong, et al. 2024. A Survey of Models for Cognitive Diagnosis: New Developments and Future Directions. *arXiv preprint arXiv:2407.05458* (2024).

[33] 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.

[34] 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).

[35] 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).

[36] Songlin Xu, Xinyu Zhang, and Lianhui Qin. 2024. EduAgent: Generative Student Agents in Learning. *CoRR* abs/2404.07963 (2024).

[37] 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).

[38] 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.

[39] 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.

[40] 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 37*. New Orleans, LA.

Appendix

A Details of Motivation Study

In this section, we provide the corresponding settings of our motivation study in Figure 1 (a) presented in Section 1.

In four CD scenarios and CAT task, we incorporate personalized textual descriptions of students proposed in Eq. 1 as textual embedding features for modeling. In zero-shot CD, this textual embedding model refers to LRCD. For zero-shot CD and inductive CD, we introduce existing models, TechCD and IDCD, respectively, as non-text embedding baselines. In transductive CD and CAT, mainstream ID embeddings are used as non-text embedding baselines. We use IRT as the CD model in CAT. All the results are reported based on AUC.

B Experiments

B.1 Details about the Datasets

In this subsection, we provide detailed introduction of the datasets and the corresponding processing details.

Source. Here we provide the dataset source we use in this paper:

• **SLP** [23]: SLP is a K-12 dataset from the online education platform SLP, recording students' performance across eight subjects over three years (7th to 9th grade). In our paper, we use two subjects: Math and Chinese.

• **NeurIPS20** [35]: NeurIPS20 comes from the NeurIPS 2020 Education Challenge, containing student response logs to Eedi math problems over two school years (2018–2020). Eedi is a widely used online learning platform that provides diagnostic multiple-choice questions for middle and high school students.

• **EDM** [4]: Derived from the EDM Cup 2023, EDM captures millions of student interactions on ASSISTments, a web-based K-12 math learning system, with concepts mainly at the elementary level.

• **MOOC** [37]: Collected from a large-scale Chinese MOOC platform, MOOC offers rich learning resources, fine-grained concepts, behavioral logs, and contextual information such as textual descriptions and annotations.

Process. To ensure sufficient response data, we exclude students with fewer than 10, 10, 30, and 30 responses in SLP, MOOC, NeurIPS20, and EDM, respectively. To reduce computational cost, we randomly sample 3000 students from MOOC. Response logs are split into 70%/10%/20% for training, validation, and testing in both stages. During Stage 1, we cap each student at 50 responses, randomly sampling when necessary. In inductive CD, students are split into existing (S_o) and new (S_u) groups at a 1:1 ratio, while in CAT, 30% of responses are used for model pre-training. To prevent information leakage, target-domain test data are excluded from training in zero-shot CD, and student textual embeddings are omitted in CAT.

B.2 Degree of Agreement (DOA)

We provide a detailed formulation of the *Degree of Agreement (DOA)* to quantify the alignment between predicted mastery and actual performance. Let $\boldsymbol{Mas} \in \mathbb{R}^{M\times K}$ denote the predicted mastery matrix for M students and K concepts. The core intuition is that if student s_a achieves higher accuracy than s_b on exercises of concept c_k, then s_a should exhibit greater mastery, i.e., $\boldsymbol{Mas}_{s_a,c_k} > \boldsymbol{Mas}_{s_b,c_k}$. The DOA for concept c_k is computed accordingly.

Table 6: The performance of EduEmbed with overlapping students. Other details are as same as Table 2.

Metric	AUC	ACC	DOA
TechCD	57.96	56.44	48.8
ZeroCD	61.77	59.07	50.81
LRCD	<u>78.56</u>	<u>72.01</u>	<u>74.96</u>
EduEmbed	**78.74***	**72.32***	**75.30***

$$\mathrm{DOA}_k = \frac{1}{Z}\sum_{a,b\in S}\delta\left(\boldsymbol{Mas}_{s_a,c_k},\boldsymbol{Mas}_{s_b,c_k}\right)\cdot\frac{\sum_{j=1}^{M}\boldsymbol{Q}_{j,k}\wedge\varphi(j,a,b)\wedge\delta\left(r_{aj},r_{bj}\right)}{\sum_{j=1}^{M}\boldsymbol{Q}_{j,k}\wedge\varphi(j,a,b)\wedge\mathbb{I}\left(r_{aj}\neq r_{bj}\right)}, \quad (9)$$

where $Z = \sum_{a,b\in S}\delta(\boldsymbol{Mas}_{s_a,c_k},\boldsymbol{Mas}_{s_b,c_k})$, $\boldsymbol{Q}_{j,k} = 1$ indicates that exercise e_j is related to concept c_k, $\varphi(j,a,b)$ determines whether both students s_a and s_b answered e_j, r_{aj} is the response of s_a to e_j, and $\mathbb{I}\left(r_{aj}\neq r_{bj}\right)$ determines whether 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 Effectiveness Analysis of Embedding Enhancement in CD scenarios and CAT

The performance of EduEmbed in CD scenarios and CAT.

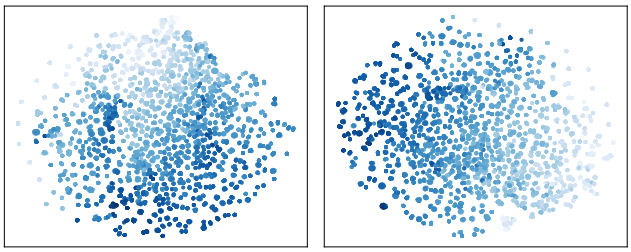
• **Transductive CD.** In transductive CD, textual embeddings offer limited benefits and can even underperform ID embeddings, as generalization demands are low and ID embeddings are well-optimized with encoders such as graph neural networks. Since textual embeddings are not further tuned during representation learning, they involve fewer trainable parameters and therefore underperform. However, EduEmbed integrates the ID paradigm to secure a strong lower bound and maintain competitiveness.

• **Inductive CD.** In inductive CD, textual embeddings yield notable gains by encoding richer information than sparse handcrafted features used in IDCD. Yet, these sparse features retain an interpretability advantage, as their structured patterns are more transparent than dense textual representations.

• **Zero-shot CD.** Textual semantics yield substantial gains in zero-shot CD across cross-domain and cross-subject settings [20]. LRCD, which fully relies on semantic features, markedly outperforms methods with limited or no semantic use (e.g., TechCD, ZeroCD). Building on this, EduEmbed fine-tunes LMs to align with CD objectives, further bridging the gap and enhancing zero-shot performance.

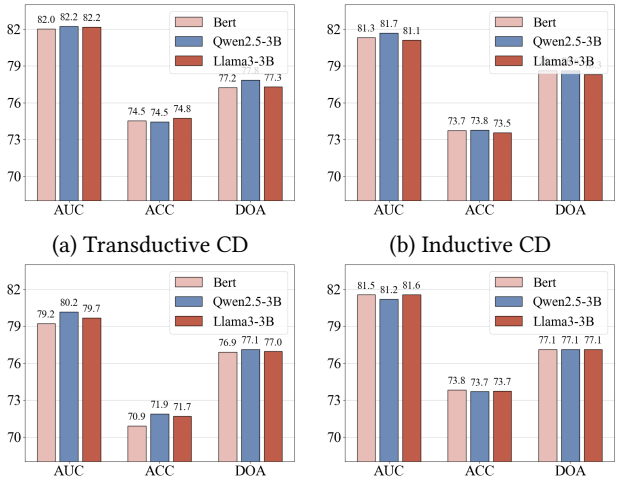
• **CAT.** In CAT, textual semantics enhance performance at all stages, with the greatest gains in early phases when ID embeddings are weak and generalize poorly. As testing progresses and ID embeddings become refined, the setting converges toward transductive CD, where ID-based methods regain superiority.

The Performance of Zero-shot CD with Overlapping Students. We construct a new dataset, SLP*, where the source and target domains share overlapping students. This dataset contains 312 students, 882 exercises, and 38 knowledge concepts, with a total of 32,996 response logs. We set SLP-CHI as the source domain and SLP-Math as the target domain. The experimental results are shown in Table 6, where EduEmbed consistently demonstrates strong performance compared to other methods.



(a) EduEmbed-w/o-TA (b) EduEmbed

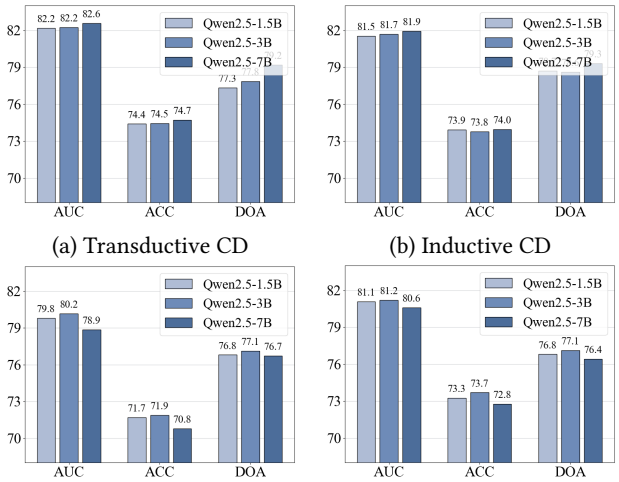
Figure 4: Visualization of students' mastery levels on SLP-Math.



(a) Transductive CD (b) Inductive CD

(c) Cross-Domain CD (d) Cross-Subject CD

Figure 5: Comparison of LMs types in four CD tasks.



(a) Transductive CD (b) Inductive CD

(c) Cross-Domain CD (d) Cross-Subject CD

Figure 6: Comparison of LMs scales in four CD tasks.

Table 7: Ablation study in transductive CD.

Metric	EduEmbed w/o IDI	EduEmbed w/o IDC	EduEmbed
AUC	82.21	82.05	**82.23**
ACC	74.40	74.27	**74.45**
DOA	77.50	77.59	**77.85**

B.4 Ablation Study

In this subsection, we provide the additional experimental results of ablation study in transductive CD and CAT in Table 7 and 8.

Settings. As for zero-shot and inductive CD, ID embeddings of new entities provide no useful information. Therefore, EduEmbed does not integrate them in these settings. Accordingly, experiments on *EduEmbed-w/o-IDI* and *EduEmbed-w/o-IDC* are omitted for inductive, cross-domain, and cross-subject CD. For inductive CD and CAT, the MLPs in the text adapter are replaced with a linear layer in *EduEmbed-w/o-TA* to satisfy the dimension transfer required by the interaction function.

Detailed Analysis.

• **Transductive CD.** *EduEmbed-w/o-RaIF* achieves strong AUC performance, suggesting limited gains since ID embeddings are already well-trained. Nevertheless, EduEmbed still offers clear advantages in both accuracy and interpretability, underscoring its effectiveness in cognitive modeling.

• **Inductive CD.** *EduEmbed-w/o-TA* also performs well, likely because MLPs add parameters but risk overfitting. These results validate the textual adapter framework, showing that even a simple linear layer ensures robust performance and offering insights for future adapter design.

• **Zero-shot CD.** EduEmbed shows weaker interpretability than *EduEmbed-w/o-RsR*, likely due to the lack of explicit semantics in role embeddings, which is a limitation more evident in cross-domain CD requiring semantic generalization. Still, its strong predictive accuracy highlights the effectiveness of the role embedding design.

• **CAT.** Similar to inductive CD, *EduEmbed-w/o-TA* achieves reasonable performance in early CAT stages. In contrast, *EduEmbed-w/o-IDI* and *EduEmbed-w/o-IDC* underperform at step 5 due to immature ID embeddings introducing noise. As CAT progresses, ID embeddings strengthen, and EduEmbed exhibits clear gains at steps 10 and 15, demonstrating the effectiveness of RaIF, as proposed in Section 4.1.

Visualization of Mastery Levels. To further evaluate the contribution of the Texual Adapter, we visualize students' mastery levels on the SLP-Math dataset via t-SNE [31], with darker shades indicating higher correct rate. Using transductive CD as a case study, as shown in Figure 4, EduEmbed demonstrates clearer clustering and smoother progression, underscoring the interpretability benefits of the Textual Adapter.

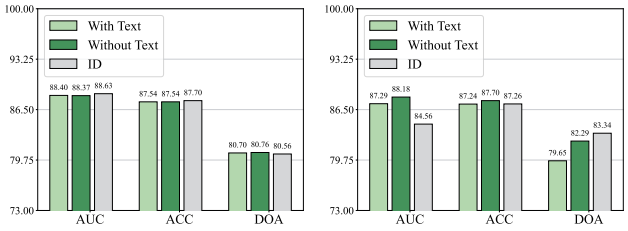
The results of the ablation study indicate that designs in both RaIF and AaRI are crucial to the overall effectiveness of EduEmbed.

B.5 Details of LMs Scales and Types

In this subsection, we provide the performance results of EduEmbed with different types and scales of LMs in transductive CD, as shown in Figure 5 and 6.

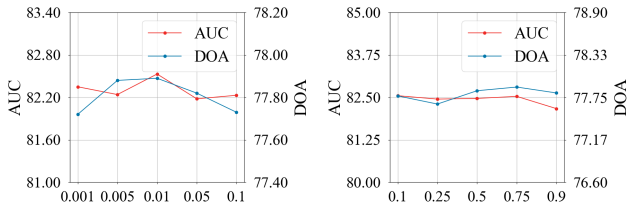
Table 8: Ablation study in CAT.

Metric		AUC / ACC (%)					
CD Model	Step	EduEmbed w/o RaIF	EduEmbed w/o RsR	EduEmbed w/o TA	EduEmbed w/o IDI	EduEmbed w/o IDC	EduEmbed
IRT	5	67.72 / 61.83	73.71 / 58.49	73.21 / 66.91	76.69 / 69.30	76.29 / 67.87	**77.45** / **70.40**
	10	73.91 / 65.54	76.00 / 69.53	73.78 / 67.34	**79.15** / **71.67**	78.20 / 70.64	79.02 / 71.48
	15	74.95 / 68.72	80.05 / 71.94	74.64 / 67.81	81.27 / **73.50**	78.92 / 69.87	**81.33** / 73.38
NCD	5	62.26 / 57.20	72.25 / 65.64	70.65 / 64.41	69.83 / 62.57	**73.37** / 64.07	72.36 / **65.74**
	10	65.30 / 61.62	75.69 / 67.12	**77.35** / **69.80**	75.92 / 69.31	73.48 / 68.05	77.26 / 69.57
	15	66.93 / 63.08	76.54 / 67.88	**78.78** / **71.55**	78.17 / 70.15	77.89 / 70.85	78.40 / 70.20



(a) Transductive CD (b) Inductive CD

Figure 7: Effect of text selection on MOOC. "OL" refer to baselines, specifically denoting ID embedding in transductive CD and IDCD for inductive CD.



(a) Effect of α (b) Effect of λ

Figure 8: Hyperparameter analysis on SLP-Math.

B.6 Text Selection Analysis

In this subsection, we provide the details of text selection experiment. We extend the exercise attribute defined in Eq. (1) in Section 4.1.1 by incorporating textual content. Since the exercise content in MOOC is in Chinese, we adopt BERT-Base-Chinese [3] as the fine-tuned LM to ensure compatibility with the dataset. As shown in Figure 7, incorporating exercise content leads to modest performance fluctuations, likely due to the trade-off between added detail and potential noise of exercise content. This suggests that in datasets lacking exercise content, deriving attributes from response logs has minimal impact on model performance, especially when ultra-high prediction precision is unnecessary.

B.7 Hyperparameter Analysis

In this subsection, We present the performance of EduEmbed with different hyperparameter settings, as shown in Figure 8. We recommend setting α to 0.01 or 0.005 and λ to 0.5 or 0.75 to generally yield relatively good performance in most cases.

Table 9: Comparison of EduEmbed and "Text-Only" on SLP-Math in transductive CD.

Metric	Text-Only	EduEmbed
AUC	75.53	**82.23**
ACC	68.93	**74.45**
DOA	76.60	**77.85**

C Discussions

Performance Robustness in Low-Generalization Scenarios. As discussed in Section 5.2.1 and Appendix B.3, LMs show limitations in transductive CD compared to traditional ID-based models. By integrating ID information, EduEmbed ensures a reliable performance lower bound while flexibly adapting to various CD scenarios. Instead of pursuing a one-size-fits-all solution, EduEmbed is designed to flexibly adapt to various CD scenarios with minimal modification, highlighting its practical extensibility. As shown in Table 9, EduEmbed achieves superior performance on SLP-Math compared to the "Text-Only" variant using raw LM embeddings without fine-tuning, highlighting that direct use of textual features alone is suboptimal in transductive CD.

Integration with Existing Learning Paradigms. Given the effectiveness of mainstream ID embeddings in cognitive modeling, this work focuses on the fusion of textual embeddings with ID embeddings, to ensure EduEmbed's compatibility across most CD tasks. Other paradigms, such as IDCD, which incorporate handcrafted interaction features as prior information, are also expected to be integrated. Notably, from a methodological perspective, EduEmbed is capable of being integrated with such paradigms. Exploring how textual embeddings can be effectively combined with increasingly diverse approaches remains an important direction for future research.

Computational Cost. Although fine-tuning LMs is generally time-consuming, our proposed decoupled EduEmbed mitigates this issue by freezing the textual embeddings by the LMs and applying them across different CD tasks. As a result, the fine-tuning process only needs to be conducted once, after which the representations can be stored locally. Therefore, in practical applications, the runtime of this component is virtually negligible, significantly improving the overall efficiency and usability of our framework.