LMCD: Language Models are Zeroshot Cognitive Diagnosis Learners

Anonymous EMNLP submission

Abstract

Cognitive Diagnosis (CD) has become a critical task in AI-empowered education, supporting personalized learning by accurately assessing students' cognitive states. However, traditional CD models often struggle in cold-start scenarios due to the lack of student-exercise interaction data. Recent NLP-based approaches leveraging pre-trained language models (PLMs) have shown promise by utilizing textual features but fail to fully bridge the gap between semantic understanding and cognitive profiling. In this work, we propose **L**anguage **M**odels as Zeroshot **C**ognitive **D**iagnosis Learners (LMCD), a novel framework designed to handle cold-start challenges by harnessing large language models (LLMs). LMCD operates via two primary phases: (1) Knowledge Diffusion, where LLMs generate enriched contents of exercises and knowledge concepts (KCs), establishing stronger semantic links; and (2) Semantic-Cognitive Fusion, where LLMs employ causal attention mechanisms to integrate textual information and student cognitive states, creating comprehensive profiles for both students and exercises. These representations are efficiently trained with off-the-shelf CD models. Experiments on two real-world datasets demonstrate that LMCD significantly outperforms state-of-the-art methods in both exercise-cold and domain-cold settings. The code is publicly available at https://anonymous.4open.science/r/LMCD-464C/

1 Introduction

Cognitive Diagnosis Models (CDMs) have become pivotal in educational technology, offering data-driven insights into students' cognitive states across different KCs,as shown in Figure 1(c). These models are indispensable for developing personalized learning systems (Huang et al., 2019a; Yu et al., 2024),computerized adaptive testing (Bi et al., 2020; Wainer et al., 2000) and so on (Huang et al., 2019b). Traditional CDMs such as Item Response Theory (IRT) (Lord, 1952), Multi-dimensional IRT (MIRT) (Reckase, 2006), Deterministic Input Noisy "And" gate model (DINA) (De La Torre, 2009), and Neural Cognitive Diagnosis Model (NCDM) (Wang et al., 2020a) have demonstrated significant success in conventional settings where abundant student-exercise interaction data is available.

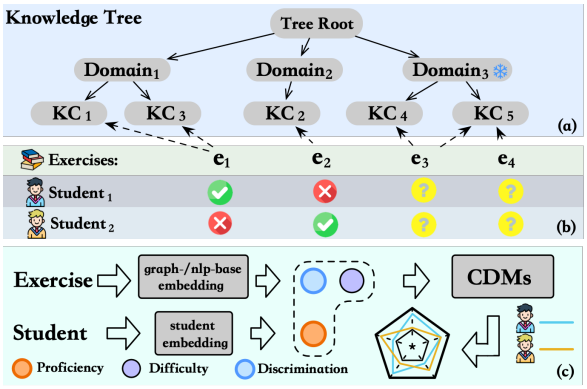

Figure 1: An illustration of the cold-start problem in cognitive diagnosis. (a) is hierarchical KC tree. (b) is sparse student-exercise interaction matrix. (c) is typical cognitive diagnosis framework for addressing cold-start problems.

However, these established approaches face substantial challenges in cold-start scenarios where there is little or no historical interaction data (Wang et al., 2024a). Cold-start refers to scenarios involving either new students or new exercises introduced into the system. Furthermore, new exercises can be classified into two types of cold-start problems based on whether their corresponding domain appears in the training data,as illustrated in Figure 1. Specifically, **exercise cold-start** occurs when the new exercises belong to the same domain as those seen in the training data, whereas **cross-domain cold-start** arises when new exercises originate from entirely unseen domains, posing a more significant challenge.

As shown in Figure 1(c), current approaches

addressing these cold-start challenges generally fall into two categories: graph-based methods and NLP-based methods. Graph-based approaches, exemplified by works like ICDM (Liu et al., 2024a) and TechCD (Gao et al., 2023), construct relationships between exercises, KCs, and students to establish connections between unseen and seen entities through graph structures. Although innovative, these methods are limited by the accuracy of KC annotations and the scarcity of high-quality educational knowledge graphs.

NLP-based methods leverage pre-trained language models(PLMs) to encode exercise texts and KCs. These approaches have shown surprising effectiveness, sometimes achieving competitive or superior results compared to SOTA models, as demonstrated in TechCD (Gao et al., 2023) and ZeroCD (Gao et al., 2024). Their ability to establish connections through semantic similarity makes them particularly promising for cross-domain scenarios. However, these methods face fundamental limitations in fully capturing cognitive complexity. For instance, exercises in NIPS34 (Wang et al., 2020b) with similar wording like *"What is 4 written as a fraction?"* and *"What is a third of a seventh?"* target entirely different KCs despite their textual similarity. Additionally, vague KC labels provide insufficient information for accurate representation through simple encoding. Most critically, these approaches struggle to effectively bridge semantic space with individual students' cognitive states. Even the latest work leveraging large language models (Liu et al., 2024b; Dong et al., 2025; Liu et al., 2025) has not adequately addressed these limitations.

Motivated by these challenges, we propose LMCD (**L**anguage **M**odels as Zeroshot **C**ognitive **D**iagnosis Learners), a novel framework that leverages large language models to address exercise-cold and domain-cold scenarios—the most prevalent challenges in online education environments. Specifically, the proposed LMCD operates in two primary phases: (1) **Knowledge Diffusion**, where LLMs generate enriched contents of exercises and KCs, creating stronger semantic links; and (2) **Semantic-Cognitive Fusion**, where we combine original exercise text with generated content and student-specific tokens as input to an LLM, using causal attention to create comprehensive representations that fuse semantic space with cognitive states. Finally, we align these representations with conventional CDM parameters, modeling discrimination, and relative difficulty (*how challenging each exercise is for each specific student*) rather than absolute difficulty, enabling more precise prediction of student performance.

The major contributions of our work are as follows:

(1) We pioneer the use of LLMs' causal attention mechanisms to capture dynamic student-exercise interactions, introducing the first explicit modeling of relative difficulty in cognitive diagnosis.

(2) We develop a flexible LLM-based framework that seamlessly integrates with off-the-shelf CDMs, enhancing their performance while maintaining their theoretical foundations.

(3) Through extensive experiments on two real-world datasets, we demonstrate LMCD's significant performance improvements in both exercise cold-start and cross-domain cold-start scenarios compared to state-of-the-art methods.

2 Preliminary

2.1 Knowledge Structure

In practical educational scenarios, KCs are typically organized into hierarchical tree structures, as illustrated in Figure 1(a). In a top-down manner, the root node $\mathcal{K}^0 = \langle \mathbb{K}, \mathcal{R} \rangle = \bigcup_{i=1}^{M^1} \mathcal{D}_i$ represents full dataset or its knowledge system (e.g., the node "Math" in NIPS34), consisting of the complete set of KCs $\mathbb{K}$ and the branch set $\mathcal{R}$. Assuming there are a total of $M^1 \equiv M_0^1$ nodes at depth 1, each of them is considered a distinct domain $\mathcal{D}_i = \langle \mathbb{K}_i^1, \mathcal{R}_i^1 \rangle$, $\forall i \in [1, M^1]$, where $\mathbb{K}_i^1 = \{k_{ij}^l \,|\, j \in [1, 2, ..., M_i^l], \, l \in [1, 2, ..., max\_depth]\}$, and $\mathcal{R}_i^1 = \{(k_{ip}, k_{iq}) \,|\, k_{ip}, k_{iq} \in \mathbb{K}_i^1\}$, (k_{ip}, k_{iq}) is a directed edge of the tree. Note that knowledge across different domains is entirely isolated, that is, $\mathbb{K}_i^1 \cap \mathbb{K}_j^1 = \emptyset$, $\forall i \neq j$. The hierarchical relationship(s) from the root to the leaf node(s) representing fine-grained terminal KCs associated with exercise v, is called its "knowledge route" $\mathcal{K}_v$.

2.2 Task Description

Based on the previous subsection, we define two cold-start scenarios as follows:

Exercise cold-start. Following TechCD (Gao et al., 2023), we define the cold-start scenario by randomly dividing data into hot ($\mathcal{H}$) and cold ($\mathcal{C}$) subsets at the exercise level. For the hot subset $\mathcal{H}$, we define students $\mathcal{U}_\mathcal{H} = \{u_1, u_2, \cdots, u_{|\mathcal{U}_\mathcal{H}|}\}$, exercises $\mathcal{V}_\mathcal{H} = \{v_1, v_2, \cdots, v_{|\mathcal{V}_\mathcal{H}|}\}$, and KCs $\mathbb{K}_\mathcal{H} = \{k_1, k_2, \cdots, k_{|\mathbb{K}_\mathcal{H}|}\}$. Student

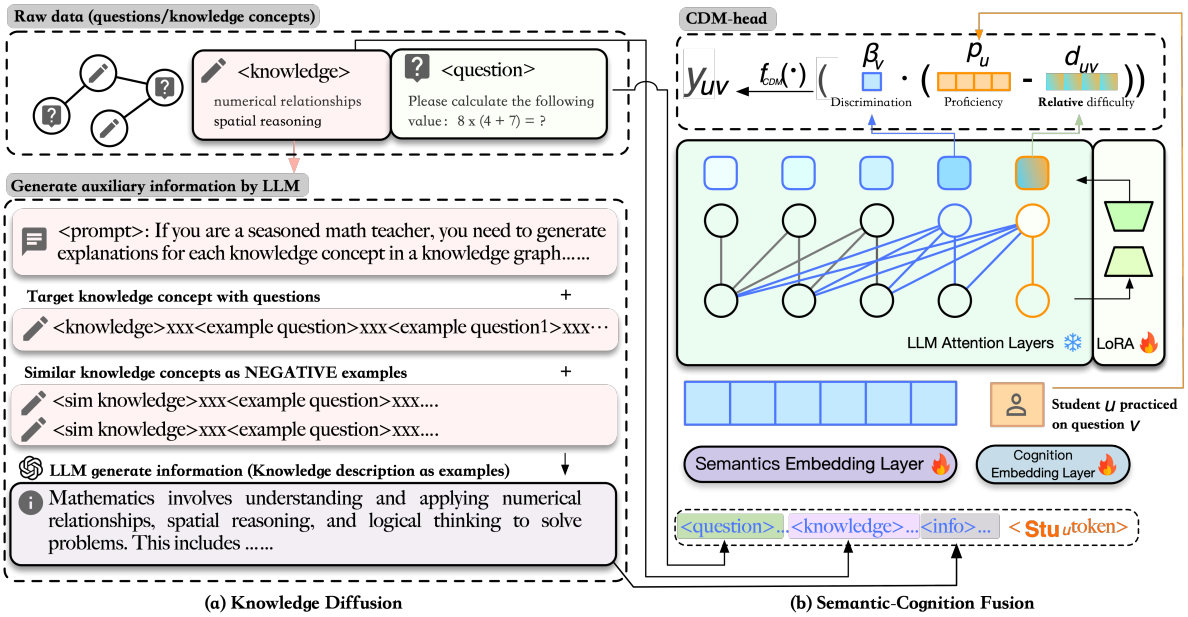

Figure 2: LMCD framework overview. (a) Knowledge Diffusion: LLMs generate enriched contents of exercises and knowledge concepts. (b) Semantic-Cognitive Fusion: Causal attention mechanisms integrate textual information with student-specific cognitive states to model relative difficulty.

exercise logs are represented as $\mathbb{R}_{\mathcal{H}} = \{(u, v, \mathcal{K}_v, y_{uv}) \mid u \in \mathcal{U}_{\mathcal{H}}, v \in \mathcal{V}_{\mathcal{H}}\}$, where $y_{uv} \in \{0, 1\}$ indicates whether student u answered exercise v correctly($y_{uv} = 1$) or not, respectively. Similarly, for the cold subset $\mathcal{C}$, we define $\mathcal{U}_{\mathcal{C}}$, $\mathcal{V}_{\mathcal{C}}$, $\mathcal{K}_{\mathcal{C}}$, and $\mathbb{R}_{\mathcal{C}} = \{(u, v, \mathcal{K}_v, y_{uv}) \mid u \in \mathcal{U}_{\mathcal{C}}, v \in \mathcal{V}_{\mathcal{C}}\}$. Importantly, while exercises are separated between subsets, KCs overlap between them ($\mathcal{K}_{\mathcal{H}} \cap \mathcal{K}_{\mathcal{C}} \neq \emptyset$), meaning exercises in $\mathcal{V}_{\mathcal{H}}$ and $\mathcal{V}_{\mathcal{C}}$ share some or all of their knowledge domains.

Cross-domain cold-start. Based on ZeroCD domain level zero-shot cognitive diagnosis (DZCD) (Gao et al., 2024), we define a more constrained scenario than exercise cold-start. We partition $\mathcal{K}^0$ at depth 1 into hot domain(s) $\mathcal{H}$ and cold domain(s) $\mathcal{C}$. The set of KCs involved in $\mathcal{H}$ and $\mathcal{C}$ are defined as $\mathbb{K}_{\mathcal{H}} = \bigcup_{i=1}^{M_{\mathcal{H}}} \mathbb{K}_i^1$ and $\mathbb{K}_{\mathcal{C}} = \bigcup_{i=M_{\mathcal{H}}}^{M} \mathbb{K}_i^1$, respectively.This scenario satisfies both $\mathbb{K}_{\mathcal{H}} \cap \mathbb{K}_{\mathcal{C}} = \emptyset$ and $\mathcal{V}_{\mathcal{H}} \cap \mathcal{V}_{\mathcal{C}} = \emptyset$, meaning $\mathcal{H}$ and $\mathcal{C}$ are isolated at both exercise and domain levels. Additionally, all students from $\mathcal{C}$ must appear in the hot subset, formalized as $\mathcal{U}_{\mathcal{C}} \subseteq \mathcal{U}_{\mathcal{H}}$, where $\mathbb{R}_{\mathcal{H}} = \{(u, v, \mathcal{K}_v, y_{uv}) \mid u \in \mathcal{U}_{\mathcal{H}}, v \in \mathcal{V}_{\mathcal{H}}\}$.

Cognitive Diagnosis Model. CDMs infer students' proficiency of specific KCs by analyzing their exercise responses. We define this generally as $y_{uv} = \mathcal{M}(u, v)$, where $y_{uv} \in \mathbb{R}^d$ represents student u's performance on exercise v. The interaction equation $\mathcal{M}$ typically follows the form $y_{uv} = \sigma(\beta(p - d))$, where $p \in \mathbb{R}^d$ represents student u's proficiency, while d and β denote the exercise v's difficulty and discrimination respectively. The dimension d varies by model: 1 for IRT, the size of $\mathbb{K}$ for NCDM, or a fixed value for MIRT. Parameters α and β are model-dependent, expressed as elements in $\mathbb{R}^d$ or as scalars.

3 Methodology

3.1 Overview

Traditional CDMs face significant challenges in cold-start scenarios due to weak connections between $\mathcal{V}_{\mathcal{H}}$ and $\mathcal{V}_{\mathcal{C}}$. The proposed LMCD overcomes these limitations with two core innovations. First, *Knowledge Diffusion*, harnesses LLMs to enrich the content of exercises and KCs, strengthening semantic links (Section 3.2). Second, *Semantic-Cognitive Fusion*, integrates the textual information of the exercise with the student's cognitive state through Large Language Models (LLMs) to generate a feedback representation of the student's response to the exercise.(Section 3.3). These enhanced representations are subsequently processed by standard CDM heads to predict response probabilities and knowledge proficiency, as illustrated in Figure 2.

3.2 Knowledge Diffusion

NLP-based CDMs are limited by brief exercise texts and vague KC labels, creating imprecise semantic relationships. Similar-looking exercises of-

ten test different KCs, while generic KC labels provide insufficient information for accurate cold-start diagnosis.

Our approach transforms concise educational texts into detailed, explicit representations through LLM-driven text-to-text generation. This process uncovers implicit knowledge within the original text, facilitating more precise differentiation and linkage across domains. We call this process "Knowledge Diffusion".

This process can be applied to both exercise texts and KCs. As illustrated in Figure 2(a), we take "KC diffusion" as an example, with the specific steps outlined as follows: (1) For each target knowledge concept k_{target} in $\mathcal{K} = \mathcal{K}_{\mathcal{H}} \cup \mathcal{K}_{\mathcal{C}}$, we identify a set of semantically similar concepts $K_{neg} = \{k^1_{neg}, k^2_{neg}, ..., k^N_{neg}\}$ as negative examples, specifically selecting the top-N semantically similar KCs from sibling nodes of k_{target}. (2) We collect exercises $\mathcal{V}_{target}$ that assess k_{target} and corresponding exercises $\mathcal{V}^i_{neg}$ for each similar KC. (3) We prompt the LLM to generate enriched KC descriptions that explicitly distinguish between the target concept and its related concepts:

$$k'_{\text{target}} = \text{LLM}\left(k_{\text{target}}, Q_{\text{target}}, K_{\text{neg}}, Q_{\text{neg}}\right) \quad (1)$$

This process yields the enriched knowledge concept set $\mathcal{K}' = \mathcal{K}'_{\mathcal{H}} \cup \mathcal{K}'_{\mathcal{C}}$, establishing stronger semantic bridges between historical and cold-start domains. Notably, negative examples are crucial for generating discriminative content, unlike other approaches (Liu et al., 2024b) that use only exercises of K_{target}. Comparative examples are provided in the Appendix A.3.

3.3 Semantic-Cognitive Fusion

As shown in Figure 2(b), we primarily focus on modifying the embedding layer of language models to incorporate student cognitive state during forward propagation, while adapting the LLM output representation to various off-the-shelf CDMs. Specifically, given exercise v and student u, we leverage the LLM to generate a personalized feedback representation unique to each student-exercise interaction. We then integrate these representations into the difficulty parameter d of CDM, and that can be considered as incorporating the student's cognitive state, representing a form of relative difficulty.

Input Embedding. We defined a special token for each student, for example, the token corresponding to student u is stu_u. We also constructed a cognitive representation embedding layer, specifically used for encoding these student tokens, so the cognitive representation of student u is as follows:

$$E_u = \text{EMBLayer}_{\text{cog}}(stu_u) \quad (2)$$

Student Feedback. To generate personalized feedback representations, we align student embedding E_u with the LLM's semantic space. This alignment enables the model to generate student-specific feedback for each exercise v. Our approach proceeds as follows: For a given exercise v, we first encode all available textual information (including LLM-generated knowledge concept descriptions) using the LLM's native embedding layer. Specifically:

$$E_v = \text{EMBLayer}_{\text{llm}}(\text{Concat}[k', v]) \quad (3)$$

The dimension of E_v is $S \times H$, where S is the length of the entire input text tokens, and H is the hidden size of the LLM. We merge E_u into the last dimension of E_v to get E_{fusion}:

$$E_{\text{fusion}} = \begin{bmatrix} E_v \\ E_u \end{bmatrix} \quad (4)$$

where $E_{\text{fusion}} \in \mathbb{R}^{(S+1)\times H}$. We feed E_{fusion} into the LLM backbone for forward propagation to obtain the final representation O_{fusion}.

$$\begin{aligned} &\mathbf{h}_0 = E_{\text{fusion}} \\ &\mathbf{h}_l = \text{FFN}_l(\text{Attn}_l(\mathbf{h}_{l-1})), \quad l = 1, 2, \ldots, N \\ &O_{\text{fusion}} = \mathbf{h}_l \end{aligned} \quad (5)$$

where Attn and FFN represent the attention layer and feed-forward network structure in the LLM based on the Transformer architecture (Vaswani et al., 2017), respectively, and N is the total number of layers in the LLM. Based on the obtained O_{fusion}, we define the following student feedback representation and exercise representation:

$$O_{\text{feedback}} = O_{\text{fusion}}[:, -1] \in \mathbb{R}^{1\times H} \quad (6)$$

$$O_{\text{v}} = O_{\text{fusion}}[:, -2] \in \mathbb{R}^{1\times H} \quad (7)$$

Leveraging the causal attention mechanism of LLMs, O_{feedback} emerges as the product of interaction between student cognitive state E_u and exercise semantic space E_v. This interaction captures how a specific student processes and responds to a particular exercise's content. Consequently, we define O_{feedback} as the personalized feedback representation for student u on exercise v. In contrast,

O_v is derived solely from the exercise's textual information processing, representing the context-independent semantic encoding of exercise v.

Output Projection. Here, we map the obtained representations to different CDM parameters and define a general form of CDMs as follows::

$$y_{uv} = f_{cdm}(\mathbf{p}, \mathbf{d}, \beta) \tag{8}$$

where $f_{cmd}(\cdot)$ represents the interaction function of different CDMs, $\mathbf{p}$ is the student's proficiency, $\mathbf{d}$ is the difficulty of exercise, β is the discrimination parameter, and y_{uv} is the predicted performance result of student u on exercise v. We projectively map the obtained O_{feedback}, E_u and O_v to get $\mathbf{d}$, $\mathbf{p}$, β as follows:

$$\mathbf{d} = \mathbf{d_{uv}} = W_d(O_{\text{feedback}}) \tag{9}$$
$$\mathbf{p} = W_p(E_u) \tag{10}$$
$$\beta = W_v(O_v) \tag{11}$$

We map O_{feedback} to the difficulty and O_v to the discrimination, which are specific parameters of CDMs. To preserve the independence of proficiency parameters ($\mathbf{p}$), we ensure they remain free from exercise-specific information by deriving them solely from student embeddings E_u. The entire framework is optimized end-to-end using cross-entropy loss:

$$\mathcal{L} = -\sum_{u,v} y_{uv} \log y_{uv} + (1 - y_{uv}) \log(1 - y_{uv}) \tag{12}$$

where $(u, v, y_{uv}) \in \mathbb{R}_{\mathcal{H}}$. We froze the LLM backbone parameters and fine-tuned it with LoRA (Hu et al., 2022). To better enable cognitive state and LLM semantic space fusion, we activate $\text{EMBLayer}_{\text{llm}}$ and $\text{EMBLayer}_{\text{cog}}$ during training.

4 Experimental

In this section, we conduct comprehensive experiments to address the following research questions:

- **RQ1** How powerful is LMCD for the exercise cold-start task?
- **RQ2** Can LMCD effectively establish links across different domains?
- **RQ3** How effective are the key components of LMCD?
- **RQ4** Is relative difficulty more reasonable compared to absolute difficulty?

4.1 Experimental Setup

Datasets. We selected two open-source real-world datasets: NIPS34 (Wang et al., 2020b) and XES3G5M (Liu et al., 2023), both with exercise text, KC identifiers and their structural relationships. For cold-start experiments, we applied 5-fold cross-validation for exercise cold-start while 3-fold for cross-domain task: In each data division, 1 fold served as cold-start data, with 20% used as test set and 80% for validation or Oracle model training, while the remaining folds were used as training set. See Appendix A.1 for specific details on data handling.

Baselines. Here we compared the following 7 methods, applicable to IRT/MIRT/NCDM prediction heads.

- Oracle: Both training and test sets are from the cold subset $\mathcal{C}$, representing the theoretical upper bound on the performance of CDMs in cold-start scenarios.
- Random: Correspondingly, the lower bound of prediction skill is measured by randomly sampling from $\text{Uniform}(0, 1)$ as the student's correct response probability.
- TechCD (Gao et al., 2023): TechCD utilizes graphical relationships between KCs to establish connections between students' practiced and unseen exercises, generating student representations through historical interactions.
- NLP-based: Unlike TechCD, which only uses BERT (Devlin et al., 2019) as a baseline, we further incorporate stronger baselines, including RoBERTa (Liu et al., 2019) and BGE (Xiao et al., 2024), to provide a more comprehensive evaluation of semantic-based methods, while ensuring fairness in the experimental setup by using the same textual content.
- KCD (Dong et al., 2025): As the current SOTA method, KCD utilizes LLMs for reasoning, transforming textual content and interaction records into prompts to extract information and generate summaries. These summaries are mapped to the behavioral space of CDMs and optimized via contrastive loss.

All the experimental details of the baseline approach can be found in Appendix A.2.

Metrics and Training Settings. We adopt commonly used metrics, namely the Area Under Curve (AUC), the Prediction Accuracy (ACC), and the Root Mean Square Error (RMSE), to validate the

effectiveness of the CDMs. For all the metrics, ↑ represents that a greater value is better, while ↓ represents the opposite. Qwen-Plus was utilized to generate descriptions information for KCs, while Qwen2.5-1.5B-base (Yang et al., 2024) serving as the foundation model for our approach. Training was conducted using the DeepSpeed (Rasley et al., 2020) framework on $8 \times A800$ GPUs, with a maximum of 10 training epochs. The learning rate was set to 0.0001 with a linear scheduler.

4.2 Performance on Exercise Cold-Start(RQ1)

We conducted exercise cold-start experiments on the NIPS34 and XES3G5M datasets, with results summarized in Table 1. The key observations are as follows: (i) Our proposed LMCD consistently outperformed other methods across most experiments. Notably, the NLP-based approach utilizing BGE embeddings achieved the second-best performance in several cases, a result attributable to BGE's strong semantic representation capabilities. This highlights the importance of LMCD's strategy to integrate exercise text with student cognitive states, which surpasses methods relying solely on textual information. (ii) The graph-based TechCD methods showed limited effectiveness in exercise cold-start scenarios. This stems from TechCD's heavy reliance on topological relationships between KCs, which are sparse in real-world datasets, limiting its practical applicability.

4.3 Performance on Cross-Domain Cold-Start(RQ2)

We conducted more challenging cross-domain experiments on the NIPS34 dataset. Due to the sparsity of knowledge embedding in NCDM, which cannot be transferred to cross-domain scenarios, we only performed experiments on IRT and MIRT. The specific results are reported in Table 2. We observe that: In cross-domain cold-start scenarios, LMCD achieved nearly optimal results in both Algebra and Geometry experiments, while KCD performed best in the Number domain. KCD works by using LLMs to generate textual summaries of both each student's problem-solving behaviors and each exercise's user-interaction statistics for transfer to CDMs during training. LMCD leverages textual description of KCs while incorporating students' cognitive information into problem representations. By contrast, TechCD failed completely in cross-domain scenarios mainly due to the fact that the cold-start data is isolated from the training set both at the KC and exercise level, and thus no transferable knowledge among domains can be obtained through the graph structure. In summary, experimental results show LMCD demonstrates more robust performance in cross-domain cold-start scenarios.

4.4 Ablation Studies (RQ3)

Impact of CDM parameter representation strategies. We conducted ablation studies on our proposed architecture to validate the effectiveness of *difficulty* and *discrimination* representation. Two LMCD variants were designed: substituting O_{feedback} with O_v in Eq.9, and conversely, replacing O_v with O_{feedback} in Eq.11. We employed IRT-head on the NIPS34 for exercise cold-start. The detailed results are reported in Table 3. The results seem to reveal that optimal performance is achieved when using O_{feedback} to represent *difficulty* and O_v to represent *discrimination*. This finding suggests that difficulty appears to be personalized.

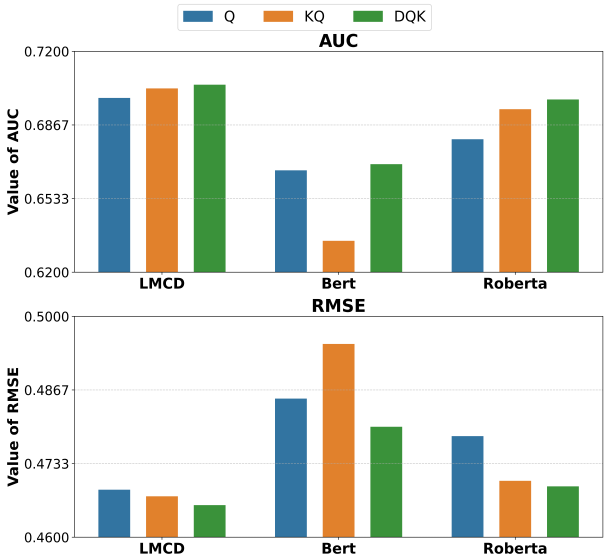

Figure 3: Impact of knowledge encoding strategies.

Impact of knowledge encoding strategies. Figure 3 demonstrates the effectiveness of our "knowledge diffusion" approach across three experimental conditions based on different inputs to Eq.3: **Q** (exercise text only), **KQ** (exercise text with knowledge concept labels), and **DKQ** (adding LLM-generated KC descriptions to KQ).

Results show that enriching encoding content generally improves performance across models, with LMCD and Roberta showing consistent gains in AUC and reductions in RMSE. BERT, however, performs worse with KQ than with Q alone, likely

Dataset	Method	IRT			MIRT			NCDM		
		ACC↑	AUC↑	RMSE↓	ACC↑	AUC↑	RMSE↓	ACC↑	AUC↑	RMSE↓
NIPS34	Oracle	0.7074	0.7738	0.4383	0.7076	0.7741	0.4381	0.7068	0.7711	0.4418
	Random	0.4957	0.4907	0.5002	0.4614	0.4947	0.5265	0.5016	0.5016	0.5205
	Bert	0.6575	0.7153	0.4651	0.6638	0.7217	0.4642	0.6653	0.7159	0.4707
	Roberta	0.6576	0.7165	0.4631	0.6631	0.7204	0.4609	0.6630	0.7204	0.4610
	Bge	0.6708	0.7285	0.4581	0.6752	0.7361	0.4581	0.6741	0.7339	0.4598
	TechCD	0.5548	0.5560	0.4964	0.5520	0.5554	0.4969	0.5414	0.5575	0.5280
	KCD	0.6504	0.7031	0.4871	0.6146	0.7044	0.4898	0.6607	0.7209	0.5819
	LMCD	**0.6813**	**0.7440**	**0.4525**	**0.6823**	**0.7431**	**0.4527**	**0.6776**	**0.7407**	**0.4545**
XES3G5M	Oracle	0.7793	0.7597	0.3937	0.7782	0.7589	0.3946	0.7644	0.7166	0.4095
	Random	0.4966	0.4998	0.5000	0.4980	0.4989	0.5001	0.4983	0.5068	0.5006
	Bert	0.7528	0.6559	0.4218	0.7583	0.6615	0.4193	0.7220	0.6271	0.4470
	Roberta	0.7578	0.6571	0.4194	**0.7603**	0.6578	0.4182	0.7100	0.6224	0.4622
	Bge	0.7520	0.6593	0.4213	0.7601	0.6710	0.4182	0.7246	0.6261	0.4478
	TechCD	0.7509	0.5232	0.4415	0.7509	0.5206	0.4448	0.7501	0.5545	0.4327
	KCD	**0.7584**	0.6430	0.4725	0.7538	0.6542	0.4208	**0.7602**	**0.6741**	**0.4162**
	LMCD	0.7560	**0.6723**	**0.4181**	0.7559	**0.6742**	**0.4174**	0.7436	0.6408	0.4284

Table 1: Performance of Exercise Cold Start on NIPS34 and XES3G5M datasets. The best performance is highlighted in bold, and the the second-best performances is underlined.

CDM	Method	Number as Target			Algerbra as Target			Geometry as Target		
		ACC↑	AUC↑	RMSE↓	ACC↑	AUC↑	RMSE↓	ACC↑	AUC↑	RMSE↓
IRT	Oracle	0.7215	0.7931	0.4278	0.7108	0.7691	0.4379	0.7172	0.7851	0.4329
	Random	0.4932	0.5022	0.5002	0.5019	0.5015	0.5001	0.4864	0.4929	0.5004
	Bert	0.6092	0.6573	0.4903	0.6251	0.6689	0.4800	0.6120	0.6462	0.4863
	Roberta	**0.6369**	0.6824	**0.4720**	0.6483	0.6982	0.4692	0.6224	0.6718	0.4816
	Bge	0.6204	0.6521	0.4957	0.5671	0.6496	0.5531	0.5979	0.6498	0.5145
	TechCD	0.4827	0.5151	0.5059	0.4936	0.5071	0.5125	0.4959	0.5046	0.5084
	KCD	0.6357	**0.6945**	0.4880	0.6505	**0.7091**	0.4857	0.6271	0.6726	0.4888
	Our Method	0.6336	0.6837	0.4726	**0.6518**	0.7049	**0.4658**	**0.6363**	**0.6843**	**0.4758**
MIRT	Oracle	0.7223	0.7942	0.4272	0.7116	0.7694	0.4377	0.7180	0.7860	0.4324
	Random	0.4354	0.4938	0.5328	0.5064	0.4951	0.5154	0.4770	0.5010	0.5224
	Bert	0.6339	0.6712	0.4814	0.6318	0.6818	0.4761	0.6212	0.6683	0.4828
	Roberta	0.6195	0.6432	0.5125	0.6524	0.7045	0.4659	0.6096	0.6339	0.5308
	Bge	0.6341	0.6681	0.4842	0.6366	0.6904	0.4819	0.6115	0.6682	0.4879
	TechCD	0.5646	0.5267	0.4956	0.4936	0.5153	0.5009	0.5231	0.5014	0.4995
	KCD	**0.6500**	**0.6974**	**0.4715**	0.5922	**0.7120**	0.4992	0.6025	0.6754	0.4994
	LMCD	0.6269	0.6888	0.4778	**0.6551**	0.7082	**0.4656**	**0.6277**	**0.6821**	**0.4783**

Table 2: Performance Comparison of Cross Domain with IRT and MIRT on NIPS34. The best performance is highlighted in bold, and the the second-best performances is underlined.

Method	**ACC**	**AUC**	**RMSE**
Eq.9$_{\mathrm{O}_{\text{feedback}} \leftarrow \mathrm{O}_v}$	0.6755	0.7387	0.4546
Eq.11$_{\mathrm{O}_v \leftarrow \mathrm{O}_{\text{feedback}}}$	0.6792	0.7404	0.4549
LMCD	0.6813	0.7440	0.4525

Table 3: Different strategies to representing $\mathbf{p}$ and β.

due to the ambiguity of concept labels without context. This pattern validates the necessity of our "knowledge diffusion" approach, which provides detailed KC descriptions rather than relying solely on potentially ambiguous concept labels.

4.5 "Difficulty" Analysis (RQ4)

Figure 4 compares exercise difficulty and student performance using data from 5 randomly selected students from NIPS34, each having completed over 50 exercises in the cold domain.The top panel displays LMCD-modeled relative difficulty, while the bottom shows BERT-based absolute difficulty. Both models demonstrate that incorrect responses (orange) typically exhibit higher difficulty levels compared to correct responses (blue), which aligns with intuitive expectations. However, LMCD's relative difficulty measure demonstrates significantly better discrimination, with less overlap between correct and incorrect response distributions. This clearer separation indicates that relative difficulty modeling provides a more effective representation for cognitive diagnosis and performance prediction.

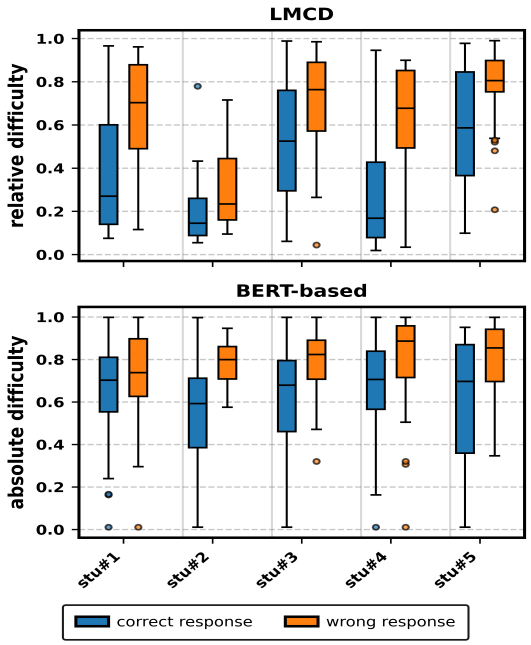

Figure 4: Relative difficulty vs Absolute difficulty.

5 Related Work

Cognitive Diagnosis. Cognitive diagnosis in education has evolved from traditional psychometric approaches to deep learning models. Early methods like IRT (Lord, 1952) and MIRT (Reckase, 2006) model student-exercise interactions through logistic functions, while DINA (De La Torre, 2009) introduce slip and guess parameters. These foundational approaches offered interpretability but limited expressiveness. NCDM (Wang et al., 2020a) mark a significant advancement by leveraging neural networks to capture more complex knowledge representations. Recent research has expanded the modeling scope by incorporating additional factors such as emotional states (Wang et al., 2024b), unlabeled data (Chen et al., 2023), and response time (Ma et al., 2025a). Despite advances, most methods struggle in cold-start scenarios with limited data.

Cold-start in Cognitive Diagnosis. Solutions primarily fall into two categories, graph-based and NLP-based approaches. Graph-based methods (Liu et al., 2024a; Gao et al., 2023, 2024) leverage structural relationships between educational entities. TechCD (Gao et al., 2023) leverage tailored knowledge concept graphs linking different domains but requires overlapping students. ZeroCD (Gao et al., 2024) utilizes early bird students in target domains to learn transferable cognitive signals, though this requirement limits practical application. NLP-based approaches show promising results by utilizing language models like BERT to encode exercise text (Gao et al., 2023, 2024), but face limits such as oversimplifying exercise difficulty and misreading unclear texts, leading to inaccuracies in cross-domain settings.

LLMs in Cognitive Diagnosis. LLMs have revolutionized natural language processing with advanced comprehension (Brown et al., 2020) and reasoning capabilities (Kojima et al., 2022; Ma et al., 2025b), but their integration with cognitive diagnosis remains nascent. Current approaches include LRCD (Liu et al., 2025), which embeds students, exercises, and concepts into a unified language space; and KCD (Dong et al., 2025), which exploits LLMs' reasoning ability to generate diagnostic information for students and exercises. However, these approaches' simplistic use of LLMs limits their effectiveness in cold-start scenarios.

6 Conclusion

In this paper, we proposed LMCD, a novel framework that harnesses LLMs to address cold-start challenges in cognitive diagnosis. Our approach innovatively leverages knowledge diffusion to establish stronger cross-domain semantic connections and employs causal attention mechanisms to model relative difficulty and student features. Experiments on two real-world datasets confirm that LMCD significantly outperforms state-of-the-art methods in both exercise-cold and domain-cold settings. To our knowledge, this is the first work that takes advantage of the inherent mechanisms of LLM to address cold-start cognitive diagnosis problems, marking a significant advance in the field.

Limitations

Despite significant improvements in exercise cold-start and cross-domain scenarios through our student token approach and causal attention interaction mechanism, our method faces limitations in diagnosing new students. Only students with training data have corresponding embedding representations, restricting new student cold-start applications. A potential workaround involves substituting new students with trained students having similar response patterns, a common approach in this field (Long et al., 2022). Additionally, our LLM-based approach contains substantially more parameters than traditional CD models, making it less suitable for time-sensitive applications. However, we consider this computational cost justified for challenging cold-start scenarios, and expect that smaller, more efficient models will become viable as LLM technology advances.

References

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 *2020 IEEE International Conference on Data Mining (ICDM)*, pages 42–51. IEEE.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Xiangzhi Chen, Le Wu, Fei Liu, Lei Chen, Kun Zhang, Richang Hong, and Meng Wang. 2023. Disentangling cognitive diagnosis with limited exercise labels. *Advances in Neural Information Processing Systems*, 36:18028–18045.

Jimmy De La Torre. 2009. Dina model and parameter estimation: A didactic. *Journal of educational and behavioral statistics*, 34(1):115–130.

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)*, pages 4171–4186.

Zhiang Dong, Jingyuan Chen, and Fei Wu. 2025. Knowledge is power: Harnessing large language models for enhanced cognitive diagnosis. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

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 AAAI Conference on Artificial Intelligence*, volume 38, pages 8417–8426.

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*, pages 983–992.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Zhenya Huang, Qi Liu, Chengxiang Zhai, Yu Yin, Enhong Chen, Weibo Gao, and Guoping Hu. 2019a. Exploring multi-objective exercise recommendations in online education systems. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 1261–1270.

Zhenya Huang, Qi Liu, Chengxiang Zhai, Yu Yin, Enhong Chen, Weibo Gao, and Guoping Hu. 2019b. Exploring multi-objective exercise recommendations in online education systems. In *CIKM '19*, CIKM '19, page 1261–1270, New York, NY, USA. Association for Computing Machinery.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Shuo Liu, Junhao Shen, Hong Qian, and Aimin Zhou. 2024a. Inductive cognitive diagnosis for fast student learning in web-based intelligent education systems. In *Proceedings of the ACM Web Conference 2024*, pages 4260–4271.

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*, Toronto, Canada.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yuanhao Liu, Shuo Liu, Yimeng Liu, Jingwen Yang, and Hong Qian. 2024b. A dual-fusion cognitive diagnosis framework for open student learning environments. *arXiv preprint arXiv:2410.15054*.

Zitao Liu, Qiongqiong Liu, Teng Guo, Jiahao Chen, Shuyan Huang, Xiangyu Zhao, Jiliang Tang, Weiqi Luo, and Jian Weng. 2023. Xes3g5m: A knowledge tracing benchmark dataset with auxiliary information. *Advances in Neural Information Processing Systems*, 36:32958–32970.

Ting Long, Jiarui Qin, Jian Shen, Weinan Zhang, Wei Xia, Ruiming Tang, Xiuqiang He, and Yong Yu. 2022. Improving knowledge tracing with collaborative information. In *Proceedings of the fifteenth ACM international conference on web search and data mining*, pages 599–607.

Frederic Lord. 1952. A theory of test scores. *Psychometric monographs*.

Haiping Ma, Yue Yao, Changqian Wang, Siyu Song, and Yong Yang. 2025a. Ad4cd: Causal-guided anomaly detection for enhancing cognitive diagnosis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 12337–12345.

Jie Ma, Zhitao Gao, Qi Chai, Wangchun Sun, Pinghui Wang, Hongbin Pei, Jing Tao, Lingyun Song, Jun Liu, Chen Zhang, and 1 others. 2025b. Debate on graph: a flexible and reliable reasoning framework for large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24768–24776.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 3505–3506.

Mark D Reckase. 2006. 18 multidimensional item response theory. *Handbook of statistics*, 26:607–642.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Howard Wainer, Neil J Dorans, Ronald Flaugher, Bert F Green, and Robert J Mislevy. 2000. *Computerized adaptive testing: A primer*. Routledge.

Fei Wang, Weibo Gao, Qi Liu, Jiatong Li, Guanhao Zhao, Zheng Zhang, Zhenya Huang, Mengxiao Zhu, Shijin Wang, Wei Tong, and 1 others. 2024a. A survey of models for cognitive diagnosis: New developments and future directions. *arXiv preprint arXiv:2407.05458*.

Fei Wang, Qi Liu, Enhong Chen, Zhenya Huang, Yuying Chen, Yu Yin, Zai Huang, and Shijin Wang. 2020a. Neural cognitive diagnosis for intelligent education systems. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 6153–6161.

Shanshan Wang, Zhen Zeng, Xun Yang, Ke Xu, and Xingyi Zhang. 2024b. Boosting neural cognitive diagnosis with student's affective state modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 620–627.

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, and 1 others. 2020b. Instructions and guide for diagnostic questions: The neurips 2020 education challenge. *arXiv preprint arXiv:2007.12061*.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 641–649.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Xiaoshan Yu, Chuan Qin, Qi Zhang, Chen Zhu, Haiping Ma, Xingyi Zhang, and Hengshu Zhu. 2024. Disco: A hierarchical disentangled cognitive diagnosis framework for interpretable job recommendation. In *2024 IEEE International Conference on Data Mining (ICDM)*. IEEE.

A Appendix

A.1 Details about the Datasets

In this research, we selected two real-world online education datasets, XES3G5M and NIPS34, which vary in size and sparsity, to comprehensively evaluate the performance of our proposed method across different practical application scenarios. The statistical information of the datasets is presented in Table 4.

NIPS34: A sub-dataset of the NeurIPS Education Challenge, containing Tasks 3 and 4, is a powerful resource tailored to the advancement of educational data analytics and machine learning applications within the education field. The dataset comprises crowdsourced diagnostic mathematics exercises collected from the Eedi educational platform between September 2018 and May 2020, targeting students from elementary through high school. Task 3 aims to accurately predict which exercises are of high quality, while Task 4 seeks to determine a personalized sequence of exercises for each student that optimally predicts their responses. NIPS34 contains a variety of features including interaction logs, the content of the exercises in picture format, the names of the KCs

and their structured annotations, which allow for a comprehensive analysis of learning behaviors and outcomes. NIPS34 contains more than 1,300,000 records, providing a wealth of high-quality data suitable for multiple types of tasks, making it an invaluable tool for conducting cognitive diagnosis research.

XES3G5M: XES3G5M is a large-scale dataset comprising over five million interactions collected from more than 18,000 third-grade students responding to approximately 8,000 math exercises. It includes extensive auxiliary information related to the exercises and their associated KCs. Sourced from a real-world online mathematics learning platform, XES3G5M not only encompasses the largest number of KCs within the mathematics domain but also provides the most comprehensive contextual information. This includes hierarchical KC relationships, exercise types, Chinese textual content and analyses, as well as timestamps of student responses. In this experiment, we use five-fold cross

	XES3G5M	NIPS34
#Student	11,453	4,918
#Exercise	7,652	948
#KC	1,175	86
#Log	5,139,044	1,382,727
#Log per student	448.7	281.2
#Log per exercise	671.6	1,458.6
#Sparsity(%)	5.1	29.7

Table 4: Statistics of datasets.

validation to ensure the reliability of the experimental results. We use the full amount of NIPS34 data, and the XES dataset, due to the fact that there are too many record exercises, has done the following treatment, and only 2000 students' corresponding records of doing the exercises are kept in each fold. Table.5 displays the average data distribution for each fold after implementing five-fold cross-validation on both datasets.

A.2 Details about Baselines

In this subsection, we provide a detailed explanation of the specific modifications made to each baseline experiment based on its original implementation to adapt it to the data split used in the specific cold-start scenario of this research:

IRT/MIRT: We adopt a three-parameter logistic (3PL) form for the interaction function. The temperature coefficient is maintained at its original setting of 1.703, and the feature dimension of MIRT is set to 4. Meanwhile, considering the specific types of the exercises, we set the upper bound of the guess coefficient to 0.5. In addition, Xavier initialization is applied to all embedding layers in the models.

Category	NIPS35	XES3G5M
Cold Exercises (test)	186.6	1055.0
Cold Exercises (oracle)	190.0	1268.2
Hot Exercises	758.0	5169.2
Train Logs	1105934.0	243309.2
Oracle Logs	221434.0	48616.8
Test Logs	55359.0	12154.8

Table 5: Mean Statistics Across Folds

NeuralCD: We employ the default settings without any modifications to ensure the consistency of the experiment results.

TechCD: The model architecture remains consistent with the original setup, with the addition of optional IRT and MIRT predict heads. Furthermore, undirected edges characterizing the similarity between fine-grained ($depth \geq 3$) KCs under the same parent node are also introduced in the construction of the Knowledge Concept Graph (KCG) for the NIPS34 and XES3G5M datasets, based on the provided KC tree structure.

NLP-base: We employ language models with frozen weights to embed the textual information of students and exercises, including descriptions generated by LLMs, thereby replacing the ID embedding layers in IRT/MIRT and NCDM. Experiments are conducted using BERT, RoBERTa, and BGE as text embedders. Specifically, we use bert-base-uncased/chinese, bge-large-en/zh-v1.5, and xlm-roberta-base for adaptation to datasets in different languages.

KCD: We use Qwen-Plus for information extraction and diagnosis and modify the prompts as follows. First, input interactions are truncated up to 20 to avoid excessively long prompts, while contents related to interactions are removed from the prompts that generate descriptions for cold-start data. Then, the text embedder employed is bge-large-en/zh-v1.5, and a Chinese version of the prompt was added for the XES3G5M dataset.

A.3 Details about Knowledge Diffusion

In order to generate descriptions for each KC, we utilize the exact format of prompt as follows:

Prompt for KC Description Generation

System Prompt: "If you are a seasoned math teacher, you need to generate explanations for each knowledge concept in a knowledge graph. I will provide you with the name of the knowledge concept and corresponding example problems, as well as names and example problems of distractor knowledge concepts that are not equivalent to it. Please carefully compare them and generate a core explanation for each knowledge concept."

KC Name: k_{target}
Example exercises for k_{target}:
Q_{target}

Distractor KCs
KC Name:k^1_{neg}
Example exercises for k^1_{neg}:
Q^1_{neg}
...
KC Name:k^n_{neg}
Example exercises for k^N_{neg}:
Q^N_{neg}

KC Descriptions: Case 1

Target KC: **Fractions, Decimals and Percentage Equivalence**

With distractor KCs:
"Fractions involve understanding parts of a whole and performing mathematical operations such as addition, subtraction, multiplication, and division using fractional numbers. This includes interpreting fractions visually and solving problems that require manipulation of fractions in equations or patterns."

Without distractor KCs:
"Fractions represent parts of a whole and are written as a numerator over a denominator. The numerator indicates how many parts are being considered, while the denominator shows the total number of equal parts the whole is divided into. Operations with fractions follow specific rules:
1. **Multiplication**: Multiply the numerators and denominators separately, then simplify if necessary.
- Example: $\frac{1}{3} \times (-9) = \frac{-9}{3} = -3$.
2. **Division**: ..."

Enrichment and robustness. As shown in Case 1, when describing the KC "Fractions, Decimals and Percentage Equivalence," the inclusion of distractors effectively prevents the LLM from mechanically enumerating definitions based solely on the target KC's name. Furthermore, it avoids an excessive focus on the content of example exercises leading to redundant information. In summary, the inclusion of distractors not only enriches the prompt but also enhances the robustness of the generated results by mitigating sensitivity to the selection of specific example exercises.

KC Descriptions: Case 2

Target KC: **Angles**

With distractor KCs:
"Angles are geometric measures that represent the amount of turn between two intersecting lines or line segments. They can be measured in degrees, estimated visually, compared in size, or determined using tools like protractors. Understanding angles involves recognizing different types (acute, obtuse, right, etc.) and performing operations such as addition, subtraction, or fraction-based calculations involving turns."

Without distractor KCs:
"An **Angle** is a measure of rotation or the amount of turn between two intersecting lines or rays. Angles are typically measured in degrees ($^\circ$), with a full turn equaling 360°. A half-turn corresponds to 180°, and a quarter-turn equals 90°, which is also called a right angle. Angles can be classified based on their size:
Acute angles are less than 90°.
Right angles are ...
To estimate or measure angles, tools such as protractors are used. Estimation involves comparing the given angle to known reference angles (e.g., 90°, 180°). In cases where direct measurement is not possible, logical reasoning or comparison may help determine relationships between angles."

Abstraction and discrimination. Besides, as shown in Case 2, which describe the KC "Angles", the incorporation of distractors results in more concise descriptions. These descriptions predominantly emphasize the knowledge and skill elements rather than specific examples. In other words, distractor KCs facilitate the abstraction of concrete exercises, thereby accentuating the differences between similar KCs at knowledge and skill level. This abstraction contributes to higher precision and discrimination between the generated descriptions.