AD4CD: Causal-Guided Anomaly Detection for Enhancing Cognitive Diagnosis

Haiping Ma[1,2*], Yue Yao[1], Changqian Wang[1], Siyu Song[1], Yong Yang[1]

[1]Institutes of Physical Science and Information Technology, Anhui University, China
[2]Department of Information Materials and Intelligent Sensing Laboratory of Anhui Province, China
hpma@ahu.edu.cn, {yuntu47, changqian.wang.dl, siyusong00, yyyangyong00}@gmail.com

Abstract

Cognitive diagnosis aimed at assessing a students' proficiency in specific knowledge concepts based on their responses to exercises. However, existing cognitive diagnosis models often overlook anomalies in students and exercises. For instance, some students might incorrectly response exercises despite having a strong grasp of the knowledge concept, or they might response correctly despite a lack of understanding. Such subtle anomalies can adversely affect the diagnostic results of the models. To address these anomalies, we conduct a qualitative analysis of how anomalous student states and exercise properties impact response outcomes using causal diagrams. We propose a framework named Anomaly Detection for Cognitive Diagnosis (AD4CD) to enhance the ability of Learning-to-Detect-Anomalous. AD4CD approaches the problem from a causal perspective, analyzing confounding paths that affect the true causal relationship between student ability and response outcomes, and designing an anomaly detection mechanism suitable for cognitive diagnostic models. Specifically, we first account for anomalous student behaviors and exercise properties and introduce response times from both students and exercises as modeling factors. By quantifying the response time distributions in high-dimensional features, we identify anomalies within skewed distributions, including both left-tail and right-tail anomalies. Using the detected anomaly scores, we comprehensively model the students' anomalous behaviors and exercise anomalies. Additionally, we reconstruct unbiased true abilities under natural conditions and use reconstruction loss as an anomaly score to assist in modeling guessing and slipping features. Lastly, AD4CD leverages a general cognitive diagnosis model as its backbone, optimizing the guessing and slipping features to provide unbiased and accurate feedback. Extensive experimental results demonstrate that AD4CD effectively captures anomalous data in the diagnostic process across three real-world datasets, enhancing the accuracy of the diagnostic results.

Introduction

Amid the development of intelligent education, cognitive diagnosis has garnered increasing attention (Zhao et al. 2023; Wu et al. 2023; Liu et al. 2023b; Yang et al. 2023, 2019). Cognitive diagnosis aims to assess students' proficiency in various knowledge concepts through response on exercises (Ma et al. 2024e; Yang et al. 2024b; Yu et al. 2024a,b). cognitive diagnosis models (CDMs) utilize students' response records and the Q-matrix, which represents the relationship between exercises and knowledge concepts, to model both students' abilities and the property of the exercises. This approach enables diagnostic feedback for students and provides an analysis of exercise characteristics, such as difficulty (Ma et al. 2024c; Yu et al. 2024c). Existing research (Yang, Qin, and Yu 2024; Shen et al. 2024) has focused on designing various neural network structures to capturethe naive binary interactions between students and exercises, such as DNN-based CDMs (Wang et al. 2022; Ma et al. 2022) represented by NCD (Wang et al. 2020) and graph-based CDMs (Wang et al. 2023; Ma et al. 2024a) represented by RCD (Gao et al. 2021). These studies uniformly treat historical response records as normal interaction behavior, which is considered to contribute equitably to model learning. However, in certain situations, the diagnostic results from CDMs indicate that the student has a deep understanding of one knowledge concept but a shallow understanding of another, with no correlation between them. Nonetheless, in practical assessments, the student incorrectly answered a question linked to the former concept but correctly responded to one related to the latter. This unobservable anomalous behavior contradicts the cognition of CDMs, leading to biased diagnostic results, which are caused by the coupling of the two entities, including the student and the exercise.

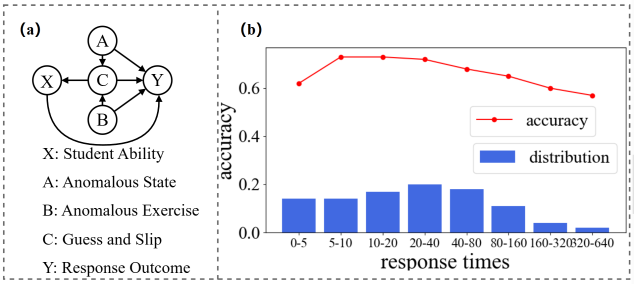

Figure 1: (a) denotes the cause-effect diagram to analyze anomalies. (b) The relationship between students' response times distribution and accuracy rate on the ASSIST09 dataset.

*Corresponding Authors.

To better understand the causes of this phenomenon, we proposed a formalized definition for qualitative analysis using a causal graph, as illustrated in Figure 1 (a). The nodes represent causal factors, and an edge $A \rightarrow B$ indicates that A directly affects B.

- Firstly, $\{X, A, B\} \rightarrow Y$ denotes that a student's inherent ability X, his anomalous state A and the anomalous properties of the exercise B all directly affect the response outcome Y. Here, A represents situations like stress or lack of concentration, while B includes errors in associating exercises with knowledge concepts, which are manually labeled (Liu et al. 2023a), or ambiguities inherent in the exercises.
- Secondly, $A \rightarrow C \rightarrow X$ suggests that a student's anomalous state increases the likelihood of guessing or slipping, thereby affecting their true ability under natural conditions. For instance, when a student is under noticeable stress, the probability of guessing increases. Similarly, when a student feels fatigued or distracted, the likelihood of careless mistakes significantly rises, leading to underestimated diagnostic feedback.
- Furthermore, $B \rightarrow C \rightarrow X$ implies that the anomalous properties of the exercise B elevate the risk of guessing or slipping. Specifically, poorly constructed exercises may compel the student to guess, while ambiguous exercises might result in careless errors.
- Lastly, $C \rightarrow Y$ indicates that even under natural conditions, there remains a non-zero probability of guessing or slipping, which can still affect the final outcome to some extent (De La Torre 2009).

We define the task of detecting anomalous behaviors within CDMs as Learning to Detect Anomalous (LDA). The challenge of the LDA task lies in the fact that anomalous behaviors are both difficult to observe and hard to quantify. This leads us to pose a counterfactual question: has a student's response been influenced by anomalous behavior? To explore this question and block paths that disrupt the true causal relationship, we re-examined the contextual information during the student's response process, providing a foundation for interpretable diagnostic feedback. As shown in Figure 1 (b), we visualized the response times of students in the ASSIST2009 dataset, which reflects the response process. The results indicate that both excessively long and short response times can affect the final accuracy. However, it is not feasible to judge the presence of anomalous interactions by setting a threshold for learning duration. This implies that each exercise has its own reasonable response time distribution, and similarly, each student has their own preferred response time. For example, students with time anxiety may answer exercises quickly. Using such method could result in most interactions with difficult or easy exercises being considered anomalous. Similarly, most interactions by students who prefer quick responses or careful deliberation would be viewed as anomalous. However, no prior work has explored the quantification of anomalous behaviors during the response process.

To address the aforementioned challenges, we designed a diagnostic framework aimed at enhancing the ability of LDA, named **A**nomaly **D**etection for **C**ognitive **D**iagnosis Framework (**AD4CD**). From a causal inference perspective, we analyzed the backdoor paths that obstruct the link between students' true abilities and their response outcomes. To satisfy the backdoor criterion, we developed an anomaly detection mechanism specifically tailored for cognitive diagnostic models. Specifically, we introduced both student response times and exercise response times. By quantifying the distribution of high-dimensional features from various perspectives, we identify biased distributions, including left-tail and right-tail anomalies. Guided by the detected anomaly scores, we integrated high-dimensional features to comprehensively model the anomalous student behavior and exercise properties. Moreover, to account for the non-zero probabilities of guessing or slipping under natural conditions, we reconstructed the unbiased true abilities by basing our model on interpretable diagnostic results. The reconstruction loss serves as an anomaly score, assisting in the modeling of guessing and slipping features. Finally, AD4CD uses general CDMs as its backbone, with the modeled guessing and slipping behaviors guiding the diagnosis of unbiased and authentic response feedback. Extensive experimental results on three real-world datasets demonstrate that AD4CD effectively captures anomalous data in the diagnostic process.

Code — https://github.com/BIMK/Intelligent-Education/

Related Work

Cognitive diagnosis

Over the past few decades, cognitive diagnosis tasks in educational psychology have achieved great success. The early IRT model (Lord 2012) and the DINA model (De La Torre 2009) are two classic cognitive diagnosis models. In IRT, students' ability states are represented as one-dimensional continuous scalar values, and a logistic function is used to predict the probability of students making a correct response to an exercise. The MIRT model (Chalmers 2012) extends students' ability states to multidimensional vectors. The DINA model uses binary vectors to represent students' mastery of various knowledge concepts. In recent years, considerable neural network-based approaches have emerged for improving the cognitive diagnosis precision (Yang et al. 2024a; Ma et al. 2024b). For example, some researches such as NeuralCD (Wang et al. 2020) and KaNCD (Wang et al. 2022) intended to utilize neural networks to capture the complex relationships between students and exercises. Meanwhile another researches such as ECD (Zhou et al. 2021), RCD (Gao et al. 2021), and HierCDF (Li et al. 2022a) utilize neural networks to learn more valuable information from contextual features, multi-layer relation graph, hierarchical structure of knowledge concepts, and so on (Ma et al. 2024d; Wang et al. 2024; Zhang et al. 2024). However, existing methods only focus on the historical interactions between students and exercises, neglecting the crucial factor of students' problem-solving states

Anomaly Detection

Anomaly detection, also known as outlier detection, is a key technology in machine learning with broad application prospects. It is widely used in fields such as anti-money laundering (Lee et al. 2020), rare disease detection (Zhao et al. 2021), and intrusion detection (Lazarevic et al. 2003). Depending on the type of data, anomaly detection can be divided into anomaly detection on tabular data, anomaly detection on sequential data, and anomaly detection on graph data. This paper primarily focuses on unsupervised anomaly detection technologies on tabular data, which mainly include distance-based approaches (Ramaswamy, Rastogi, and Shim 2000; Breunig et al. 2000), clustering-based approaches(Ester et al. 1996), tree-based approaches(Liu, Ting, and Zhou 2008), and neural network based approaches(Zong et al. 2018; Wang et al. 2019). For example, as the representatives of distance-based approaches, KNN (Ramaswamy, Rastogi, and Shim 2000) calculates the distance to its nearest k neighbors and judges whether a point is an outlier based on the size of this distance. Similar to KNN, LOF (Breunig et al. 2000) also performs measurements, but instead of using distance, it evaluates the local density deviation relative to its neighbors. As the representatives of tree-based approaches, iForest (Liu, Ting, and Zhou 2008), isolates data points by randomly cutting the data. Normal data points are difficult to isolate, while anomalous points are more easily isolated. However, these anomaly detection algorithms cannot be directly applied to the field of cognitive diagnosis.

Preliminary

In this section, we formally define the cognitive diagnosis task. We define three entity sets: the student set $S = (s_1, s_2, \ldots, s_N)$ of size N; The exercise set $E = (e_1, e_2, \ldots, e_M)$ of size M; the knowledge concept set $K = (k_1, k_2, \ldots, k_C)$ of size C. The exercise-concept matrix is defined as $Q \in R^{M \times C}$. If $Q_{i,j} = 1$, then exercise e_i contains knowledge concept k_j; otherwise, $Q_{i,j} = 0$. We also define the interaction record as a quadruple $(s_i, e_j, r_{i,j}, t_{i,j}) \in R$, where e_j is the exercise interacted with by student s_i, $r_{i,j} = 1$ indicates that student s_i answered exercise e_j correctly, and $r_{i,j} = 0$ otherwise. $t_{i,j}$ represents the time taken by student s_i to answer exercise e_j. R represents the entire set of interactions. So, we can formalize the research problem in this paper as follows:

Problem Definition: *Given the students' response logs R and the expert-defined Q matrix, our goal is to diagnose students' proficiency after blocking confounding factors.*

Method

Causal Inspired Analysis

As illustrated in Figure 2(a), we present the causal graph of the LDA task. It reveals that introducing certain confounding paths leads to uncontrollable bias and weak interpretability. To explore the causal effect between X and Y, while considering the confounding variables $\{A, B, C\}$, we first analyze all paths between X and Y. These include four paths: (1) $X \rightarrow Y$, (2) $X \leftarrow C \leftarrow A \rightarrow Y$, (3) $X \leftarrow C \leftarrow B \rightarrow Y$, and (4) $X \leftarrow C \rightarrow Y$. We utilize the backdoor criterion to condition on these paths. The backdoor criterion requires that for a pair of ordered variables (P, Q) and a variable set Z, the following conditions must be satisfied: (1)Z contains no descendants of P, and (2)Z blocks all backdoor paths between P and Q. Accordingly, we identify the backdoor paths as those represented by (2), (3), and (4). Thus, our goal is to block these confounding paths in order to discover the true causal relationship between X and Y.

Anomaly Detection

To block the confounding paths $X \leftarrow C \leftarrow A \rightarrow Y$ and $X \leftarrow C \leftarrow B \rightarrow Y$, we conditionally adjust for the sets of confounding variables $\{A, C\}$ and $\{B, C\}$, respectively. The adjustment formula based on the backdoor criterion is:

$$P(Y \mid do(X)) = \sum_{\{A,C\} \cup \{B,C\}} P(Y \mid X, A, C) P(A, C) + P(Y \mid X, B, C) P(B, C), \tag{1}$$

where $P(Y \mid X, \cdot)$ represents the influence of X on Y after controlling for confounders, and $P(\cdot)$ denotes the joint probability distribution of the confounders. To condition on the confounders, we will separately detect abnormal student states and anomalous exercise properties, which correspond to blocking the first and second confounding paths.

Since abnormal data exist within the interaction data, and students' response times can serve as a basis for anomaly detection, we first transform the response times in the response log into response vectors to quantify the detection. For student i, let $x_i^s = (x_{i1}, x_{i2}, \ldots, x_{iM})^T$ represent his response vector. Similarly, for exercise j, let $x_j^e = (x_{1j}, x_{2j}, \ldots, x_{Nj})^T$ denote its response vector. Given the set of response times $T = \{t_{ij} \mid (s_i, e_j, r_{i,j}, t_{i,j}) \in R\}$ in the response log, x_{ij} is defined as follows:

$$x_{ij} = \begin{cases} t_{ij}, & \text{if } (s_i, e_j, r_{ij}) \in R, \\ -1, & \text{else .} \end{cases} \tag{2}$$

Due to the data collection mechanisms of the platform, the response times include both discrete and continuous values. For the former, we map the features using a parameter matrix E, while for the latter, we apply an affine transformation to project them into the same dimensional space:

$$x_{ij}^{enc} \in \mathbb{R}^{1 \times d} = \begin{cases} W_j * x_{ij} + b_j, & t_{ij} \text{ is continuous,} \\ \text{lookup}(E_j, x_{ij}), & \text{else,} \end{cases} \tag{3}$$

where W and b represent learnable parameters, and d denotes the feature dimension. By using this approach, we obtain encoded representations for the student and exercise, denoted as $x_i^{enc(s)} \in \mathbb{R}^{1 \times d_{s_i} \times d}$ and $x_j^{enc(e)} \in \mathbb{R}^{1 \times d_{s_e} \times d}$, respectively. It is important to note that this encoding is only applied when $x_{ij} \neq -1$, so d_{s_i} and d_{s_e} represent the actual number of questions the student responded to and the number of responses the exercise received, respectively.

Next, we propose an anomaly scoring mechanism to capture potential anomalous signals from the high-dimensional

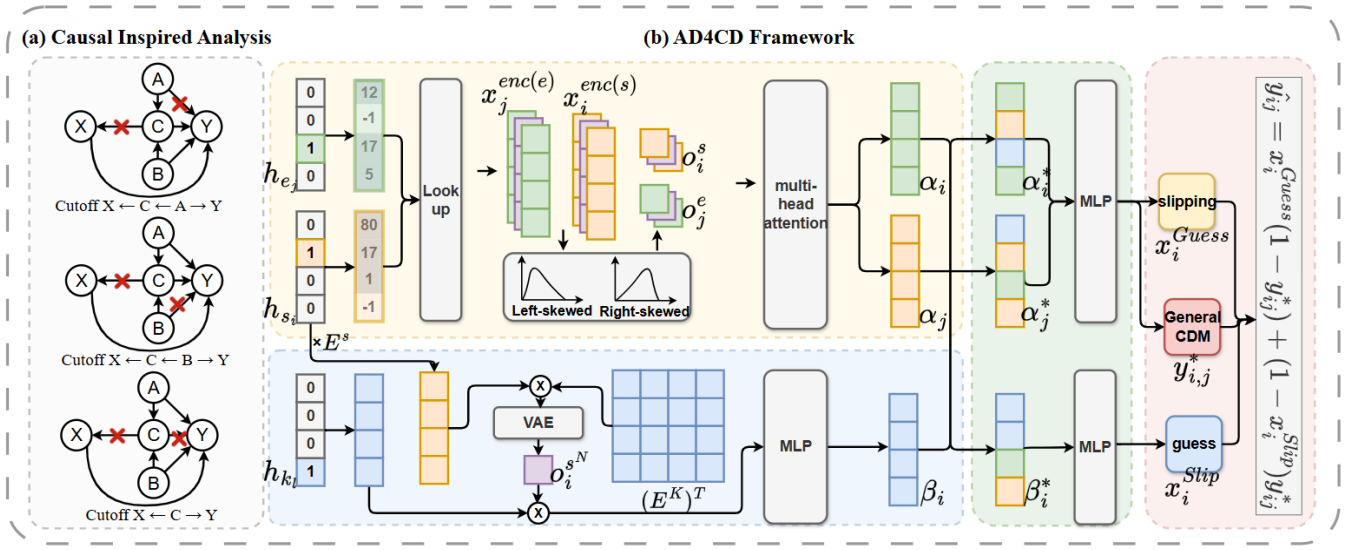

Figure 2: (a) the causal inspired analysis, and (b) the schematic diagram of AD4CD framework

representations. The anomaly score leverages the multi-dimensional relationships between features to quantify deviations from expected behavior within the feature space. For each encoded high-dimensional data point x_{ij}^{enc}, we define its anomaly score o_{ij} based on the point's position relative to the overall distribution in the embedding space. Instead of relying on predefined cumulative distribution functions, we evaluate the behavior of data points across multiple dimensions to determine whether they exhibit anomalous characteristics. Specifically, we assume that the high-dimensional representations of both students and exercises are independently and identically sampled from a distribution with a empirical cumulative distribution function (ECDF) (Li et al. 2022b) $\hat{F} : \mathbb{R}^d \to [0, 1]$. For any $a \in \mathbb{R}^d$, we define:

$$\hat{F}_{left}^{(h)}(a) := \frac{1}{n}\sum_{i,j=1}^{n}\left\{x_{ij}^{enc^{(h)}} \le a\right\}$$
$$\text{and } \hat{F}_{right}^{(h)}(a) := \frac{1}{n}\sum_{i,j=1}^{n}\left\{x_{ij}^{enc^{(h)}} \ge a\right\}. \tag{4}$$

Here, $x_{ij}^{enc^{(h)}}$ denotes the h-th feature of the multidimensional data points. This function provides a measure of extremeness by evaluating how small $\hat{F}(x_{ij}^{enc(s)})$ and $\hat{F}(x_{ij}^{enc(e)})$ are across all dimensions. If the results are small, this suggests that the data point lies in the far left tail of the joint distribution, indicating that the point may be an outlier. Therefore, we estimate the left and right tails of the CDF across all d dimensions under the independence assumption:

$$\hat{F}_{left}(x_{ij}^{enc}) := \prod_{h=1}^{d}\hat{F}_{left}(x_{ij}^{enc^{(h)}})$$
$$\text{and } \hat{F}_{right}(x_{ij}^{enc}) := \prod_{h=1}^{d}\hat{F}_{right}(x_{ij}^{enc^{(h)}}). \tag{5}$$

The sample skewness sparsity γ_h for h-th dimension is calculated as follows:

$$\gamma_h = \frac{\mathbb{E}_{i=1,\dots,n}(x_{nj}^{enc^{(h)}} - \mathbb{E}_{i=1,\dots,n}x_{nj}^{(h)})^3}{\left[\frac{1}{n-1}\mathbb{E}_{i=1,\dots,n}(x_{nj}^{enc^{(h)}} - \mathbb{E}_{i=1,\dots,n}x_{nj}^{(h)})^2\right]^{\frac{3}{2}}}. \tag{6}$$

When $\gamma_h < 0$, it indicates left-skewed distribution, and vice versa. To adaptively learn the anomaly scores, we integrate the above steps to obtain the complete computation:

$$o(x_{ij}^{enc}) = -\sum_{h=1}^{d}\Big[\{\gamma_h < 0\}\log(\hat{F}_{left}^{h}(x_{ij}^{enc^h})) + \{\gamma_h \ge 0\}\log(\hat{F}_{right}^{h}(x_{ij}^{enc^h}))\Big]. \tag{7}$$

Through this mechanism, we calculate anomaly scores for each student and exercise, denoted as $o_i^s \in \mathbb{R}^{d_s \times 1}$ and $o_j^e \in \mathbb{R}^{d_e \times 1}$, respectively. Furthermore, we utilize an attention mechanism to model the impact of anomalous states guided by the anomaly scores. Specifically, the anomaly scores serve as the query and key, while the high-dimensional response representations $x_i^{enc(s)}$ and $x_j^{enc(e)}$ act as the value. This process aims to capture potential anomalous signals from the high-dimensional representations. We then obtain the anomaly influence representations from both the student and exercise perspectives:

$$\alpha_i = softmax(\frac{o_i^s \times o_i^{sT}}{\sqrt{d}})x_i^{enc(s)},$$
$$\alpha_j = softmax(\frac{o_j^e \times o_j^{eT}}{\sqrt{d}})x_i^{enc(e)}, \tag{8}$$

where $\alpha_i \in \mathbb{R}^{1\times d_s \times d}$ and $\alpha_j \in \mathbb{R}^{1\times d_e \times d}$ are used to control the anomalous states of students A and the anomalous properties of exercises B, respectively.

Detection in Natural State

Even in the absence of anomalous student states or exercise preperties, under natural conditions, students may still guess or slip with a certain probability (corresponding to the path $X \leftarrow C \rightarrow Y$). To block this confounding path, we aim to explore the natural state of student ability, capturing subtle deviations in their abilities. Given that such deviations directly affect diagnostic results and representation learning, we multiply the natural ability of the student x_i^N with the representations of all knowledge concepts. This approach enables the diagnostic results of different knowledge concepts to be interpreted through intrinsic knowledge relationships. For a given student i and the one-hot vector h_{s_i} we formulate it as:

$$x_i^K = h_{s_i} \times (E^K)^T \tag{9}$$

where $E^K \in \mathbb{R}^{C \times d}$ refers to the learnable embedding matrix for knowledge concepts (with N and C representing the number of students and knowledge concepts, respectively).

Given that the obtained student representations of mastery across all knowledge states may be influenced by noisy data, we propose to model the uncertainty in these inferred abilities based on the observed data. We map the observed ability representations to a latent space, and then capture the unbiased latent ability by learning the probability distribution of the latent variables, which is used to reconstruct the true ability of the students. Our goal is to maximize the evidence lower bound (ELBO) based on the observed x_i^K:

$$\begin{aligned}\mathcal{L}_{ELBO}(\theta, \phi; x_i^K) =& \lambda_1 \mathbb{E}_{q_\phi(z|x_i^K)} \left[\log p_\theta(x_i^K \mid z)\right] \\ &- \lambda_2 \mathrm{KL}(q_\phi(z \mid x_i^K) \| p(z)),\end{aligned} \tag{10}$$

where the first term is the reconstruction term, ensuring that the latent representations can accurately reconstruct unbiased abilities, thereby mitigating uncertainty. The second term encourages the variational posterior $q_\phi(\cdot)$ to approximate the prior distribution $p(\cdot)$, which is a standard normal distribution. We focus specifically on the reconstruction term. First, we model the observed ability representations x_i^K as a distribution in the latent space:

$$z \sim q_\phi(z \mid x_i^K) = \mathcal{N}(z; \mu(x_i^K), \sigma^2(x_i^K)), \tag{11}$$

which captures the student's latent cognitive ability. This latent variable allows to faithfully reconstruct unbiased ability:

$$p_\theta(x_i^K \mid z) \sim \mathcal{N}(x_i^K; f_\theta(z), \mathbf{I}), \tag{12}$$

where $f_\theta(\cdot)$ is a decoder that maps the latent representations back to the observation space. The difference between the reconstructed and observed abilities is considered as a natural state anomaly score, which can be calculated as:

$$o_i^{s^N} = \mathbb{E}_{h=1,\ldots,d}(x_i^{K^{(h)}} - f_\theta(z)^{(h)}), \tag{13}$$

where h denotes the h-th representation feature. To model this natural state anomaly score, we fuse it with the specific knowledge concept corresponding to the current response, denoted as knowledge concept k_l. This fusion is formulated as:

$$\beta_i = f_\varphi(\sigma(o_i^{s^N} \times h_{k_l} \times E^K)), \tag{14}$$

where $f_\varphi(\cdot)$ represents a multi-layer perceptron, h_{k_l} is the one-hot vector of knowledge concept k_l, and $\sigma(\cdot)$ is the activation function.

Data	ASSIST09	ASSIST17	Junyi
#Students	2,954	1,708	115,886
#Exercises	17,707	3162	721
#Knowledge concepts	123	102	39
#Responses logs	331,106	882,075	14,767,179

Table 1: Dataset Statistics

AD4CDM: An Implementation of AD4CD

In this section, we introduce AD4CDM as an implementation of AD4CD to demonstrate its feasibility. Since the three confounding paths mentioned earlier all involve guessing and slipping, represented by the variable C, we conditionally control the three anomaly representations modeled in the first two subsections, conditioned on C. These three types of anomaly representations collectively influence the modeling of guessing and slipping scenarios. Thus, we integrate these three representations using an attention mechanism:

$$\begin{aligned}Q_i = K_i = V_i &= \alpha_i \oplus \alpha_j \oplus \beta_i, \\ \alpha_i^*, \alpha_j^*, \beta_i^* &= \mathrm{softmax}(\frac{Q \times K^T}{\sqrt{3 * d}})V,\end{aligned} \tag{15}$$

where the three representations serve as queries, keys, and values, respectively, and $\oplus$ denotes the concatenation operation. Subsequently, we extract the integrated representations to obtain two representations within the variable C:

$$\begin{aligned}x_i^{\text{Guess}} \in \mathbb{R}^{1\times 1} &= f_{\psi_G}(\sigma(\alpha_i^* \oplus \alpha_j^* \oplus \beta_i^*)), \\ x_i^{\text{Slip}} \in \mathbb{R}^{1\times 1} &= f_{\psi_S}(\sigma(\alpha_i^* \oplus \alpha_j^* \oplus \beta_i^*)),\end{aligned} \tag{16}$$

where f_{ψ_G} and f_{ψ_S} represent linear layers with different parameter weights. To block the influence of variable C on the true response results, we model these learned representations together with a general CDM to obtain unbiased response results:

$$\begin{aligned}y_{ij}^* &= \mathcal{M}(h_{s_i}, h_{e_j}, h_{c_l}), \\ \hat{y}_{ij} &= x_i^{\text{Guess}}(1 - y_{ij}^*) + (1 - x_i^{\text{Slip}})y_{ij}^*,\end{aligned} \tag{17}$$

where $\mathcal{M}(\cdot)$ denotes the prediction of the general CDMs, h_{e_j} is the one-hot vector for exercise j, and $\hat{y}$ represents the unbiased diagnostic feedback predicted by AD4CD. We use the cross-entropy loss function to supervise the learning of the entire AD4CD framework:

$$\mathcal{L}_{CD} = -\sum_{(s_i, e_j, r_{i,j}, t_{i,j}) \in R} r_{ij} \log \hat{y}_{ij} + (1 - r_{ij}) \log (1 - \hat{y}_{ij}). \tag{18}$$

Finally, the complete loss function for training the entire AD4CD is defined as:

$$\mathcal{L} = \mathcal{L}_{CD} + \mathcal{L}_{ELBO}. \tag{19}$$

Experiment

Experimental Settings

Dataset Introduction We conducted experiments on three real-world datasets: ASSISTments09, ASSISTments17, and Junyi. These datasets all include a field for students' response times. ASSISTments (Feng, Heffernan,

MODEL	ASSIST09			ASSIST17			Junyi		
	ACC↑	RMSE↓	AUC↑	ACC↑	RMSE↓	AUC↑	ACC↑	RMSE↓	AUC↑
IRT	0.7020	0.4558	0.7152	0.6720	0.4608	0.6884	0.7571	0.4153	0.6834
IRT+	**0.7239**	**0.4291**	**0.7568**	**0.6751**	**0.4547**	**0.7133**	**0.7733**	**0.3985**	**0.7319**
DINA	0.6922	0.4604	0.6967	0.6498	0.4753	0.6636	0.7541	0.4299	0.6737
DINA+	**0.7204**	**0.4318**	**0.7412**	**0.6604**	**0.4619**	**0.6940**	**0.7761**	**0.4007**	**0.7225**
MIRT	0.7187	0.4506	0.7473	0.6838	0.4543	0.7083	0.7695	0.4031	0.7286
MIRT+	**0.7343**	**0.4232**	**0.7655**	**0.6851**	**0.4507**	**0.7254**	**0.7839**	**0.3944**	**0.7467**
NCD	0.7262	0.4302	0.7525	0.6705	0.4600	0.6923	0.7716	0.4024	0.7303
NCD+	**0.7367**	**0.4216**	**0.7672**	**0.6762**	**0.4524**	**0.7195**	**0.7816**	**0.3881**	**0.7557**
KSCD	0.7340	0.4232	0.7750	0.6856	0.4521	0.7135	0.7732	0.4004	0.7367
KSCD+	**0.7416**	**0.4182**	**0.7823**	**0.6898**	**0.4453**	**0.7370**	**0.7782**	**0.3914**	**0.7508**
KANCD	0.7357	0.4225	0.7778	0.6858	0.4530	0.7124	0.7739	0.4002	0.7373
KANCD+	**0.7403**	**0.4182**	**0.7815**	**0.6895**	**0.4465**	**0.7340**	**0.7838**	**0.3872**	**0.7646**

Table 2: In the overall performance comparison across the three datasets, the "+" symbol indicates that the AD4CD framework has been added to the original base model, and all improvements are statistically significant.(i.e., two-sided t-test with $p < 0.01$.

and Koedinger 2009) is a publicly available dataset collected from the online tutoring system ASSISTments, while Junyi (Chang et al. 2015) is from the online learning platform "Junyi Academy," established in 2012. For these three datasets, we filtered out students with fewer than 10 response logs to ensure sufficient data for model training. After processing, the statistical results of the three datasets are shown in Table 1.

Evaluation Method We use three metrics, namely Root Mean Square Error (RMSE), Prediction Accuracy (ACC), and the Area Under the ROC Curve (AUC), are used to assess the prediction performance.

Backbone Models To verify the effectiveness of our model, we conducted comparative experiments based on the mainstream cognitive diagnosis models. We first selected three statistical-based cognitive diagnosis models: IRT (Lord 2012), MIRT (Chalmers 2012), and DINA (De La Torre 2009). Additionally, we chose three neural network-based models: NCD (Wang et al. 2020), KSCD (Ma et al. 2022), and KANCD (Wang et al. 2022).

Parameter Setup The loss function ratios was set to 1, 0.01 and 1. The hyperparameters of the comparison methods were tuned on the validation set according to the original papers. All models were implemented in Pytorch, and all experiments were conducted on a Linux server with an RTX4090.

Overall Performance

To validate the effectiveness of our framework, we conducted prediction experiments on the three datasets mentioned above. Table 2 shows a comparison of the performance of the six base cognitive diagnosis models with and without the AD4CD framework. Firstly, it is evident that models equipped with AD4CD outperform the original models across all datasets, reflecting the effectiveness of anomaly detection in enhancing cognitive diagnosis performance. Secondly, we observed that the performance improvement of our framework on the ACC is relatively lower for some models, while the AUC metric shows a higher improvement. This may be due to the common issue of data distribution imbalance across datasets. We also performed statistical hypothesis testing and found that the p-value was less than 0.01, which further demonstrates the validity of our proposed framework.

Models	ACC	RMSE	AUC
w/o slipping	0.7140	0.4318	0.7524
w/o guess	0.7282	0.4271	0.7577
w/o fusion	0.7254	0.4307	0.7467
AD4CD	**0.7367**	**0.4216**	**0.7672**

Table 3: Ablation study based on ASSIST09 dataset.

Ablation Study

In order to study the impact of different modules on the overall performance within the framework, we conducted ablation experiments. The results are shown in Table 3. We removed guess state modeling, slipiing modeling, and anomalous state fusion modeling, denoted as "w/o guess", "w/o slipping", and "w/o fusion", respectively. The results show that all our frameworks achieve the best results, which effectively illustrates the effectiveness of our proposed modules. It also shows that we are able to model the anomalous states effectively, thus improving the accuracy of the model.

Case Study

Representation Visualization In order to better visualize the outliers identified by our framework, we conducted visualization experiments on the model. First, we displayed the outliers learned from the interaction duration data in the model. We used the ASSIST09 dataset and extracted a portion of the G graph from the model, where the horizontal axis represents different student IDs, the vertical axis represents different problem IDs, and the numbers indicate the

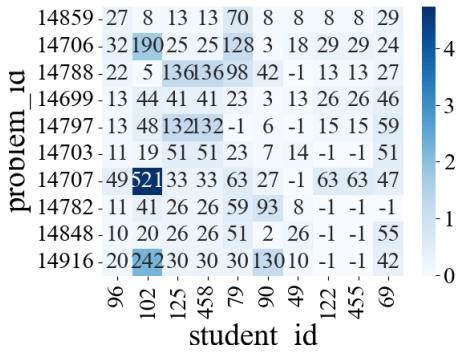

Figure 3: Abnormal information heat map.

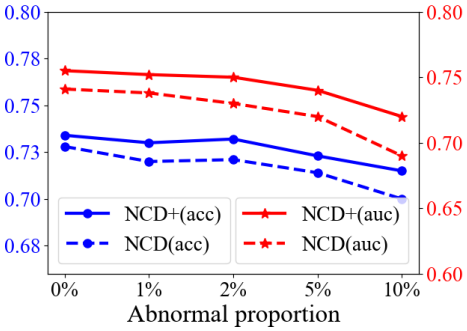

Figure 4: Robustness Experiment.

interaction durations for the students with the problems. A value of -1 represents no interaction, and the colors indicate the outliers we learned, with darker colors representing more significant outliers. It is worth noting that due to the sparsity of the dataset, we extracted the densest 10×10 matrix from the G graph for better visualization. As shown in Figure 3, our anomaly detection method successfully learned the abnormal status values of students, which will aid in the subsequent modeling of the impact of these abnormal statuses.

Robustness Experiment We constructed simulated anomalous learning interaction data on the ASSIST09 dataset for additional robustness testing of our model. Specifically, we randomly reversed each student's interaction data within different answer time intervals. Figure 4 show the performance differences between our NCD+AD4CD model and the original NCD after introducing different amounts of noise data points for each student. We can observe that as the amount of noise data increases, the performance of both models decreases. However, the decline in NCD+AD4CD is slower, validating that our model is more robust. It is worth noting that even without adding any simulated noise data, our model still performs well as it can capture potential anomalies present in the original dataset.

Anomalous Interaction Analyze To demonstrate the role of anomaly detection in our model, as shown in the Figure 5, we present an example of diagnosing students using the NCD and NCD-AD4CD models on the Junyi dataset. It

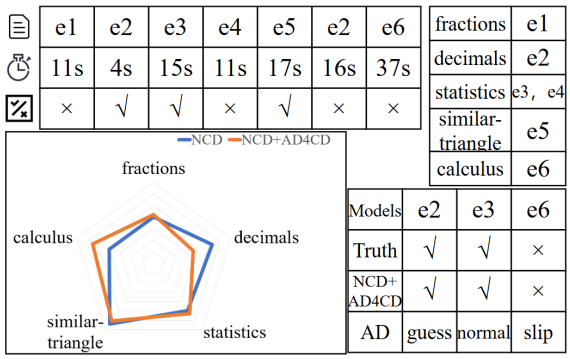

Figure 5: An example of student anomalous interaction on the Junyi dataset.

can be observed that the diagnostic results for the knowledge concepts "calculus" and "decimals" differ between the two models, which is mainly due to the AD4CD model taking into account the impact of students' anomalous states. For the knowledge concepts "decimals," the student correctly answered the question in a short time but then answered incorrectly on a subsequent question of the same knowledge concepts within a normal time frame. Considering the impact of the student's state, AD4CD determined that the student had not mastered the "decimals" knowledge concepts, predicting a high probability of guessing in the first attempt, and subsequently judged the first response as a guess in the Bayesian posterior calculations. Similarly, for "calculus," the student answered incorrectly after taking an excessively long time. The AD4CD model considered that this might be due to the student's overthinking or loss of concentration, leading to a slip, and therefore predicted a higher slip probability, ultimately classifying this interaction as a slip in the Bayesian posterior calculations. This example shows that our framework can identify students' anomalous interaction and produce interpretable diagnostic results.

Conclusion

In this paper, we analyze the impact of students' anomalous behaviors and exercise properties on response outcomes using a causal diagram. We propose a new task, Learning to Detect Anomalies (LDA), and introduce a framework called AD4CD (Anomaly Detection for Cognitive Diagnosis), which blocks confounding paths and uncovers the causal relationship between students' abilities and response outcomes. To achieve this, we develop an anomaly detection mechanism tailored to cognitive diagnostic tasks, identifying anomalous signals from response time distributions and reconstructing students' true abilities. Experiments on real-world datasets demonstrate the effectiveness and scalability of our framework.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (NO. 62107001)and in part by the Enterprise Research Transformation and Industrialization Special Project (NO. 2023-GX-C13)

References

Breunig, M. M.; Kriegel, H.-P.; Ng, R. T.; and Sander, J. 2000. LOF: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, 93–104.

Chalmers, R. P. 2012. mirt: A multidimensional item response theory package for the R environment. *Journal of statistical Software*, 48: 1–29.

Chang, H.-S.; Hsu, H.-J.; Chen, K.-T.; et al. 2015. Modeling Exercise Relationships in E-Learning: A Unified Approach. In *EDM*, 532–535.

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

Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X.; et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, 226–231.

Feng, M.; Heffernan, N.; and Koedinger, K. 2009. Addressing the assessment challenge with an online system that tutors as it assesses. *User modeling and user-adapted interaction*, 19: 243–266.

Gao, W.; Liu, Q.; Huang, Z.; Yin, Y.; Bi, H.; Wang, M.-C.; Ma, J.; Wang, S.; and Su, Y. 2021. RCD: Relation map driven cognitive diagnosis for intelligent education systems. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, 501–510.

Lazarevic, A.; Ertoz, L.; Kumar, V.; Ozgur, A.; and Srivastava, J. 2003. A comparative study of anomaly detection schemes in network intrusion detection. In *Proceedings of the 2003 SIAM international conference on data mining*, 25–36. SIAM.

Lee, M.-C.; Zhao, Y.; Wang, A.; Liang, P. J.; Akoglu, L.; Tseng, V. S.; and Faloutsos, C. 2020. Autoaudit: Mining accounting and time-evolving graphs. In *2020 IEEE International Conference on Big Data (Big Data)*, 950–956. IEEE.

Li, J.; Wang, F.; Liu, Q.; Zhu, M.; Huang, W.; Huang, Z.; Chen, E.; Su, Y.; and Wang, S. 2022a. Hiercdf: A bayesian network-based hierarchical cognitive diagnosis framework. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, 904–913.

Li, Z.; Zhao, Y.; Hu, X.; Botta, N.; Ionescu, C.; and Chen, G. H. 2022b. Ecod: Unsupervised outlier detection using empirical cumulative distribution functions. *IEEE Transactions on Knowledge and Data Engineering*, 35(12): 12181–12193.

Liu, F. T.; Ting, K. M.; and Zhou, Z.-H. 2008. Isolation forest. In *2008 eighth ieee international conference on data mining*, 413–422. IEEE.

Liu, S.; Qian, H.; Li, M.; and Zhou, A. 2023a. QCCDM: A q-augmented causal cognitive diagnosis model for student learning. In *ECAI 2023*, 1536–1543. IOS Press.

Liu, S.; Yu, X.; Ma, H.; Wang, Z.; Qin, C.; and Zhang, X. 2023b. Homogeneous Cohort-Aware Group Cognitive Diagnosis: A Multi-grained Modeling Perspective. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 4094–4098.

Lord, F. M. 2012. *Applications of item response theory to practical testing problems*. Routledge.

Ma, H.; Li, M.; Wu, L.; Zhang, H.; Cao, Y.; Zhang, X.; and Zhao, X. 2022. Knowledge-sensed cognitive diagnosis for intelligent education platforms. In *Proceedings of the 31st ACM international conference on information & knowledge management*, 1451–1460.

Ma, H.; Song, S.; Qin, C.; Yu, X.; Zhang, L.; Zhang, X.; and Zhu, H. 2024a. DGCD: An Adaptive Denoising GNN for Group-level Cognitive Diagnosis. IJCAI.

Ma, H.; Song, S.; Qin, C.; Yu, X.; Zhang, L.; Zhang, X.; and Zhu, H. 2024b. DGCD: An Adaptive Denoising GNN for Group-level Cognitive Diagnosis. In *The 33rd International Joint Conference on Artificial Intelligence (IJCAI-24)*.

Ma, H.; Wang, C.; Zhu, H.; Yang, S.; Zhang, X.; and Zhang, X. 2024c. Enhancing cognitive diagnosis using uninteracted exercises: A collaboration-aware mixed sampling approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 8877–8885.

Ma, H.; Wang, C.; Zhu, H.; Yang, S.; Zhang, X.; and Zhang, X. 2024d. Enhancing cognitive diagnosis using uninteracted exercises: A collaboration-aware mixed sampling approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 8877–8885.

Ma, H.; Xia, A.; Wang, C.; Wang, H.; and Zhang, X. 2024e. Diffusion-Inspired Cold Start with Sufficient Prior in Computerized Adaptive Testing. *arXiv preprint arXiv:2411.12182*.

Ramaswamy, S.; Rastogi, R.; and Shim, K. 2000. Efficient algorithms for mining outliers from large data sets. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, 427–438.

Shen, J.; Qian, H.; Zhang, W.; and Zhou, A. 2024. Symbolic Cognitive Diagnosis via Hybrid Optimization for Intelligent Education Systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 14928–14936.

Wang, F.; Liu, Q.; Chen, E.; Huang, Z.; Chen, Y.; Yin, Y.; Huang, Z.; and Wang, S. 2020. Neural cognitive diagnosis for intelligent education systems. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 6153–6161.

Wang, F.; Liu, Q.; Chen, E.; Huang, Z.; Yin, Y.; Wang, S.; and Su, Y. 2022. NeuralCD: a general framework for cognitive diagnosis. *IEEE Transactions on Knowledge and Data Engineering*.

Wang, H.; Pang, G.; Shen, C.; and Ma, C. 2019. Unsupervised representation learning by predicting random distances. *arXiv preprint arXiv:1912.12186*.

Wang, S.; Zeng, Z.; Yang, X.; Xu, K.; and Zhang, X. 2024. Boosting neural cognitive diagnosis with student's affective state modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 620–627.

Wang, S.; Zeng, Z.; Yang, X.; and Zhang, X. 2023. Self-supervised graph learning for long-tailed cognitive diagnosis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 110–118.

Wu, L.; Zhao, H.; Li, Z.; Huang, Z.; Liu, Q.; and Chen, E. 2023. Learning the explainable semantic relations via unified graph topic-disentangled neural networks. *ACM Transactions on Knowledge Discovery from Data*, 17(8): 1–23.

Yang, S.; Chen, M.; Wang, Z.; Yu, X.; Zhang, P.; Ma, H.; and Zhang, X. 2024a. DisenGCD: A Meta Multigraph-assisted Disentangled Graph Learning Framework for Cognitive Diagnosis. *arXiv preprint arXiv:2410.17564*.

Yang, S.; Qin, L.; and Yu, X. 2024. Endowing Interpretability for Neural Cognitive Diagnosis by Efficient Kolmogorov-Arnold Networks. *arXiv preprint arXiv:2405.14399*.

Yang, S.; Yu, X.; Tian, Y.; Yan, X.; Ma, H.; and Zhang, X. 2024b. Evolutionary neural architecture search for transformer in knowledge tracing. *Advances in Neural Information Processing Systems*, 36.

Yang, Y.; Fu, Z.-Y.; Zhan, D.-C.; Liu, Z.-B.; and Jiang, Y. 2019. Semi-supervised multi-modal multi-instance multi-label deep network with optimal transport. *IEEE Transactions on Knowledge and Data Engineering*, 33(2): 696–709.

Yang, Y.; Zhang, C.; Song, X.; Dong, Z.; Zhu, H.; and Li, W. 2023. Contextualized knowledge graph embedding for explainable talent training course recommendation. *ACM Transactions on Information Systems*, 42(2): 1–27.

Yu, X.; Qin, C.; Shen, D.; Ma, H.; Zhang, L.; Zhang, X.; Zhu, H.; and Xiong, H. 2024a. Rdgt: enhancing group cognitive diagnosis with relation-guided dual-side graph transformer. *IEEE Transactions on Knowledge and Data Engineering*.

Yu, X.; Qin, C.; Shen, D.; Yang, S.; Ma, H.; Zhu, H.; and Zhang, X. 2024b. Rigl: A unified reciprocal approach for tracing the independent and group learning processes. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 4047–4058.

Yu, X.; Qin, C.; Zhang, Q.; Zhu, C.; Ma, H.; Zhang, X.; and Zhu, H. 2024c. DISCO: A Hierarchical Disentangled Cognitive Diagnosis Framework for Interpretable Job Recommendation. *arXiv preprint arXiv:2410.07671*.

Zhang, Z.; Wu, L.; Liu, Q.; Liu, J.; Huang, Z.; Yin, Y.; Zhuang, Y.; Gao, W.; and Chen, E. 2024. Understanding and improving fairness in cognitive diagnosis. *Science China Information Sciences*, 67(5): 152106.

Zhao, H.; Zhao, C.; Zhang, X.; Liu, N.; Zhu, H.; Liu, Q.; and Xiong, H. 2023. An ensemble learning approach with gradient resampling for class-imbalance problems. *INFORMS Journal on Computing*, 35(4): 747–763.

Zhao, Y.; Hu, X.; Cheng, C.; Wang, C.; Wan, C.; Wang, W.; Yang, J.; Bai, H.; Li, Z.; Xiao, C.; et al. 2021. Suod: Accelerating large-scale unsupervised heterogeneous outlier detection. *Proceedings of Machine Learning and Systems*, 3: 463–478.

Zhou, Y.; Liu, Q.; Wu, J.; Wang, F.; Huang, Z.; Tong, W.; Xiong, H.; Chen, E.; and Ma, J. 2021. Modeling context-aware features for cognitive diagnosis in student learning. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, 2420–2428.

Zong, B.; Song, Q.; Min, M. R.; Cheng, W.; Lumezanu, C.; Cho, D.; and Chen, H. 2018. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *International conference on learning representations*.