Noise-Aware Graph-based Cognitive Diagnostic Framework Through Low-Rank Alignment

Guixian Zhang[1], Yanmei Zhang[1*], Guan Yuan[1,2*], Shang Liu[1], Xiaojing Du[3], Debo Cheng[4]

[1]School of Computer Science and Technology/School of Artificial Intelligence, China University of Mining and Technology
[2]Mine Digitization Engineering Research Center of the Ministry of Education, China University of Mining and Technology
[3]UniSA STEM, University of South Australia
[4]School of Computer Science and Technology, Hainan University
guixian@cumt.edu.cn, ymzhang@cumt.edu.cn, yuanguan@cumt.edu.cn

Abstract

Graph Neural Networks (GNNs) have effectively improved the performance of Cognitive Diagnosis Models (CDMs). Existing works have proposed a series of Graph-based Cognitive Diagnosis Frameworks (GCDFs) to enhance robustness to noise. However, these robust designs are often general methods for GNNs and are not designed for cognitive diagnosis, which undermines real cognitive information during the denoising process. Interestingly, a noteworthy phenomenon has been overlooked: even without robustness designs, GCDFs can still learn correct information in noisy environments. In this paper, we conduct a comprehensive empirical analysis of this issue. We found that noise primarily accumulates in lower singular components. Even in noisy environments, the principal subspaces of representations still remain stable. Based on these findings, we propose a **N**oise-aware **C**ognitive **D**iagnostic framework based on **L**ow-rank **A**lignment, named **NCDLA**. The framework first performs low-rank reconstruction of the interaction matrix between students and exercises, retaining only larger singular values to achieve noise reduction. Then, the reconstructed interaction matrix and the original interaction matrix are combined with the Q matrix to form a noise-reduced heterogeneous graph and an original heterogeneous graph. In order to distinguish between the interaction patterns of correct and incorrect responses, we decompose the heterogeneous graph according to the type of response. NCDLA achieves denoising of student representations and exercises representations through a self-supervised strategy based on low-rank reconstruction and a spectral anchor regularisation method. Extensive experiments on three datasets demonstrate that NCDLA achieves optimal prediction performance and robustness.

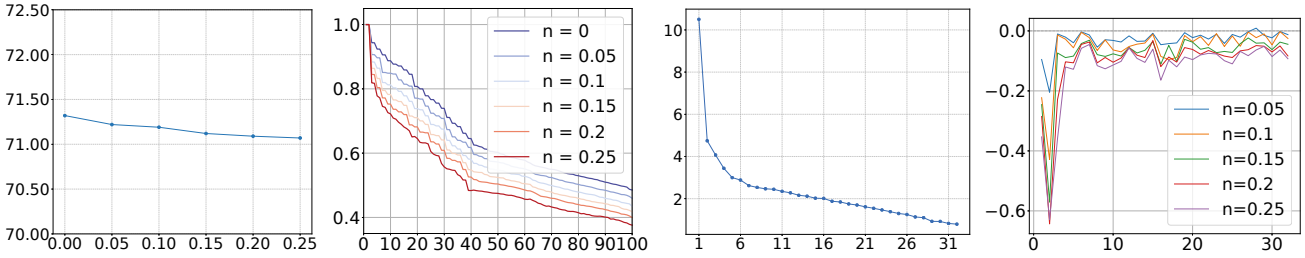
Introduction

With the development of artificial intelligence (Du, Liu, and Zhang 2025; Xu et al. 2023), intelligent education has become a hot topic, which can improve the quality of education and promote education equity (Wang et al. 2024a). Cognitive Diagnosis Model (CDM), as an fundamental tool for assessing students' knowledge mastery (Liu et al. 2023; Zhao et al. 2025), has become a support for building an intelligent education system. In the CDMs, students with similar levels of knowledge mastery will have similar performance patterns on exercises that require similar combinations of knowledge concepts (Chiu, Douglas, and Li 2009; Liu and Cheng 2018). However, data in real-world scenarios are often noisy (Zhang et al. 2025b; Wei et al. 2024; Si et al. 2025; Ma et al. 2024; Yao et al. 2024; Song et al. 2025; Du et al. 2025), guesses or slips in students' reponse logs often lead the model to incorrectly assess the students' knowledge mastery, which reduces the trustworthiness of CDMs (Liu et al. 2018; Zhao et al. 2024).

Recent studies (Qian et al. 2024; Shao et al. 2025) have found that representation learning using Graph Neural Networks (GNNs) (Lin et al. 2023; Wu et al. 2024; Zhang, Zhang, and Yuan 2024; Guan et al. 2025b; Liu and Lu 2025) before cognitive diagnosis can achieve better results. To enhance the robustness of CD, existing works have proposed different frameworks. ORCDF (Qian et al. 2024) injected noise during the training process, allowing the model to adapt itself to the noise during training to improve generalisation. However, such randomly added interactions may introduce non-realistic correlation patterns that conflict with interactions triggered by the true level of knowledge mastery, leading to the enhancement of possible misinformation. ISGCN (Shao et al. 2025) calculates the reliability of edges based on node representations and deletes edges with low reliability, but it cannot guarantee the trustworthiness of node representations and may mistakenly delete correct information. It is worth noting that the robustness designs of these frameworks are essentially general methods for GNNs and are not designed specifically for cognitive diagnostic, which undermines real cognitive information.

A noteworthy phenomenon is that even without robustness designs, Graph-based Cognitive Diagnostic Frameworks (GCDFs) can still learn correct information in noisy environments as shown in Fig. 1(a). However, all the existing works have ignored this phenomenon and failed to analyze its causes. In this paper, we conduct a comprehensive empirical analysis of this issue using singular values. Noise affects both the adjacency matrix and the representation learned by the model. We first studied how different levels of noise affect the adjacency matrix and found that

*Corresponding Author.



(a) Accuracy at different levels of noise without robustness designs.

(b) Singular values of the interaction matrix.

(c) Singular value of student representation at no noise.

(d) Variations in singular values of student representation.

Figure 1: Analysis of experimental results on the NeurIPS2020 dataset. Simply using LightGCN can achieve certain robustness. $n = 0$ represents no noise, and $n = 0.05$ represents 5% of the interactions were modified. Specifically, we randomly added or deleted interactions, and this modification may be in the right responses or in the wrong responses.

the enhancement of noise further suppresses singular values, forcing information to be homogenised. Then, we analysed the changes in student representations under different noise levels. We found that the main information in student representations is concentrated in a small portion of the principal subspace and has low rank. The principal subspaces of the representations remain stable under any level of noise. However, it is worth noting that as noise increases, the principal subspaces undergo subtle changes, forcing information to become uniform. Meanwhile, the singular value at the tail shows unstable fluctuations, which indicates that the noise intensifies the error information. This ultimately leads us to conclude that CDMs can learn correct information from noisy environments because the principal subspaces remain relatively stable all the time. To enhance the stability of the principal subspaces without destroying cognitive information, we propose a **N**oise-aware **C**ognitive **D**iagnostic framework based on **L**ow-rank **A**lignment, named **NCDLA**. NCDLA first performs low-rank reconstruction on the interaction matrix between students and exercises. The denoised matrix obtained from the reconstruction only has large singular values representing the master subspace. To ensure that the true cognitive information is not destroyed during representation learning, we employ a self-supervised alignment strategy to align the representations on the denoised and original graphs. Meanwhile, to further ensure the stability of the principal subspace and prevent the information from being homogenized by noise, we propose a spectral anchor regularization method that constraining the attenuation of the principal singular value. After obtaining the representation with stable principal subspaces, we utilized the existing CDMs for response prediction and ultimately achieved robust and accurate cognitive diagnosis results. The main contributions of this paper are as follows:

- We have ascertained the robustness of GCDFs stems from the stability of their principal subspaces through empirical analysis. To the best of our knowledge, this is the first investigation into why GCDFs can learn correct information in noisy environments.
- We propose **NCDLA** to enhance robustness without destroying real cognitive information, which includes a self-supervised alignment strategy based on low-rank reconstruction and a spectral anchor regularisation method.
- We conducted extensively comparative experiments and robustness experiments with different levels of noise on three real-world datasets. The experimental results demonstrate that NCDLA achieves the best precision and robustness.

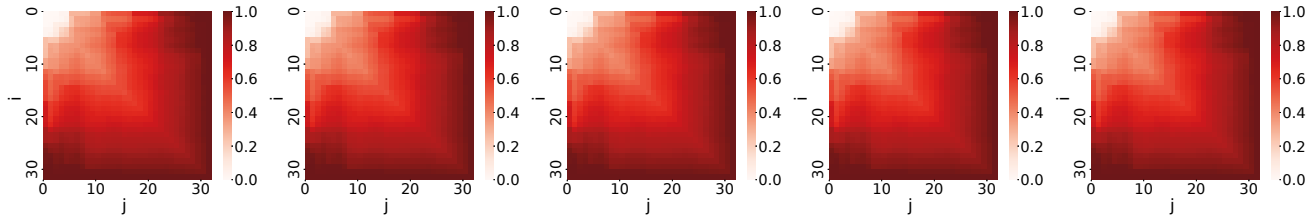
Empirical Analysis in Noisy Environments

As mentioned in the **Introduction**, although GCDFs exhibit natural robustness in noisy environments, the reason why they can maintain the ability to learn correct information remains unclear. Existing works lack a fundamental understanding of how models acquire this ability. In order to analyse this phenomenon, we perform Singular Value Decomposition (SVD) of the interaction matrix and the student representation under different levels of noises on the NeurIPS2020 dataset (Wang et al. 2021). All experimental results are obtained based on the combination of LightGCN (He et al. 2020) and KaNCD (Wang et al. 2023). The introduction to the datasets are described in Section **Experiments**. To avoid the influence of extremely large or small singular values on observation, we use a normalized adjacency matrix for analysis. As the Theorem 1 shows, the singular values of the normalized matrix will remain between 0 and 1. Fig. 1(b) represents the variation of the singular values of the interaction matrix under different noises, and we can see that as the noise increases the singular values are suppressed, forcing the signal energy to be homogenised and weakening the strength of the principal components.

Theorem 1 *Let* $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r$ *be the singular values of* $\tilde{\mathbf{R}}$ *given by* $\tilde{\mathbf{R}} = \mathbf{D}_S^{-\frac{1}{2}}\mathbf{R}\mathbf{D}_E^{-\frac{1}{2}}$, *where* $\mathbf{D}_S = \mathrm{diagMat}(\mathbf{R} \cdot \mathbf{1})$ *and* $\mathbf{D}_E = \mathrm{diagMat}\left(\mathbf{1}^\top \mathbf{R}\right)$ *are degree matrices. Then*

$$1 \geq \sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r \geq 0. \tag{1}$$

To analyse the performance of CDM in noisy environments, we artificially created noisy environments of different degrees. Specifically, we randomly added or deleted interactions, and this modification may be in the right responses or in the wrong responses. Fig. 1(c) represents the



(a) $n = 0.05$ (b) $n = 0.1$ (c) $n = 0.15$ (d) $n = 0.2$ (e) $n = 0.25$

Figure 2: The subspace similarity between representations in different noise environments and those without noise on the NeurIPS2020 dataset. No matter how large the noise is, the principal subspace maintains an extremely high similarity. As the noise increases, the representation is forced to homogenize.

variation of the singular value of the student representation under different noises. The singular values show an overall decreasing trend, with a significant decrease in the principal components and a slower decrease in the tailed singular values. In particular, there is a partial amplification of the tailed singular values, which indicates that there is some noise that enhancing the misinformation. Meanwhile, the singular values vary greatly, and the principal components have low rank, which is determined by the properties of LightGCN as described in Theorem 2. LightGCN magnifies the singular values at the head and reduces those at the tail, making the learned representation have a low-rank tendency. Notably, the fluctuations of the larger singular values are always consistent, indicating that the information of the dominant component is well preserved and the model is still able to accurately learn the interaction patterns.

Theorem 2 *For student-exercise interactions with normalized matrix* $\hat{\mathbf{R}} = \mathbf{D}_S^{-\frac{1}{2}}\mathbf{R}\mathbf{D}_E^{-\frac{1}{2}}$ *and SVD* $\hat{\mathbf{R}} = \mathbf{P}\,\mathrm{diag}(\sigma_k)\mathbf{Q}^\top$*, LightGCN's output with L layers satisfies:* $\mathbf{H}_S = \mathbf{P}\,\mathrm{diag}\left(\psi_k^{(L)}\right)\mathbf{W}_S$, $\mathbf{H}_E = \mathbf{Q}\,\mathrm{diag}\left(\psi_k^{(L)}\right)\mathbf{W}_E$, *where* $\psi_k^{(L)} = \frac{1}{L+1}\sum_{l=0}^{L}\sigma_k^l$ *is strictly increasing in* σ_k *and satisfies* $\lim_{L\to\infty}\psi_k^{(L)} = 0$*, for* $\sigma_k < 1$ *and* $\lim_{L\to\infty}\psi_k^{(L)} = 1$*, for* $\sigma_k = 1$.

While the analysis of singular values provides important insights into how noise affects cognitive diagnosis, we do not yet fully understand how much important information is retained by student representations. To further explore this issue, we analysed the subspace similarity between student representations under different levels of noise and student representations under no noise. We conducted the analysis based on the Frobenius-norm (Peng et al. 2016) on the NeurIPS dataset, with the experimental results shown in the Fig. 2. Given $\mathbf{W} \in \mathbb{R}^{d\times d}$ and $\mathbf{W}' \in \mathbb{R}^{d\times d}$ with singular value decompositions: $\mathbf{W} = \mathbf{U}_W\mathbf{\Sigma}_W\mathbf{V}_W^\top$, $\mathbf{W}' = \mathbf{U}_{W'}\mathbf{\Sigma}_{W'}\mathbf{V}_{W'}^\top$,where singular vectors are ordered by descending singular values. The principal subspace similarity metric for $i, j \in \{1, 2, \ldots, k\}$ $(k \le d)$ is:

$$\mathcal{S}(i,j) = \frac{\left\| \left(\mathbf{U}_W^{\langle i\rangle}\right)^\top \mathbf{U}_{W'}^{\langle j\rangle} \right\|_F^2}{\min(i,j)}. \tag{2}$$

The similarity of student representations decreases as noise increases, indicating that noise interferes with the CD. As noise increases, the similarity of the principal subspaces decreases, but the range of high-similarity spaces expands, further illustrating that noise forces homogenisation, weakening the intensity of principal components. However, regardless of the level of noise, the principal subspaces maintain high similarity. According to Theorem 2, the representations learned by LightGCN naturally have a low-rank tendency, which explains the previously mentioned phenomenon and enables CDMs to achieve a certain degree of accuracy even under noise interference. Therefore, how to enhance the stability of the principal subspaces without destroying the real cognitive information has become the research focus of this paper.

Method

In this section, we provide a comprehensive overview of NCDLA. The overall framework is shown in the Fig. 3.

Graph Construction and Representation Learning

Initial representations for each node are obtained by first encoding students, exercises, and knowledge using trainable embeddings $\mathbf{H}_S \in \mathbf{R}^{N\times d}$, $\mathbf{H}_X \in \mathbf{R}^{M\times d}$, and $\mathbf{H}_K \in \mathbf{R}^{Z\times d}$ respectively. In CD, the input data consists of two parts: the exercise-concept relationship matrix (also known as the Q matrix) and the response logs L. To better utilise interaction information and facilitate modelling, we deconstructs these complex data into a heterogeneous graph. We divide the heterogeneous graph into two subgraphs based on whether students answered exercises correctly:

$$\tilde{\mathbf{A}}_R = \begin{bmatrix} \mathbf{O} & \mathbf{I}_R & \mathbf{O} \\ \mathbf{I}_R^\top & \mathbf{O} & \mathbf{Q} \\ \mathbf{O} & \mathbf{Q}^\top & \mathbf{O} \end{bmatrix}, \tilde{\mathbf{A}}_W = \begin{bmatrix} \mathbf{O} & \mathbf{I}_W & \mathbf{O} \\ \mathbf{I}_W^\top & \mathbf{O} & \mathbf{Q} \\ \mathbf{O} & \mathbf{Q}^\top & \mathbf{O} \end{bmatrix}, \tag{3}$$

where $\mathbf{I}_R$ denotes right response matrix, $\mathbf{I}_W$ denotes the wrong response matrix, $\mathbf{O}$ denotes the zero matrix, $\mathbf{Q}$ denotes the Q matrix, and $\top$ indicates matrix transpose.

We employ LightGCN (He et al. 2020) as the GNN method. We first perform symmetric normalisation to avoid the influence of degree distribution imbalance, enabling more stable gradient propagation:

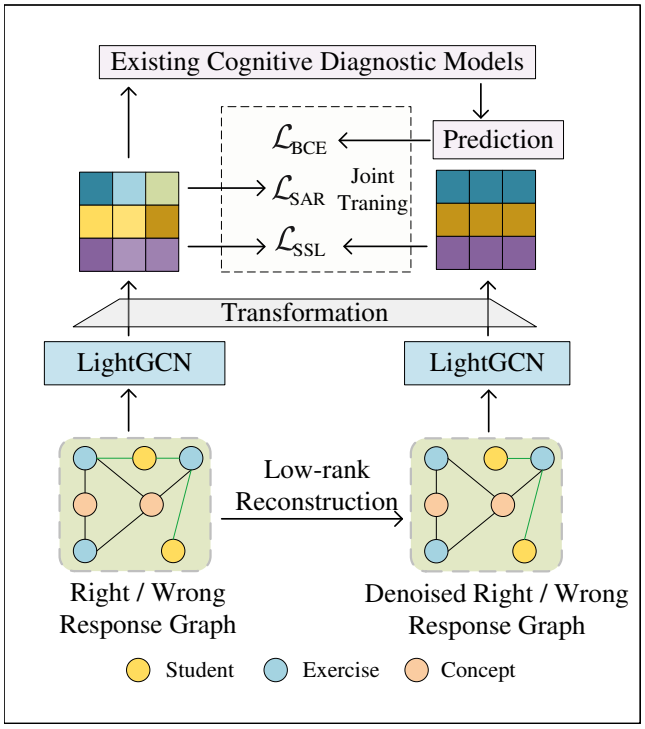


Figure 3: The proposed NCDLA framework. In the figure, CDMs refers to any existing cognitive diagnostic models.

$$\mathbf{A} = \mathbf{D}^{-\frac{1}{2}}\tilde{\mathbf{A}}\mathbf{D}^{-\frac{1}{2}}, \tag{4}$$

where $\tilde{\mathbf{A}} \in \{\tilde{\mathbf{A}}_R, \tilde{\mathbf{A}}_W\}$ is the original adjacency matrix, and $\mathbf{D} \in \mathbb{R}^{(N+M+Z)\times(N+M+Z)}$ is the degree matrix.

Subsequently, layered information propagation and fusion are performed based on the normalised interaction graph:

$$\mathbf{H}^{(l)} = \mathbf{A}\mathbf{H}^{(l-1)}, \tag{5}$$

where $\mathbf{H}^{(l)} \in \mathbb{R}^{(N+M+Z)\times d}$ denotes the node representation at the l-th layer, and d is the dimensionality.

Node representations are learned separately on the correct and incorrect interaction graphs. To prevent information confusion caused by directly superimposing response signals, a dual-channel adaptive aggregator is employed to fuse the node representations from both graphs:

$$\mathbf{H}_{WR} = \phi\left(\mathbf{W}_R^{(l)}\mathbf{H}_R^{(l)} + \mathbf{W}_W^{(l)}\mathbf{H}_W^{(l)}\right), \tag{6}$$

where $\mathbf{H}_R^{(l)}, \mathbf{H}_W^{(l)} \in \mathbb{R}^d$ represent the node representations from the correct and incorrect interaction graphs respectively, $\mathbf{W}_R^{(l)}, \mathbf{W}_W^{(l)} \in \mathbb{R}^{d\times d}$ are trainable parameter matrices, and $\phi(\cdot)$ denotes a Multi-Layer Perceptron (MLP). The final representation is obtained by applying average pooling to the layer-wise representations:

$$\mathbf{H} = \frac{1}{L}\sum_{l=0}^{L}\gamma^{(l)}\mathbf{H}_{WR}^{(l)}, \tag{7}$$

where $\gamma^{(l)}$ are identical layer weights satisfying $\sum_{l=0}^{L}\gamma^{(l)} = 1$.

Low-rank Reconstruction and Alignment

As discussed in **Empirical Analysis**, the key to the robustness of CDMs lies in the stability of the primary subspace. To this end, we propose a low-rank reconstruction method based on using SVD to obtain the denoised adjacency matrix. By the Theorem 3 (Eckart and Young 1936), SVD provides the unique minimiser of the Frobenius norm among all rank-r approximations, ensuring maximal retention of latent cognitive signals while discarding noise perturbations.

Theorem 3 (Eckart-Young-Mirsky Theorem) *For a rank-r matrix $\mathbf{P}$ with SVD $\mathbf{P} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, if $k < r$:*

$$\arg\min_{\tilde{\mathbf{P}}:\ \text{rank}(\tilde{\mathbf{P}})=k} \|\mathbf{P} - \tilde{\mathbf{P}}\|_{\mathbf{F}} = \mathbf{U}\mathbf{\Sigma}_k\mathbf{V}^T, \tag{8}$$

where $\mathbf{\Sigma}_k$ retains $\mathbf{\Sigma}$'s k largest singular values with others zeroed.

Specifically, we reconstruct a denoised matrix $\hat{\mathbf{A}}_{se}$ from observed interactions $\mathbf{A}_{se} \in \mathbb{R}^{n\times m}$ through a sequence of spectral operations. First, the target rank k is determined via dimensional thresholding: $k = \rho \cdot \min(n, m)$. In this paper, ρ is set to 0.04. A random projection matrix $\Omega \in \mathbb{R}^{m\times k}$ with $\mathcal{N}(0,1)$ entries is generated, $\mathbf{Y}_0 = \mathbf{A}_{se}\Omega$, initiating the iterative refinement process:

$$\mathbf{Y}_t = \text{QR}\left(\mathbf{A}_{se}^{\mathsf{T}} \cdot \text{QR}(\mathbf{A}_{se}\mathbf{Y}_{t-1})\right), \tag{9}$$

where $\text{qr}(\cdot)$ refers to QR decomposition.

After three power iterations, an orthonormal basis $\mathbf{Q_3} = \text{QR}(\mathbf{Y}_3) \in \mathbb{R}^{n\times l}$ is extracted, projecting the original matrix into a noise-robust subspace. The reduced matrix $\mathbf{B} = \mathbf{Q_3}^\top\mathbf{A}_{se} \in \mathbb{R}^{k\times n}$ is decomposed via SVD:

$$\mathbf{B} = \mathbf{U}_B\mathbf{\Sigma}_k\mathbf{V}_B^\top, \tag{10}$$

where $\mathbf{\Sigma}_k = \text{diag}(\sigma_1, \ldots, \sigma_k, 0, \ldots, 0)$. Finally, we can get the reconstructed student-exercise interaction matrix:

$$\hat{\mathbf{A}}_{se} = \mathbf{Q_3}\mathbf{U}_B\mathbf{\Sigma}_k\mathbf{V}_B^\top. \tag{11}$$

Following Eq.(3), we can get the denoised graph $\hat{\mathbf{A}}$. Our goal is to have similar representations from the original graph $\mathbf{A}$ and the denoised graph $\hat{\mathbf{A}}$, ensuring that CDM is still effective even when there is noise on the graph. We propose to use Self-Supervised alignment Loss (SSL) to enhance the robustness of the representation through $\mathcal{L}_{\text{SSL}}$:

$$-\sum_{s\in S}\log\left(\exp\left(\mathbf{h}_s'\mathbf{h}_s^\top/\tau\right)\right) - \sum_{e\in E}\log\left(\exp\left(\mathbf{h}_e'\mathbf{h}_e^\top/\tau\right)\right), \tag{12}$$

where $\mathbf{h}_{s_a}'$ denotes the corresponding representation via $\hat{\mathbf{A}}$.

Training and Optimization

NCDLA framework enables plug-and-play combination with arbitrary existing CDMs. First, for models requiring

fixed dimensionality, a feature transformation layer is designed to resolve dimension matching issues:

$$\mathbf{H}_t = \mathrm{ReLU}\left(\mathbf{H}\mathbf{W}_t + \mathbf{b}_t\right), \tag{13}$$

where $\mathbf{W}_t$ is the weight matrix and $\mathbf{b}_t$ is the bias vector.

As pointed out in the **empirical analysis**, the rapid decay of high singular values in learned representations remains a significant challenge, particularly as it increases susceptibility to noise and degrades feature discriminability. To address this limitation, we propose a Spectral Anchor Regularisation (SAR) method, a novel loss component that explicitly preserves the spectral integrity of latent representations by counteracting the disproportionate attenuation of dominant singular components. SAR operates by first decomposing the student representation $\mathbf{H}_S$ via SVD:

$$\mathbf{H}_S = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top, \tag{14}$$

which generates ordered singular values σ_i where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d$. Then, we designed an exponential weighted penalty function for singular values. Exponential weighting creates a soft thresholding effect that disproportionately penalises attenuation of early singular values, with λ modulating the strength of spectral prioritisation:

$$\mathcal{L}_{\text{SAR}} = -\frac{1}{a}\sum_{i=1}^{a} e^{\lambda \cdot i} \cdot \sigma_i. \tag{15}$$

The denoised representations serve as inputs to downstream diagnosis models. The prediction of student-exercise interactions can be formalised as:

$$\hat{Y}_{SX} = \mathrm{CDM}\left(\mathbf{H}_S, \mathbf{H}_X, \mathbf{H}_K\right). \tag{16}$$

During NCDLA framework training, end-to-end optimisation is achieved through a joint training mechanism. Specifically, the negative log-likelihood function is first employed to measure response prediction capability:

$$\begin{aligned}\mathcal{L}_{\text{BCE}} = -\sum_{(S,X)\in L} &\left[y_{sx}\log\sigma\left(\hat{y}_{sx}\right) + \right. \\ &\left.\left(1-y_{sx}\right)\log\left(1-\sigma\left(\hat{y}_{sx}\right)\right)\right],\end{aligned} \tag{17}$$

where L denotes response logs, and y_{sx} indicates whether the current student answered correctly.

We then combine prediction loss, self-supervised alignment loss, and SAR loss to obtain the final loss function:

$$\mathcal{L} = \mathcal{L}_{\text{BCE}} + \alpha \cdot \mathcal{L}_{\text{SSL}} + \beta \cdot \mathcal{L}_{\text{SAR}}, \tag{18}$$

where α and β are the hyperparameters.

Experiments

In this section, we conducted comprehensive experiments to verify the predictability and robustness of NCDLA.

Setups

Datasets and Metrics. Experiments were performed on three public datasets: Assist17 (Feng, Heffernan, and Koedinger 2009), NeurIPS2020 (Wang et al. 2021), and Junyi (Chang, Hsu, and Chen 2015). Table 1 presents detailed statistics. Sparsity Rate represents the density of interactions in the student-exercise interaction matrix. Average Correctness Rate (ACR) reflects the average difficulty level of items across the dataset. The datasets were partitioned into training, validation, and test sets using a 7:1:2 ratio. Three metrics were employed to measure classification model performance: Accuracy (ACC), Area Under the receiver operating characteristic Curve (AUC), and F1-score. Additionally, we adopt the Degree of Agreement (DOA) (Qian et al. 2024) as the interpretability metric.

Dataset	Assit2017	NeurIPS2020	Junyi
#Students	1,709	2,840	10,000
#Exercises	3,162	6,000	734
#Concepts	102	268	734
#Response Logs	390,311	214,328	408,057
Sparsity	0.072	0.012	0.055
ACR	0.815	0.631	0.687
Q Density	1.22	4.14	1.00

Table 1: Statistics of three datasets.

Baselines and Implementation Details. We integrate various CDMs with different frameworks. We selected two types of CDMs as backbones: IRT (Hambleton and Swaminathan 2013) and MIRT (Chalmers 2012) based on statistical methods, and NCD (Wang et al. 2020), MFKC (Li et al. 2022b) and KaNCD (Wang et al. 2023) based on neural networks. For frameworks, we adopted LightGCN (LGCN) (He et al. 2020), ORCDF (Qian et al. 2024), and ISGCD (Shao et al. 2025) for comparison. It is worth noting that both ORCDF and ISGCD have made improvements for robustness. For all methods, we use Xavier to initialize parameters and Adam for optimization. For fair comparison, the embedding size d of MIRT and KaNCD is set to 32, and d of NCDM and MFKC is uniformly set to the number of knowledge concepts K. The batch size of all datasets is set to 4096. For the hyperparameters in NCDLA, the learning rate is set to 0.04, α to 0.001, β to 0.1, λ to 0.2, and a is set to 2. Each experimental result is the average under five random seeds.

Comparative Experiments

The comparative experimental results across three datasets are presented in Table 2. It can be observed that NCDLA consistently improves all baseline CDMs across all datasets. This confirms that NCDLA effectively removes noise from the data and achieves more accurate cognitive diagnosis. MIRT demonstrates enhanced diagnostic capabilities by modelling multiple latent dimensions of students, overcoming the limitation of traditional IRT's unidimensional ability assumption. MFKC achieves diagnostic improvement beyond KaNCD attains superior diagnostic outcomes by implicitly learning inter-knowledge relationships through latent vectors, leveraging mastered knowledge components to infer proficiency in uncovered ones. LightGCN significantly enhances cognitive diagnosis models by utilising GNNs to model heterogeneous interaction graphs while incorporating existing models as response prediction modules. This effi-

Dataset	Assit2017				Neurips2020				Junyi			
Method	ACC	AUC	F1	DOA	ACC	AUC	F1	DOA	ACC	AUC	F1	DOA
IRT	85.56	88.27	91.66	-	69.95	75.22	76.85	-	76.27	80.37	83.49	-
LGCN-IRT	86.75	89.43	92.11	-	71.43	76.56	78.63	-	77.46	81.74	84.31	-
ORCDF-IRT	86.83	89.43	92.13	-	71.45	76.68	78.24	-	77.57	81.68	84.42	-
ISGCD-IRT	86.78	89.45	92.03	-	71.52	76.76	78.70	-	77.50	81.42	84.46	-
NCDLA-IRT	**86.85**	**89.49**	**92.15**	-	**71.65**	**84.95**	**78.78**	-	**77.64**	**81.50**	**84.68**	-
MIRT	86.38	89.32	91.98	-	70.09	74.51	77.84	-	77.09	80.95	84.25	-
LGCN-MIRT	86.07	88.58	91.89	-	70.49	75.47	78.25	-	77.29	81.17	84.28	-
ORCDF-MIRT	86.76	89.52	92.11	-	71.54	76.58	78.35	-	77.51	81.33	84.25	-
ISGCD-MIRT	87.02	90.08	92.26	-	71.55	76.98	78.07	-	77.68	81.49	84.67	-
NCDLA-MIRT	**87.28**	**90.15**	**92.40**	-	**71.71**	**77.25**	**78.47**	-	**77.79**	**81.44**	**84.64**	-
NCDM	82.22	79.56	89.39	55.54	69.73	74.19	76.80	66.55	74.27	78.07	82.59	50.41
LGCN-NCD	83.29	82.94	89.87	56.94	70.55	75.58	77.00	67.57	75.00	74.80	82.43	55.44
ORCDF-NCD	85.63	87.59	91.27	60.27	71.36	76.68	77.72	69.25	76.86	80.89	83.78	59.43
ISGCD-NCD	85.76	87.91	91.37	59.43	71.43	76.86	78.19	69.23	76.51	80.46	84.27	55.36
NCDLA-NCD	**87.82**	**90.78**	**92.74**	**68.15**	**72.07**	**77.39**	**78.80**	**69.84**	**77.35**	**81.45**	**84.05**	**60.09**
MFKC	83.01	84.34	89.93	55.98	70.86	76.05	77.21	68.35	74.87	78.43	82.51	49.24
LGCN-MFKC	85.77	88.51	91.34	59.65	71.62	76.82	78.42	69.90	76.19	80.83	82.47	59.33
ORCDF-MFKC	86.94	89.58	92.08	62.70	71.66	76.72	78.55	72.15	77.26	81.05	84.37	60.64
ISGCD-MFKC	86.87	89.45	92.05	64.23	71.82	77.20	78.17	71.75	76.96	80.98	83.73	61.27
NCDLA-MFKC	**87.34**	**89.91**	**92.38**	**68.25**	**71.86**	**77.37**	**78.87**	**72.46**	**77.34**	**81.39**	**84.21**	**61.36**
KaNCD	85.03	86.45	90.98	57.90	71.01	75.97	77.91	68.67	75.27	75.21	83.32	53.86
LGCN-KaNCD	85.49	87.05	91.25	58.38	71.32	76.59	77.33	69.23	76.61	79.89	83.60	58.88
ORCDF-KaNCD	87.14	89.73	92.27	62.42	71.59	76.92	78.18	70.09	77.54	81.43	84.35	60.58
ISGCD-KaNCD	87.04	89.62	92.21	64.42	71.87	77.20	77.69	70.37	77.57	81.80	84.61	60.60
NCDLA-KaNCD	**87.66**	**90.19**	**92.62**	**68.07**	**71.96**	**77.53**	**78.56**	**70.75**	**77.78**	**82.13**	**84.66**	**61.37**

Table 2: The comparative experimental results on three datasets. **Bold** indicates the best result.

cacy stems from the inherently heterogeneous topological relationships among core educational entities. The model's higher-order relational learning capability enables extraction of deeper cognitive patterns from sparse interaction data.

ORCDF, ISGCD, and NCDLA divide the heterogeneous graph into right-response and wrong-response graphs, effectively mitigating confusion between correct and erroneous patterns. However, although ORCDF improves performance through stochastic response reversal as data augmentation during training, this random inversion may inadvertently reinforce erroneous information present in the original data. ISGCD selectively removes edges based on uncertainty and heterogeneity metrics, but such operations risk erroneous elimination of valid information, resulting in ISGCD underperforming ORCDF under some scenarios. NCDLA's empirically designed low-rank reconstruction methodology effectively circumvents these limitations.

Robustness Analysis

In this paper, noise refers to either adding new interactions or deleting existing interactions. For these robustness experiments, modified interactions constituted 5%, 15%, and 25% of the total interactions respectively. KaNCD was employed as the backbone model to evaluate performance under four frameworks. Results on the NeurIPS2020 dataset are presented in the Fig. 4. Experimental findings demonstrate that while all frameworks exhibit gradually declining accuracy with increasing noise levels, they maintain reasonable predictive capability. Crucially, NCDLA framework consistently preserves superior accuracy across all noise conditions. Although ORCDF enhances model robustness through random interaction-type inversion, this randomised operation primarily targets generalisation improvement. This enhancement of generalisation mainly stems from the fact that when there is a significant distribution bias in the interaction pattern, the random inversion strategy can create meaningful virtual negative samples, effectively increase the diversity of negative samples, and alleviate the overfitting of the model. However, such random inversions may inadvertently reinforce erroneous information, potentially misleading the model. ISGCD performs edge deletion based on computed reliability metrics, but cannot guarantee the resulting representations are both accurate and trustworthy, risking loss of valid information. NCDLA's low-rank reconstruction methodology achieves precise noise detection and targeted mitigation, effectively eliminating noise impact. By circumventing the interference issues inherent in alternative approaches, NCDLA delivers optimal robustness in cognitive diagnosis.

Ablation Studies

To clarify each module's contribution, ablation studies were conducted by respectively removing the low-rank reconstruction alignment (w/o ALI) and spectral anchor regularisation (w/o SAR) components. Similar to previous robustness testing, varying noise levels were artificially introduced

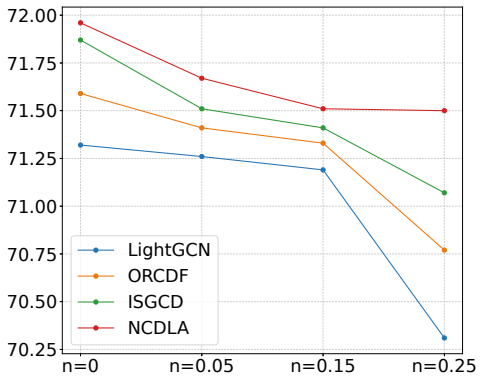


Figure 4: Accuracy under different noises on NeurIPS2020.

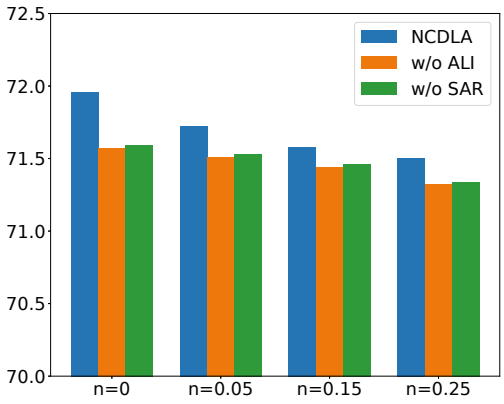


Figure 5: Ablation experiments on NeurIPS2020.

to assess the robustness of each variant. Experimental results on the NeurIPS2020 dataset are shwon in the Fig. 5. Results demonstrate that the removal of either module significantly diminishes diagnostic accuracy, confirming each component's effectiveness in enhancing cognitive diagnosis performance. Concurrently, framework robustness consistently decreases with module removal. The decrease in results is more pronounced when the low-rank reconstruction module is removed. This is because low-rank reconstruction undertakes the task of structural noise filtering, and SAR performs refined representation constraints on this basis. Without the SAR module, this variant can still learn the denoised representation through the denoised graph structure and self-supervised alignment. Crucially, ALI and SAR stabilise the primary subspace from complementary perspectives: the former operates on the adjacency matrix structure while the latter regulates node representations.

Related Works

In this section, we introduce the related works in two parts: conventional methods and GNN-based methods.

Conventional Cognitive Diagnosis Methods. Statistical methods such as IRT (Hambleton and Swaminathan 2013; Xu et al. 2025) and MIRT (Chalmers 2012) effectively infer students' performance. With advances in computational hardware, deep neural networks have been introduced into cognitive diagnosis methodologies (Zhang et al. 2023; Wang et al. 2024b; Shen et al. 2024; Liu et al. 2025). NCDM (Wang et al. 2020) recognised the limitations of expert-designed interaction functions and proposed a novel data-driven cognitive diagnosis framework. MFKC (Li et al. 2022b) designed and modeled the difficulty and discrimination of knowledge concepts based on neural networks. KaNCD (Wang et al. 2023) uses matrix factorization layers to reflect students' proficiency in grasping concepts. However, these methods typically treat students, exercises, and concepts as independent entities, rendering them inadequate for learning the intricate interaction patterns among these three elements. This limitation results in the loss of valuable interaction structural information during the inference of cognitive states.

GNN-based Cognitive Diagnosis Methods. With the extensive research and application of GNNs (Chen, Wang, and He 2025; Luo et al. 2024; Guan et al. 2025a; Cheng et al. 2025; Zhang et al. 2025a), some researchers (Gao et al. 2021; Hou et al. 2025) have recognised that the relationships can form graph structures, leading them to incorporate GNN into their CDMs (Li et al. 2022a). However, existing works (Qian et al. 2024; Yao et al. 2024; Shao et al. 2025) have found that frameworks that first conduct representation learning and then concatenate CDMs as the backbone often have better results. ORCDF (Qian et al. 2024) employed data augmentation and self-supervised alignment to acquire relatively robust representations. ISGCD (Shao et al. 2025) quantifies the uncertainty of edges, achieving robust results by removing unreliable edges. However, these methods do not notice the natural robustness of LightGCN. In this paper, we conduct empirical analysis to identify the sources of CDMs' robustness and propose an accurate and robust graph-based cognitive diagnosis framework.

Conclusion

In this paper, we conducted extensive empirical analyses to investigate the robustness in CDMFs. Our findings reveal that stabilisation of the principal subspace is pivotal for maintaining the robustness of cognitive diagnosis, as noise induces information homogenisation. Building upon these findings, we propose a novel framework named NCDLA, which first denoises interaction graphs via low-rank reconstruction and aligns representations through self-supervised learning. Concurrently, a spectral anchor regularisation loss function is introduced to stabilise the singular values of node representations to avoid homogenization. Comprehensive experimental evaluations demonstrate that NCDLA achieves optimal diagnostic accuracy and noise resistance. This paper represents the first systematic investigation into why graph-based cognitive diagnosis frameworks can effectively learn correct information in noisy environments. These findings advance the theoretical understanding of graph-based diagnostic frameworks and offer valuable insights for future graph-based cognitive diagnosis framework design.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 6250071514, Xuzhou K&D Program under Grant KC23296, the Science and Technology Program of Xuzhou under Grant No.KC22047, the Graduate Innovation Program of China University of Mining and Technology 2024WLKXJ183, the Fundamental Research Funds for the Central Universities 2024-10949, and the Postgraduate Research & Practice Innovation Program of Jiangsu Province KYCX24_2781.

References

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.; and Chen, K.-T. 2015. Modeling exercise relationships in E-learning: A unified approach. In *Proceedings of the 8th International Conference on Educational Data Mining*, 532–535.

Chen, J.; Wang, M.; and He, K. 2025. Hybrid Long-Range Dependency-Aware Graph Convolutional Network for Node Classification. *Knowledge and Information Systems*.

Cheng, J.; Liang, K.; Feng, P.; Liu, W.; Tang, Y.; and He, C. 2025. Clustering Diffusion Model with Frequency-Signal Modulation for Variational Graph Autoencoders. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–18.

Chiu, C.-Y.; Douglas, J. A.; and Li, X. 2009. Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika*, 74: 633–665.

Du, E.; Li, X.; Jin, T.; Zhang, Z.; Li, R.-H.; and Wang, G. 2025. Graphmaster: Automated graph synthesis via llm agents in data-limited environments. *arXiv preprint arXiv:2504.00711*.

Du, E.; Liu, S.; and Zhang, Y. 2025. Mixture of length and pruning experts for knowledge graphs reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 432–453.

Eckart, C.; and Young, G. 1936. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3): 211–218.

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.

Guan, R.; Liu, T.; Tu, W.; Tang, C.; Luo, W.; and Liu, X. 2025a. Sampling Enhanced Contrastive Multi-View Remote Sensing Data Clustering with Long-Short Range Information Mining. *IEEE Transactions on Knowledge and Data Engineering*, 1–15.

Guan, R.; Tu, W.; Wang, S.; Liu, J.; Hu, D.; Tang, C.; Feng, Y.; Li, J.; Xiao, B.; and Liu, X. 2025b. Structure-Adaptive Multi-View Graph Clustering for Remote Sensing Data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 16933–16941.

Hambleton, R. K.; and Swaminathan, H. 2013. *Item response theory: Principles and applications*. Springer Science & Business Media.

He, X.; Deng, K.; Wang, X.; Li, Y.; Zhang, Y.; and Wang, M. 2020. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, 639–648.

Hou, M.; Li, X.; Guo, T.; Liu, Z.; Tian, M.; Luo, R.; and Luo, W. 2025. Cognitive Fluctuations Enhanced Attention Network for Knowledge Tracing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 14265–14273.

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, S.; Guan, Q.; Fang, L.; Xiao, F.; He, Z.; He, Y.; and Luo, W. 2022b. Cognitive diagnosis focusing on knowledge concepts. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 3272–3281.

Lin, Y.; Yang, M.; Yu, J.; Hu, P.; Zhang, C.; and Peng, X. 2023. Graph matching with bi-level noisy correspondence. In *Proceedings of the IEEE/CVF international conference on computer vision*, 23362–23371.

Liu, C.; and Cheng, Y. 2018. An application of the support vector machine for attribute-by-attribute classification in cognitive diagnosis. *Applied Psychological Measurement*, 42(1): 58–72.

Liu, F.; Zhang, Y.; Liu, S.; Ji, S.; Yu, K.; and Wu, L. 2025. Prompt Transfer for Dual-Aspect Cross-Domain Cognitive Diagnosis. *IEEE Transactions on Computational Social Systems*, 1–13.

Liu, Q.; Wu, R.; Chen, E.; Xu, G.; Su, Y.; Chen, Z.; and Hu, G. 2018. Fuzzy cognitive diagnosis for modelling examinee performance. *ACM Transactions on Intelligent Systems and Technology*, 9(4): 1–26.

Liu, Y.; Zhang, T.; Wang, X.; Yu, G.; and Li, T. 2023. New development of cognitive diagnosis models. *Frontiers of Computer Science*, 17(1): 171604.

Liu, Z.; and Lu, W. 2025. MDN: Modality Decomposition Network for Multimodal Recommendation. In *Proceedings of the 2025 International Conference on Multimedia Retrieval*, 871–879.

Luo, R.; Huang, H.; Yu, S.; Han, Z.; He, E.; Zhang, X.; and Xia, F. 2024. FUGNN: Harmonizing Fairness and Utility in Graph Neural Networks. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2072–2081.

Ma, H.; Song, S.; Qin, C.; Yu, X.; Zhang, L.; Zhang, X.; and Zhu, H. 2024. DGCD: an adaptive denoising GNN for group-level cognitive diagnosis. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 2261–2269.

Peng, X.; Lu, C.; Yi, Z.; and Tang, H. 2016. Connections between nuclear-norm and frobenius-norm-based representations. *IEEE Transactions on Neural Networks and Learning Systems*, 29(1): 218–224.

Qian, H.; Liu, S.; Li, M.; Li, B.; Liu, Z.; and Zhou, A. 2024. ORCDF: An Oversmoothing-Resistant Cognitive Diagnosis Framework for Student Learning in Online Education Systems. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2455–2466.

Shao, P.; Yang, Y.; Gao, C.; Chen, L.; Zhang, K.; Zhuang, C.; Wu, L.; Li, Y.; and Wang, M. 2025. Exploring Heterogeneity and Uncertainty for Graph-based Cognitive Diagnosis Models in Intelligent Education. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 1233–1243.

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*, 14928–14936.

Si, C.; Cui, Y.; Yang, F.; Yang, X.; and Shen, W. 2025. Why Can Accurate Models Be Learned from Inaccurate Annotations? *arXiv preprint arXiv:2505.16159*.

Song, Y.; Wei, Y.; Lu, Y.; Sun, Q.; Shao, M.; Wang, L.-e.; Hu, C.; Li, X.; and Fu, X. 2025. Mitigating Message Imbalance in Fraud Detection with Dual-View Graph Representation Learning. *arXiv preprint arXiv:2507.06469*.

Wang, F.; Gao, W.; Liu, Q.; Li, J.; Zhao, G.; Zhang, Z.; Huang, Z.; Zhu, M.; Wang, S.; Tong, W.; et al. 2024a. A survey of models for cognitive diagnosis: New developments and future directions. *arXiv preprint arXiv:2407.05458*.

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*, 6153–6161.

Wang, F.; Liu, Q.; Chen, E.; Huang, Z.; Yin, Y.; Wang, S.; and Su, Y. 2023. NeuralCD: A General Framework for Cognitive Diagnosis. *IEEE Transactions on Knowledge and Data Engineering*, 35(8): 8312–8327.

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

Wang, Z.; Lamb, A.; Saveliev, E.; Cameron, P.; Zaykov, J.; Hernandez-Lobato, J. M.; Turner, R. E.; Baraniuk, R. G.; Barton, C.; Jones, S. P.; et al. 2021. Results and insights from diagnostic questions: The neurips 2020 education challenge. In *NeurIPS 2020 Competition and Demonstration Track*, 191–205. PMLR.

Wei, Y.; Yuan, H.; Fu, X.; Sun, Q.; Peng, H.; Li, X.; and Hu, C. 2024. Poincaré differential privacy for hierarchy-aware graph embedding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 9160–9168.

Wu, Z.; Mo, Y.; Zhou, P.; Yuan, S.; and Zhu, X. 2024. Self-training based few-shot node classification by knowledge distillation. In *Proceedings of the AAAI conference on artificial intelligence*, 15988–15995.

Xu, Z.; Cheng, D.; Li, J.; Liu, J.; Liu, L.; and Wang, K. 2023. Disentangled representation for causal mediation analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 10666–10674.

Xu, Z.; Kandanaarachchi, S.; Ong, C. S.; and Ntoutsi, E. 2025. Fairness evaluation with item response theory. In *Proceedings of the ACM on Web Conference 2025*, 2276–2288.

Yao, F.; Liu, Q.; Yue, L.; Gao, W.; Li, J.; Li, X.; and He, Y. 2024. Adard: An adaptive response denoising framework for robust learner modeling. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 3886–3895.

Zhang, G.; Yuan, G.; Cheng, D.; Liu, L.; Li, J.; and Zhang, S. 2025a. Mitigating propensity bias of large language models for recommender systems. *ACM Transactions on Information Systems*, 43(6): 1–26.

Zhang, G.; Zhang, S.; and Yuan, G. 2024. Bayesian graph local extrema convolution with long-tail strategy for misinformation detection. *ACM Transactions on Knowledge Discovery from Data*, 18(4): 1–21.

Zhang, H.; Tang, H.; Sun, Y.; He, S.; and Li, Z. 2025b. Modality-Specific Interactive Attack for Vision-Language Pre-Training Models. *IEEE Transactions on Information Forensics and Security*, 20: 5663–5677.

Zhang, Y.; Qin, C.; Shen, D.; Ma, H.; Zhang, L.; Zhang, X.; and Zhu, H. 2023. Relicd: A reliable cognitive diagnosis framework with confidence awareness. In *Proceedings of the 2023 IEEE International Conference on Data Mining*, 858–867.

Zhao, G.; Huang, Z.; Cheng, C.; Zhuang, Y.; Mao, Q.; Li, X.; Wang, S.; and Chen, E. 2025. Multi-Perspective Consolidation Enhanced Cognitive Diagnosis via Conditional Diffusion Model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 1174–1182.

Zhao, G.; Huang, Z.; Zhuang, Y.; Bi, H.; Wang, Y.; Wang, F.; Ma, Z.; and Zhao, Y. 2024. A Diffusion-Based Cognitive Diagnosis Framework for Robust Learner Assessment. *IEEE Transactions on Learning Technologies*, 17: 2281–2295.