Toward Fair and Efficient Intelligent Learning: A Generative Cognitive Diagnosis Approach

Jiatong Li, Yingying Ni, Minjuan Wang, *Senior Member, IEEE*, John Chi-Kin Lee, Qi Liu, *Member, IEEE*, Mengxiao Zhu, *Member, IEEE*, Zhenya Huang, *Associate Member, IEEE*, Shijin Wang, and Enhong Chen, *Fellow, IEEE*

***Abstract*—Intelligent learning (iLearning) represents a holistic approach to e-learning that dynamically aligns instructional materials with individual learner needs. Within the iLearning framework, the cognitive diagnosis model (CDM) serves as a cornerstone technology by inferring human cognitive states from diagnostic test response logs. Traditional CDMs typically treat cognitive states as learnable parameters, optimizing them through a score prediction-driven paradigm. However, this approach falls short in the nonidentifiability problem of diagnostic outputs and lacks efficiency in instant diagnosis for incoming learners. These limitations pose a significant challenge to the individual fairness and utility of cognitive assessments. In this study, we introduce a novel generative diagnosis paradigm designed to resolve these limitations. By enabling the inductive inference of cognitive states without the need for parameter reoptimization, our generative approach ensures strict identifiability and addresses the low efficiency of optimization-based diagnosis. We present two streamlined yet powerful instantiations of this paradigm: generative item response theory and the generative neural cognitive diagnosis model. Extensive experiments on real-world datasets confirm that our methodology significantly outperforms traditional models, offering superior scalability and reliability. Ultimately, this generative framework provides a robust foundation for the development of next-generation iLearning environments.**

***Index Terms*—Cognitive diagnosis (CD), fairness, intelligent learning (iLearning), psychometrics, user modeling.**

Received 1 January 2026; revised 24 February 2026; accepted 26 February 2026. Date of publication 16 March 2026; date of current version 11 May 2026. This work was supported in part by the National Key Research and Development Program of China under Grant 2024YFC3308200, in part by the National Natural Science Foundation of China under Grant 62525606, in part by the Key Technologies R&D Program of Anhui Province under Grant 202423k09020039, and in part by the Fundamental Research Funds for the Central Universities. *(Jiatong Li and Yingying Ni contributed equally to this work.) (Corresponding author: Qi Liu.)*

Jiatong Li, Qi Liu, Zhenya Huang, and Enhong Chen are with the State Key Laboratory of Cognitive Intelligence, University of Science and Technology of China, Hefei 230026, China (e-mail: cslijt@mail.ustc.edu.cn; qiliuql@ustc.edu.cn; huangzhy@ustc.edu.cn; cheneh@ustc.edu.cn).

Yingying Ni is with the School of Media & Communication, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: niinin7@sjtu.edu.cn).

Minjuan Wang is with The Education University of Hong Kong, Hong Kong SAR, China, and also with the San Diego State University, San Diego, CA 92182 USA (e-mail: mwang@sdsu.edu).

John Chi-Kin Lee is with the Academy for Applied Policy Studies and Education Futures (AAPSEF), The Education University of Hong Kong, Hong Kong SAR, China (e-mail: jcklee@eduhk.hk).

Mengxiao Zhu is with the Department of Science and Technology Communication and School of Humanities and Social Sciences, University of Science and Technology of China, Hefei 230026, China (e-mail: mxzhu@ustc.edu.cn).

Shijin Wang is with the State Key Laboratory of Cognitive Intelligence Hefei 230026, China, and also with the iFLYTEK AI Research (Central China), iFLYTEK Company Ltd., China (e-mail: sjwang3@iflytek.com).

Digital Object Identifier 10.1109/TLT.2026.3674350

I. Introduction

INTELLIGENT learning (iLearning) encompasses a multifaceted approach to modern education [1], [2], [3], [4], as shown in Fig. 1. It can be seen as one type of holistic e-learning that matches learning materials with user needs [5]. At its core, iLearning is characterized by allocating resources according to the demands of learners with the aid of information and communication technology, providing learners with better learning management to meet the needs of mobile and ubiquitous learning, and promoting personalized and adaptive learning [4]. This pedagogical model operates through a collaborative framework that distinguishes itself from traditional instructional methods. Specifically, iLearning employs "pull learning," a student-centered paradigm in which learners actively "pull" learning content according to their own schedules and locations, contrasting sharply with conventional "push learning" models where educators transmit predetermined content to passive recipients [6]. The cornerstone of "pull learning" lies in fostering collaboration and mutual resource sharing among the student community, a capability substantially enhanced through the implementation of Web 2.0 technologies that form the foundation of Education 3.0 [5]. This is partly based on the assumption that learners are aware of their own needs, as well as motivated and capable of engaging with ICT tools.

Along this line, cognitive diagnosis (CD) is a significant task in iLearning, which aims to model latent and complex cognitive states of human learners via mining their response data on diagnostic tests [7], [8], [9]. Stemming from educational measurement, CD has been widely applied in various evaluation scenarios, such as student knowledge proficiency modeling [10], [11], computerized adaptive testing [12], [13], and model evaluation [14], [15]. Fig. 2 presents two typical cognitive diagnosis models (CDMs), the item response theory (IRT) model [16], [17] and the neural cognitive diagnosis model (NCDM) [18]. Given question data and response data, CDMs utilize parameter estimation algorithms to fit response scores and optimize learner and item parameters. Learner parameters are then presented as diagnostic outputs. The IRT models learner ability as a single variable, whereas NCDM models learner ability as knowledge mastery degrees. Compared to explicit evaluation techniques, such as the classical testing theory [19], [20], CD enables a deeper understanding of the cognition construction of test-takers, providing clearer guidance on the ability development of test-takers. Therefore, how to design fair and efficient CDMs

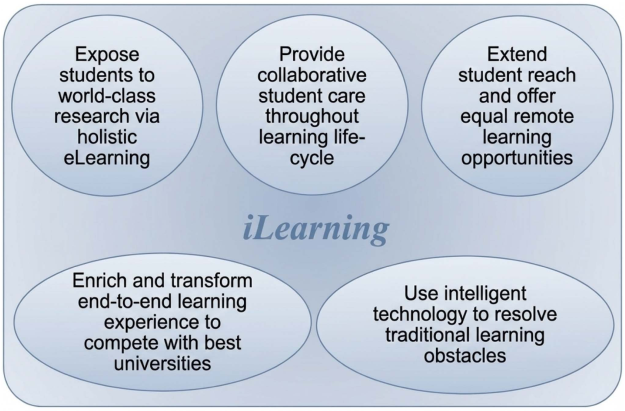


Fig. 1. Overview of intelligent learning (iLearning), from [1].

that yield consistent diagnoses for students with similar abilities has become a significant research question.

Challenges: The design of CDMs has focused on appropriately modeling latent cognitive states via accurate fitting of response data. Researchers emphasized designing an appropriate interaction function [or item response function (IRF) in IRT] for CDMs to predict response data as accurately as possible [18]. For instance, the IRT uses a logistic function to predict response data given the random variable representation of learner traits and item attributes. In this study, this paradigm is called the *transductive prediction paradigm*. However, the transductive prediction paradigm of CD confronts the following inevitable challenges.

1) *A lack of fairness for diagnostic outputs due to nonidentifiability:* The diagnostic results of CDMs violate the basic principle of fairness due to the nonidentifiability problem [21]. For example, Alex and Paul have identical response score records, as presented in Fig. 2. However, CDMs cannot provide equivalent diagnostic results for these learners despite their equivalent performance. This inequitable treatment creates unjustified advantages or disadvantages in educational decisions and interventions. This systematic unfairness directly undermines the reliability of CD.
2) *Unable to diagnose instantly for incoming learners:* When new learners come for CD, the CDM has to be retrained from scratch to update the cognitive states [22], [23]. This can lead to a long time latency ($\approx$ 5.5 h for 832 sequentially arrived incoming learners; see Fig. 3) and high computational cost (requires retraining the whole model) and can lead to inconsistency for existing learner cognitive states before and after retraining, potentially introducing unfair assessment disparities across time.

Beyond these structural limitations, the effectiveness of cognitive modeling is further complicated by the diverse profiles of the learners. Individual differences, such as the distinction between deep learners who seek underlying principles and surface learners who focus on rote memorization, necessitate diagnostic models that can capture nuanced learning approaches [24]. Furthermore, the actual receptivity and usage of iLearning platforms are influenced by a complex interplay of learner readiness, performance, trust, interactivity, and ethical awareness [25]. Integrating these human-centric factors into CDMs is essential for ensuring that automated diagnoses are not only accurate but also aligned with the psychological and ethical realities of the modern learner.

Motivation and Contribution: In this study, we seek to advance the field by transitioning from traditional predictive CD to a *generative CD* paradigm. We extend the identifiable cognitive diagnosis model (ID-CDM) [26] into a flexible generative framework that fundamentally disentangles cognitive state inference from response data prediction. Within this architecture, the standard transductive prediction paradigm is repurposed as a modular component of a broader generative process, effectively addressing the aforementioned challenges. Specifically, to mitigate fairness concerns, our framework integrates identifiability and monotonicity conditions to ensure more reliable and equitable diagnostic outcomes. Furthermore, the paradigm facilitates instant diagnosis by enabling inductive inference of cognitive states, thereby bypassing the need for computationally expensive parameter reoptimization for new learners. We provide two concrete instantiations of this approach: *generative item response theory (G-IRT)* and the *generative neural cognitive diagnosis model (G-NCDM)*. Extensive experiments on real-world datasets demonstrate the effectiveness of our methods, which we believe will serve as a robust foundation for future iLearning environments.

II. Related Work

A. Cognitive Diagnosis

Traditional CDMs predominantly follow a transductive paradigm, treating learner cognitive states as learnable embeddings optimized through response score prediction. Within iLearning ecosystems, these models serve as the engine for personalization, transforming raw interaction logs into the actionable insights required to align educational content with learner needs. Early discrete models, such as DINA [27], assume binary knowledge mastery, while continuous models, such as IRT [28], [29], represent learner ability θ_i and question parameters (a_j, b_j) via logistic interaction functions, typically defined as $P(r_{ij} = 1|\theta_i, a_j, b_j) = \frac{1}{1+\exp\{-a_j(\theta_i - b_j)\}}$. Parameters in these frameworks are estimated using statistical inference methods, such as Monte Carlo Markov Chain (MCMC) [30], [31] or variational inference (VI) [32]. Multidimensional extensions (MIRT) [16] and deep learning-based approaches have further enhanced diagnostic precision. For instance, NCDM [18] utilizes neural networks to model complex learner–item interactions, while recent advancements [33], [34], [35] incorporate context-aware features and knowledge graphs. Despite their accuracy, these models rely on a transductive bottleneck: cognitive states are tied to parameter optimization during score fitting. This often results in non-identifiability and necessitates computationally expensive retraining for new learners. Such limitations in fairness and scalability pose significant hurdles for the deployment of reliable, real-time CDMs for iLearning environments.

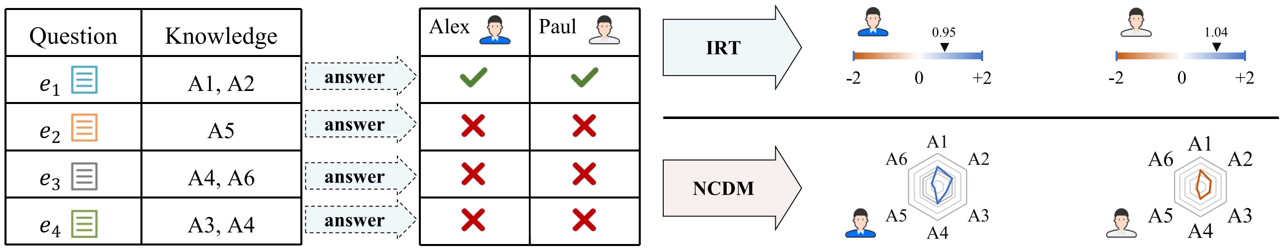


Fig. 2. Overview of CD. Given question data and response data, diagnostic models optimize parameters to fit response scores. User parameters are presented as diagnostic outputs.

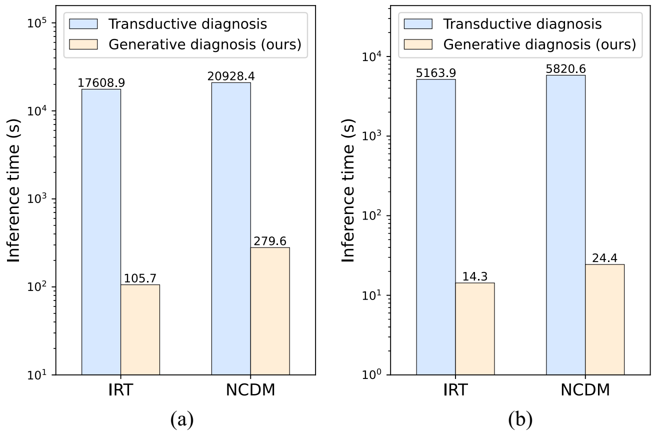


Fig. 3. Total running time for diagnosis of new learners using transductive and generative CDMs. Y-ticks are on a logarithmic scale for convenience. n denotes the number of new learners. (a) ASSIST ($n = 832$). (b) Math1 ($n = 842$).

B. Generative Model

Generative modeling [36], [37], [38], [39] has revolutionized diverse domains by learning underlying data distributions to synthesize novel, high-fidelity samples. Fundamental frameworks, such as variational autoencoders [40] and generative adversarial networks [41], established the groundwork for probabilistic latent variable modeling and adversarial training. In computer vision, diffusion models [42], [43] have achieved state-of-the-art performance by formulating generation as a reverse denoising process: $p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2 I)$. Simultaneously, natural language processing has been transformed by Transformer-based architectures, such as BERT [44], which leverage masked language modeling to encode bidirectional context via $P(x_i|\text{mask}) = \text{softmax}(W \cdot h_i + b)$. In the field of recommender systems, architectures, such as U-AutoRec [45] and collaborative denoising autoencoders (CDAE) [46], reconstruct user–item interaction matrices to capture personalized preferences, often employing user-specific noise terms. Despite these cross-domain successes, the generative paradigm remains largely underutilized within the CD community. In the context of iLearning, generative models offer profound potential: by treating cognitive states as latent variables within a generative framework, systems can perform inductive inference rather than transductive optimization. This transition allows iLearning environments to achieve instantaneous, identifiable, and fair diagnostic outputs for new learners without model retraining.

III. Preliminaries

In CD, we use S to represent the learner set, E to represent the item set, and $D = \{(s_i, e_j, y_{ij}) | s_i \in S, e_j \in E, y_{ij} \in \mathbb{R}\}$ to represent the response score set, respectively. y_{ij} represents the response score of s_i on e_j. In the classical setting of CD, response scores are binary. This is common in objective test items, such as multiple-choice questions. For subjective test items, response scores are usually on an ordinal scale.

Meanwhile, the learner latent trait set is represented by Θ. The item attribute set is represented by Ψ. One basic assumption in CD is that Θ and Ψ decide response scores. Unfortunately, Θ and Ψ are unobservable. So, CD models first assume the interaction between learner latent traits and item attributes via a well-defined IRF, then find the optimal latent variables by fitting response score data. Formally, the CD task is defined as follows.

Definition 3.1. Cognitive diagnosis: Given the learner set S, item set E, and response score set D, the goal of CD is to find the actual learner cognitive states Θ^* and test item attributes Ψ^* that can accurately reconstruct the response scores, i.e.,

$$\Theta^*, \Psi^* = \arg\min_{\Theta, \Psi, \Omega} \mathbf{E}_{(s_i, e_j, y_{ij}) \sim D} \left[\mathcal{L}\left(y_{ij}, f_\Omega(\theta_i; \psi_j)\right) \right] \quad (1)$$

where $\mathcal{L}$ is the loss function and f_Ω is the IRF.

For instance, in two-parameter item response theory (2PL-IRT) models, each learner trait θ_i is a random variable representing its overall ability, while each item attribute $\psi_j = (a_j, b_j)$ is a pair of random variables representing its discrimination level a_j and difficulty level b_j. The IRF is

$$f^{(\mathrm{IRT})}(\theta_i; \psi_j) = \frac{1}{1 + \exp(-a_j(\theta_i - b_j))}. \quad (2)$$

The IRF output is viewed as the probability of a successful binary response score (i.e., $y_{ij} = 1$). Therefore, the distance metric $\mathcal{L}^{(\mathrm{IRT})}$ is the negative log-likelihood function. That is

$$\mathcal{L}^{(\mathrm{IRT})}(y, \hat{y}) = -\left[y \log \hat{y} + (1 - y) \log(1 - \hat{y})\right]. \quad (3)$$

In previous research, the estimation of latent variables depends on directly applied parameter estimation algorithms, such as the MCMC for distribution estimation of cognitive states or gradient descent (GD) for point estimation of cognitive states.

A. Fairness Requirements of Diagnostic Results

Definition 3.2. Identifiability in CD: Let $B = \{\Theta, \Psi\}$ be diagnostic results, where Θ and Ψ denote learner cognitive state and item attribute, respectively. Let $\{f_{\mathrm{IRF}}(\boldsymbol{\theta}; \boldsymbol{\psi}) : \Theta \times \Psi \to \{0, 1\} | \boldsymbol{\theta} \in \Theta, \boldsymbol{\psi} \in \Psi\}$ be a response function family that

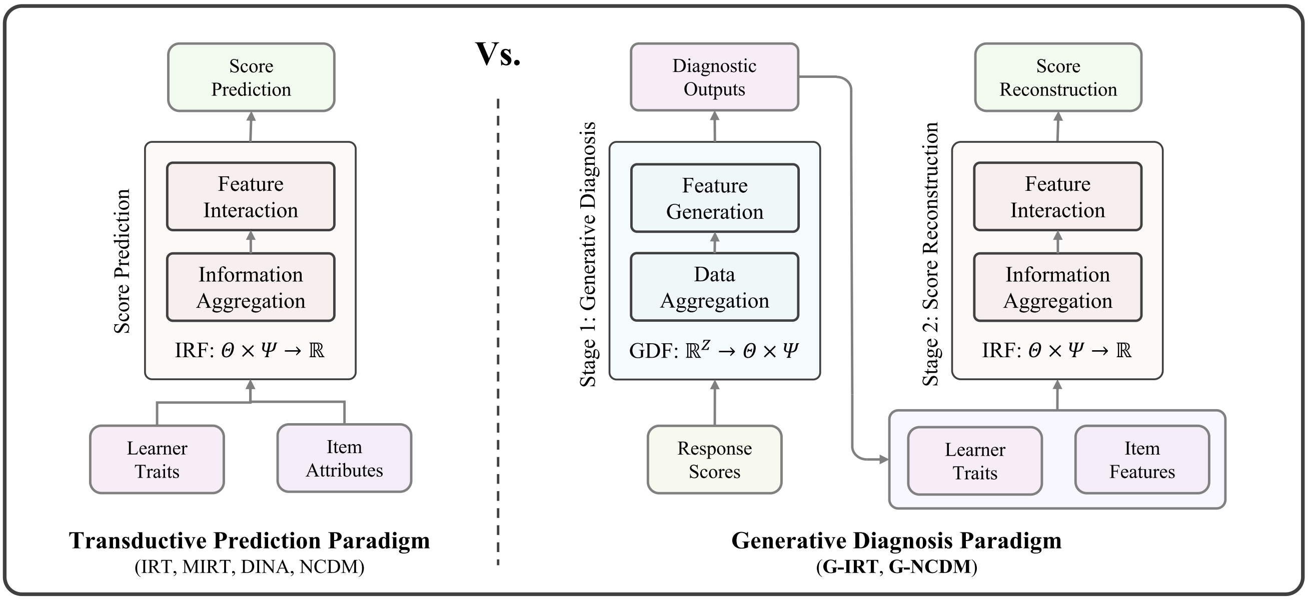


Fig. 4. Overview of (left) the classic transductive prediction paradigm and (right) the new generative diagnosis paradigm.

simulates the learner–item interaction. Then, $\boldsymbol{y}_i^{(s)} = f_{\text{IRF}}(\boldsymbol{\theta}_i; \cdot)$ denotes the score distribution of learner with trait θ_i. $\boldsymbol{y}_k^{(e)} = f_R(\cdot; \boldsymbol{\psi}_k)$ denotes the score distribution of item with feature ψ_k. Then, the diagnostic results are identifiable if and only if their distinct values lead to a distinct distribution of scores. Specifically, this is formalized as follows:

$$\boldsymbol{y}_i^{(s)} = \boldsymbol{y}_j^{(s)} \to \boldsymbol{\theta}_i = \boldsymbol{\theta}_j \tag{4}$$

$$\boldsymbol{y}_k^{(e)} = \boldsymbol{y}_l^{(e)} \to \boldsymbol{\psi}_k = \boldsymbol{\psi}_l. \tag{5}$$

Diagnostic results are identifiable if both (4) and (5) are satisfied.

For example, as presented in Fig. 2, Alex and Paul have the same response score distribution. According to the identifiability requirement, Alex and Paul should have equivalent cognitive states. However, none of the diagnostic reports of IRT and NCDM show equivalent cognitive states for them. As a result, the diagnostic reports are unidentifiable. In real-world scenarios, the nonidentifiability problem makes diagnostic results unfair to individuals.

Definition 3.3. Explainability in CD: We define the explainability of CD as the degree to which diagnostic results accurately represent a learner's genuine cognitive state.

To illustrate, if a learner is proficient in the concept of "*Inequality*," their diagnosed trait value for that concept should be sufficiently high to reflect their actual mastery. However, ensuring explainability directly is challenging since true mastery levels remain latent and unobservable. Consequently, in CD-based modeling, explainability is typically addressed indirectly through the monotonicity assumption [16], [18].

Definition 3.4. Monotonicity Assumption: This principle posits that a learner's probability of correctly solving a problem is a monotonically nondecreasing function of their proficiency in the required knowledge components [18]. Formally, the assumption is expressed as follows:

$$\boldsymbol{\theta_i}^{(l)} \succeq \boldsymbol{\theta_j}^{(l)} \Leftrightarrow y_{il} \geq y_{jl} \quad \forall s_i, s_j \in S,\ e_l \in E \tag{6}$$

where $\boldsymbol{\theta_i}^{(l)}$ represents the specific mastery level of learner s_i regarding the knowledge concepts associated with question e_l.

In transductive CDMs, satisfying this assumption typically relies on the inherent monotonicity of the interaction function [18]. Traditional models, such as DINA [27] and IRT [29], utilize linear interaction functions that naturally adhere to this property. Conversely, in neural-based CDMs, such as NCDM [18], monotonicity is enforced by constraining the interaction function's weights to be nonnegative.

IV. Generative Diagnosis Paradigm

To address the limitations of transductive models, we propose a generative diagnosis paradigm that reframes CD as a representation learning problem (see Fig. 4). This framework consists of two stages: a *generative diagnosis stage*, where a learnable generative diagnosis function (GDF) maps raw response scores directly to cognitive states, and a *score reconstruction stage*, which utilizes an IRF to recover the original scores. By establishing a functional mapping rather than relying on iterative parameter fitting, the GDF ensures structural identifiability and facilitates inductive inference. This allows the model to maintain diagnostic consistency across identical performance profiles and provide real-time results for new learners without the need for computational reoptimization.

A. Generative Diagnosis Function

The generative diagnosis is accomplished via a well-defined GDF. The GDF takes response scores as input, then directly generates cognitive states according to the input and its inner parameters. Formally, GDF itself is represented by $g_\phi : \mathbb{R}^Z \to \Theta \times \Psi$, parameterized by ϕ. Here, Z denotes the scale of the input data. GDF outputs the latent traits via a generation process from response score data

$$\boldsymbol{\theta}_i, \boldsymbol{\psi}_j = g_\phi\left(\boldsymbol{y}_i^{(s)}; \boldsymbol{y}_j^{(e)}\right) \tag{7}$$

here $y_i^{(s)}$ represents all response scores related to learner s_i, and $y_j^{(e)}$) represents all response scores related to item e_j. Overall, the GDF includes two steps: 1) data aggregation and 2) feature generation. The first step aggregates the sparse, huge response score data into a denser and smaller representation for the next step. The second step then extracts latent trait and item feature information from the condensed representation to generate diagnostic outputs. We will show how this process works using a specific instantiation of the generative diagnosis paradigm in the following sections.

By shifting cognitive state estimation from parameter estimation to diagnostic generation, the goal of CD shifts from finding the optimal Θ and Ψ to finding the optimal GDF, parameterized by ϕ. *Moreover, the cognitive state diagnosis process is disentangled from the score prediction process in the inference stage.* This enables two significant advantages of the generative diagnosis paradigm compared to the transductive prediction paradigm.

1) *Efficacy in diagnosing new-coming learners:* When new learners come, their cognitive state estimation could be instantly obtained by directly inputting their response scores to the GDF and running the generation process, without retraining the whole model.
2) *Fairness in latent trait estimation:* The stable generative diagnosis process ensures that diagnostic outputs are rigorously identifiable. In addition, this benefits the explainability and causality between response data and cognitive states generated by GDF.

B. Item Response Function

The IRF defines the interaction relationship between learner traits and item features, which is mathematically identical to the transductive prediction paradigm. Formally, IRF is represented by $f_\Omega : \Theta \times \Psi \to \mathbb{R}$, parameterized by Ω. IRF then outputs the prediction of the response score from latent traits

$$\hat{y}_{ij} = f_\Omega(\boldsymbol{\theta}_i; \boldsymbol{\psi}_j). \tag{8}$$

Within the whole generative CD paradigm, the IRF plays an assistant role in helping the learning of the GDF via response score reconstruction, rather than dominating the latent trait estimation.

C. Optimization of Generative CDMs

Given the GDF and the IRF above, the training of a generative CDM becomes a parameter optimization task, similar to the classical CD in Definition 3.1. Differently, the training and diagnosis are disentangled in generative CDMs—the model does not need to retrain to get new diagnostic outputs. Here we give a formal definition of the optimization goal.

Definition 4.1 (Optimization Goal of Generative Diagnosis Paradigm): Given the GDF g_ϕ, the IRF f_Ω, the learner set S, the item set E, and the response score set D, the goal of the generative diagnosis paradigm is to optimize the generative diagnosis process $\boldsymbol{\theta}_i, \boldsymbol{\psi}_j = g_\phi(\boldsymbol{y}_i^{(s)}; \boldsymbol{y}_j^{(e)})$ by finding optimal function parameter ϕ^* and Ω^* such that

$$\phi^*, \Omega^* = \arg\min_{\phi,\omega} \mathbf{E}_{(s_i,e_j,y_{ij})\sim D}\left[\mathcal{L}\left(y_{ij}, f_\Omega(\boldsymbol{\theta}_i; \boldsymbol{\psi}_j)\right)\right]. \tag{9}$$

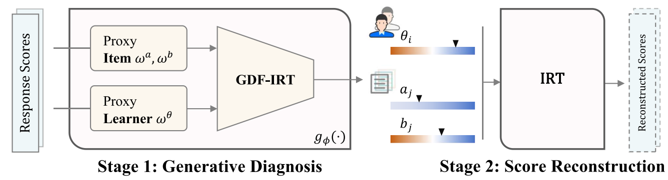


Fig. 5. Overview of G-IRT model.

To demonstrate the universality of the proposed paradigm, we instantiate it within two distinct and representative CD frameworks: the classical IRT and the modern NCDM. These implementations, denoted as G-IRT and G-NCDM, respectively, serve to verify the paradigm's adaptability across different modeling architectures (probabilistic versus neural). In the following sections, we detail the specific formulation of these instantiations, beginning with the integration of the generative mechanism into the IRT framework.

V. G-IRT: Generative IRT

In this section, we introduce the G-IRT, which could obtain learner and item traits via feature generation without changing the original IRT structure. An overview of G-IRT is presented in Fig. 5.

A. Challenges in Latent Trait Estimation of IRT

IRT is a classical CDM that utilizes machine learning techniques to optimize learner and item parameters. Challenges in latent trait estimation include *controllability and efficiency*. The *controllability* challenge lies in the interdependence of latent traits during the estimation process. To jointly estimate respondent and item parameters, various estimation algorithms, such as MCMC, VI and GD, were applied to the estimation of IRT parameters [32], [47], [48]. However, it is nontrivial to interpret and control the relationship between latent traits and specific score distributions because parameters are deductively learned. The *efficiency* challenge lies in the incremental estimation of respondent traits, which is common in the deployment phase of an IRT model. Every time a new learner arrives, the IRT model needs to be retrained. The accumulated time cost could be high when the number of new respondents is huge.

B. GDF of G-IRT

We demonstrate the GDF of G-IRT using 2PL-IRT. In the traditional 2PL-IRT, the IRF is defined as

$$P(y_{ij}|\theta_i, a_j, b_j) = \frac{1}{1 + \exp(-a_j(\theta_i - b_j))}. \tag{10}$$

Ideally, if we know the values of two of the three latent traits θ_i, a_j, and b_j, together with the correct probability, we could induce the last latent trait. Formally, we first use the sigmoid function symbol $\sigma(x) = 1/(1+e^{-x})$ to simplify the IRF representation. Through equivalent transformation of the IRF, we can represent any of the latent traits by the rest of the latent traits with the scoring probability, as shown in the following equations:

$$\theta_i = b_j + \frac{\sigma^{-1}(P_{ij})}{a_j}, \quad \text{if } P_{ij}, a_j, b_j \text{ are known} \tag{11}$$

$$a_j = \frac{\sigma^{-1}(P_{ij})}{\theta_i - b_j}, \quad \text{if } P_{ij}, b_j, \theta_i \text{ are known} \tag{12}$$

$$b_j = \theta_i - \frac{\sigma^{-1}(P_{ij})}{a_j}. \quad \text{if } P_{ij}, \theta_i, a_j \text{ are known.} \tag{13}$$

Practically, the calculation of (11)–(13) is nontrivial because we have no information about the latent traits. The only evidence we have is the learners' score matrix. However, an approximate induction of latent traits can be efficiently achieved using G-IRT. The core intuition is to *resolve the interdependence* of latent traits by introducing "proxy parameters" in the GDF. These proxies serve as the learnable inputs that drive the feature generation process. Specifically, they act as stand-ins for the unobserved priors required in the analytical inversions [see (11)–(13)]. By replacing unknown true traits with these learnable proxies, we decouple the estimation, allowing parameters to be generated via a direct forward pass. Note that these proxies are auxiliary variables for the generative process and are distinct from the final diagnostic learner/item traits.

In G-IRT, we first define the latent trait estimation conditioned on proxy parameters as follows:

$$\theta_i(\omega_j^a, \omega_j^b) = \omega_j^b + \frac{\lambda R_{ij}}{\omega_j^a} \tag{14}$$

$$a_j(\omega_i^\theta, \omega_j^b) = \left| \frac{\lambda R_{ij}}{\omega_i^\theta - \omega_j^b} \right| \tag{15}$$

$$b_j(\omega_i^\theta, \omega_j^a) = \omega_i^\theta - \frac{\lambda R_{ij}}{\omega_j^a} \tag{16}$$

where $\lambda \in \mathbb{R}_+$ is a hyperparameter. In this definition, $\sigma^{-1}(P_{ij})$ is approximated by λR_{ij}. $\omega_\cdot^\theta$, $\omega_\cdot^a$, and $\omega_\cdot^b$ are the *proxy parameters* for estimating latent traits. Now the problem is, for each observed $R_{ij} \neq 0$, we can calculate an estimated value, namely, $\theta_i(\omega_j^a, \omega_j^b)$. We aim find a general θ_i that can minimize its distance between every $\theta_{ij}^{\omega}, j = 1, 2, \ldots, M$. Our goal is to

$$\hat{\theta}_i = \arg\min_{\theta} \sum_{j:R_{ij}\neq 0} (\theta - \theta_i(\omega_j^a, \omega_j^b))^2. \tag{17}$$

We can establish similar goal for $\hat{a}_j$ and $\hat{b}_j$. The solution of them is

$$\theta_i = \frac{1}{Z_i^{(s)}} \sum_{j:R_{ij}\neq 0} \theta_i(\omega_j^a, \omega_j^b) \tag{18}$$

where $Z_i^{(s)} = \sum_{j=1}^{M} \mathbb{I}(R_{ij} \neq 0)$. Similarly, for item parameters, we have the following optimal estimation:

$$a_j = \frac{1}{Z_j^{(e)}} \sum_{i:R_{ij}\neq 0} a_j(\omega_i^\theta, \omega_j^b) \tag{19}$$

Algorithm 1: Training of G-IRT.

Input: $D = \{(s_i, e_j, y_{ij}) | s_i \in S, e_j \in E, y_{ij} \in \mathbb{R}\}$, the training dataset
Output: g_ϕ, the GDF of a trained G-IRT
Initialize g_ϕ;
for epoch in $1 \ldots T$ **do**
  for $(s_i, e_j, y_{ij}) \in D$ **do**
    $\boldsymbol{y}_i^{(s)} \leftarrow [y_{ij'} | j' : (s_i, e_{j'}, y_{ij'}) \in D]$;
      ▷ Obtain the learner response vector for s_i
    $\boldsymbol{y}_j^{(e)} \leftarrow [y_{i'j} | i' : (s_{i'}, e_j, y_{i'j}) \in D$;
      ▷ Obtain the item response vector for e_j
    $\theta_i, a_j, b_j \leftarrow g_\phi(\boldsymbol{y}_i^{(s)}; \boldsymbol{y}_j^{(e)})$;
      ▷ Stage 1: Feature generation for learner and item
    $\hat{y}_{ij} \leftarrow f^{(IRT)}(\theta_i, a_j, b_j)$;
      ▷ Stage 2: Score reconstruction using IRT
  end for
  $g_\phi \leftarrow$ update$(\mathcal{L}(y, \hat{y}), g_\phi)$;
      ▷ Update GDF using gradient descent
end for
return g_ϕ;

Algorithm 2: Instant Diagnosis Using G-IRT.

Input: g_ϕ, the GDF of a trained G-IRT; D_i, the response scores of a new-coming learner s_i
Output: θ_i, the learner ability value
$\boldsymbol{y}_i^{(s)} \leftarrow$ transform(D_i); ▷Transform the D_i into response score vector, which scores indexed by item IDs.
$\theta_i \leftarrow g_\phi^{(s)}(\boldsymbol{y}_i^{(s)})$; ▷Generative diagnosis
return θ_i;

$$b_j = \frac{1}{Z_j^{(e)}} \sum_{i:R_{ij}\neq 0} b_j(\omega_i^\theta, \omega_j^a). \tag{20}$$

To formalize these definitions to GDF, we first replace the response log $R_{ij} \in \{-1, 0, 1\}$ with $y_{ij} \in \{0, 1\}$. Since $R_{ij} = 1$ denotes correct response, $R_{ij} = 0$ denotes no response, and $R_{ij} = -1$ denotes incorrect response, we have $R_{ij} = 2y_{ij} - 1 \forall R_{ij} \neq 0$. $R_{ij} = 0$ means y_{ij} does not exist in the log. As a result, we can define the GDF of G-IRT as follows:

$$g_\phi(y_i^{(s)}; y_j^{(e)}) \equiv \begin{cases} \theta_i = \frac{1}{|\boldsymbol{y}_i^{(s)}|} \sum_{j:y_{ij}\in y_i^{(s)}} \left[\omega_j^b + \frac{\lambda(2y_{ij}-1)}{\omega_j^a} \right] \\ a_j = \frac{1}{|\boldsymbol{y}_j^{(e)}|} \sum_{i:y_{ij}\in y_j^{(e)}} \left| \frac{\lambda(2y_{ij}-1)}{\omega_i^\theta - \omega_j^b} \right| \\ b_j = \frac{1}{|\boldsymbol{y}_j^{(e)}|} \sum_{i:y_{ij}\in y_j^{(e)}} \left[\omega_i^\theta - \frac{\lambda R_{ij}}{\omega_j^a} \right] \end{cases} \tag{21}$$

where $\phi = (\omega^\theta, \omega^a, \omega^b)$ is also the optimization target of the GDF. The training and diagnosis algorithms of G-IRT are shown in Algorithms 1 and 2, respectively.

C. Analysis of G-IRT

1) Controllability of Parameter Estimation: A common assumption of IRT is that latent traits follow a specific but unobserved distribution. For traditional IRT models, setting an appropriate prior distribution for latent traits (e.g., standard Gaussian distribution for θ) could control the posterior distribution of them to control their scale. Controllable scale help learners to understand their relative ability compared to the whole learner group. For G-IRT, however, the scale of latent traits is decided by the GDF parameter ϕ rather than their prior distribution. Here, we analyze how to control the scale of ϕ to control the parameter scale of θ, a, and b.

For convenience, we assume that every R_{ij} is nonzero. We further assume the following hyperparameters: let $\omega_j^b \in (\beta_{\min}, \beta_{\max}), j = 1, \ldots, M$, let $\omega_i^\theta \in (\beta_{\min}, \beta_{\max}), i = 1, \ldots, N$, and let $\omega_j^a \in (\epsilon, \zeta), j = 1, \ldots, M$.

Parameter scale of θ_i: We start with discussing the range of the respondent parameter θ_i. For the upper bound of θ_i, we have

$$\theta_i \leq \overline{\omega_b} + \lambda \left(\frac{1}{M} \sum_{j=1}^{M} \frac{1}{\omega_j^a} \right) < \beta_{\max} + \frac{\lambda}{\epsilon}. \tag{22}$$

We aim to limit the range of θ_i within $(\theta_{\min}, \theta_{\max})$. Similarly, we can obtain the lower bound of θ. By applying them to the range (p, q), we get the constraint of λ

$$\lambda \leq \min\{\epsilon(\theta_{\max} - \beta_{\max}), \epsilon(\beta_{\min} - \theta_{\min})\}. \tag{23}$$

Parameter scale for cold-start setting: A significant advantage of G-IRT is its utility in cold-start learner modeling. Consider a new learner with its response score that has never been observed in training data. The G-IRT could directly input its response score vector to the GDF to obtain its ability. However, the cold-start performance of G-IRT is not only decided by its training procedure, but the hyperparameter setting. Considering this, we aim to appropriately set the hyperparameters of G-IRT so that it can at least handle some special cases. Specifically, we discuss a special case of all-correct response data, i.e., $R = \mathbf{1}_{N\times M}$. The anticipated learner ability should always be greater than the item difficulty. By computing these values, we have

$$\theta_i = \frac{1}{M} \sum_{j=1}^{M} \omega_j^b + \frac{\lambda}{M} \sum_{j=1}^{M} \frac{1}{\omega_j^a} \tag{24}$$

$$b_j = \frac{1}{N} \sum_{i=1}^{N} \omega_i^\theta - \lambda \frac{1}{\omega_j^a} \tag{25}$$

$$\theta_i - b_j = \overline{\omega_b} - \overline{\omega_\theta} + \lambda \left(\frac{1}{\omega_j^a} + \frac{1}{M} \sum_{j'=1}^{M} \frac{1}{\omega_{j'}^a} \right). \tag{26}$$

We expect that $(\theta_i - b_j)$ is greater than zero so that G-IRT has a good cold-start performance. Therefore, we limit

$$\lambda > \frac{\max(\overline{\omega_b} - \overline{\omega_\theta})}{\min(\frac{1}{\omega_j^a} + \frac{1}{M} \sum_{j'=1}^{M} \frac{1}{\omega_{j'}^a})} \geq \frac{\zeta}{2}(\beta_{\max} - \beta_{\min}). \tag{27}$$

Surprisingly, the constraint for cold-start setting leads to a lower bound of λ, which jointly works with the upper bounds of λ mentioned above to decide its region.

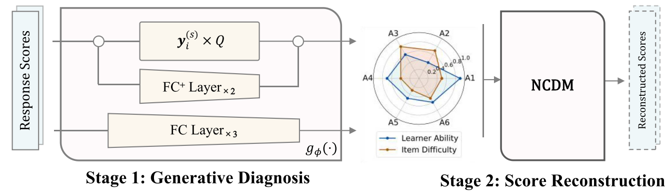


Fig. 6. Overview of the G-NCDM.

Make lower and upper bounds compatible: We aim to let the lower bound and upper bound of λ be compatible to simultaneously satisfy the two properties. Therefore, we have

$$\frac{\zeta}{\epsilon} \leq \min \left\{ \frac{2(\beta_{\min} - \theta_{\min})}{\beta_{\max} - \beta_{\min}}, \frac{2(\theta_{\max} - \beta_{\max})}{\beta_{\max} - \beta_{\min}} \right\}. \tag{28}$$

VI. G-NCDM: Generative NCDM

In this section, we introduce the application of the generative diagnosis paradigm in deep learning-based CDMs. We propose the G-NCDM,[1] a simple yet effective method that captures the complex mapping between response scores and features. An overview of G-NCDM is presented in Fig. 6.

A. Challenges in Deep Learning-Based CDMs

As illustrated in Fig. 2, current deep learning-based CDMs characterize learner proficiency through knowledge mastery levels. *A primary limitation of these models is the coupling of model training with trait diagnosis, which prevents real-time diagnostic capabilities and compromises diagnosis fairness.* Regarding the latter, deep learning-based CDMs are particularly susceptible to issues of nonidentifiability and explainability overfitting. *Identifiability*, a cornerstone of CD-based modeling extensively discussed in the literature [49], [50], requires that diagnostic outcomes distinguish between learners who exhibit different response distributions. Essentially, distinct latent traits must produce distinct observable data patterns. Furthermore, *explainability* refers to the extent to which diagnostic results accurately mirror a learner's *true* cognitive state. In DL-CDMs, this is typically quantified by the monotonic relationship between mastery levels and actual response scores. However, our empirical analysis reveals that existing CDMs are prone to "explainability overfitting." In this phenomenon, diagnostic results maintain high explainability on training data but fail to generalize to unobserved test data.

B. GDF of G-NCDM

The design of the GDF within the G-NCDM framework aims to maintain diagnostic precision while satisfying both identifiability and monotonicity requirements. To achieve this, it utilizes fully connected layers with specific parameter constraints, allowing the model to learn the diagnostic process from data while strictly adhering to these two pivotal conditions. The GDF

[1] G-NCDM is developed from our proposed ID-CDM [26].

is formally defined as follows:

$$g_\phi\left(\boldsymbol{y}_i^{(s)};\boldsymbol{y}_j^{(e)}\right) \equiv \left\{g_\phi^{(s)}\left(\boldsymbol{y}_i^{(s)}\right);g_\phi^{(e)}\left(\boldsymbol{y}_j^{(e)}\right)\right\} \tag{29}$$

$$\boldsymbol{\theta}_i^{(\text{imp})} = \text{FC}_{\times 2}^{+}\left(2\boldsymbol{y}_i^{(s)}-1\right) \tag{30}$$

$$\boldsymbol{\theta}_i^{(\text{exp})} = \sigma\left(\frac{\left(2\boldsymbol{y}_i^{(s)}-1\right)\times Q}{\sqrt{K}}\right) \tag{31}$$

$$\boldsymbol{\theta}_i = (1-\alpha)\cdot\boldsymbol{\theta}_i^{(\text{imp})}+\alpha\cdot\boldsymbol{\theta}_i^{(\text{exp})} \tag{32}$$

$$\boldsymbol{\psi}_j = \text{FC}_{\times 3}\left(2\boldsymbol{y}_j^{(e)}-1\right) \tag{33}$$

where all activation functions are sigmoid. The $\text{FC}^{+}(\cdot)$ denotes a fully connected layer with nonnegative weight parameters, which is designed to ensure the monotonicity between response scores and knowledge mastery degrees.

C. IRF of G-NCDM

Within the IRF, we employ neural networks to capture the complex interactions between learners and items. Specifically, we first apply single-layer perceptrons to aggregate conceptwise diagnostic results into low-dimensional latent features, thereby obtaining more robust representations of both learners and questions. These aggregated features are then processed through fully connected layers to reconstruct the final response scores. The diagnostic output aggregation layer is formally defined as follows:

$$\boldsymbol{\theta}_{i,j}^{(\text{dense})} = \text{FC}^{+}(\boldsymbol{\theta}_i\odot\boldsymbol{q}_j) \tag{34}$$

$$\boldsymbol{\psi}_j^{(\text{dense})} = \text{FC}(\boldsymbol{\psi}_j\odot\boldsymbol{q}_j) \tag{35}$$

where q_j is the jth row of the expert-labeled Q-matrix [51]. q_j is for masking unnecessary knowledge dimensions to ensure that the optimization only targets for knowledge dimensions required by the item. For $\boldsymbol{q}_j = (q_{j,1},\ldots,q_{j,K}), q_{j,l}, l = 1,2,\ldots,K$, is a binary value denotes whether knowledge dimension l is required by item j.

Subsequently, the aggregated representations of s_i and e_j are fed into a three-layer fully connected neural network to predict the response probability, defined as follows:

$$\hat{y}_{ij} = \text{FC}_{\times 3}\left(\boldsymbol{\theta}_{i,j}^{(\text{dense})}-\boldsymbol{\psi}_j^{(\text{dense})}\right). \tag{36}$$

Finally, neural network parameters of the GDF and the IRF are learnable parameters. The loss function of G-NCDM is the cross-entropy loss between actual response scores and the reconstructed correct probabilities. The model can be trained via GD algorithms. Specifically, the training and diagnosis algorithms of G-NCDM are shown in Algorithms 3 and 4, respectively.

Algorithm 3: Training of G-NCDM.

Input: $D = \{(s_i,e_j,y_{ij})|s_i\in S, e_j\in E, y_{ij}\in\mathbb{R}\}$, the training dataset; $Q = (q_{jl})_{M\times K}$, the item-knowledge Q-matrix
Output: g_ϕ, the GDF of a trained G-NCDM

```
Initialize g_φ and f_ω;
for epoch in 1...T do
  for (s_i, e_j, y_ij) ∈ D do
    y_i^(s) ← [y_ij' | j' : (s_i, e_j', y_ij') ∈ D];   ▷Obtain the learner response vector for s_i
    y_j^(e) ← [y_i'j | i' : (s_i', e_j, y_i'j) ∈ D;    ▷Obtain the item response vector for e_j
    θ_i, ψ_j ← g_φ(y_i^(s); y_j^(e));                  ▷Stage 1: Feature generation for learner and item
    ŷ_ij ← f_ω(θ_i, ψ_j; Q);                           ▷Stage 2: Score reconstruction using neural network
  end for
  g_φ ← update(L(y, ŷ), g_φ);                          ▷Update GDF using gradient descent
end for
return g_φ;
```

Algorithm 4: Instant Diagnosis Using G-NCDM.

Input: g_ϕ, the GDF of a trained G-NCDM; D_i, the response scores of a new-coming learner
Output: $\boldsymbol{\theta}_i$, the learner ability value

```
y_i^(s) ← transform(D_i);      ▷Transform the D_i into a response score vector, which scores indexed by item IDs.
θ_i ← g_φ^(s)(y_i^(s));        ▷Generative diagnosis
return θ_i;
```

VII. Experiments

A. Experiment Overview

We aim to answer the following research questions via experiments.

1) *RQ1 (Instant diagnosis for new learners):* How is the score reconstruction performance of generative CDMs?
2) *RQ2 (Offline diagnosis for existing learners):* How is the score prediction performance of generative and transductive CDMs?
3) *RQ3 (Reliability of diagnosis):* How is the identifiability and explainability of diagnostic outputs of different CDMs?
4) *RQ4 (Statistical features of diagnosis):* How is the statistical feature of diagnostic outputs of generative CDMs?
5) *RQ5 (Utility of diagnosis):* How can generative CDMs be used in real-world applications?

B. Experiment Setup

1) Datasets: In this study, we evaluate the performance of G-IRT and G-NCDM on two representative intelligent education datasets, the ASSIST (ASSISTments 2009-2010 "skill builder") [52] and the Math1 [53] dataset. These datasets reflect two classical scenarios of CD, including 1) *online learning platforms*, which have a large number of learners, various test items, and sparse response scores, and 2) *offline ability tests*, which have a large number of learners, fewer but well-designed test items and dense response scores. Details of the datasets are available in Table I. In experiments, we split datasets by

TABLE I
SUMMARY OF EXPERIMENT DATASETS

Statistics	ASSIST-0910	Math1
Scenario	Online learning platform	Offline ability test
# Learners	4163	4209
# Items	17 746	20
# KCs	123	11
# KCs per item	1.19±0.47	3.35±1.31
# Scores	324,572	84,180
# Scores per learner	107.26±155.25	20.0±0.0
Sparsity rate	99.6%	0.0%
Correct rate	65.4%	42.4%

$D_{\text{train}} : D_{\text{valid}} : D_{\text{test}} = 70\% : 10\% : 20\%$. We split datasets using two methods: the *user split* for score reconstruction and the *random split* for score prediction. The user split ensures that the train, test, and validation sets have disjoint learner groups. The random split ensures that every learner is shared by the train, test, and validation sets.

2) Baselines: We compare generative CDMs with two categories of baselines. 1) *Transductive CDMs* for score prediction. The baseline includes DINA [27], IRT [16], MIRT [16], and NCDM [18]. Transductive CDMs model learner and item features via score prediction. These models can predict response scores for existing learners and items, but *cannot instantly diagnose new learners' cognitive stats and reconstruct response scores*. For the fairness of comparison, we further choose 2) *encoder–decoder user modeling methods* for score reconstruction. The baseline includes U-AutoRec [45] and CDAE [46].

3) Implementation Setting: We implement all models with PyTorch using Python.[2] For deep learning models, the parameters are initialized with Xavier normal [54]. The optimization algorithm is Adam [55]. All experiments are run on a Linux server with one NVIDIA Tesla-A100 GPU and two 2.10 GHz Intel Xeon E5-2620 v4 CPUs.

4) Hyperparameter Setting: For G-IRT, we set $\beta_{\min} = 2.5$, $\beta_{\max} = 3.5$, $\epsilon = 1$, $\zeta = 3$, $\theta_{\min} = 1$, $\theta_{\max} = 5$, and $\lambda = 1.5$. This hyperparameter setting satisfies constraints in (23), (27), and (28). For G-NCDM, we set $\alpha = 0.9$.

C. Experiment Analysis

1) Score Reconstruction Performance (RQ1): Table II presents the score reconstruction performance results for RQ1, comparing our proposed generative CDMs against traditional encoder–decoder baselines on two real-world datasets. The results demonstrate the superior performance of the generative paradigm across multiple evaluation metrics. On the ASSIST dataset, both G-IRT and G-NCDM achieve substantial improvements over the encoder–decoder methods, with G-NCDM reaching the highest accuracy (0.735) and best RMSE (0.433), while G-IRT attains the best F1-Score (0.827). Notably, even G-IRT without training achieves a competitive F1-Score performance (0.816), highlighting the inherent effectiveness of the generative modeling approach. On the Math1 dataset, G-IRT demonstrates consistent superiority with the highest accuracy (0.782) and best RMSE (0.408), while G-NCDM shows comparable performance to the best encoder–decoder baseline CDAE. The performance gains are particularly pronounced on the ASSIST dataset, where generative models achieve accuracy improvements of approximately 3% –4% and F1-Score improvements of 2% –3% over traditional methods. These results validate that the generative paradigm effectively addresses the score reconstruction task while maintaining the capability for instant diagnosis of new learners without requiring parameter re-optimization, thus confirming the practical advantages of our proposed approach.

2) Score Prediction Performance (RQ2): Table III presents the score prediction performance results for RQ2, evaluating both generative and transductive CDMs on existing learners using random data splits. The results reveal that our proposed generative models achieve remarkable performance, often surpassing traditional transductive methods that were specifically designed for offline diagnosis scenarios. On the ASSIST dataset, G-NCDM demonstrates superior performance across all metrics, achieving the highest accuracy (0.734), F1-score (0.811), and best RMSE (0.428), outperforming the best transductive baseline NCDM by 1.4% in accuracy and 1.9% in F1-Score. Similarly, on the Math1 dataset, both G-IRT and G-NCDM achieve identical and superior accuracy (0.734) compared to the best transductive method NCDM (0.727), with G-NCDM showing the most balanced performance with the highest F1-Score (0.710) and best RMSE (0.413). Notably, even G-IRT without training demonstrates competitive performance, particularly on the ASSIST dataset, where it achieves 0.700 accuracy and 0.810 F1-Score, highlighting the inherent strength of the generative modeling paradigm. These results are particularly significant as they demonstrate that generative models not only excel at instant diagnosis for new learners (as shown in RQ1) but also maintain competitive or superior performance in traditional offline diagnosis scenarios, effectively bridging the gap between inductive and transductive learning paradigms in CD.

3) Reliability of Diagnostic Outputs (RQ3): We analyze the reliability of diagnostic outputs from their identifiability and explainability.

Identifiability of diagnostic results: To quantitatively assess the identifiability of CDMs on augmented data, we utilize the identifiability score (IDS) [26]. This metric serves as a proxy for identifiability: a higher IDS indicates greater consistency between the diagnostic results of original and shadow learners (or questions). A maximum IDS signifies rigorous identifiability. Specifically, the IDS for learner traits Θ is defined as follows:

$$\text{IDS}(\Theta) = \frac{1}{Z_{\text{IDS}}} \sum_{i \in S} \sum_{j \in S} \frac{I(\boldsymbol{r}_i = \boldsymbol{r}_j) \wedge I(i \neq j)}{[1 + \text{dist}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)]^2} \tag{37}$$

where $Z_{\text{IDS}} = \sum_{i \in S} \sum_{j \in S} I(\boldsymbol{r}_i = \boldsymbol{r}_j) \wedge I(i \neq j)$. The $\text{dist}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)$ is the distance measurement of diagnostic results. In instantiation, we choose the Manhattan distance [56] as the measurement. $\text{IDS}(\Theta)$ is monotonically decreasing at $\text{dist}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)$. Furthermore, *learner traits are identifiable if and only if* $\text{IDS}(\Theta) = 1$. Along this line, the identifiability of question features Ψ is also evaluated by calculating $\text{IDS}(\Psi)$. Table IV presents the identifiability analysis results for RQ3, revealing a fundamental

[2] Our code is available at https://github.com/CSLiJT/Generative-CD.git.

TABLE II
SCORE RECONSTRUCTION PERFORMANCE OF GENERATIVE USER/LEARNER MODELING (RQ1)

Paradigm	Model	ASSIST (User Split)			Math1 (User Split)		
		ACC↑	F1-Score↑	RMSE↓	ACC↑	F1-Score↑	RMSE↓
Encoder –decoder	U-AutoRec	0.707±0.001	0.793±0.001	0.438±0.001	0.749±0.003	0.728±0.006	0.419±0.005
	CDAE	0.701±0.001	0.796±0.001	0.449±0.001	0.758±0.005	0.746±0.003	0.425±0.003
Generative CD	G-IRT w/o train	0.707±0.001	0.816±0.001	0.533±0.002	0.781±0.001	0.728±0.001	0.411±0.001
	G-IRT	0.734±0.001	**0.827±0.001**	0.451±0.001	**0.782±0.002**	0.731±0.002	**0.408±0.003**
	G-NCDM	**0.735±0.001**	0.822±0.001	**0.433±0.001**	0.749±0.003	**0.747±0.002**	0.420±0.002

The bold highlights the highest (optimal) metric score among methods. The underlined highlights the second-highest (suboptimal) metric score among methods.

TABLE III
SCORE PREDICTION PERFORMANCE OF TRANSDUCTIVE/GENERATIVE CD (RQ2)

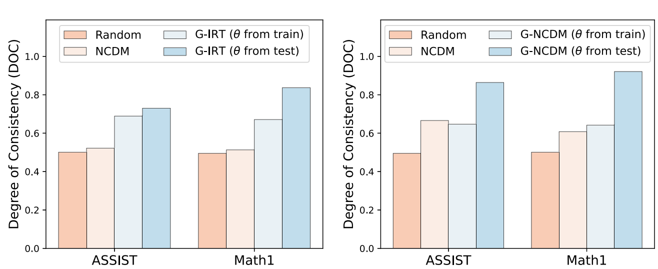
Paradigm	Model	ASSIST (Random Split)			Math1 (Random Split)		
		ACC↑	F1-Score↑	RMSE↓	ACC↑	F1-Score↑	RMSE↓
Transductive CD	DINA	0.665±0.001	0.483±0.001	0.800±0.002	0.588±0.001	0.682±0.001	0.475±0.001
	IRT	0.673±0.002	0.792±0.001	0.462±0.002	0.703±0.002	0.702±0.001	0.443±0.001
	MIRT	0.697±0.002	0.773±0.001	0.472±0.001	0.708±0.001	0.667±0.001	0.440±0.001
	NCDM	0.720±0.008	0.792±0.001	0.433±0.002	0.727±0.001	0.668±0.002	0.416±0.001
Generative CD	G-IRT w/o train	0.700±0.001	0.810±0.001	0.530±0.001	0.731±0.001	0.649±0.001	0.427±0.001
	G-IRT	0.719±0.001	0.800±0.001	0.440±0.001	**0.734±0.001**	0.657±0.001	0.425±0.001
	G-NCDM	**0.734±0.002**	**0.811±0.001**	**0.428±0.002**	**0.734±0.001**	**0.710±0.001**	**0.413±0.001**

The bold highlights the highest (optimal) metric score among methods. The underlined highlights the second-highest (suboptimal) metric score among methods.

TABLE IV
IDS ↑ OF DIAGNOSTIC RESULTS (RQ3)

Paradigm	Model	IDS ↑ of Learner Diagnostic Result Θ			IDS ↑ of Question Diagnostic Result Ψ		
		ASSIST	Math1	$\mathcal{I}(\Theta)$	ASSIST	Math1	$\mathcal{I}(\Psi)$
Transductive CD	DINA	0.550±0.003	0.451±0.006	✗	0.208±0.001	0.193±0.019	✗
	IRT	0.691±0.004	0.690±0.004	✗	0.376±0.001	0.543±0.035	✗
	MIRT	0.047±0.001	0.046±0.000	✗	0.041±0.000	0.085±0.005	✗
	NCDM	0.857±0.001	0.662±0.005	✗	0.616±0.000	0.420±0.012	✗
	NCDM-Const	0.897±0.001	0.688±0.003	✗	0.968±0.000	0.915±0.010	✗
Generative CD	G-IRT	**1.000±0.000**	**1.000±0.000**	✔	**1.000±0.000**	**1.000±0.000**	✔
	G-NCDM	**1.000±0.000**	**1.000±0.000**	✔	**1.000±0.000**	**1.000±0.000**	✔

$\mathcal{I}(X)$ indicates whether X is identifiable.
The bold highlights the highest (optimal) metric score among methods. The underlined highlights the second-highest (suboptimal) metric score among methods.



(a) (b)

Fig. 7. Degree of consistency (DOC↑) of diagnostic results (RQ3). (a) Overall abilities. (b) Knowledge proficiencies.

advantage of generative CDMs in terms of diagnostic reliability. The results demonstrate a stark contrast between traditional transductive and our proposed generative approaches in terms of identifiability guarantees. All traditional transductive methods (DINA, IRT, MIRT, NCDM, and NCDM-Const) fail to achieve identifiability for both learner diagnostic results (Θ) and question diagnostic results (Ψ), as indicated by the "✗" symbols, with their IDS ranging from as low as 0.041 (MIRT) to 0.968 (NCDM-Const) across different datasets. In sharp contrast, both G-IRT and G-NCDM achieve perfect identifiability with IDS scores of 1.000 for both learner and question diagnostics on both ASSIST and Math1 datasets, as confirmed by the "✔" symbols. This perfect identifiability ensures that the diagnostic outputs produced by our generative models are theoretically guaranteed to be unique and reliable, addressing a critical limitation of traditional CD methods. The superior identifiability performance of generative models stems from their well-designed generation process that explicitly incorporates identifiability conditions, providing practitioners with diagnostic results they can trust

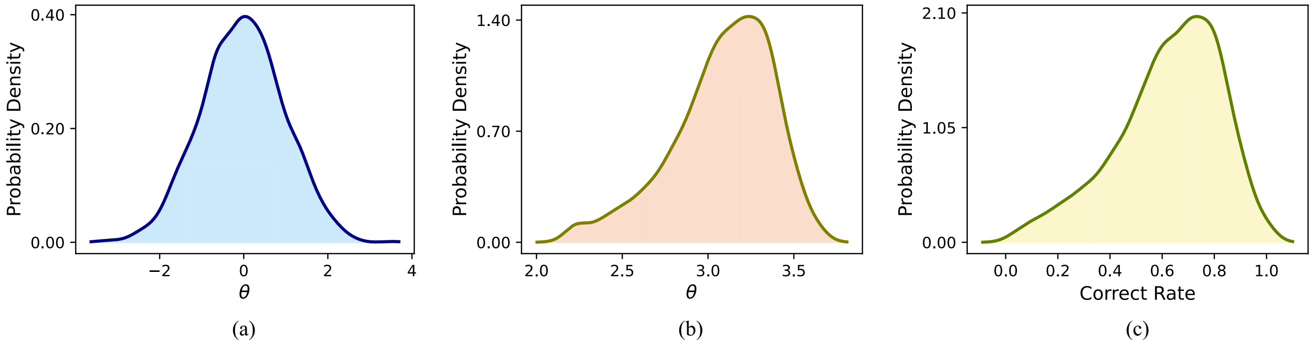


Fig. 8. Visualization of diagnostic results using KDE. Learners w/o response logs are removed from the visualization. (a) IRT@ASSIST. (b) G-IRT@ASSIST. (c) Correct@ASSIST.

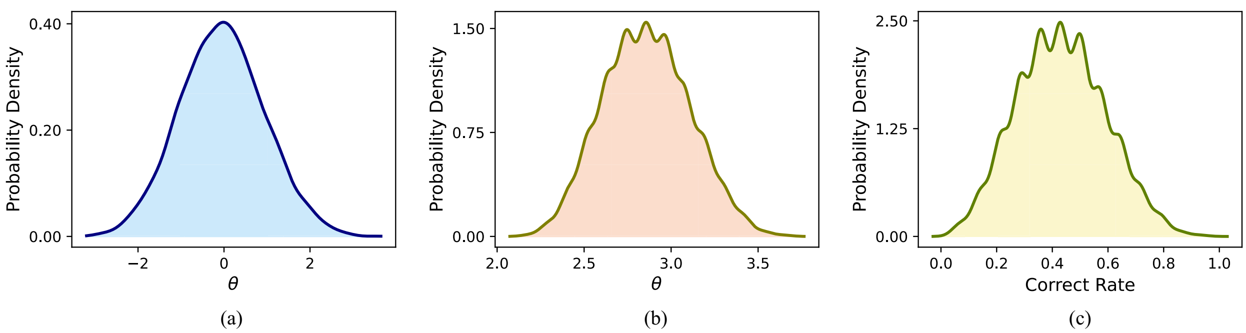


Fig. 9. Visualization of diagnostic results using KDE. Learners w/o response logs are removed from the visualization. (a) IRT@Math1. (b) G-IRT@Math1. (c) Correct@Math1.

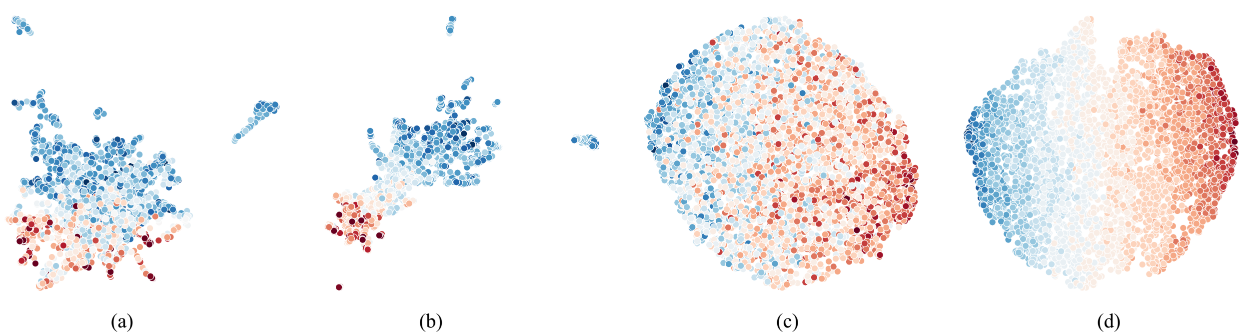


Fig. 10. Visualization of diagnostic results using u-map, colored with learner score rates. Learners w/o response logs are removed from the visualization. (a) NCDM@ASSIST. (b) G-NCDM@ASSIST. (c) NCDM@Math1. (d) G-NCDM@Math1.

TABLE V
Raw Response Score Cases of Learners in Math1(RQ5)

Learner	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1737	✓	✓	✓	✓	✗	✗	✓	✓	✓	✗	✓	✓	✗	✓	✗	✓	✗	✗	✗	✗
2094	✓	✗	✗	✓	✗	✗	✓	✓	✗	✗	✗	✗	✗	✓	✗	✗	✓	✗	✗	✗

Each column denotes response scores on an item.

TABLE VI
Knowledge Correct Rates in Math1(RQ5)

Learner	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11
1737	0.67	1.0	0.43	1.0	0.75	0.33	0.0	0.40	0.50	0.42	0.59
2094	0.33	1.0	0.0	0.0	1.0	0.17	0.0	0.0	0.50	0.42	0.35

Each column denotes average correct rates on items requiring the knowledge.

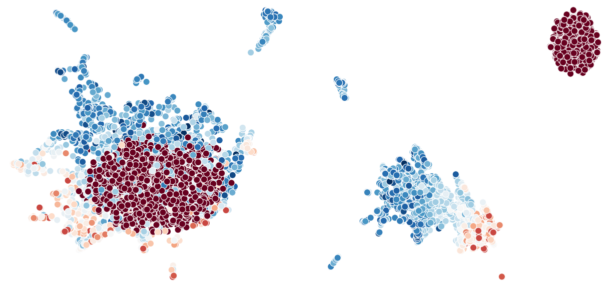
for high-stakes educational assessment and decision-making scenarios.

Explainability of diagnostic results: Fig. 7 presents the explainability analysis results for RQ3. Our approach is motivated by the principle that an explainable learner's knowledge proficiency should be monotonic with respect to their performance; specifically, higher proficiency levels should result in higher response scores on relevant questions. Drawing inspiration from prior research [57], we adopt the degree of consistency (DOC) [26] as our evaluation metric. For a given question e_l, the DOC is defined as follows:

$$\mathrm{DOC}(e_l) = \frac{\sum_{i,j} \delta(r_{il}, r_{jl}) \sum_{k=1}^{K} q_{lk} \wedge J(l,i,j) \wedge \delta(\theta_{ik}, \theta_{jk})}{\sum_{i,j} \delta(r_{il}, r_{jl}) \sum_{k=1}^{K} q_{lk} \wedge J(l,i,j) \wedge I(\theta_{ik} \neq \theta_{jk})}. \tag{38}$$

The DOC measures the monotonicity between diagnostic results and response scores for learners answering the same item. We present experiment results for IRTs and NCDMs, respectively, because the former diagnose learner cognitive states as overall ability (single-dimensional value), while the latter diagnose learner cognitive states as knowledge proficiencies (multidimensional value). Here, "θ from train/test" denotes that the evidence data for generative diagnosis is from the training/test dataset. Response scores for calculating DOC are from the test dataset. Regarding the result analysis, we have two findings. First, it can be observed that the DOC of generative CDMs (i.e., G-IRT and G-NCDM) with θ from the training dataset is mostly higher than that of the corresponding transductive CDMs. This observation indicates that generative CDMs can mostly capture the monotonicity between explicit response scores and implicit cognitive states, demonstrating the psychometric explainability of diagnostic results for learners in various scenarios. Second, it can be observed that the DOC of generative CDMs with θ from the test dataset is always higher than other results. This essentially demonstrates the explainability of *generative diagnosis*. As both the evidence data for diagnosing learner cognitive states and the response scores for calculating DOC are from the same dataset, generative CDMs can effectively capture the complex mapping between response scores and cognitive states, generating reliable diagnostic results while maintaining excellent score reconstruction accuracy (see Table II).

4) Distribution Analysis of G-IRT (RQ4.1): To explore the utility of G-IRT, we visualize the histogram of estimated learner ability θ and correct rates. Figs. 8 and 9 present the visualization results, with the probability density estimated by the kernel density estimation. The evaluation motivation is that the distribution of appropriate learner diagnosis results should present an association with the distribution of observed learner correct rates. From Fig. 8, we can observe that the distribution of learner correct rates is skewed. This property is well preserved by the distribution of diagnostic results of G-IRT. From Fig. 9, we can observe that the distribution of learner correct rates is multimodal. This property is also well preserved by G-IRT. These observations demonstrate the advantage of the generative diagnosis paradigm. The reason is that, according to (14) and (18), the generative process of learner ability θ is equivalent to calculating a weighted average of response scores. Therefore, the calculation of score rate can be viewed as a special case of the generative diagnosis process. On the contrary, transductive IRT relies on parameter estimation to diagnose learner ability, which loses the information of the distribution of the overall correct rate during parameter estimation of learner ability.



(a) (b)

Fig. 11. Visualization of diagnostic results in ASSIST using u-map, colored with learner score rates. All learners in the raw data, including those without response logs (colored with *dark red*), are included in the visualization. (a) NCDM. (b) G-NCDM.

5) Clustering Analysis of G-NCDM (RQ4.2): To examine the statistical feature of diagnostic traits, we visualize the diagnostic outputs of the CDMs. For NCDM and G-NCDM, which represent cognitive states as multidimensional knowledge proficiencies, we employ UMAP [58] for dimensionality reduction. In these visualizations, learner data points are colored according to their overall correct rates to determine if the diagnostic results effectively distinguish between different performance levels. Figs. 10 and 11 display these results, from which we derive two primary findings.

Finding 1—Cognitive states diagnosed by G-NCDM are better clustered with respect to correct rates: Fig. 10 presents visualization results that have removed learners without response scores. From the visualization on ASSIST, we can observe that cognitive states generated by G-NCDM are linearly separable, as red points (i.e., learners with a correct rate less than 0.5) aggregate in the bottom left of the figure, while blue points (i.e., learners with a correct rate greater than 0.5) spread in the center of the figure. From the visualization on Math1, we can also observe that points with similar color (i.e., learners with similar correct rate) aggregate tightly for G-NCDM@Math1, while points with different colors usually mix within a small region for NCDM@Math1. Both the linear separability and the tight aggregation of similar color demonstrate the statistical reliability of G-NCDM.

Finding 2—G-NCDM can effectively recognize empty learners: Online learning platforms inevitably include empty learners, such as the ASSIST dataset introduced in Table I. We present in Fig. 11 the cognitive states of all learners, including those empty. It could be observed that empty learners (dark red points) are well separated and aggregated in the top right of the visualization of G-NCDM, while they overlap with other learners in the visualization of NCDM. Essentially, this ability of G-NCDM originates from the generative diagnosis mechanism. Since the score vectors of all empty learners are the zero vector, their diagnostic results are also the same as and different from other

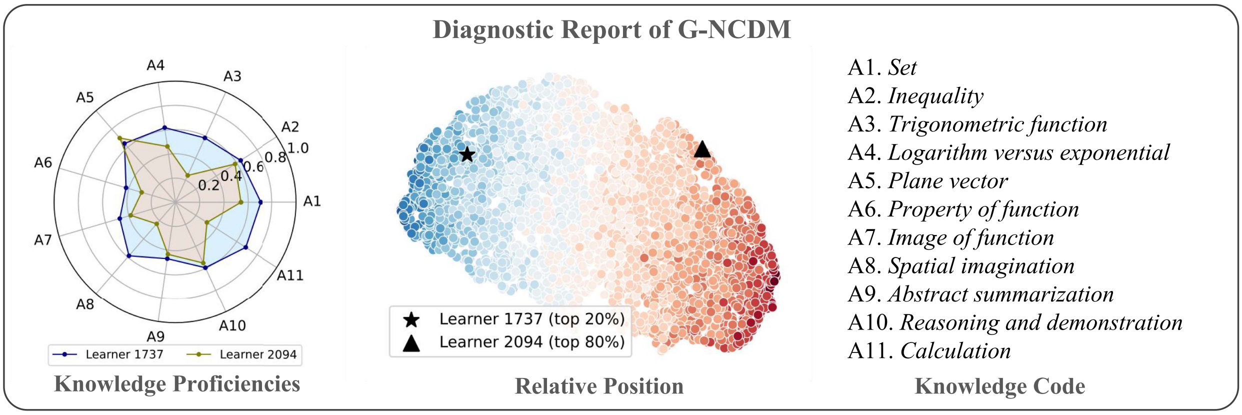


Fig. 12. Case of diagnostic report generated by G-NCDM. The middle term shows the visualization of all learners' diagnostic outputs using u-map, with each point colored by the learner's correct rate (blue denotes higher correct rates while red denotes lower correct rates).

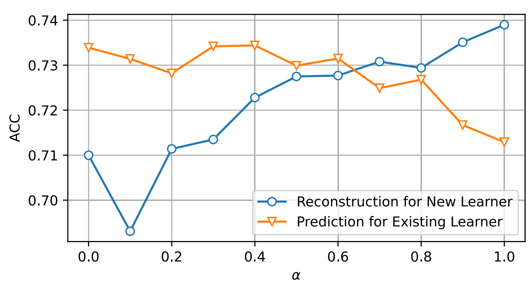


Fig. 13. Score reconstruction and prediction performance of G-NCDM w.r.t. hyperparameter α in ASSIST. The higher the α, the large the weight of $\boldsymbol{\theta}_i^{(\text{exp})}$ in generating $\boldsymbol{\theta}_i$.

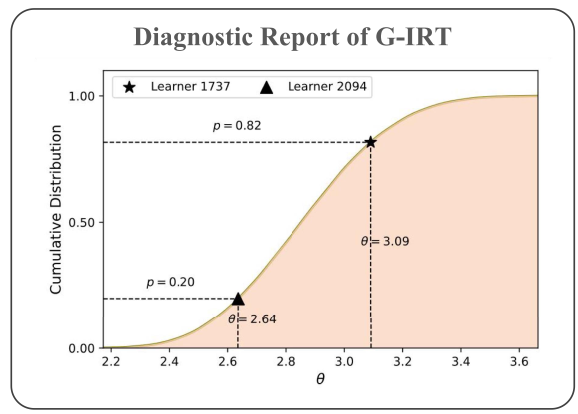


Fig. 14. Case of diagnostic report generative by G-IRT. The orange curve presents the CDF of generated learner ability. The term p denotes the CDF value of learners, which is equivalent to the ratio of learners whose correct rates are lower than the target learner.

learners. This finding validates the generalization ability and outlier detection ability of G-NCDM.

6) A Case Study of G-IRT and G-NCDM (RQ5): We explore the utility of generative CDMs in real-world scenarios via a case study. Specifically, we select Math1 for the case study because the dataset comes from the representative offline test scenario and the response scores are dense. We then select two learners, including one ranked top 20% and one ranked top 80% by correct rates. We present their raw response scores and knowledge correct rates, as shown in Tables V and VI. We then input their data into G-IRT and G-NCDM, respectively, to obtain and visualize the diagnostic reports. Each diagnostic report consists of two components: *1) the learner's cognitive state* and *2) the learner's relative position in the learner group*. We make a crossover comparison between the two learners' response data and diagnostic reports to see if G-IRT and G-NCDM work for this scenario. We have the following two findings.

Finding 1—G-NCDM can well reflect the real knowledge proficiency of learners: By comparing observed knowledge correct rates (see Table VI) and generated knowledge proficiencies (see the left of Fig. 12), we find that the latter is consistent with the former. For instance, for learners 1737 and 2094, only the knowledge correct rate on A5 "*plane vector*" of the former is lower than that of the latter. This is exactly reflected in their generated knowledge proficiencies. In addition, both learners perform poorly on knowledge $A6$ "*property of function*" and A7 "*image of function*," with correct rates less than 0.33. This is also well reflected in their knowledge proficiencies, with values always less than 0.4. These observations essentially demonstrate the effectiveness of the forked design of the GDF of G-NCDM [see (30)]. Technically, the direct, parameter-free generation of $\boldsymbol{\theta}_i^{(\text{exp})}$ ensures that diagnostic results largely reflect knowledge correct rates, while the neural network-based generation of $\boldsymbol{\theta}_i^{(\text{imp})}$ calibrates diagnostic results for more fine-grained diagnosis and better score prediction performance (see Fig. 13). The center of Fig. 12 presents the learners' relative position in the learner group, which also reflects their different cognitive states. Overall, G-NCDM can well reflect knowledge conceptwise cognitive states in this scenario.

Finding 2—G-IRT is beneficial to learner ranking and overall cognitive state diagnosis: As shown in Fig. 14, the diagnostic report of G-IRT is presented as a cumulative distribution function (CDF) of generated learner abilities. We can observe from the report that learner 1737 is ranked top $(1-p)\times 100\% = 17\%$ $(\theta = 3.09)$, and learner 2094 is ranked top $(1-p)\times 100\% = 78\%$ $(\theta = 2.64)$. These results are consistent with observed response data, demonstrating the effectiveness of G-IRT in real-world scenarios.

VIII. Conclusion And Future Work

A. Conclusion

In this study, we proposed the generative CD paradigm in iLearning, with a core focus on enhancing diagnostic fairness and efficiency. The generative diagnosis paradigm overcomes two significant challenges of traditional CDMs that lead to inequitable assessment. The first challenge is a lack of identifiability, which directly violates the fundamental fairness principle. Diagnostic outputs provided by CDMs are unidentifiable and lack psychometric reliability, resulting in different diagnoses for students with identical response patterns. This inequity systematically disadvantages certain learners based solely on algorithmic artifacts rather than actual ability differences. The second challenge is their unavailability for instant diagnosis of new learners, which limits the efficiency of CDMs in realistic iLearning systems. CD for new-coming learners requires retraining transductive CDMs, which is resource-consuming and can lead to inconsistent assessments of existing learners—creating an unfair temporal bias in educational evaluation. The generative diagnosis paradigm effectively overcomes these fairness challenges via the generative diagnosis procedure, which generates learner cognitive states using a GDF rather than parameter optimization, ensuring that students with the same response patterns receive equivalent diagnoses. Technically, we further proposed two instantiations of the generative diagnosis paradigm, the G-IRT and the G-NCDM. Respecting G-IRT, we established the GDF by substituting unobservable parameters in the inverse of the IRF of IRT with proxy parameters, creating a more equitable assessment framework. We then analyzed mathematical properties of G-IRT and provided parameter constraints for better cold-start performance and controllability, ensuring consistent diagnosis across diverse learner populations. Regarding G-NCDM, we propose a neural network-driven GDF based on the previous ID-CDM. Notably, G-NCDM introduced Q-matrix mapping in the generative process, which enhanced the relationship between diagnostic outputs and knowledge dimensions while maintaining fairness across student groups. Next, we demonstrated the effectiveness of generative CDMs via extensive experiments on two real-world datasets. Experiment results revealed that generative CDMs not only have excellent score reconstruction (for instant diagnosis) and score prediction (for offline diagnosis) performances, but can generate reliable and fair diagnostic outputs that treat students with similar abilities consistently. These results demonstrate the accuracy, reliability, fairness and utility of generative CDMs, demonstrating their potential to contribute to the construction of future iLearning environments.

B. Future Work

The primary limitation of this study lies in the current lack of a continual learning ability in the proposed methods. There remain some questions to be ascertained and enriched with future findings. The nature of causation perhaps needs more time series analysis as well as response-time analysis [59], [60]. The connections between cognitive states and knowledge proficiencies still need more advanced and follow-up research related to neuroscience to provide insights into G-NCDM. In addition, there might be possibilities of outlier cases of student learning that call for in-depth analysis.

Specifically, response data, test items, and learners would continually accumulate in a learning platform, which requires the CDM to continually learn from incoming data and update its diagnosis ability. Although current generative CDMs can instantly diagnose new-coming learners, response data of new-coming learners and items still cannot be used for model updating. Along this line, we hope to develop mechanisms or algorithms that enable generative CDMs to continually learn from incoming response data. Continual learning ability of generative CDMs is beneficial for a host of online learning/evaluation applications, such as student cognitive ability tests and large language model evaluation.

Acknoweldgement

The views presented in this paper are personal only and do not necessarily represent those of the Education University of Hong Kong and do not commit the respective organization.

References

[1] J. W. P. Ng, N. Azarmi, M. Leida, F. Saffre, A. Afzal, and P. D. Yoo, “The intelligent campus (iCampus): End-to-end learning lifecycle of a knowledge ecosystem,” in *Proc. 6th Int. Conf. Intell. Environ*, 2010, pp. 332–337.

[2] H.-Y. Wu, H.-S. Wu, I.-S. Chen, and Y.-P. Su, “Toward better intelligent learning (iLearning) performance: What makes ilearning work for students in a university setting?,” *Behav. Inf. Technol.*, vol. 42, no. 1, pp. 60–76, 2023.

[3] S. Jayasiriwardene and D. Meedeniya, “An adaptive and interactive learning toolkit (iLearn),” *Softw. Impacts*, vol. 15, 2023, Art. no. 100471.

[4] M. Wang, “An intelligent learning model based on cloud computing,” *EBTIC's Int. iCampus Initiative Intern. Rep.*, vol. 1, pp. 1–16, 2011.

[5] Q. Zhu and M. Wang, “Team-based mobile learning supported by an intelligent system: Case study of STEM students,” *Interact. Learn. Environ.*, vol. 28, no. 5, pp. 543–559, 2020.

[6] J. Xiao, M. Wang, and W. N. Jason, “The design and implementation of a cloud-based intelligent learning (C-iLearning) system,” in *Proc. Int. Symp. Inf. Technol. Med. Educ.*, 2012, pp. 417–421.

[7] X. Yang and S. E. Embretson, “Construct validity and cognitive diagnostic assessment,” in *Cognitive Diagnostic Assessment for Education: Theory and Applications*. Cambridge, U.K.: Cambridge Univ. Press, 2007, pp. 119–145.

[8] F. Wang et al., “A survey of models for cognitive diagnosis: New developments and future directions,” 2024, *arxiv:2407.05458*.

[9] H. Ma, Z. Huang, W. Tang, H. Zhu, H. Zhang, and J. Li, “Predicting student performance in future exams via neutrosophic cognitive diagnosis in personalized e-learning environment,” *IEEE Trans. Learn. Technol.*, vol. 16, no. 5, pp. 680–693, Oct. 2023.

[10] F. Wang et al., “Dynamic cognitive diagnosis: An educational priors-enhanced deep knowledge tracing perspective,” *IEEE Trans. Learn. Technol.*, vol. 16, no. 3, pp. 306–323, Jun. 2023.

[11] G. Zhao et al., “A diffusion-based cognitive diagnosis framework for robust learner assessment,” *IEEE Trans. Learn. Technol.*, vol. 17, pp. 2281–2295, 2024.

[12] H. Liu, X. You, W. Wang, S. Ding, and H. Chang, “The development of computerized adaptive testing with cognitive diagnosis for an English achievement test in China,” *J. Classif.*, vol. 30, no. 2, pp. 152–172, 2013.

[13] H. Huang, “Effects of item calibration errors on computerized adaptive testing under cognitive diagnosis models,” *J. Classif.*, vol. 35, no. 3, pp. 437–465, 2018.

[14] F. M. Polo, L. Weber, L. Choshen, Y. Sun, G. Xu, and M. Yurochkin, “tinyBenchmarks: Evaluating LLMs with fewer examples,” in *Proc. 41st Int. Conf. Mach. Learn.*, 2024, pp. 34303–34326.

[15] H. Zhou, et al., “Lost in benchmarks? Rethinking large language model benchmarking with item response theory,” in *Proc. 40th AAAI Conf. Artif. Intell.*, vol. 40, no. 41, pp. 35085–35093, 2026.
[16] M. D. Reckase, “Multidimensional item response theory models,” *Multidimensional Item Response Theory. Statistics for Social and Behavioral Sciences*. New York, NY, USA: Springer, 2009, pp. 79–112.
[17] M. Uto and M. Ueno, “Item response theory for peer assessment,” *IEEE Trans. Learn. Technol.*, vol. 9, no. 2, pp. 157–170, Jan. 2016.
[18] F. Wang et al., “NeuralCD: A general framework for cognitive diagnosis,” *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 8, pp. 8312–8327, Aug. 2023.
[19] J. Algina and R. D. Penfield, “Classical test theory,” in *The Sage Handbook of Quantitative Methods in Psychology*. Thousand Oaks, CA, USA: Sage Publications Ltd., 2009, pp. 93–122.
[20] C. E. DeMars, “Classical test theory and item response theory,” *The Wiley Handbook of Psychometric Testing*. Hoboken, NJ, USA: Wiley, 2018, ch. 2, pp. 49–73.
[21] Y. Gu and G. Xu, “The sufficient and necessary condition for the identifiability and estimability of the DINA model,” *Psychometrika*, vol. 84, no. 2, pp. 468–483, 2019.
[22] C. Qin et al., “Model identification and Q-matrix incremental inference in cognitive diagnosis,” *Knowl. Based Syst.*, vol. 86, pp. 66–76, 2015.
[23] S. Tong et al., “Incremental cognitive diagnosis for intelligent education,” in *Proc. 28th ACM SIGKDD Conf. Knowl. Discov. Data Mining*, 2022, pp. 1760–1770.
[24] M. A. McDaniel, C. M. Wally, R. F. Frey, and H. K. Bates, “Individual differences in student learning: A comparison between the student approaches to learning and concept-building frameworks,” *Behav. Sci.*, vol. 15, 2025, Art. no. 1055.
[25] S. H. Alshammari, M. E. Alrashidi, M. H. Alshammari, A. E. A. Alshammari, and A. F. Alkhwaldi, “Determinants of student adoption of artificial intelligence applications in higher education,” *Sci. Rep.*, vol. 15, 2025, Art. no. 35921.
[26] J. Li et al., “Towards the identifiability and explainability for personalized learner modeling: An inductive paradigm,” in *Proc. ACM Web Conf. 2024*, 2024, pp. 3420–3431.
[27] J. de la Torre, “DINA model and parameter estimation: A didactic,” *J. Educ. Behav. Statist.*, vol. 34, no. 1, pp. 115–130, 2009.
[28] G. H. Fischer, “Derivations of the Rasch model,” in *Rasch Models*. New York, NY, USA: Springer, 1995, pp. 15–38.
[29] J. Brzezinska, “Item response theory models in the measurement theory,” *Commun. Statist. Simul. Comput.*, vol. 49, no. 12, pp. 3299–3313, 2020.
[30] A. E. Gelfand and A. F. M. Smith, “Sampling-based approaches to calculating marginal densities,” *J. Amer. Statist. Assoc.*, vol. 85, no. 410, pp. 398–409, 1990.
[31] W. K. Hastings, “Monte Carlo sampling methods using Markov chains and their applications,” *Biometrika*, vol. 57, no. 1, pp. 97–109, 1970.
[32] M. Wu, R. L. Davis, B. W. Domingue, C. Piech, and N. D. Goodman, “Variational item response theory: Fast, accurate, and expressive,” in *Proc. 13th Int. Conf. Educ. Data Mining*, 2020. pp. 257–268.
[33] Y. Zhou et al., “Modeling context-aware features for cognitive diagnosis in student learning,” in *Proc. 27th ACM SIGKDD Conf. Knowl. Discov. Data Mining*, 2021, pp. 2420–2428.
[34] W. Gao et al., “RCD: Relation map driven cognitive diagnosis for intelligent education systems,” in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retriev.*, 2021, pp. 501–510.
[35] J. Li et al., “HierCDF: A Bayesian network-based hierarchical cognitive diagnosis framework,” in *Proc. 28th ACM SIGKDD Conf. Knowl. Discov. Data Mining*, 2022, pp. 904–913.
[36] B. Jing, H. Stärk, T. S. Jaakkola, and B. Berger, “Generative modeling of molecular dynamics trajectories,” in *Proc. 38th Annu. Conf. Neural Inf. Process. Syst.*, 2024, pp. 40534–40564
[37] Y. Song and S. Ermon, “Generative modeling by estimating gradients of the data distribution,” in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2019, pp. 11895–11907.
[38] Q. Sun et al., “Generative multimodal models are in-context learners,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 14398–14409.
[39] R. Bhattacharjee, S. Dasgupta, and K. Chaudhuri, “Data-copying in generative models: A formal framework,” in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 2364–2396.
[40] D. P. Kingma and M. Welling, “Auto-encoding variational Bayes,” in *Proc. 2nd Int. Conf. Learn. Represent*, 2014.
[41] I. Goodfellow et al., “Generative adversarial networks,” *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.
[42] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2020, pp. 6840–6851.
[43] P. Dhariwal and A. Nichol, “Diffusion models beat GANs on image synthesis,” in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2021, pp. 8780–8794
[44] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proc. Conf. North Amer. Ch. Assoc. Comput. Linguist. Hum. Lang. Technol.*, 2019, pp. 4171–4186
[45] S. Sedhain, A. K. Menon, S. Sanner, and L. Xie, “AutoRec: Autoencoders meet collaborative filtering,” in *Proc. 24th Int. Conf. World Wide Web Companion*, 2015, pp. 111–112.
[46] Y. Wu, C. DuBois, A. X. Zheng, and M. Ester, “Collaborative denoising auto-encoders for top-n recommender systems,” in *Proc. 9th ACM Int. Conf. Web Search Data Mining*, 2016, pp. 153–162.
[47] K. F. Colvin and L. A. Keller, “Markov chain Monte Carlo item response theory estimation,” in *Wiley StatsRef: Statistics Reference Online*. Chichester, UK: Wiley, Aug. 2016, pp. 1–9.
[48] N. Saxena, V. Lodaya, and T. Thakur, “IRT : Improving student response prediction with Gaussian initialisation and other modifications,” in *Proc. 21st Int. Conf. Adv. Learn. Technol.*, 2021, pp. 166–167.
[49] P. Xu and M. C. Desmarais, “An empirical research on identifiability and Q-matrix design for DINA model,” in *Proc. 11th Int. Conf. Educ. Data Mining*, 2018, pp. 438–443.
[50] G. Xu, “Identifiability and cognitive diagnosis models,” in *Handbook of Diagnostic Classification Models*. Cham, Switzerland: Springer, 2019, pp. 333–357.
[51] K. K. Tatsuoka, “Rule space: An approach for dealing with misconceptions based on item response theory,” *J. Educ. Meas.*, vol. 20, no. 4, pp. 345–354, 1983.
[52] M. Feng, N. T. Heffernan, and K. R. Koedinger, “Addressing the assessment challenge with an online system that tutors as it assesses,” *User Model. User Adapt. Interact.*, vol. 19, no. 3, pp. 243–266, 2009.
[53] Q. Liu et al., “Fuzzy cognitive diagnosis for modelling examinee performance,” *ACM Trans. Intell. Syst. Technol.*, vol. 9, no. 4, pp. 48:1–48:26, 2018.
[54] X. Glorot and Y. Bengio, “Understanding the difficulty of training deep feedforward neural networks,” in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, 2010, pp. 249–256.
[55] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Proc. 3rd Int. Conf. Learn. Represent.*, 2015.
[56] S. Craw, “Manhattan distance,” in *Encyclopedia of Machine Learning*. Boston, MA, USA: Springer, 2010, pp. 639–639.
[57] F. Fouss, A. Pirotte, J. Renders, and M. Saerens, “Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation,” *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 3, pp. 355–369, Mar. 2007.
[58] L. McInnes and J. Healy, “UMAP: Uniform manifold approximation and projection for dimension reduction,” *J. Open Source Softw.*, vol. 3, no. 29, 2018, Art. no. 861.
[59] A. H. İnce and S. Özbay, “Enhancing ability estimation with time-sensitive IRT models in computerized adaptive testing,” *Appl. Sci.*, vol. 15, no. 13, 2025, Art. no. 6999.
[60] J. Mo, L. Hao, H. Yuan, and Z. Shou, “A cognitive diagnosis model using convolutional neural networks to predict student scores,” *Appl. Sci.*, vol. 15, no. 6, 2025, Art. no. 2875.

Jiatong Li received the B.S. degree in data science from the University of Science and Technology of China (USTC), Hefei, China, in 2022, and the M.S. degree in data science from the School of Artificial Intelligence and Data Science, USTC, in 2025. He is currently working toward the Ph.D. degree in computer science with the Department of Computer Science, University of Wisconsin-Madison, Madison, WI, USA.

He has authored and coauthored relevant papers on conferences, including Conference and Workshop on Neural Information Processing Systems, ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Web Conference, and International Meeting of the Psychometric Society. His research interests include intelligent learning and cognitive diagnosis.

Yingying Ni received the B.S. degree from Anhui University, Hefei, China, and the M.A. degree in communication from the University of Science and Technology of China, Hefei. She is currently working toward the doctoral degree in communication with Shanghai Jiao Tong University, Shanghai, China.

She has authored and coauthored relevant papers on conferences, such as International Communication Association. She is dedicated to understanding how digital technologies reshape information flows and social relationships, as well as the power relations and social impacts inherent in these processes. Her research interests include information distribution, algorithmic inequality, and intelligent learning.

Minjuan Wang (Senior Member, IEEE) received the B.A. degree in Chinese literature from Beijing/Peking University, Beijing, China, the M.A. degree in comparative literature from Penn State University, State College, PA, USA, and the Ph.D. degree in information science and learning technologies from the University of Missouri–Columbia, Columbia, MO, USA.

She is a Chair Professor of emerging technologies and future education, Co-Director of the Global Institute for Emerging Technologies, and Assistant Vice President with The Education University of Hong Kong, Hong Kong SAR, China. She is an internationally recognized scholar and has keynoted numerous international conferences. Her research interests include futuristic mindset and talent development, new and emerging technologies in various educational settings, XR and immersive learning, and the design and implementation of artificial intelligence and robotics for education, training, health, and well-being.

Dr. Wang is a Senior Member of IEEE Education Society. She is recognized internationally for her research, publishing, and dedicated service to IEEE and other scholarly communities.

John Chi-Kin Lee received the B.A. degree in education from the University of Hong Kong, Hong Kong SAR, China, the M.Sc. degree from the University of Oxford, Oxford, U.K., and the M.A. and Ph.D. degrees in education from the Chinese University of Hong Kong, Hong Kong SAR, China.

He is currently a Distinguished Academic and Leader in education, and the President and Chair Professor with The Education University of Hong Kong, Hong Kong SAR, China, where he was the Dean. His career is marked by significant contributions to curriculum and instruction, including roles as Vice President (Academic) and Provost. He has an extensive research background with a focus on various educational areas. He is a top-cited scientist and has secured numerous grants for his projects. He is an active member of several editorial boards, has authored and coedited more than 25 books, and has been involved in education and social services locally, nationally, and globally.

Dr. Lee was the recipient of the Vice-Chancellor's Exemplary Teaching Award at the Chinese University of Hong Kong. His projects have won awards, and he has held various prestigious positions, including Changjiang Scholar Chair Professor, Chair of the United Nations Educational, Scientific and Cultural Organization, and membership in Chinese People's Political Consultative Conference.

Qi Liu (Member, IEEE) received the Ph.D. degree in computer science from the University of Science and Technology of China (USTC), Hefei, China, in 2013.

He is currently a Professor with USTC. His research is supported by the National Science Fund for Excellent Young Scholars and the Youth Innovation Promotion Association of Chinese Academy of Sciences. He has authored or coauthored more than 100 papers in refereed journals and conference proceedings, such as TKDE, TOIS, TNNLS, NeurIPS, ICML, ICLR, and KDD. His general research interests include data mining, knowledge discovery, and artificial intelligence.

Dr. Liu was the recipient of the KDD 2018 Best Student Paper Award (Research) and the ICDM 2011 Best Research Paper Award.

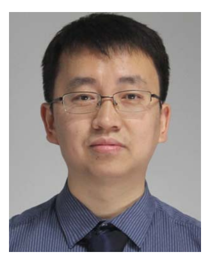
Mengxiao Zhu (Member, IEEE) received the B.E. degree in computer science, the B.S. degree in English and science and the M.E. degree in computer science from the University of Science and Technology of China (USTC), Hefei, China, in 2001 and 2004, respectively, the M.A. degree in communication from the University of Illinois at Urbana-Champaign, Urbana-Champaign, IL, USA, in 2007, and the Ph.D. degree in industrial engineering and management sciences from Northwestern University, Evanston, IL, USA, in 2012.

Since 2020, she has been a Professor with the School of Humanities and Social Sciences at USTC. Before joining USTC, she worked as a Research Scientist in the Research and Development Division at Educational Testing Service for more than seven years. She has been leading and involved in multiple NSFC, NSF and NIH-funded projects in the past 20 years focusing on AI-enabled learning and assessment, learning analytics using big data, and the dynamics of social relations and collaborations both in real world, such as research institutions, and in virtual worlds, such as large-scale online role-playing games. Her current research interests include psychometric models for collaborative problem solving, simulations and games in education, personalized learning and learning analytics, and the interactions of AI and human in education.

Zhenya Huang (Associate Member, IEEE) received the Ph.D. degree in computer science from the University of Science and Technology of China (USTC), Hefei, China, in 2020.

He is currently an Associate Professor with USTC. He has authored or coauthored more than 50 papers in refereed journals and conference proceedings, including TKDE, TOIS, TNNLS, AAAI, KDD, SIGIR, and ICDM. His research interests include artificial intelligence, knowledge reasoning, and intelligent education.

Dr. Huang has served regularly on the program committee of numerous conferences and is a reviewer for the leading academic journals.



Shijin Wang received the B.E. degree in electronic science and technology from the University of Science and Technology of China, Hefei, China, in 2003, and the Ph.D. degree in pattern recognition and intelligent system from the Institute of Automation, Chinese Academy of Science, Beijing, China, in 2008.

He is currently the President of IFLYTEK Beijing Research, Vice President of IFLYTEK AI Research, Vice Leader of technology and industry working group of AIIA, and Young Expert of Internet Association of China. He has authored or coauthored more than 30 papers in refereed conferences, such as ICASSP, ACL, KDD, SIGIR, and AAAI. His research interests include speech processing, natural language processing, and intelligent education.

Dr. Wang led the team that won more than ten championships in international technical evaluation, such as Blizzard Challenge and CHiME.

Enhong Chen (Fellow, IEEE) received the Ph.D. degree in computer science from the University of Science and Technology of China, Hefei, China, in 1996.

He is currently a Professor and the Vice Director with the State Key Laboratory of Cognitive Intelligence, Hefei. His research interests include data mining, machine learning, and artificial intelligence. He has authored or coauthored more than 200 papers in refereed conferences and journals, including TPAMI, TKDE, TNNLS, TOIS, ICML, NeurIPS, KDD, ICLR, and AAAI.

Dr. Chen was the recipient of the Best Application Paper Award on KDD 2008, the Best Research Paper Award on ICDM 2011, the Best Student Paper Award on KDD 2018 (Research), and the Best Student Paper Award on KDD 2024 (Research). His research is supported by the National Science Foundation for Distinguished Young Scholars of China. He is an Associate Editor for IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE TSMCS, ACM TIST, and WWWJ.