Exploiting Non-Interactive Exercises in Cognitive Diagnosis

Fangzhou Yao[1,2] , **Qi Liu**[1,2 *] , **Min Hou**[1,2] , **Shiwei Tong**[1,2] , **Zhenya Huang**[1,2] , **Enhong Chen**[1,2] , **Jing Sha** [3] , **Shijin Wang** [2,3]

[1]Anhui Province Key Laboratory of Big Data Analysis and Application,
University of Science and Technology of China
[2]State Key Laboratory of Cognitive Intelligence
[3]iFLYTEK AI Research (Central China), iFLYTEK Co., Ltd.
{fangzhouyao, minho, tongsw}@mail.ustc.edu.cn, {huangzhy, qiliuql, cheneh}@ustc.edu.cn,
{jingsha, sjwang3}@iflytek.com

Abstract

Cognitive Diagnosis aims to quantify the proficiency level of students on specific knowledge concepts. Existing studies merely leverage observed historical students-exercise interaction logs to access proficiency levels. Despite effectiveness, observed interactions usually exhibit a power-law distribution, where the long tail consisting of students with few records lacks supervision signals. This phenomenon leads to inferior diagnosis among few records students. In this paper, we propose the Exercise-aware Informative Response Sampling (EIRS) framework to address the long-tail problem. EIRS is a general framework that explores the partial order between observed and unobserved responses as auxiliary ranking-based training signals to supplement cognitive diagnosis. Considering the abundance and complexity of unobserved responses, we first design an Exercise-aware Candidates Selection module, which helps our framework produce reliable potential responses for effective supplementary training. Then, we develop an Expected Ability Change-weighted Informative Sampling strategy to adaptively sample informative potential responses that contribute greatly to model training. Experiments on real-world datasets demonstrate the supremacy of our framework in long-tailed data.

1 Introduction

Cognitive diagnosis has been increasingly needed to assess and improve individual development in intelligent educational applications [Liu, 2021; Zhou *et al.*, 2021]. Given the historical interaction logs of students, it aims to discover their latent cognitive states (proficiency levels) on knowledge concepts and reveal some exercise features such as difficulty and discrimination [Pandey and Karypis, 2019; Wu *et al.*, 2020; Tong *et al.*, 2021].

Existing cognitive diagnosis models (CDMs) mainly focus on assessing students' proficiency level based on historical student-exercise interaction logs like Item Response Theory (IRT) [Lord, 1980] and NeuralCD [Wang *et al.*, 2020]. However, in real scenarios, a large number of students interact with very few exercises [Lu *et al.*, 2022], which leads to the fact that the distribution over students is quite imbalanced and even long-tailed. In Figure 1(a), we sort 10000 students randomly sampled from the well-known dataset Junyi (math practicing logs, description in Section 5.1) by the number of interaction times in a descending order. We notice a heavy long-tailed distribution that nearly 90% students' interaction times are less than 50. The proposed works pay little attention to this problem. They validate model's performance on datasets by filtering students with few interactions to ensure enough logs for executing diagnostic tasks, which runs counter with the task of diagnosing each student's knowledge level. In Figure 1(b), we evaluate recent advanced CDMs' performance with accuracy on different students grouped by interaction times. It shows that all models exhibit lower accuracy in students group with limited interactions compared to students with plentiful interactions (more than 50 times). From this observation, it is reasonable to presume that limited interactions lead to inaccurate and uncertain diagnosis results. In other words, insufficient supervised signals result in poor robustness of CDMs in students with few interactions.

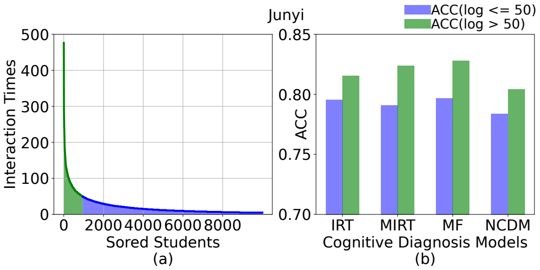


Figure 1: (a) Sorted interaction times for students and (b) CDMs performances comparison on different student groups.

To tackle this problem, we naturally come up with incorporating auxiliary training signals from non-interactive exercises, and the relationship between student-exercise-knowledge can be used to infer potential signals. Specifically, a student probably performs similarly on exercises that share

[*] Corresponding Author.

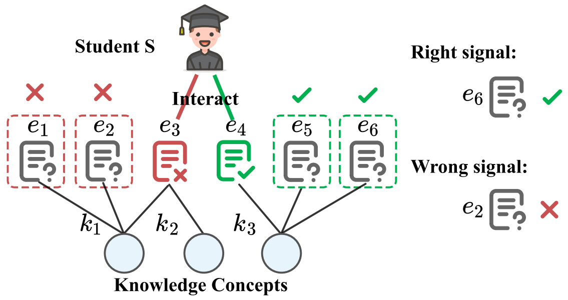


Figure 2: A simple example of distilling response signals through student-exercise-knowledge.

the same knowledge concept. An example is shown in Figure 2, student s has a correct response to exercise e_4 and an incorrect response to exercise e_3, s may have a higher proficiency level towards the knowledge concept k_3 behind e_4 than k_1 behind e_3. Thus, we may infer a partial order relationship that the probability of correctly answering e_6 with k_3 is much higher than e_2 with k_1. The partial order relationship has great potential in constructing additional training signals. Unfortunately, there are still many challenges in designing an effective solution to exploit this relationship and guide the training signal sample on unobserved data. On one hand, how to identify *reliable signals* from the massive potential partial order pairs is still an open issue. Due to the absence of ground-truth responses, partial order pairs bring information meanwhile noise. For example, a student's response to an exercise with the same knowledge concept as a previously correctly responded one may be deemed as correct. However, the actual signal may be incorrect due to the new exercise being significantly more difficult than the previously responded one. On the other hand, how to effectively select *informative signals* is a non-trivial problem. When exploring the partial order pairs towards possible training signals, the scale of training samples increases dramatically. A common solution is randomly sampling instances from unobserved data [Rendle *et al.*, 2012]. However, prior works [Rendle and Freudenthaler, 2014; Lian *et al.*, 2020; Qin *et al.*, 2020] demonstrate that it makes the model converge slowly, especially when the pool of instances is large. In order to speed up convergence, we need to design an efficient sampling strategy.

In this work, we propose a general framework called Exercise-aware Informative Response Sampling (EIRS) to overcome the long-tailed distribution problem in cognitive diagnosis. EIRS provides reliable and informative auxiliary training signals that can be seamlessly incorporated into existing cognitive diagnosis models. To achieve this goal, we first explore the partial-order relationship between student-exercise-knowledge in depth. We then design the Exercise-aware Candidates Selection (ECS) module to ensure the reliability of the training signals. Two indicators that measure the difficulty or discrimination of exercises are proposed to help identify reliable training pairs. For discriminating informative signals, we introduce Expected Ability Change weighted Informative Sampling strategy (EIS). EIS adaptively selects samples from two perspectives including contribution to model training and student ability change. Experiments on two real-world datasets demonstrate that the proposed EIRS improves CDMs' robustness on long-tailed data and speeds up model convergence.

2 Related Work

2.1 Cognitive Diagnosis

The task of evaluating students' knowledge level from historical response logs has been studied since 1950s. Item Response Theory (IRT) [Lord, 1980] is a standard statistical model cognitive diagnosis, which uses single-dimension variables to represent the trait features and logistic function. Later, Multidimensional Item Response Theory (MIRT) [Reckase, 2009] proposed to use multidimensional trait features instead of single dimension. Another classic model, Noisy "And" gate model (DINA) [De La Torre, 2009] diagnoses the mastery state by binary variables and considerates students' slip and guessing factors. Recently, the prevalence of deep learning motivates a rich line of work on cognitive diagnosis. Recently, some researchers introduce the deep learning into cognitive diagnosis [Wang *et al.*, 2020; Ma *et al.*, 2022; Gao *et al.*, 2021]. Wang *et al.* proposed NeuralCD framework to learn the interaction function between students and responses with neural networks. Furthermore, Gao *et al.* designed a multi-layer student-exercise-concept relation map to model the interactive and structural relations. The methods described above learn trait parameters from entire historical response logs, which can suffer from a long-tailed problem and worsen across students' diagnoses when there are insufficient supervision signals

2.2 Sampling Strategy

One key component of our framework is to sample potential responses for the observed response anchor, which is most relevant to sampling strategy technology applied in some domains like natural language processing [Mikolov *et al.*, 2013], recommendation [Rendle and Freudenthaler, 2014; Qin *et al.*, 2019], etc. Static sampling strategies sample unobserved data based on a predefined distribution, such as uniform and popularity distribution corresponding to random sampling [Rendle *et al.*, 2009] and popularity-based sampling [Caselles-Dupré *et al.*, 2018; Mikolov *et al.*, 2013] respectively. However, static methods cannot adjust to model training, suffering from low quality of samples. Adaptive sampling was proposed later, such as DNS [Zhang *et al.*, 2013] which dynamically selects hard samples that are difficult for current model to discriminate. Inspired by generative adversarial learning [Goodfellow *et al.*, 2014], some researchers have studied adversarial training between the sampling model (the generator) and the training model (the discriminator) [Wang *et al.*, 2018; Park and Chang, 2019]. For example, Park [2019] proposed AdvIR to generate hard negatives by adding adversarial perturbations to them. However, since student-exercise responses depend on complex features such as knowledge concepts, difficulty and discrimination of exercises [Liu *et al.*, 2021], sampling task in cognitive diagnosis is more challenging than sampling in other scenarios.

3 Problem Definition

Here we give a formal definition of cognitive diagnosis. Let $S = \{s_1, s_2, ..., s_N\}$ be the set of N students, $E = \{e_1, e_2, ..., e_M\}$ represent the set of M exercises, and $K = \{k_1, k_2, ..., k_L\}$ denote the set of L knowledge concepts. We define the student-exercise interaction set of the entire space as $R = S \times E$. The observed interaction logs are a triplet set $R_O = \{(s, e, r_{se}) | (s, e) \in R, r_{se} \in \{0, 1\}\}$ where r_{se} represents a student's response to an exercise (i.e., 0 indicates wrong answer while 1, otherwise). The number of interactive logs is much smaller than that of $|R|$, that is $|R_O| \ll |R|$. Besides, we have Q-matrix [Tatsuoka, 1995] labeled by experts, $Q = \{Q_{ij}\}_{M \times L}$, where $Q_{ij} = 1$ indicates the exercise e_i relates to the knowledge concept j and $Q_{ij} = 0$ otherwise.

Given: Students' interactions logs R_O and Q-matrix Q.

Goal: Quantify students' knowledge level on specific knowledge concepts by modeling the student performance prediction process.

4 Methodology

In this section, we introduce Exercise-aware Informative Response Sampling (EIRS) framework which could be applied to all existing CDMs. In the following parts, we will first introduce the backbone cognitive module with the optimization task. Then we will explain the shortcomings of the current optimization task and show how to leverage the partial order to formulate a new ranking optimization task. After that, we will dive into the details of our proposed partial-order response sampling strategy and the learning algorithm.

4.1 Basic Cognitive Diagnosis Model

Cognitive diagnosis model (CDM) is for assessing students' proficiency level according to their observed responses to exercises. Generally, CDM contains two steps: (1) the embedding layer to obtain the diagnostic factors of students and exercises, (2) the interactive layer to learn the interaction function among the factors and output the probability of correctly answering the exercises. After training, we get students' proficiency vectors from the first step as diagnostic results.

Formally, given the student set S, exercise set E, and knowledge concept set K, through corresponding embedding-lookup layer, we represent them as $H_S \in \mathbb{R}^{N \times d}, H_E \in \mathbb{R}^{M \times d}, H_K \in \mathbb{R}^{L \times d}$. Each row of trainable metrics represents the representation of trait features (e.g., h_s is the s^{th}-row of H_S that represents the student s's proficiency). For an exercise e, its difficulty h_e^{diff} and discrimination h_e^{dis} are two important characteristics, we further denote $h_e = [h_e^{diff}, h_e^{dis}]$. Here, we diagnose the cognitive state of student s as h_s and the characteristic of exercise e as h_e. To verify the diagnosis, an interaction function f_C is used to predict whether the student can answer the exercise correctly:

$$\hat{y}_{se} = f_C(h_s, h_e), \tag{1}$$

where $\hat{y}_{se}$ is the probability of the student s correctly answering the exercise e. The architecture of the embedding layer and the interaction function f_C can be arbitrary, all existing CDMs can be chosen, such as IRT [Lord, 1980], MIRT [Reckase, 2009], NeuralCD [Wang *et al.*, 2020], etc. When training the CDM, for each record in the observed response logs set $R_O = \{(s, e, r_{se}) | s \in S, e \in E_O, r_{se} \in \{0, 1\}\}$, we calculate the loss function of basic cognitive diagnosis model as the cross-entropy loss between the prediction score $\hat{y}_{se}$ and the true label r_{se}:

$$\mathcal{L}_{\mathcal{P}} = -\sum_{(s,e,r_{se}) \in R_O} (r_{se}\, log\, \hat{y}_{se} + (1 - r_{se})\, log(1 - \hat{y}_{se})). \tag{2}$$

The model is fit to predict the observed correct responses with value 1 and the incorrect responses with value 0. However, this can be problematic when there are not enough observed interactions, particularly for long-tailed students.

4.2 Partial-Order Ranking

In fact, optimizing the backbone cognitive diagnosis model only through traditional prediction tasks may not provide sufficient training signals, resulting in inferior diagnostic performance. Therefore, we aim to exploit the massive unobserved data. We propose creating item pairs from unobserved interactions as auxiliary training data and optimizing for correctly ranking item pairs instead of only scoring single observed item in cognitive diagnosis. In this section, we give the formulation of partial-order ranking learning.

For each student, we denote E_O, E_U as the interactive and non-interactive exercises set. Thus, the interactive (non-interactive) exercises can be divided into positives E_O^+ (E_U^+) and negatives E_O^- (E_U^-) based on the responses or potential responses to them, where +, - represent correct and incorrect responses respectively. According to the monotonicity theory [Rosenbaum, 1984] declaring that a learner's proficiency is monotonic with the probability of correctly responding to a test item. It can be inferred that a student's proficiency on a correct response is higher than an incorrect one.

Formally, for any exercises e_o^+, e_u^+, e_o^-, e_u^- taken from E_O^+, E_U^+, E_O^-, E_U^-, respectively, we have the following partial order between interactive and non-interactive exercises:

$$y_{so}^+ > y_{su}^-, y_{su}^+ > y_{so}^-, \tag{3}$$

where $y_{so}^+ > y_{su}^-$ means that CDM should give a higher score to the observed correct response to exercise e_o^+ than the potential incorrect response to exercise e_u^-, $y_{su}^+ > y_{so}^-$ similarly. Inspired by the great success of the BPR loss [Rendle *et al.*, 2012] in recommender systems which is defined as maximizing the difference between the predicted probability of a positive pair and a negative pair, we formulate the following constraint based on this theory:

$$\mathcal{L}_{\mathcal{R}} = -\sum_{s, e_o^+, e_u^-} ln\sigma(y_{so}^+ - y_{su}^-) - \sum_{s, e_o^-, e_u^+} ln\sigma(y_{su}^+ - y_{so}^-). \tag{4}$$

4.3 Partial-Order Response Sampling

A key concern in optimizing cognitive models by ranking-based auxiliary training is how to construct reasonable partial-order pairs. The pivotal step is to screen out effective exercises from a large number of non-interactive exercises. If we randomly select non-interactive exercises, the corresponding partial order pairs will bring noise and limited information to the model, which can results in inaccurate

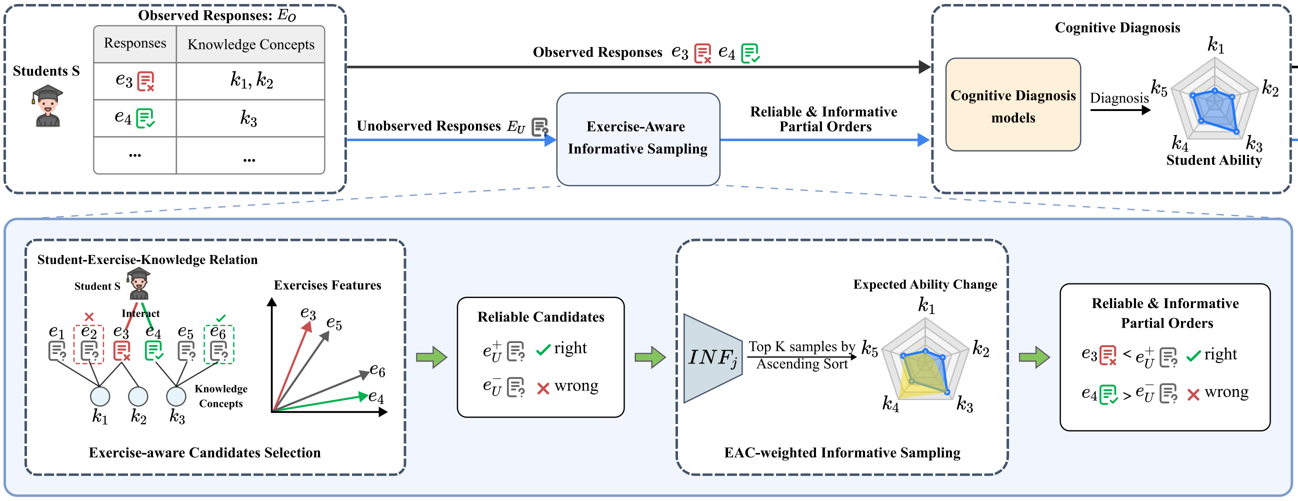

Observed Responses: E_O
Students S
Responses
Knowledge Concepts
e_3
k_1, k_2
e_4
k_3
...
...
Observed Responses e_3 e_4
Unobserved Responses E_U
Exercise-Aware Informative Sampling
Reliable & Informative Partial Orders
Cognitive Diagnosis
Cognitive Diagnosis models
Diagnosis
k_1
k_2
k_3
k_4
k_5
Student Ability
L_P
L_R
Student-Exercise-Knowledge Relation
Student S
Interact
e_1 e_2 e_3 e_4 e_5 e_6
k_1 k_2 k_3
Knowledge Concepts
Exercises Features
e_3 e_5 e_6 e_4
Exercise-aware Candidates Selection
Reliable Candidates
e_U^+ ✓ right
e_U^- ✗ wrong
INF_j
Top K samples by Ascending Sort
Expected Ability Change
EAC-weighted Informative Sampling
Reliable & Informative Partial Orders
e_3 < e_U^+ ✓ right
e_4 > e_U^- ✗ wrong

Figure 3: Exercise-aware Informative Sampling Framework

diagnosis and slow convergence. Therefore, we propose a two-stage response sampling strategy to ensure the reliability and informativeness of the auxiliary signals. As shown in Figure 3, in the first stage, we design a novel Exercise-aware Candidates Selection module (ECS). This module can produce non-interactive candidates set C_U by considering the knowledge relations between exercises and the similarity of their features to responded exercises. In the second stage, we further design EAC-weighted informative sampling (EIS) module, which adaptively samples informative unobserved responses to make significant changes to current model parameters. We will describe these two modules in detail.

Exercise-Aware Candidates Selection Module

Given a large number of non-interactive logs set R_U for a particular student, ECS is to obtain a reliable candidate responses set C_U from R_U. We first describe how to obtain reliable candidate responses and then present a hypothesis test to formally demonstrate the reliability of the selected samples based on the student-exercise-knowledge relationship.

To evaluate the reliability level of the unobserved responses, we consider two aspects. The first is the relationship between student-exercise-knowledge. Intuitively, if a student is skilled at one knowledge concept such as *geometry*, there is a high probability of giving correct responses to exercises related to *geometry*. Therefore, we suppose that a student performs similarly on exercises sharing the same knowledge concept. Using the student-exercise-knowledge relations, we can preliminarily infer reliable correct or incorrect responses from the unobserved responses. Formally, given an interactive log (s, e, r_{se}), we define e's similar knowledge concepts set $N(e) = \{e_u | e_u \in E_U, K(e) \cap K(e_u) \neq \emptyset\}$, where $E_U \in E$ is the set of exercises that the student s has never interacted with and $K(e)$ returns knowledge concepts set that e relates to. Based on the assumption that a student similarly performs on exercises that share same knowledge concepts, we can construct the set of unobserved response logs:

$$R_u = \{(s, e_u, r_{se'}) | e_u \in N(e)\}, \quad (5)$$

where $r_{se'}$ is consistent with the student's response in interactive logs (s, e, r_{se}) with overlapping knowledge concept.

Secondly, other vital properties of exercises like difficulty and discrimination have a huge impact on students' responses. As we discussed previously in Figure 2, both e_5 and e_6 have overlapping knowledge concept k_3 with e_4, while e_5 may be too difficult for the student to answer correctly based on his current knowledge states. We take into account the inherent properties (e.g., difficulty and discrimination) of exercises, then calculate the similarity between exercises as the sampling probability for each non-interactive exercise:

$$p(e_u|s, e) = \frac{h_e^T \cdot h_{e_u}}{\sum_{e_i \in N(e)} h_e^T \cdot h_{e_i}}, \quad (6)$$

where $h_e = [h_e^{diff}, h_e^{dis}]$ represents exercise e's characteristic factor from the basic CDM. A higher probability $p(e_u|s, e)$ score reflects a higher reliability level of the potential response based on observed responses. In this way, we sample some unobserved logs R_{su} according to the probability and combine the unobserved logs set generated by each log to obtain a reliable candidate logs set $C_U = \bigcup_{i=1}^{|R_O|} R_u^i$. After acquiring the reliable candidate response logs set, we can select samples from the set C_U to pair with observed responses for partial-order ranking learning.

It is worth noting that we used the key knowledge concept of the exercise as a criterion for candidate selection. Thus we suppose Consistency Assumption by student-exercise interaction and validate it on two real datasets.

Consistency Assumption. *Students' responses to similar exercises with overlapping knowledge concept are consistent.*

To validate this assumption, we conduct a hypothesis test on two datasets ASSISTments and Junyi (data description in Section 5.1). Specifically, we first give some important notations used in our testing without loss of generality. Student's response results (right or wrong) of exercises e_a, e_b are notated with r_a and r_b, and K_a, K_b are the knowledge concepts

Datasets	$\bar{P}_u$ (s)	α	p-value
ASSISTments	0.638	0.05	1.89-74
Junyi	0.707	0.05	1.49-60

Table 1: One-sided test on real datasets.

that e_a, e_b contain. Then, $P_{um}(r_a = r_b | K_a \cap K_b \neq \emptyset)$ represents the probability that student u_m responses to e_a and e_b with overlapped knowledge are consistent. Let $P_{u_1}, ..., P_{u_M}$ be i.i.d. from the $N(\mu, \sigma^2)$ distribution, where μ is unknown and σ^2 is known.

Generally, if our assumption is valid, the probability P_{um} should be over 0.5. Therefore, we perform a one-tailed test of significance with the null and alternative hypothesis:

$$H_0 : \mu \leq 0.5; H_1 : \mu > 0.5, \tag{7}$$

The mean of P_{um} is notated as $\bar{P}_u$, and α is the level of significance. The p-value represents the probability of observing a given event under the null hypothesis. The testing results are reported in Table 1, and we observe that the p-values are much smaller than α on both datasets, so we reject the null hypothesis and accept the alternative hypothesis. Consequently, it can be concluded that students' responses to exercises with overlapping knowledge are consistent.

Expected Ability Change Weighted Informative Sampling Module

Although we acquire reliable unobserved response candidates set C_U, the size of candidates is still large, which causes the inefficiency problem and it is impractical to traverse over the whole data to obtain the gradients. In that, we design the Expected Ability Change-weighted Informative Sampling strategy (EIS) to speed up the convergence with informative samples. This strategy evaluates the informativeness of potential responses from two perspectives: 1) contribution to model training, 2) student ability change.

1) Contribution to Model Training. In partial order ranking, we need to the sample potential exercises e_u based on their responses from the candidates set C_U to pair with observed responses to exercise e_o, while it is highly possible to sample low-quality instances. Since these responses already have a large gap with observed response, sampling them as potential responses hardly changes model parameters. Here, we want to select informative instances which will significantly change model parameters through partial order ranking. Following previous work [Rendle and Freudenthaler, 2014; Lian *et al.*, 2020], we measure a sample's contribution to ranking task by the gradient magnitude based on the objective function of pairwise ranking (Eq.4):

$$\Delta_{s,e_o,e_u} = 1 - \sigma(\hat{y}_{so} - \hat{y}_{su}), e_o \in E_O, e_u \in C_U, \tag{8}$$

which indicates that a response difficultly distinguished from the opposite response (i.e. $\hat{y}_{so} - \hat{y}_{su} \to 0$) contributes much to gradient (i.e. $\Delta_{s,e_o,e_u} \to 1$). For example, if the current observed response is correct, incorrect samples with higher prediction scores make a greater contribution to the optimization. Then we define candidates' difficulty level by the gap of their prediction scores in CDMs:

$$INF_u = |\hat{y}_{so} - \hat{y}_{su}|. \tag{9}$$

A smaller INF_u for e_u indicates a higher difficulty for CDMs to identify it from the known response to e_o. During training, we can reserve the top *k* informative samples by their INF values to faster training process.

2) Expected Ability Change. Beside significant contribution to partial ranking, we also aim to diagnose students' knowledge states as soon as possible. Therefore, we hypothesize that if student feature H_s undergoes a significant change by adding the unobserved response, the responded exercise can be considered informative.

However, it is challenging to calculate because the response to non-interactive exercise is unknown. Inspired by recent works [Bi *et al.*, 2020], we calculate each sample's importance by expected ability change (EAC). To formulate this, let ΔH_{su} be the ability change of the target student s, $H_s(R_o)$ denote the student's current ability with observed response R_o and $H_s(R_o \cup r_{su})$ represent adding the unobserved response r_{su}. As such, the EAC weight is defined as follows:

$$\begin{aligned} \Delta H_{se_u} &= E_{r_{se_u} \sim f_C(h_s, h_{e_u})} |H_s(R_o \cup r_{se_u}) - H_s(R_o)|, \\ w_u &= \frac{exp(\Delta H_{se_u})}{\sum\limits_u exp(\Delta H_{se_u})}. \end{aligned} \tag{10}$$

By considering the weight w_u of response sample e_u, we reformulate the ranking loss in the following way:

$$\mathcal{L}_{\mathcal{R}} = -\sum_{s,e_o^+,e_u^-} w_u \cdot ln\sigma(y_{so}^+ - y_{su}^-) - \sum_{s,e_o^-,e_u^+} w_u \cdot ln\sigma(y_{su}^+ - y_{so}^-), \tag{11}$$

which enables our framework to pay more attention to instances that bring about a greater change to student ability.

4.4 Learning Algorithm

The observed responses can help optimize CDMs' parameters by the objective of the basic cognitive model. In addition, we propose a response sampling strategy to sample reliable and informative unobserved responses. Thus, the partial order training signals can alleviate the long tail problem. Combining the prediction loss $\mathcal{L}_{\mathcal{P}}$ of the *Basic Cognitive Diagnosis Model* and the ranking loss $\mathcal{L}_{\mathcal{R}}$ of the *Partial-Order Response Sampling Strategy*, we obtain the complete loss function:

$$\mathcal{L} = \mathcal{L}_{\mathcal{P}} + \lambda \cdot \mathcal{L}_{\mathcal{R}}, \tag{12}$$

Algorithm 1 Exercise-aware Informative Response Sampling

```
Input: Training Set $R_O = \{(s, e_i, r_{ui})\}$, Q-matrix
Output: CDMs Parameters $\Theta_C$
Initialize parameters randomly;
while not converge do
    Sample a mini-batch $R_B \in R_O$ of size B.
    for each observed response logs $(s, e, r_e)$ do
        Get candidates set $C_U$ for student $s$ based on Eq.6
        Select top k instances from $C_U$ based on Eq.9
        Evaluate instances' quality $w_u$ based on Eq.10
        Update parameters $\Theta_C$ w.r.t. Eq.12
    end for
end while
```

Metrics		ASSISTments						Junyi					
		Long-tailed data			Whole data			Long-tailed data			Whole data		
		AUC% ↑	ACC%↑	RMSE% ↓	AUC% ↑	ACC% ↑	RMSE%↓	AUC% ↑	ACC% ↑	RMSE%↓	AUC%↑	ACC%↑	RMSE% ↓
IRT	Origin	70.88	68.07	46.34	72.84	70.78	44.34	80.15	79.53	38.14	80.41	80.14	37.62
	PopRS	73.62	70.07	45.33	75.07	71.97	43.60	80.71	79.58	38.13	80.95	80.13	37.61
	EIRS	**75.60**	**69.34**	**44.41**	**76.16**	**72.04**	**42.77**	**81.74**	**79.87**	**37.54**	**81.67**	**80.44**	**37.10**
MF	Origin	72.41	68.79	45.24	72.95	70.62	43.97	79.68	79.10	38.70	80.06	79.81	38.06
	PopRS	75.93	71.64	44.18	75.66	72.60	**43.17**	80.61	79.07	38.14	80.65	79.66	37.71
	EIRS	**76.32**	**72.14**	**44.20**	**76.25**	**73.19**	43.21	**82.33**	**80.27**	**37.46**	**82.38**	**80.76**	**36.99**
MIRT	Origin	71.22	68.05	48.08	72.96	70.53	45.55	79.28	79.67	37.90	79.52	80.18	37.51
	PopRS	75.25	71.82	43.95	74.11	69.47	45.63	80.76	79.44	38.23	80.87	80.07	37.73
	EIRS	**77.03**	**72.33**	**43.03**	**76.91**	**73.29**	**42.11**	**82.07**	**80.18**	**37.34**	**82.14**	**80.66**	**36.92**
NeuralCD	Origin	74.95	**71.74**	49.61	75.65	**71.15**	47.52	78.59	78.36	39.67	78.68	**79.55**	39.28
	PopRS	75.16	71.10	45.21	74.43	70.25	44.73	80.14	78.60	38.47	80.44	79.49	38.59
	EIRS	**75.63**	71.57	**44.76**	**75.36**	71.03	**44.12**	**80.85**	**79.23**	**38.30**	**80.60**	79.43	**38.02**

Table 2: Experimental results on student performance prediction.

where λ is a trade-off hyper-parameter that balances the two losses. The learning process is shown in Algorithm 1.

Since the model is easier to learn the original signals provided by observed data, while auxiliary ranking signal is relatively difficult to learn although they contain rich information. Inspired by curriculum learning [Bengio *et al.*, 2009], which allows the training process to start with simpler tasks and gradually increase the difficulty, we propose to linearly increase λ as the epoch number t increases. In this fashion, our framework is more stable and effective for diagnosing students' knowledge proficiency.

5 Experiments

In this section, we first introduce the datasets and our experimental setups. Then, we conduct extensive experiments to answer the following questions:

RQ1: Can EIRS improve performance of existing CDMs?

RQ2: Can EIRS perform well on long-tailed students with few interactions?

RQ3: Can EIRS accelerate the convergence of CDMs?

RQ4: How does the number of samples influence the performance of the EIRS framework?

5.1 Experimental Setup

Dataset Description. We conduct experiments on two real-world datasets, i.e., ASSISTments[1] and Junyi dataset [2]. ASSISTments (ASSISTments 2009-2010 "skill builder") is an open dataset collected by the ASSISTments online tutoring systems and Junyi is collected from the E-learning website Junyi Academy. Most proposed works validate model's performance on datasets by filtering students with few interactions to ensure enough logs to accomplish diagnosis tasks. We keep all response logs in both datasets, excluding students with interaction times below 5. The number 5 is to ensure that the dataset can be split into train and test sets at an 8:2 ratio.

Baselines. To evaluate the performance of our EIRS framework, we use four well-known CDMs as baseline methods: IRT [Lord, 1980], MF [Toscher and Jahrer, 2010], MIRT [Reckase, 2009] and NeuralCD [Wang *et al.*, 2020]. In multidimensional models (i.e., MIRT and NeuralCD), we set the dimension of latent trait features of both student and item unitedly as the number of knowledge concepts, i.e., 112 in ASSISTments and 39 in Junyi. Furthermore, we compare popularity-based sampling (PopRS) method [Mikolov *et al.*, 2013] with EIRS which calculates each exercise's popularity based on the response rate in all students.

Experimental Setup. In our framework, we set the sample number from [1,2,3,4,5]. For the curriculum coefficient λ, its initial value λ_0 is chosen from the interval (0, 1], and λ linearly increases from λ_0 to 1 as the number of epochs increases. We employ the Adam algorithm [Kingma and Ba, 2015] for optimization, and all the hyper-parameters are tuned in the validation datasets. Our code is available at https://github.com/fannazya/EIRS.

Evaluation Metrics. Because the true knowledge proficiency is unknown, to directly evaluate the performance of a CDM is difficult. Following previous works, a reasonable solution is to measure performance by the prediction scores in diagnosis models as the diagnostic results can be acquired through learners performance prediction task. Here, we evaluate the model based on some classification and regression metrics such as *AUC*, *Accuracy* and *RMSE*.

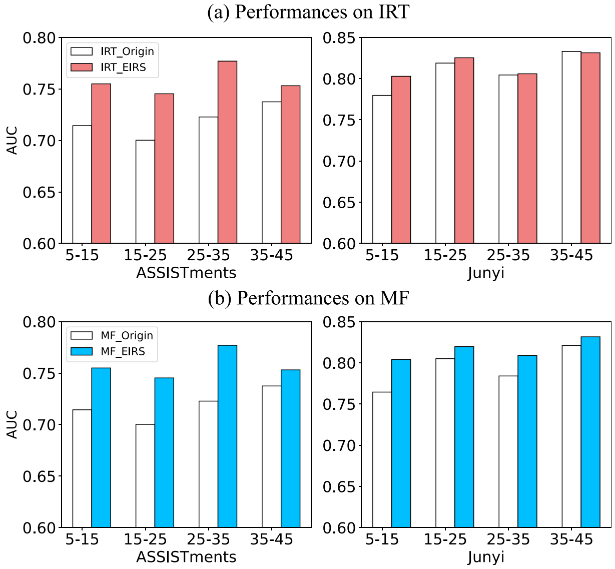
5.2 Experimental Results

Performance Comparison (RQ1)

In order to validate the generality and effectiveness of the EIRS framework, we incorporate it into different existing CDMs and compare the performances both on the long-tailed data and whole data with baseline methods. The long-tail data extracted from test data consists of response logs of students whose interaction times are less than 50. Table 2 shows the results of EIRS with baselines, where 'Origin', 'EIRS' represent the baseline and baseline incorporating our framework respectively. The best results are shown in bold. There are the following findings. First, almost all the baselines'

[1] https://sites.google.com/site/assistmentsdata/home/2009-2010-assistment-data

[2] https://pslcdatashop.web.cmu.edu/Files?datasetId=1198



(a) Performances on IRT

(b) Performances on MF

Figure 4: Performance comparison on long-tailed data. The horizontal axis is interval of students interaction times.

results on the long-tailed data are inferior to those on the whole data, which indicates it is difficult to diagnose with few interaction logs. Second, both PopRS and EIRS perform better than the Origin baseline, showing that incorporating ranking-based training signals alleviates the long-tailed problem. Specifically, EIRS significantly outperforms both the original model and popularity sampling method on the long-tailed data, and has a fair contribution to improving performance on the whole data. Some results in the long-tailed data are even as good as those in the whole data. Hence, we are able to conclude from these observations that our framework is effective to alleviate the long-tail problem by adding auxiliary ranking-based training signals. Simultaneously, the non-interactive exercises selected by EIRS are reliable enough to provide proper training signals.

Performance on Long-Tailed Data (RQ2)

Our framework focuses on strengthening the robustness of CDMs, especially in long-tailed data where students interact with few exercises. We further divide them by interactions times into several groups to see the detailed improvements in different groups. The results are reported in Figure 4. We can see that the baseline MF and IRT applied with our framework was improved a lot, especially in students whose interaction times are very few (from 5 to 15). Therefore, we can conclude that EIRS's contribution on diagnosing students knowledge level with limited response logs is prominent.

Efficiency of Informative Sampling (RQ3)

The above experiments show the effectiveness of samples generated by EIRS. Additionally, we compare the performance in terms of efficiency by the convergence speed of model training (the number at which an early stop occurs). We can observe that there is not much difference in convergence speed at the beginning. The reason for this is that we take a curriculum learning way (Section 4.4), where basic cognitive diagnosis is the dominant task at the beginning of training. As the epoch number increases, the test loss converges quickly when partial order ranking plays a vital role. The fast convergence speed indicates that the efficacy of EIS module to sample informative responses that bring much change to model learning.

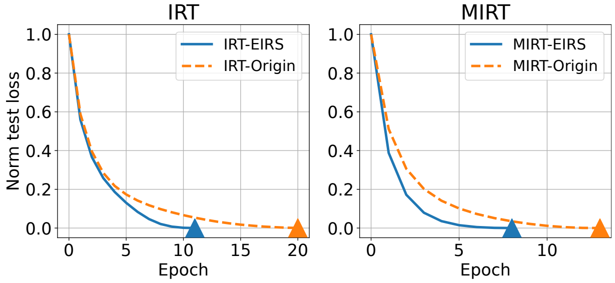


Figure 5: Convergence against training epoch on Junyi. The triangle marker represents the epoch number at which early stopping occurs.

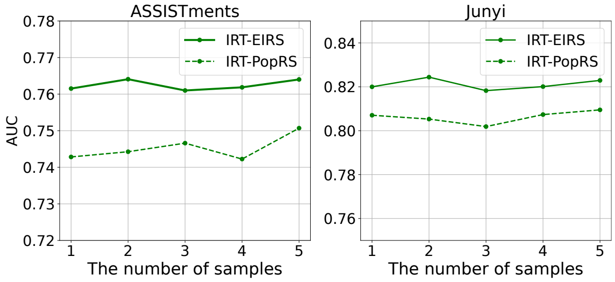


Figure 6: Effects of the number of samples.

Effects of Sampling Number (RQ4)

The key component in EIRS is the two-stage sampling consisting of ECS and EIS modules (Section 4.3) to generate reliable and informative exercises. To verify our sampling strategy's superiority, we compare it with popularity-based sampling (PopRS). We vary the number of samples K from set $\{1, 2, 3, 4, 5\}$ to see the effects on the whole data. As shown in Figure 6, EIRS achieves high performance when the sample number is small and consistently outperforms PopRS at different sample numbers. Besides, our framework basically holds steady though the number changes. These give credit to the ECS module which generates reliable candidate logs through knowledge relations and exercises similarity, effectively avoiding data noise in unobserved data.

6 Conclusion

In this paper, we designed a general sample framework EIRS that exploits reliable and informative non-interactive data and can be seamlessly incorporated to existing cognitive diagnosis methods. Then we did experiments on two real datasets to validate the performance. The results showed the essence of non-interactive data and the superiority of our proposed framework on long-tailed data. We hope this work can stimulate more studies in the future leading to a prolonged period.

Acknowledgements

This research was partially supported by grants from the National Key Research and Development Program of China (Grant No. 2021YFF0901003), the National Natural Science Foundation of China (Grants No. 62106244), the Students' Innovation and Entrepreneurship Foundation of USTC (No. XY2022X12), the Open Research Fund of the State Key Laboratory of Cognitive Intelligence (iED2022-002), and the Iflytek joint research.

References

[Bengio *et al.*, 2009] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009.

[Bi *et al.*, 2020] Haoyang Bi, Haiping Ma, Zhenya Huang, Yu Yin, Qi Liu, Enhong Chen, Yu Su, and Shijin Wang. Quality meets diversity: A model-agnostic framework for computerized adaptive testing. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 42–51. IEEE, 2020.

[Caselles-Dupré *et al.*, 2018] Hugo Caselles-Dupré, Florian Lesaint, and Jimena Royo-Letelier. Word2vec applied to recommendation: Hyperparameters matter. *CoRR*, abs/1804.04212, 2018.

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

[Gao *et al.*, 2021] Weibo Gao, Qi Liu, Zhenya Huang, Yu Yin, Haoyang Bi, Mu-Chun Wang, Jianhui Ma, Shijin Wang, and Yu Su. 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*, pages 501–510, 2021.

[Goodfellow *et al.*, 2014] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2672–2680, 2014.

[Kingma and Ba, 2015] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[Lian *et al.*, 2020] Defu Lian, Qi Liu, and Enhong Chen. Personalized ranking with importance sampling. In *Proceedings of The Web Conference 2020*, pages 1093–1103, 2020.

[Liu *et al.*, 2021] Qi Liu, Zhenya Huang, Yu Yin, Enhong Chen, Hui Xiong, Yu Su, and Guoping Hu. Ekt: Exercise-aware knowledge tracing for student performance prediction. *IEEE Transactions on Knowledge and Data Engineering*, 33(1):100–115, 2021.

[Liu, 2021] Qi Liu. Towards a new generation of cognitive diagnosis. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 4961–4964. ijcai.org, 2021.

[Lord, 1980] Frederic M Lord. *Applications of item response theory to practical testing problems*. Routledge, 1980.

[Lu *et al.*, 2022] Yu Lu, Penghe Chen, Yang Pian, and Vincent W. Zheng. Cmkt: Concept map driven knowledge tracing. *IEEE Transactions on Learning Technologies*, 15(4):467–480, 2022.

[Ma *et al.*, 2022] Haiping Ma, Jingyuan Wang, Hengshu Zhu, Xin Xia, Haifeng Zhang, Xingyi Zhang, and Lei Zhang. Reconciling cognitive modeling with knowledge forgetting: A continuous time-aware neural network approach. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 2174–2181. ijcai.org, 2022.

[Mikolov *et al.*, 2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

[Pandey and Karypis, 2019] Shalini Pandey and George Karypis. A self-attentive model for knowledge tracing. *International Educational Data Mining Society*, 2019.

[Park and Chang, 2019] Dae Hoon Park and Yi Chang. Adversarial sampling and training for semi-supervised information retrieval. In *The World Wide Web Conference*, pages 1443–1453, 2019.

[Qin *et al.*, 2019] Chuan Qin, Hengshu Zhu, Chen Zhu, Tong Xu, Fuzhen Zhuang, Chao Ma, Jingshuai Zhang, and Hui Xiong. Duerquiz: A personalized question recommender system for intelligent job interview. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2165–2173, 2019.

[Qin *et al.*, 2020] Chuan Qin, Hengshu Zhu, Tong Xu, Chen Zhu, Chao Ma, Enhong Chen, and Hui Xiong. An enhanced neural network approach to person-job fit in talent recruitment. *ACM Transactions on Information Systems (TOIS)*, 38(2):1–33, 2020.

[Reckase, 2009] Mark D Reckase. Multidimensional item response theory models. In *Multidimensional item response theory*, pages 79–112. Springer, 2009.

[Rendle and Freudenthaler, 2014] Steffen Rendle and Christoph Freudenthaler. Improving pairwise learning for item recommendation from implicit feedback. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 273–282, 2014.

[Rendle *et al.*, 2009] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. BPR: bayesian personalized ranking from implicit feedback. In

UAI 2009, Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, Montreal, QC, Canada, June 18-21, 2009, pages 452–461. AUAI Press, 2009.

[Rendle *et al.*, 2012] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618*, 2012.

[Rosenbaum, 1984] Paul R Rosenbaum. Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika*, 49(3):425–435, 1984.

[Tatsuoka, 1995] Kikumi K Tatsuoka. Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. *Cognitively diagnostic assessment*, pages 327–359, 1995.

[Tong *et al.*, 2021] Shiwei Tong, Qi Liu, Runlong Yu, Wei Huang, Zhenya Huang, Zachary A Pardos, and Weijie Jiang. Item response ranking for cognitive diagnosis. In *IJCAI*, pages 1750–1756, 2021.

[Toscher and Jahrer, 2010] Andreas Toscher and Michael Jahrer. Collaborative filtering applied to educational data mining. *KDD cup*, 2010.

[Wang *et al.*, 2018] J Wang, L Yu, W Zhang, Y Gong, Y Xu, B Wang, P Zhang, and D Zhang. A minimax game for unifying generative and discriminative information retrieval models. *SIGIR Proc*, 2018.

[Wang *et al.*, 2020] Fei Wang, Qi Liu, Enhong Chen, Zhenya Huang, Yuying Chen, Yu Yin, Zai Huang, and Shijin Wang. Neural cognitive diagnosis for intelligent education systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6153–6161, 2020.

[Wu *et al.*, 2020] Mike Wu, Richard L Davis, Benjamin W Domingue, Chris Piech, and Noah Goodman. Variational item response theory: Fast, accurate, and expressive. *arXiv preprint arXiv:2002.00276*, 2020.

[Zhang *et al.*, 2013] Weinan Zhang, Tianqi Chen, Jun Wang, and Yong Yu. Optimizing top-n collaborative filtering via dynamic negative item sampling. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 785–788, 2013.

[Zhou *et al.*, 2021] Yuqiang Zhou, Qi Liu, Jinze Wu, Fei Wang, Zhenya Huang, Wei Tong, Hui Xiong, Enhong Chen, and Jianhui Ma. Modeling context-aware features for cognitive diagnosis in student learning. In *KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021*, pages 2420–2428. ACM, 2021.