Disentangled Knowledge Tracing for Alleviating Cognitive Bias

Anonymous Author(s)

Abstract

In the realm of Intelligent Tutoring System (ITS), the accurate assessment of students' knowledge states through Knowledge Tracing (KT) is crucial for personalized learning. However, due to data bias, *i.e.*, the unbalanced distribution of question groups (*e.g.*, concepts), conventional KT models are plagued by cognitive bias, which tends to result in cognitive underload for overperformers and cognitive overload for underperformers. More seriously, this bias is amplified with the exercise recommendations by ITS. After delving into the causal relations in the KT models, we identify the main cause as the confounder effect of students' historical correct rate distribution over question groups on the student representation and prediction score. Towards this end, we propose a Disentangled Knowledge Tracing (DisKT) model, which separately models students' familiar and unfamiliar abilities based on causal effects and eliminates the impact of the confounder in student representation within the model. Additionally, to shield the contradictory psychology (*e.g.*, guessing and mistaking) in the students' biased data, DisKT introduces a contradiction attention mechanism. Furthermore, DisKT enhances the interpretability of the model predictions by integrating a variant of Item Response Theory. Experimental results on 11 benchmarks and 3 synthesized datasets with different bias strengths demonstrate that DisKT significantly alleviates cognitive bias and outperforms 14 baselines in evaluation accuracy.

Keywords

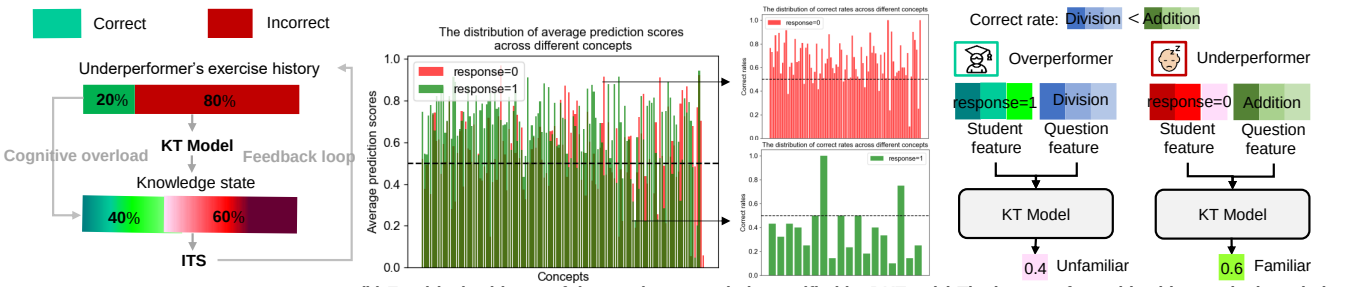
Knowledge Tracing, Educational Data Mining

1 Introduction

In recent years, especially with the explosion of large language models (*e.g.*, GPT-4o), AI for Education has received widespread attention [3, 15, 64, 74]. Intelligent Tutoring System (ITS), as a component of this field, has also seen rapid development [30, 72]. The success of ITS lies in its ability to recognize each student's knowledge state and recommend personalized learning resources (*e.g.*, exercises) based on the large-scale learning data obtained from online learning environments [33]. *Knowledge Tracing (KT), an essential task in ITS, aims to assess the evolution of each student's knowledge state over time based on previous learning interactions and predict their future performance.* Conventional KT models typically assess students' knowledge states based on their interaction history, which usually exhibits data bias[1], *i.e.*, the distribution of question groups (*e.g.*, concepts) is unbalanced. Therefore, KT models often face the issue of **cognitive bias, which usually manifests as cognitive underload on overperformers and cognitive overload on underperformers.** Figure 1(a) illustrates the issue of cognitive overload on underperformers with the example of exercise recommendation. ITS recommends exercises to an underperformer, who gets 80% incorrect despite a considerable portion of simple questions. However, KT model still assesses that the student is familiar with 20% of the questions, causing the ITS to overestimate the student's abilities and recommend exercises that are difficult to respond to. Meanwhile, cognitive underload will eventually prompt the ITS to recommend low-value exercises to overperformers. Clearly, ITS recommendations, based on the evaluation results of the KT model, that deviate from the current knowledge state of students do not meet the requirements of intelligent education for adaptive learning [29, 36, 44, 50]. What's worse, due to the feedback loop [10, 11], cognitive bias of KT model will be amplified over time (*e.g.*, when an underperformer responds to simple questions incorrectly, ITS may recommend more difficult questions, making it more likely to respond incorrectly) until it reaches a critical point between easy and difficult questions for the student [21, 67], causing the student's real knowledge state to gradually deviate from the model prediction, losing the effectiveness of recommendation, and seriously affecting the students' experience with the ITS.

After scrutinizing the causal relations in the KT model, we attribute cognitive bias to a confounder [48], *i.e.*, the student historical correct rate distribution over question groups. The question features (usually the joint representation of the questions and the concepts) and the student features (*e.g.*, the binary responses to the questions) are usually embedded in the vector chronologically, and then encoded by the KT model to predict the evaluation scores for different concepts. In other words, the KT model evaluates the conditional probability of the student's knowledge state given the question features and student features. From the perspective of causality, the question features and student features can be regarded as the cause of the prediction score, and the KT model performs causal modeling on them. But through the observation of causal relations, we find that the hidden confounder, *i.e.*, the student historical correct rate distribution over question groups, affects both the student representation and the prediction score. Meanwhile, through structured probability modeling, the conventional KT models are affected by the confounder, thereby causing a spurious correlation between the student representation and the prediction score: for questions with higher correct rates, underperformers will get higher prediction scores, even if the students are obviously incorrect, similarly, for questions with lower correct rates, overperformers will get lower prediction scores, even if the students have responded correctly. Figure 1(b) is the empirical evidence of the spurious correlation verified by DKT [51] applied to the assist09 dataset [16]. We calculate the average prediction scores of the model when students respond incorrectly (response=0) and correctly (response=1) across different concepts (Figure 1(b) left). According to the classical test theory [9], we calculate the correct rate of the concept (*i.e.*, the concept difficulty) with average prediction scores ≥ 0.5 for incorrect responses and with average prediction scores < 0.5 for correct responses (Figure 1(b) right). As shown in Figure 1(b) upper right, when the correct rate of almost all concepts is high, although the students respond incorrectly, the model still predicts higher scores, and vice

[1]Data bias refers to the over- or under-representation of certain categories, features, or labels relative to others in a dataset. In this work, we focus on measurement bias [18], a common type of data bias, which is consistent with the class imbalance in pattern recognition, *e.g.*, significant differences in the correct rates among different concepts.



(a) An example of cognitive overload on underperformers.

(b) Empirical evidence of the spurious correlation verified by DKT applied to the assist09 dataset.

(c) The impact of cognitive bias on the knowledge state assessment of different performers.

Figure 1: Illustration of cognitive bias.

versa (Figure 1(b) lower right), which makes the KT model exhibit cognitive bias, undermining the effectiveness of the knowledge state assessment for overperformers and underperformers (see the example in Figure 1(c)).

In order to eliminate the spurious correlation, we propose Disentangled Knowledge Tracing (DisKT), a novel approximate causal model based on causal effects. DisKT models simple and difficult questions separately, thereby modeling the abilities that students are familiar and unfamiliar with, and eliminating the impact of the confounder in student representation within the model. In addition, from Figure 1(b) right, we notice that students make mistakes in questions with extremely high correct rates, and sometimes they can correctly respond to questions with extremely low correct rates. We attribute this to the contradictory psychology (*e.g.*, guessing and mistaking) [7, 13, 17, 34, 75, 78] which are not conducive to the modeling of students' actual knowledge state, which inspires us to design a contradictory attention to shield these factors. Finally, we design a variant of Item Response Theory (IRT) [55, 70], integrating the abilities that students are familiar and unfamiliar with, to enhance the interpretability of the model prediction layer.

In summary, this work contributes in four aspects:

- We analyze the causal relations in the conventional KT model through a causal graph, and reveal the cause of cognitive bias from the perspective of causal probability.
- Based on causal effects, we propose a novel approximate causal model, DisKT, which eliminates the impact of the confounder to alleviate cognitive bias. We also propose a contradiction attention to shield the contradictory psychology (*e.g.*, guessing and mistaking). In addition, we design a variant of IRT to enhance the interpretability of model predictions.
- We construct three datasets with different bias strengths, and design a metric to measure the effectiveness of DisKT in alleviating cognitive bias. In addition, we propose two contradictory metrics to determine the potential of our proposed contradictory attention in shielding guessing and mistaking.
- Extensive experiments on 11 benchmarks from 10 different subjects show that DisKT not only effectively alleviates cognitive bias, but also has superior evaluation accuracy compared to other 14 baselines.

2 Related Work

Knowledge Tracing (KT) has been a cornerstone in the development of intelligent tutoring system, enabling personalized education by assessing and predicting students' knowledge states over time [2, 23, 40, 53]. Early attempts at KT, such as the Bayesian Knowledge Tracing (BKT) [13], are based on the Hidden Markov Model, using a binary variable to represent knowledge states. Item Response Theory (IRT) [54] is a factor analysis method designed to model the relationship between students' abilities and their responses by measuring the gap between student ability and question difficulty. With the rise of deep learning, recent studies have utilized deep learning models to address KT issues [2, 51, 59]. DKT [51] first applies LSTM [25] to KT, followed by the introduction of Memory-Augmented Neural Networks (MANN) [57], and DKVMN [76] based on MANN, which uses the key matrix and the value matrix to dynamically store students' mastery of concepts. SKVMN [1] combines the recurrent modeling capability of DKT with the memory network structure of DKVMN. Deep-IRT [71] integrates IRT and DKVMN to make deep learning based KT explainable. Later, GKT [41] uses a graph to model the relationships between knowledge concepts. With the success of attention [5], especially the Transformer [66] in the NLP field, SAKT [42] first introduces self-attention networks to capture the relevance between knowledge concepts and student interactions, followed by state-of-the-art models or frameworks with variant attention structures: AKT [20], DTransformer [73], FoLiBi [28], sparseKT [27]. Meanwhile, some popular training techniques are also applied in KT, such as ATKT [22] and CL4KT [33], which respectively use adversarial training and contrastive learning to enhance student interaction representation. However, despite these studies attempting to address issues in KT and achieving impressive results in evaluation accuracy, there is a lack of comprehensive and reasonable explanation, even Deep-IRT is limited in prediction. In contrast, DisKT is based on causal effect modeling, where the calculation of each interpretable parameter is transparent.

Surprisingly, previous KT research has lacked attention to such an important topic as data bias, and the closest to our DisKT is the Core framework [14]. The Core framework considers that existing models tend to remember answer bias, *i.e.*, the unbalanced distribution of correct and incorrect answers for each question, thus providing a shortcut for achieving good predictive performance in KT and mitigating the answer bias by subtracting the direct causal effect from the total causal effect captured during training in testing [6, 46, 65]. Compared to DisKT, the Core framework has two obvious disadvantages. The Core framework does not recognize the impact of learned bias on different populations, *i.e.*, different cognitive bias exists for overperformers and underperformers, and does not realize that such bias will be amplified. Meanwhile, the Core

Table 1: **Summary of the primary notations.**

Symbols	Description
$\mathcal{S}$,$\mathcal{Q}$,$\mathcal{C}$	The set of students, questions,knowledge concepts.
s, q, c, r	The student, the question, the knowledge concept and the response to the question.
$q_t, Concept_{q_t}, r_t$	The question at time t, the concepts corresponding to the question at time t, the response to the question at time t.
X_t	The whole interaction sequence of the student at time t.
S, Q, D, M, Y	The student features (*e.g.*, $\{r\}$), the question features (*e.g.*, $[\{q\}, \{c\}]$), the student historical correct rate distribution over question groups, the concept-level student representation, the prediction scores given by KT model.
s,m,d,y	The student feature, the hidden representation of the student, the correct rate distribution of the student, the prediction score given by KT model.
s^*, m^*, d^*	The student feature in the counterfactual world, the hidden representation of the student in the counterfactual world, the correct rate distribution of the student in the counterfactual world.
$\mathcal{D}$, $\mathcal{M}$	The sample space of D, the sample space of M.
$e_{c_t}, r_{r_t}, d_{q_t}, \mu_{c_t}, g_{r_t}$	The embedding of concept c_t, the embedding of response r_t, the scalar representing the difficulty of the question q_t, the variation of the question containing the concept c_t, the variation embedding of the response r_t.
d, N, h	The dimension of embedding, the number of Transformer encoders, the number of attention heads.
$Q_{1:t}, S_{1:t}, S^+_{1:t}, S^-_{1:t}$	The Rasch embeddings of the student or interactions at time t, the Rasch embeddings of the questions at time t, the interaction embeddings responding to the familiar abilities at time t, the interaction embeddings responding to the unfamiliar abilities at time t.
$W^1, W^2, W_1, W_2, b^1, b^2, b_1, b_2$	The trainable matrix and parameters.
$H^i_{t+1}, H^+_{t+1}, H^-_{t+1}$	The knowledge states predicted by the i-th Transformer encoder at time t, the knowledge states responding to the familiar abilities at time t, the knowledge states responding to the unfamiliar abilities at time t.
λ_t, β, γ	The random value at time t, the lower bound of λ_t, the initial value decided by the student.
$\mathcal{L}_{DisKT}, \mathcal{L}_{bce}, \mathcal{L}_{cl}$	The total loss, the binary cross-entropy loss, the regularization loss, respectively.
$p(c), diff(c), 1(\cdot)$	The correct rate of concept c, the correct rate of c obtained from the training set, the indicator function.

framework does not delve into the contradictory psychology [7] in biased data, *e.g.*, guessing and mistaking, and their adverse impact on the real knowledge state. In contrast, DisKT provides a more detailed causal graph, alleviates bias within the model, and designs a contradictory attention to shield contradictory psychology, alleviating bias while also improving evaluation accuracy.

3 Methodology

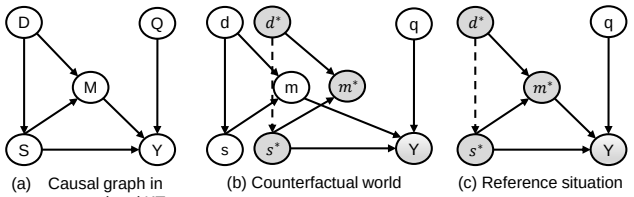
To ensure clarity and comprehension for readers, the important notations used in this session are meticulously presented in Table 1.

3.1 KT Formulation

Formally, we define a set of students $\mathcal{S}$, a set of questions $\mathcal{Q}$, and a set of concepts $\mathcal{C}$. Historical interactions of a student $s \in S$ are represented as $X_t = \{(q_1, Concept_{q_1}, r_1), (q_2, Concept_{q_2}, r_2), ..., (q_t, Concept_{q_t}, r_t)\}$, where $q_t \in \mathcal{Q}$ refers to the question responded to by the student at time t, $Concept_{q_t} \subset \mathcal{C}$ denotes the set of concepts related to q_t, and $r_t \in \{0, 1\}$ indicates the student's response to the question (0 for incorrect, 1 for correct). The aim of KT is to predict the probability $p(r_{t+1} \mid X_t, q_{t+1}, Concept_{q_{t+1}})$ of a student correctly responding to the next question q_{t+1}.

3.2 Causality Perspective on Cognitive Bias

In this section, we construct a causal graph for the conventional KT model. Based on the causal graph, we probabilistically model the conventional KT model and find that the confounder in student representation is the main culprit leading to cognitive bias. Consequently, we build a counterfactual world to eliminate the influence of the confounder in the real world based on causal effect.



(a) Causal graph in conventional KT (b) Counterfactual world (c) Reference situation

Figure 2: The causal graphs for conventional and counterfactual KT. $*$ denotes the reference values.

3.2.1 Causal Graph

As shown in Figure 2(a), we use a causal graph to describe the causal relations in KT, which includes five variables: S, Q, D, M, Y. Note that we use capital letters (*e.g.*, D) to represent variables and lowercase letters (*e.g.*, d) to represent specific values of these variables. Specifically,

- S represents student features. For a student, $s = r_{1:t}$ represents the binary response sequence up to time t.
- Q represents question features, which is usually a joint representation of the questions and concepts [20, 27].
- D denotes the student historical correct rate distribution over question groups (*e.g.*, concept). Given n concepts $\{c_1, ..., c_n\}$, $d_s = [p_s(c_1), p_s(c_2), ..., p_s(c_n)]$ represents the correct rate distribution of student s across different concepts, where $p_s(c_n)$ refers to the correct rate on concept c_n of student s in history.
- M is the concept-level student representation. A vector m represents the hidden representation of the student, learned by a KT model (*e.g.*, DKT), for different concepts. m is determined only by s and d, that is, m can be represented by a function $M(s, d)$ with s and d.
- Y with specific value $y \in [0, 1]$ is the prediction score.
- $D \rightarrow S$ indicates that the student historical correct rate distribution over question groups influences the student's representation, tending to overfit unbalanced historical data and exhibiting cognitive bias [56, 67].
- $(D, S) \rightarrow M$ indicates that D and S determine the concept-level student representation.
- $(S, M, Q) \rightarrow Y$ indicates that S affects Y by two paths: the direct path $S \rightarrow Y$, representing the student's actual knowledge state, and the indirect path $S \rightarrow M \rightarrow Y$, which reflects the polarization of the prediction score caused by the bias in the concept-level student representation, *i.e*, simpler question groups are more likely to have higher prediction scores, and more difficult question groups are more likely to be predicted with lower scores.

According to the causal theory [48], D is associated with both S and Y, and is a confounder between S and Y. Next, through structured probability modeling, we explore how the student historical distribution leads to the polarization of prediction score via biased student representation.

3.2.2 Probability Modeling

Due to the confounder D between S and Y, there exists an issue of cognitive bias when the existing KT models predict the conditional probability $P(Y \mid S = s, Q = q)$. Given $S = s$ and $Q = q$, $P(Y \mid S = s, Q = q)$ is formalized as follows:

$$
\begin{aligned}
& P(Y \mid S = s, Q = q) \\
&= \frac{\sum_{d \in \mathcal{D}} \sum_{m \in \mathcal{M}} P(Y, s, q, d, m)}{P(s, q)} && \text{(1a)} \\
&= \frac{\sum_{d \in \mathcal{D}} \sum_{m \in \mathcal{M}} P(d)\, P(s \mid d)\, P(m \mid d, s) P(q)\, P(Y \mid s, q, m)}{P(s)\, P(q)} && \text{(1b)} \\
&= \sum_{d \in \mathcal{D}} \sum_{m \in \mathcal{M}} P(d \mid s)\, P(m \mid d, s) P(Y \mid s, q, m) && \text{(1c)} \\
&= \sum_{d \in \mathcal{D}} P(d \mid s)\, P(Y \mid s, q, M(d, s)) && \text{(1d)} \\
&= P(d_s \mid s)\, P(Y \mid s, q, M(d_s, s)), && \text{(1e)}
\end{aligned}
$$

where $\mathcal{D}$ and $\mathcal{M}$ are the sample spaces of D and M, respectively. Eq. (1a), Eq. (1b), and Eq. (1c) are derived from the total probability rule, causal graph, and Bayes formula, respectively. And m is determined by certain d and s, so we get Eq. (1d). We only study the sample space d_s of student s, thus obtaining Eq. (1e).

From Figure 2(a) and Eq. (1e), we find that D not only affects S but affects Y through $M(d_s, s)$, causing a spurious correlation: for questions in simpler or more difficult question group (*e.g.*, concept c_n), the prediction scores are higher or lower, *i.e.*, the high or low prediction scores are caused by the student historical distribution rather than the questions themselves. In Eq. (1e), a higher or lower correct rate $p_s(c_n)$ in d_s will make $M(d_s, s)$ show a better or worse knowledge state of concept c_n, and then increase or decrease the prediction scores of questions in c_n through $P(Y \mid s, q, M(d_s, s))$. This ultimately leads to the polarization of prediction scores for questions in easy and difficult concepts, *i.e.*, the cognitive bias towards overperformers and underperformers.

3.2.3 Counterfactual Counterfactual technology is a method of estimating causal effects by considering events that may occur under different conditions and analyzing "how the results would change if the situation were different" [52]. As shown in Figure 2(b), we construct a counterfactual world: D does not affect Y through S but only affects Y through M, that is, the counterfactual can estimate how much the prediction score would be if D had no effect on S. The key to the counterfactual is to intervene causally on S, also called the do-operator [47, 49, 69], *i.e.*, $do(S = s^*)$ forcibly cuts off the edge $D \longrightarrow S$ in Figure 2(a), replaces s in Eq. (1e) with s^*, and obtains $P(Y \mid do(S = s^*), Q = q) = p(d)p(s^*)P(Y \mid s^*, q, M(d, s^*))$, where d **can be considered a constant distribution.**

3.2.4 Causal Effect In causal effects, the Total Effect (TE) of $S = s$ on Y denotes the change in Y caused by the S when it changes from the reference value s^* in Figure 2(c) to the expected value s in Figure 2(a). Given $Q = q$, the TE of $S = s$ on Y is formalized as:

$$TE = Y_{s,m,q} - Y_{s^*,m^*,q}, \tag{2}$$

where $Y_{s^*,m^*,q}$ represents the reference state of Y when $S = s^*$, and S is not affected by D. Typically, TE can be decomposed into $TE = NDE + TIE$, where NDE and TIE respectively represent the natural direct effect and total indirect effect [6, 46, 65].

Specifically, given $Q = q$, the NDE of $M = m$ on Y refers to the change in the prediction score Y when the M changes from the reference value m^* to the expected value m and $do(S = s^*)$ is enforced. The NDE is formalized as follows:

$$NDE = Y_{s^*,m,q} - Y_{s^*,m^*,q}, \tag{3}$$

where $Y_{s^*,m,q}$ is the prediction score in the counterfactual world, and S is not affected by D and M remains unchanged (as shown in Figure 2(b)).

Therefore, given $Q = q$, the TIE of $S = s$ can be obtained by subtracting NDE from TE:

$$TIE = TE - NDE = Y_{s,m,q} - Y_{s^*,m,q}. \tag{4}$$

Therefore, the TIE of $S = s$ on Y is the change in the Y caused by the S when it changes from the reference value s^* to the expected value s without affecting the M.

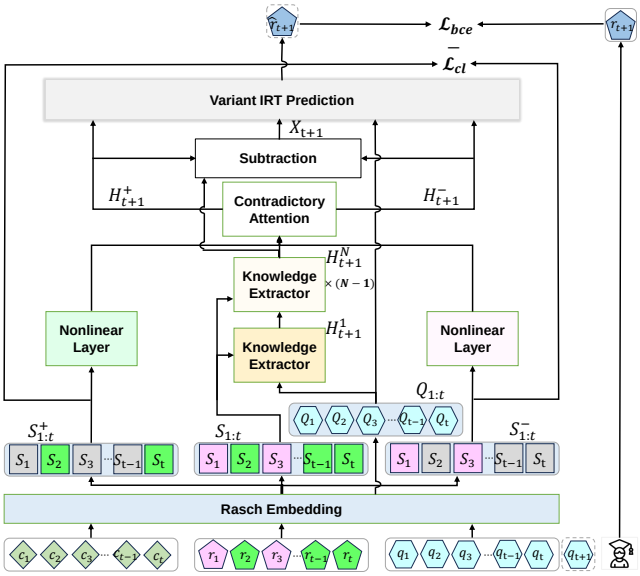


Figure 3: The architecture of the Disentangled Knowledge Tracing model (DisKT).

In Eq. 4, NDE estimates how much the prediction score would be in the counterfactual world if the KT model only had the student historical distribution and did not track the student's knowledge state. Intuitively, TIE represents the final prediction score, which reduces the NDE of the student historical distribution [14, 67]. Therefore, **the prediction score for underperformers on questions of high correct rate would be largely suppressed, conversely, the prediction score for overperformers on questions of high correct rate would be liberated.**

3.3 Disentangled Knowledge Tracing

Through the analysis of causal effects, the key to eliminating the influence of the confounder lies in how to causally intervene on S so that the student representation after $do(S = s^*)$ only represents the student historical distribution. In addition, the trouble caused by the factual and counterfactual inference of the traditional causal model is also a problem worth considering [14, 67, 68, 77]. To this end, DisKT we propose is an approximate causal model. Considering that the student historical distribution is fundamentally indicating that students are familiar with simple concepts but not good at difficult concepts, DisKT models the responses of incorrect and correct responses separately along with the concepts, roughly classifying the concepts responded to correctly as simple concepts, and the concepts responded to incorrectly as difficult concepts, thus modeling the students' familiar and unfamiliar abilities. **Due to separate modeling, it is difficult to track knowledge of either side, thus achieving that the student representation after intervention approximately represents the student historical distribution**. In addition, to avoid double causal inference, DisKT executes the process of causal intervention within the model, and approximates M by assigning contradictory attention weights considering the contradictory psychology (*e.g.*, guessing and mistaking) [7, 13, 17, 34, 75, 78] of the factual student representation to both sides, thereby reducing the burden of re-inference.

The architecture of DisKT is shown in Figure 3, with details as follows.

3.3.1 Rasch Embedding KT models often describe multiple concepts as a single concept, *i.e.*, $c_t = Concept_{q_t}$, and due to data sparsity, they use concepts instead of questions as the subject of assessment [20, 33, 51]. We use the Rasch model in psychology [20, 37, 39, 54] to construct the t-th embeddings of question (*i.e.*, Q_t) and interaction (*i.e.*, S_t):

$$\begin{aligned} &Q_t = c_{c_t} + d_{q_t} \cdot \mu_{c_t}, S_t = e_{(c_t,r_t)} + d_{q_t} \cdot \nu_{(c_t,r_t)}, \\ &e_{(c_t,r_t)} = c_{c_t} + r_{r_t}, \nu_{(c_t,r_t)} = c_{c_t} + g_{r_t}, \end{aligned} \tag{5}$$

where $c_{c_t} \in \mathbb{R}^d$ is the embedding of concept c_t, $d_{q_t} \in \mathbb{R}$ is a scalar representing the difficulty of question q_t, $\mu_{c_t} \in \mathbb{R}^d$ summarizes the variation of questions containing concept c_t, $r_{r_t} \in \mathbb{R}^d$ is the embedding of response r_t, $g_{r_t} \in \mathbb{R}^d$ is the variation embedding of response r_t, $e_{(c_t,r_t)} \in \mathbb{R}^d$ and $\nu_{(c_t,r_t)} \in \mathbb{R}^d$ are the embedding and variation embedding of the concept-response interaction (c_t, r_t). d denotes the dimension of these embeddings.

Therefore, given the student's historical interaction sequence $\{q_{1:t}, c_{1:t}, r_{1:t}\}$, the factual embeddings of questions (*i.e.*, $Q_{1:t}$) and interactions (*i.e.*, $S_{1:t}$) are represented as follows:

$$S_{1:t}, Q_{1:t} = \text{Rasch Embedding}(q_{1:t}, c_{1:t}, r_{1:t}). \tag{6}$$

Through artificial intervention, that is, in order to obtain the correct interaction embeddings, DisKT masks the elements corresponding to the incorrect response positions in $q_{1:t}, c_{1:t}$ and $r_{1:t}$, and vice versa, the obtained counterfactual interaction embeddings $S^+_{1:t}$ and $S^-_{1:t}$ are:

$$\begin{aligned} &S^+_{1:t,=} \text{Rasch Embedding}(r_{1:t} \cdot q_{1:t}, r_{1:t} \cdot c_{1:t}, mask + (1 - mask) \cdot \\ &r_{1:t}), \\ &S^-_{1:t,=} \text{Rasch Embedding}((1 - r_{1:t}) \cdot q_{1:t}, (1 - r_{1:t}) \cdot c_{1:t}, mask \cdot r_{1:t}), \end{aligned} \tag{7}$$

where $mask$ is the mask value (*e.g.*, 2) of the response sequence, while the question and concept sequences are masked by 0.

3.3.2 Knowledge Extractor In order to effectively encode the embeddings of questions and interactions, the knowledge extractor employs N Transformer encoders [66]. For the first encoder, the knowledge extractor takes the question and interaction embeddings $Q_{1:t}$ and $S_{1:t}$ as input and outputs the knowledge state H^1_{t+1} extracted from the current questions and interactions:

$$\begin{cases} H^1_{t+1} = \text{LN}(\text{Dropout}(\text{Res}(\text{FFN}(S = \text{Multihead}^1)))), \\ \text{FFN}(S) = \text{GeLU}(SW^1 + b^1)W^2 + b^2, \\ \text{Multihead}^1 = \text{Concat}(\text{head}^1_1, \cdots, \text{head}^1_h)W^1_h, \\ \text{head}^1_i = \text{Attention}(Q = Q^{/h}_{1:t}, K = Q^{/h}_{1:t}, V = S^{/h}_{1:t}), \\ \text{Attention}(Q, K, V) = \text{Concat}(\textbf{zero}, \text{Softmax}(\dfrac{QK^T}{\sqrt{d/h}})[1:, :])V, \end{cases} \tag{8}$$

where LN, Dropout, Res, FFN refer to layer normalization [4], dropout technique [60], residual connection [24] and fully-connected feed-forward, respectively, $W^1 \in \mathbb{R}^{d\times d}$, $W^2 \in \mathbb{R}^{d\times d}$, $b^1 \in \mathbb{R}^d$, $b^2 \in \mathbb{R}^d$ are learnable parameters and GeLU$(\cdot)$ is the activation function, h is the number of attention heads (*e.g.*, 2) and $W^1_h \in \mathbb{R}^{d\times d}$, $Q^{/h}_{1:t}$ and $S^{/h}_{1:t}$ represent splitting the d dimensions of $Q_{1:t}$ and $S_{1:t}$ into h parts, respectively, and $\textbf{zero} \in \mathbb{R}^d$ is a zero vector indicating that the historical interaction before the first question is not available. Eq. 8 can be abbreviated as

$$H^1_{t+1} = \text{Encoder}(Q = Q_{1:t}, K = Q_{1:t}, V = S_{1:t}), \tag{9}$$

so the output of the last encoder is

$$H^N_{t+1} = \text{Encoder}(Q = H^{N-1}_{t+1}, K = H^{N-1}_{t+1}, V = S_{1:t}). \tag{10}$$

The design of the knowledge extractor enables DisKT to summarize the performance of students over multiple time scales and extract comprehensive knowledge.

3.3.3 Contradictory Attention Considering the burden of re-reasoning and the impact of contradictory psychology (*e.g.*, guessing and mistaking), the contradictory attention we designed assigns the knowledge learned by the factual student representation from the knowledge extractor to the student representations in the counterfactual world, which previously perform feature extraction through a nonlinear layer (*e.g.*, FFN). Meanwhile, it shields the weights of the contradictory psychology through a selective Softmax function (Softmax*). Finally, it forms the knowledge states H^+_{t+1} and H^-_{t+1} representing familiar and unfamiliar abilities in the counterfactual world:

$$\begin{cases} H^+_{t+1} = \text{Attention}^*(Q = H^N_{t+1}, K = H^N_{t+1}, V = \text{FFN}(S^+_{1:t})), \\ H^-_{t+1} = \text{Attention}^*(Q = H^N_{t+1}, K = H^N_{t+1}, V = \text{FFN}(S^-_{1:t})), \\ \text{Attention}^*(Q, K, V) = \text{Concat}(\textbf{zero}, \text{Softmax}^*(X = \dfrac{QK^T}{\sqrt{d/h}})[1:, \\ :])V, \\ \text{Softmax}^*(X) = \text{Softmax}(\textbf{one} - exp(CV(c_{1:t}, r_{1:t}), d) \cdot \text{Softmax}(X)), \\ CV(c_t, r_t) = \mathbb{1}(\max(\lambda_t, \beta)(1 - r_t + (2r_t - 1) \cdot \text{diff}(c_t)) < \alpha^2_t), \end{cases} \tag{11}$$

where $\textbf{one} \in \mathbb{R}^{d\times d}$ is a vector of ones, $exp(\cdot, d)$ represents column expansion by d dimensions, $CV(c_{1:t}, r_{1:t})$ represents the contradictory variable sequence determined by $c_{1:t}$ and $r_{1:t}$, and $\mathbb{1}(\cdot)$ denotes the indicator function. λ_t is a random value, representing the probability that the student is not affected by the contradictory psychology at time t, and the smaller it is, the more likely it is to be affected by the contradictory psychology. β is the lower bound of λ_t (*e.g.*, 0.1), preventing the dominant position of the uncertain λ_t. α_t is a determined value, representing the degree threshold of a student's contradictory psychology at time t. diff(c_t) refers to the correct rate obtained from the training set through c_t according to the classical test theory [9].

The form of $CV(c_t, r_t)$ is intuitive: the more difficult c_t is, the more likely the student is to guess; the simpler c_t is, and the student responses incorrectly, the more likely the student is to experience a psychology of mistaking. α_t can be determined as follows:

$$\begin{cases} \alpha_t = \gamma, & \text{if } t = 1, \\ \alpha_t = \sqrt{\dfrac{\sum_{i=1}^{t-1}(\max(\lambda_i, \beta)(1 - r_i + (2r_i - 1) \cdot \text{diff}(c_i)))}{t-1}}, & \text{else}, \end{cases} \tag{12}$$

where γ (*e.g.*, 0.2) is the initial value of the student's contradiction. As can be seen from Eq. 12, if a student has more and more severe contradictory psychology in the past, the contradiction threshold should be higher, and vice versa.

3.3.4 Variant IRT Prediction Since the prediction scores range from 0 to 1, DisKT subtracts the NDE from Eq. 4 in terms of features:

$$X_{t+1} = H_{t+1} - (H^{+}_{t+1} + H^{-}_{t+1}). \tag{13}$$

Then, DisKT explicitly indicates the questions to be predicted and generates final prediction scores through an MLP:

$$\hat{r} = \sigma(\text{ReLU}(\left[(X_{t+1} - d_{q_{1:t}}) \oplus (H^{+}_{t+1} - H^{-}_{t+1}) \oplus Q_{1:t}\right] W_1 + b_1)W_2 + b_2), \tag{14}$$

where $\oplus$ denotes the concatenation operation. $W_1 \in \mathbb{R}^{3d\times d}$, $W_2 \in \mathbb{R}^{d\times 1}$, $b_1 \in \mathbb{R}^{d}$, $b_2 \in \mathbb{R}^{1}$ are learnable parameters in the MLP. $\sigma(\cdot)$ denotes the sigmoid function and $\text{ReLU}(\cdot)$ is the activation function. $d_{q_{1:t}}$ can be obtained from Eq. 5. Eq. 14 not only considers the student's overall ability and the question difficulty, but also integrates the abilities that the student is familiar and unfamiliar with, making the prediction more interpretable.

3.3.5 Model Training DisKT applies a binary cross-entropy loss to directly optimize the assessment of knowledge state:

$$\mathcal{L}_{bce} = -\sum_{i=1}^{t}(r_i \log \hat{r}_i + (1 - r_i)\log(1 - \hat{r}_i)). \tag{15}$$

In addition, in order to expedite model convergence and ensure that the model learns two different types of abilities, familiar and unfamiliar, DisKT introduces an additional regularization term to constrain model learning:

$$\mathcal{L}_{cl} = \|S^{+}_{1:t} - S^{-}_{1:t}\|. \tag{16}$$

Therefore, the final objective function of DisKT is:

$$\mathcal{L}_{DisKT} = \mathcal{L}_{bce} - \mathcal{L}_{cl}. \tag{17}$$

4 Experiments

We conduct extensive experiments, aiming to answer the following four research questions to demonstrate the effectiveness of DisKT:

- **RQ1:** How does DisKT perform compare to various state-of-the-art KT models?
- **RQ2:** How does DisKT alleviate cognitive bias compared to the most advanced KT models?
- **RQ3:** How effective is DisKT in shielding guessing and mistaking?
- **RQ4:** What are the impacts of the components (*e.g.*, contradictory attention) on DisKT?

4.1 Experimental Setup

4.1.1 Datasets We evaluate the performance of DisKT on 11 public datasets: assist09, algebra05, algebra06, statics, ednet, prob, linux, comp, database, spanish, slepemapy. The introduction and detailed processing of the datasets can be found in Appendix A, and Table 4 presents the statistics of the processed datasets.

4.1.2 Baselines We compare DisKT with 14 state-of-the-art models as follows: DKT [51], DKVMN [76], SKVMN [1], Deep-IRT [71], GKT [41], SAKT [42], AKT [20], ATKT [22], CL4KT [33], CoreKT [14], DTransformer [73], simpleKT [37], FoLiBiKT [28], sparseKT [27], Mamba4KT [8], MIKT [62]. Their introductions can be found in Appendix B.

4.1.3 Implementation Details We adopt 5-fold cross-validation and folds are split based on the students. 10% of the training set is used for validation, which is not only used for parameter tuning but for early stopping strategy, that is, if AUC does not improve within 10 epochs, the training is halted. We focus on the most recent 100 interactions per student, as this recent information is crucial for future predictions [33]. The models are optimized by Adam [31] with the following settings: the batch size is 512, the learning rate is set to 0.001, the dropout rate is 0.05, and the embedding dimension is 64. All models are trained in PyTorch [45] on a Linux server with two GeForce RTX 3090s. Following the previous works [28, 33, 58], the evaluation metrics include Area Under the ROC Curve (AUC), Accuracy (ACC) and Root Mean Square Error (RMSE). Our code and datasets are available at https://anonymous.4open.science/r/DisKT.

4.2 Comparison with SOTA (RQ1 & RQ2)

4.2.1 Overall Performance *w.r.t.* Accuracy Table 2 presents the evaluation performance of the compared models in terms of AUC, ACC, and RMSE. Overall, our DisKT consistently outperforms other baselines on all metrics across all datasets. The main observations are as follows:

- DisKT has quite substantial evaluation performance on almost all datasets. Specifically, in terms of AUC, DisKT's average relative improvement over the strongest baseline on all datasets is 3.2%. These impressive results demonstrate the effectiveness of our causal modeling, enabling DisKT to eliminate the influence of confounder (*i.e.*, student historical correct rate distribution) in the student representation, and therefore, the model can generate correct cognition for underperformers and overperformers, thereby improving the accuracy of the evaluation.
- Interpretable models like Deep-IRT and CoreKT show poorer performance in biased data due to sacrificing performance for interpretability. For instance, CoreKT instantiated based on AKT generally performs worse than AKT. In contrast, DisKT eliminates the spurious correlation between user representation and prediction score within the model, thereby alleviating the amplification of cognitive bias, bringing performance improvements while also enhancing interpretability.
- Compared to other neural network structures, the attention mechanism, especially Transformer, significantly affects the performance of KT models. This is consistent with the research results in [37, 38]. This gap is more pronounced on larger datasets (the statistical information of the datasets is shown in Table 4), indicating the superiority of the attention mechanism for large-scale real-world datasets, *i.e.*, it is expert in capturing long-term dependencies, thus it can extract rich information from large-scale data as [20, 37] described. And DisKT significantly outperforms other Transformer-based models in larger datasets, proving that our DisKT can better exploit the powerful potential of Transformer.

4.2.2 Performance on Alleviating Cognitive Bias Due to data bias, the KT model tends to exhibit cognitive bias in different student populations: overperformers can't solve difficult questions, while underperformers tend to get simple questions right. To further evaluate the effectiveness of DisKT in alleviating cognitive bias, we conduct experiments on synthetic data. Specifically, according to the classical test theory [9], we build three datasets of different bias strengths based on ednet according to the frequency of correct

Table 2: Comparison of DisKT and 14 KT models on 11 datasets. The averages across five test folds are reported. Best results in bold, next best underlined. %Improv. denotes the relative performance improvement achieved by DisKT over the strongest baseline. * and ** indicate that the improvements over the strongest baseline are statistically significant, with p <0.05 and p <0.01, respectively. A model with ✓indicates that it is interpretable.

Dataset	Metric	DKT	DKVMN	SKVMN	Deep-IRT✓	GKT	SAKT	AKT	ATKT	CL4KT	CoreKT✓	DTransformer	simpleKT	FoLiBiKT	sparseKT	Mamba4KT	MIKT✓	DisKT✓	%Improv.
assist09	AUC ↑	0.7591	0.7570	0.7434	0.7566	0.7484	0.7348	0.7705	0.7543	0.7597	0.7415	0.7508	0.7709	0.7710	0.7670	0.7632	0.7693	**0.7923****	2.76%
	ACC ↑	0.7166	0.7172	0.7084	0.7180	0.7127	0.7000	0.7192	0.7171	0.7194	0.7020	0.7042	0.7209	0.7165	0.7092	0.7163	0.7182	**0.7275**	0.92%
	RMSE ↓	0.4333	0.4333	0.4391	0.4334	0.4374	0.4442	0.4349	0.4348	0.4339	0.4436	0.4505	0.4372	0.4356	0.4396	0.4344	0.4339	**0.4298**	-0.81%
	E_{KL} ↓	2.2937	2.7656	1.6577	1.5838	3.1287	1.9080	2.3615	3.2528	2.5203	0.9663	2.6926	2.6387	2.4640	3.3186	2.1881	1.2552	**0.8272**	-14.40%
algebra05	AUC	0.7780	0.7713	0.6260	0.7698	0.7806	0.7493	0.7932	0.7624	0.7864	0.7579	0.7694	0.7874	0.7923	0.7806	0.7793	0.7912	**0.8033****	1.27%
	ACC	0.7893	0.7896	0.7570	0.7886	0.7889	0.7800	0.7938	0.7882	0.7940	0.7643	0.7877	0.7927	0.7925	0.7856	0.7896	0.7940	**0.8009****	0.87%
	RMSE	0.3854	0.3862	0.4212	0.3865	0.3851	0.3942	0.3811	0.3899	0.3833	0.4038	0.3906	0.3842	0.3818	0.3875	0.3837	0.3824	**0.3786***	-0.66%
	E_{KL}	1.5390	1.7926	1.5537	1.1326	2.7290	1.0647	3.3858	3.3612	2.9318	1.0453	2.8382	2.6958	2.4516	3.6268	2.1229	1.1556	0.6731	-35.61%
algebra06	AUC	0.7598	0.7685	0.6294	0.7663	0.7578	0.7330	0.7740	0.7426	0.7714	0.7509	0.7391	0.7695	0.7731	0.7694	0.7652	0.7721	**0.7846****	1.37%
	ACC	0.7934	0.7989	0.7762	0.7977	0.7874	0.7845	0.7953	0.7878	0.7968	0.7710	0.7871	0.7923	0.7967	0.7940	0.7963	0.8001	**0.8033****	0.6%
	RMSE	0.3823	0.3784	0.4089	0.3794	0.3851	0.3921	0.3797	0.3891	0.3798	0.3946	0.3892	0.3814	0.3787	0.3804	0.3801	0.3784	**0.3755**	-0.77%
	E_{KL}	1.6260	1.8720	1.7570	1.4124	2.1218	1.7843	3.0552	3.4929	3.1842	1.2619	2.9758	2.1577	2.6260	2.3219	3.0274	0.9790	0.4366	-55.40%
statics	AUC	0.7611	0.7596	0.6692	0.7515	0.7528	0.7304	0.7999	0.7421	0.7783	0.7894	0.7690	0.7888	0.7988	0.7887	0.7732	0.7812	**0.8086****	1.09%
	ACC	0.7643	0.7688	0.7211	0.7661	0.7657	0.7560	0.7761	0.7658	0.7691	0.7598	0.7732	0.7723	0.7794	0.7775	0.7713	0.7805	**0.7883**	1.14%
	RMSE	0.4027	0.4019	0.4359	0.4059	0.4017	0.4148	0.3889	0.4130	0.4000	0.3990	0.3966	0.3924	0.3894	0.3909	0.5004	0.3903	**0.3823***	-1.7%
	E_{KL}	1.2465	1.0537	1.3327	1.0583	1.9220	1.0112	2.1257	2.3292	2.2731	0.9330	2.3550	2.0695	1.8791	2.8956	1.5234	0.9990	0.3634	-61.05%
ednet	AUC	0.6589	0.6657	0.6531	0.6656	0.6569	0.6499	0.7003	0.6544	0.6651	0.6697	0.6978	0.7048	0.6995	0.7006	0.6654	0.6978	**0.7384****	4.8%
	ACC	0.6289	0.6346	0.6259	0.6337	0.6193	0.6240	0.6592	0.6243	0.6349	0.6284	0.6559	0.6573	0.6582	0.6557	0.6372	0.6563	**0.6863****	4.11%
	RMSE	0.4771	0.4756	0.4784	0.4759	0.4788	0.4809	0.4746	0.4797	0.4759	0.4762	0.4799	0.4730	0.4756	0.4727	0.4783	0.4742	**0.4592****	-2.86%
	E_{KL}	1.5759	1.8922	1.8992	1.4677	2.8940	1.4293	3.5788	3.5569	2.9641	1.3239	2.9710	2.9339	2.7848	3.0593	2.2640	1.3282	1.0915	-17.82%
prob	AUC	0.7159	0.7192	0.7038	0.7190	0.7098	0.7100	0.7376	0.7062	0.7213	0.7293	0.7354	0.7265	0.7270	0.7437	0.7153	0.7386	**0.7731****	3.95%
	ACC	0.6786	0.6888	0.6768	0.6900	0.6806	0.6818	0.7015	0.6849	0.6877	0.6889	0.6922	0.6971	0.6974	0.7057	0.6792	0.7016	**0.7215****	2.24%
	RMSE	0.4543	0.4520	0.4564	0.4521	0.4555	0.4562	0.4491	0.4585	0.4524	0.4537	0.4518	0.4498	0.4522	0.4483	0.4518	0.4493	**0.4353****	-2.9%
	E_{KL}	2.2592	1.9730	2.1217	1.9318	3.0130	1.9626	3.2336	2.9985	3.0381	1.5459	2.9324	3.1385	2.6544	3.3783	2.6266	1.4151	1.1539	-18.45%
linux	AUC	0.7421	0.7470	0.6991	0.7441	0.7402	0.7348	0.8225	0.7532	0.7580	0.7837	0.8211	0.8221	0.8216	0.8249	0.7638	0.8215	**0.8622****	4.52%
	ACC	0.7625	0.7643	0.7535	0.7634	0.7612	0.7595	0.7977	0.7657	0.7674	0.7576	0.7979	0.7968	0.7945	0.7983	0.7724	0.7974	**0.8152****	2.12%
	RMSE	0.4042	0.4029	0.4151	0.4038	0.4067	0.4075	0.3741	0.4010	0.4002	0.4047	0.3742	0.3746	0.3756	0.3734	0.3816	0.3742	**0.3601****	-3.56%
	E_{KL}	1.3783	1.6537	1.3766	1.0416	2.5219	1.1474	2.8558	3.0674	2.7366	1.2180	2.0694	2.7254	2.0221	3.0517	2.5250	1.3688	1.1726	-3.72%
comp	AUC	0.7239	0.7170	0.6631	0.7150	0.7132	0.7082	0.7986	0.7256	0.7243	0.7420	0.7988	0.8000	0.7979	0.7964	0.7412	0.8004	**0.8324****	4.05%
	ACC	0.8037	0.8017	0.7991	0.8015	0.7914	0.8000	0.8203	0.8048	0.8040	0.7825	0.8217	0.8191	0.8196	0.8197	0.8087	0.8184	**0.8264***	0.57%
	RMSE	0.3788	0.3808	0.3899	0.3813	0.3878	0.3833	0.3589	0.3784	0.3805	0.3915	0.3582	0.3585	0.3592	0.3591	0.3703	0.3592	**0.3510***	-2.01%
	E_{KL}	1.4752	1.8255	1.3531	1.4114	2.5373	1.2133	3.1581	3.2651	2.8347	0.9743	2.5250	2.7168	2.5230	3.8315	2.3617	1.7811	0.6139	-36.99%
database	AUC	0.7490	0.7531	0.6924	0.7498	0.7497	0.7419	0.8263	0.7546	0.7587	0.7839	0.8184	0.8272	0.8253	0.8367	0.7653	0.8342	**0.8769****	4.8%
	ACC	0.8336	0.8340	0.8278	0.8330	0.8327	0.8317	0.8485	0.8346	0.8328	0.7883	0.8478	0.8497	0.8500	0.8531	0.8392	0.8490	**0.8690****	1.86%
	RMSE	0.3530	0.3522	0.3644	0.3528	0.3538	0.3553	0.3310	0.3518	0.3523	0.3830	0.3328	0.3301	0.3302	0.3266	0.3492	0.3295	**0.3090****	-5.39%
	E_{KL}	1.5489	1.7857	1.6278	1.4311	2.6658	1.3567	2.8020	2.9272	2.8333	1.1717	2.5254	2.5921	2.3290	2.7767	2.2149	1.3156	1.0446	-10.85%
spanish	AUC	0.8029	0.8081	0.7277	0.8032	0.8114	0.7950	0.8391	0.8047	0.8202	0.8273	0.8170	0.8408	0.8399	0.8395	0.8154	0.8374	**0.8529****	1.44%
	ACC	0.7443	0.7508	0.6952	0.7461	0.7496	0.7417	0.7745	0.7545	0.7550	0.7613	0.7513	0.7734	0.7735	0.7718	0.7562	0.7703	**0.7847**	1.45%
	RMSE	0.4156	0.4145	0.4482	0.4177	0.4155	0.4236	0.3968	0.4186	0.4119	0.4053	0.4108	0.3963	0.3962	0.3959	0.4127	0.3972	**0.3872****	-2.2%
	E_{KL}	1.6210	2.1518	1.6243	1.2779	2.8264	1.0295	2.7911	3.3182	3.0481	1.1414	2.5848	2.4377	2.5896	3.0693	2.1793	1.7428	1.0554	-7.53%
slepemapy	AUC	0.6861	0.6989	0.6473	0.6935	0.6539	0.6709	0.7258	0.6952	0.7097	0.7135	0.7217	0.7269	0.7230	0.7255	0.7052	0.7293	**0.7632****	4.99%
	ACC	0.7780	0.7789	0.7786	0.7782	0.7844	0.7739	0.7835	0.7790	0.7857	0.7238	0.7865	0.7868	0.7826	0.7876	0.7796	0.7875	**0.7944****	0.86%
	RMSE	0.3991	0.3978	0.4046	0.3998	0.4009	0.4050	0.3906	0.3983	0.3938	0.4205	0.3892	0.3889	0.3916	0.3903	0.3954	0.3874	**0.3808****	-2.08%
	E_{KL}	1.5677	1.5361	1.8969	1.4273	2.4737	1.2688	2.7324	3.1252	2.5348	1.2598	2.7514	2.7453	2.6287	3.5356	2.3774	1.2623	0.9866	-21.68%

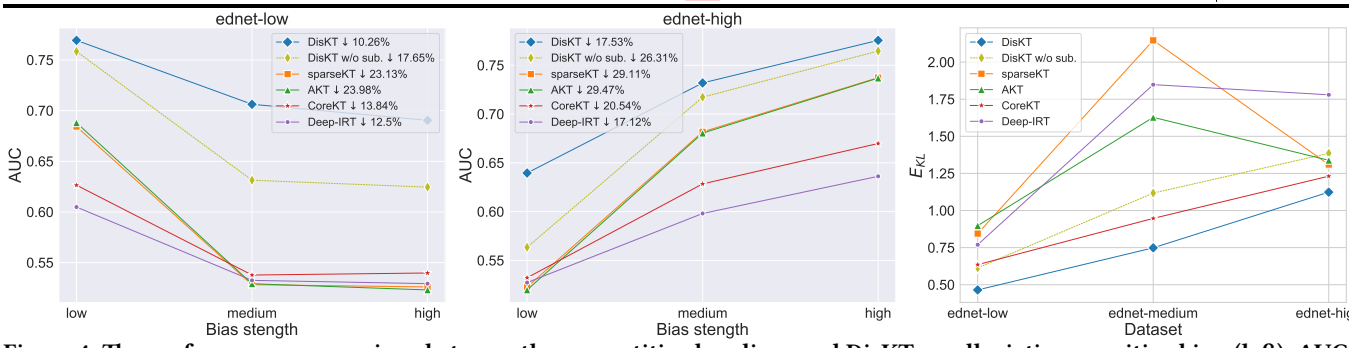


Figure 4: The performance comparison between the competitive baselines and DisKT on alleviating cognitive bias. (left): AUC performance changes of DisKT and baselines optimized with different bias strengths on ednet, tested with ednet-low and ednet-high respectively. (right): E_{KL} scores of several representative models on three datasets with different bias strengths.

responses (< 60%, 60%~80%, and ≥ 80%): ednet-low, ednet-medium, and ednet-high. These datasets are consistent with the settings in Section 4.1.3. We choose two interpretable models, Deep-IRT and CoreKT, and two best-performing models, AKT and sparseKT, as baselines for studying cognitive bias. We test the models optimized by the three datasets with ednet-low and ednet-high, respectively, and their AUC performance changes are shown in Figure 4 left. We make the following observations. 1) The AUC performance of all models trained by ednet-high and tested with the ednet-low has decreased, which indicates that models that have only seen high-accuracy, *i.e.*, simple questions, overload the cognition of underperformers. Similarly, the AUC performance of all models trained by ednet-low and tested with the ednet-high has also decreased, reflecting the model's cognition underload for overperformers. 2) Compared to other strong competitors, DisKT effectively alleviates the two cognitive biases while maintaining optimal evaluation performance. In contrast, even though interpretable models like Deep-IRT and CoreKT achieve good results, this comes at the cost of sacrificing evaluation performance. 3) Cognitive bias greatly affects the evaluation performance of the baselines, even rendering

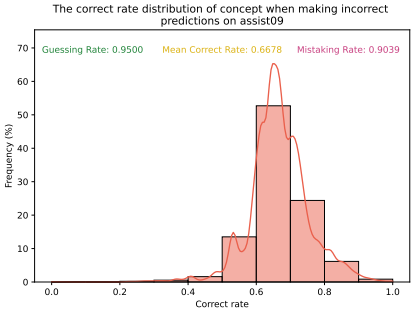


Figure 5: Empirical evidence of guessing and mistaking verified by DKT applied to the assist09 dataset.

their evaluation ineffective, *i.e.*, AUC performance is around 0.5, while DisKT still maintains AUC performance above 0.6. However, the AUC performance of DisKT drops significantly after removing the causal effect (DisKT $w/o.$ sub.), indicating the correctness of our modeling based on causal effect.

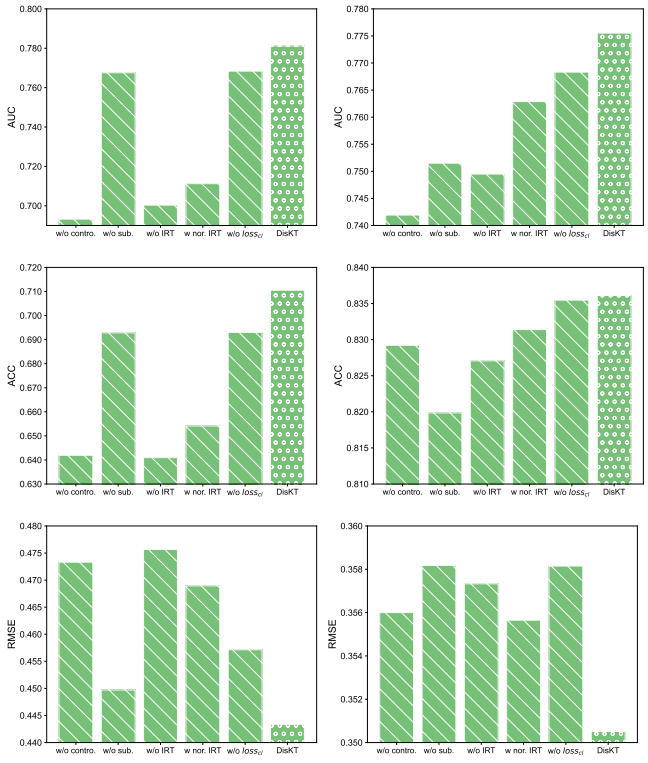
However, some metrics like AUC can only indirectly reflect the impact of cognitive bias on the model. In KT, there is a lack of a direct measure that can reflect the amplification degree of cognitive bias in the model. To fill this gap, inspired by [63], we have designed a calibration metric E_{KL}, which measures the gap between the actual and predicted correct rate distributions when the model makes incorrect predictions. The calculation method can be found in Appendix D. Higher E_{KL} scores suggest a more serious issue of cognitive bias. The E_{KL} scores of several representative models on three datasets with different bias strengths are shown in Figure 4 right. We can see that, regardless of the bias strength of the dataset, DisKT always maintains the lowest E_{KL} scores. Meanwhile, CoreKT achieves sub-optimal results due to its mitigation of answer bias. Deep-IRT, due to its simple structure, cannot adapt to datasets with larger bias strengths. AKT and sparseKT learn data biases but are unable to suppress the amplification of cognitive bias within the model, leading to the worst results. Additionally, Table 2 provides the E_{KL} scores of DisKT compared to other baselines under the dataset with different unseen biases, and DisKT generally has lower E_{KL} scores. This further proves the effectiveness of DisKT in alleviating cognitive bias.

4.3 In-depth Analysis (RQ3 & RQ4)

4.3.1 Effect of Shielding Guessing and Mistaking In Section 1, we find that there existing the contradictory psychology (*e.g.*, guessing and mistaking) in the students' biased data. We have provided empirical evidence of guessing and mistaking using DKT on the assist09 dataset. Specifically, we have analyzed the correct rate distribution of concept when DKT makes incorrect predictions. We consider concepts with a correct rate of less than or equal to 0.3 as difficult concepts and those with a correct rate of greater than or equal to 0.7 as easy concepts. We then calculate the proportion of correct responses by students for difficult concepts (Guessing Rate) and the proportion of incorrect responses for easy concepts (Mistaking Rate). The results, as shown in Figure 5, indicate that the Guessing Rate and Mistaking Rate are surprisingly high at 0.9500 and 0.9039, respectively. This suggests that contradictory psychology (*e.g.*, guessing and mistaking) can easily have a negative impact on the assessment of the real knowledge state. Intuitively, the Guessing Rate and Mistaking Rate measure the adverse effects of guessing and mistaking on the model, respectively, with lower values indicating that the model is better at shielding the impact of these inevitable psychological factors. Table 3 presents the contradictory metrics (Guessing Rate and Mistaking Rate) of DisKT and several representative baselines on three datasets with different bias strengths. We note that Deep-IRT achieves good results in the Guessing Rate on all datasets, indicating that purely introducing question difficulty helps reduce the adverse impact of guessing. Meanwhile, DisKT achieves the best results in Mistaking Rate on all datasets while generally outperforming other baselines in Guessing Rate. However, the results of DisKT without the contradictory attention (DisKT $w/o.$ con.) are generally pessimistic, confirming the effectiveness of our designed contradictory attention in shielding guessing and mistaking.

Table 3: Performance comparison in terms of shielding guessing and mistaking.

Dataset	Metric	Deep-IRT	CoreKT	AKT	sparseKT	DisKT w/o. con.	DisKT
ednet-low	Guessing Rate↓	**0.7372**	0.7968	0.9633	0.9577	0.9417	0.7515
	Mistaking Rate↓	0.9129	0.8292	0.8631	0.9295	0.8710	**0.7406**
ednet-medium	Guessing Rate	**0.3627**	0.8000	0.9853	0.9855	0.9559	0.7778
	Mistaking Rate	0.9557	0.9097	0.8298	0.8304	0.9241	**0.6904**
ednet-high	Guessing Rate	0.3796	0.5612	0.4397	0.4427	0.3985	**0.2797**
	Mistaking Rate	0.9596	0.9233	0.9314	0.9304	0.9851	**0.8895**



(a) ednet-low **(b) ednet-high**

Figure 6: Ablation study on ednet-low and ednet-high.

4.3.2 Ablation Study We have constructed five variants of DisKT to explore the impact of different components on DisKT, as shown in Figure 6. Specifically, in addition to the previously mentioned

"w/o. sub." and "w/o. con.", "w/o. IRT" removes the variant IRT module, "w. nor. IRT" uses a normal IRT module, and "w/o. $loss_{cl}$" omits the loss function $loss_{cl}$. The following observations are made: (1) "w/o. con." shows a similar degree of performance decline across both datasets, highlighting the importance of contradiction attention in shielding guessing and mistaking and effectively tracking knowledge states. (2) "w/o. sub." exhibits a slight performance decline on ednet-low and a sharp decline on ednet-high, indicating that DisKT, which models based on causal effects, is more adaptable to data with high bias. (3) "w. nor. IRT" experiences a significant performance decline, and "w/o. IRT" even more so, emphasizing not only the effectiveness of the IRT module but also the advantages of our proposed the variant IRT module. Additionally, we have provided the interpretable prediction process of variant IRT in the Appendix C. (4) "w/o. $loss_{cl}$" shows a slight performance decline, indicating that the regularization term $loss_{cl}$ accelerates convergence while also effectively learning the abilities of familiarity and unfamiliarity.

5 Conclusion and Future Work

In this work, we elucidate that cognitive bias within KT models stems from the confounder from a causal perspective. In response, we propose the DisKT model based on causal effect, effectively nullifying the impact of the confounder in student representation. Moreover, DisKT incorporates a contradiction attention to shield the contradictory psychology (*e.g.*, guessing and mistaking) in the students' biased data. Meanwhile, we innovate a variant of IRT to enhance the interpretability of model predictions. Our findings, supported by rigorous experiments across 11 benchmarks and 3 synthesized datasets, reveal that DisKT not only significantly alleviates cognitive bias but also surpasses 14 baseline models in terms of evaluation accuracy.

The avenues of future work include *i*) further exploration of educational psychology, such as forgetting, *ii*) investigation of the critical points between simple and difficult questions for different students and *iii*) discovery of more fine-grained causal relations.

References

[1] Ghodai Abdelrahman and Qing Wang. 2019. Knowledge tracing with sequential key-value memory networks. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*. 175–184.

[2] Ghodai Abdelrahman, Qing Wang, and Bernardo Nunes. 2023. Knowledge tracing: A survey. *Comput. Surveys* 55, 11 (2023), 1–37.

[3] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[4] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer Normalization. arXiv:1607.06450

[5] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[6] Andrea Bellavia and Linda Valeri. 2017. Decomposition of the Total Effect in the Presence of Multiple Mediators and Interactions. *American Journal of Epidemiology* 187, 6 (11 2017), 1311–1318. https://doi.org/10.1093/aje/kwx355 arXiv:https://academic.oup.com/aje/article-pdf/187/6/1311/29159894/kwx355.pdf

[7] Nicholas C Burbules and Marcia C Linn. 1988. Response to contradiction: Scientific reasoning during adolescence. *Journal of Educational Psychology* 80, 1 (1988), 67.

[8] Yang Cao and Wei Zhang. 2024. Mamba4KT: An Efficient and Effective Mamba-based Knowledge Tracing Model. *arXiv preprint arXiv:2405.16542* (2024).

[9] Joseph C. Cappelleri, J. Jason Lundy, and Ron D. Hays. 2014. Overview of Classical Test Theory and Item Response Theory for the Quantitative Assessment of Items in Developing Patient-Reported Outcomes Measures. *Clinical Therapeutics* 36, 5 (2014), 648–662. https://doi.org/10.1016/j.clinthera.2014.04.006

[10] David Carless. 2019. Feedback loops and the longer-term: towards feedback spirals. *Assessment & Evaluation in Higher Education* 44, 5 (2019), 705–714. https://doi.org/10.1080/02602938.2018.1531108 arXiv:https://doi.org/10.1080/02602938.2018.1531108

[11] Allison J. B. Chaney, Brandon M. Stewart, and Barbara E. Engelhardt. 2018. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *Proceedings of the 12th ACM Conference on Recommender Systems* (Vancouver, British Columbia, Canada) *(RecSys '18)*. Association for Computing Machinery, New York, NY, USA, 224–232. https://doi.org/10.1145/3240323.3240370

[12] Youngduck Choi, Youngnam Lee, Dongmin Shin, Junghyun Cho, Seoyon Park, Seewoo Lee, Jineon Baek, Chan Bae, Byungsoo Kim, and Jaewe Heo. 2020. Ednet: A large-scale hierarchical dataset in education. In *Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part II 21*. Springer, 69–73.

[13] Albert T Corbett and John R Anderson. 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction* 4 (1994), 253–278.

[14] Chaoran Cui, Hebo Ma, Chen Zhang, Chunyun Zhang, Yumo Yao, Meng Chen, and Yuling Ma. 2023. Do We Fully Understand Students' Knowledge States? Identifying and Mitigating Answer Bias in Knowledge Tracing. *arXiv preprint arXiv:2308.07779* (2023).

[15] Yuhao Dan, Zhikai Lei, Yiyang Gu, Yong Li, Jianghao Yin, Jiaju Lin, Linhao Ye, Zhiyan Tie, Yougen Zhou, Yilei Wang, et al. 2023. Educhat: A large-scale language model-based chatbot system for intelligent education. *arXiv preprint arXiv:2308.02773* (2023).

[16] Mingyu Feng, Neil Heffernan, and Kenneth Koedinger. 2009. Addressing the assessment challenge with an online system that tutors as it assesses. *User modeling and user-adapted interaction* 19 (2009), 243–266.

[17] Robert B Frary. 1988. Formula scoring of multiple-choice tests (correction for guessing). *Educational measurement: Issues and practice* 7, 2 (1988), 33–38.

[18] Wayne A Fuller. 2009. *Measurement error models*. John Wiley & Sons.

[19] Theophile Gervet, Ken Koedinger, Jeff Schneider, and Tom Mitchell. 2020. When is Deep Learning the Best Approach to Knowledge Tracing? *Journal of Educational Data Mining* 12, 3 (Oct. 2020), 31–54. https://doi.org/10.5281/zenodo.4143614

[20] Aritra Ghosh, Neil Heffernan, and Andrew S Lan. 2020. Context-aware attentive knowledge tracing. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. 2330–2339.

[21] Jose Gonzalez-Brenes, Yun Huang, and Peter Brusilovsky. 2014. General Features in Knowledge Tracing to Model Multiple Subskills, Temporal Item Response Theory, and Expert Knowledge, In The 7th International Conference on Educational Data Mining. *The 7th International Conference on Educational Data Mining*, 84 – 91. http://d-scholarship.pitt.edu/26017/

[22] Xiaopeng Guo, Zhijie Huang, Jie Gao, Mingyu Shang, Maojing Shu, and Jun Sun. 2021. Enhancing Knowledge Tracing via Adversarial Training. arXiv:2108.04430 [cs.CY]

[23] Mithun Haridas, Nirmala Vasudevan, Surya Gayathry, Georg Gutjahr, Raghu Raman, and Prema Nedungadi. 2019. Feature-Aware knowledge tracing for generation of concept-knowledge reports in an intelligent tutoring system. In *2019 IEEE Tenth International Conference on Technology for Education (T4E)*. IEEE, 142–145.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *arXiv preprint arXiv:1512.03385* (2015).

[25] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[26] Liya Hu, Zhiang Dong, Jingyuan Chen, Guifeng Wang, Zhihua Wang, Zhou Zhao, and Fei Wu. 2023. PTADisc: A Cross-Course Dataset Supporting Personalized Learning in Cold-Start Scenarios. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 44976–44996. https://proceedings.neurips.cc/paper_files/paper/2023/file/8cf04c64d1734e5f7e63418a2a4d49de-Paper-Datasets_and_Benchmarks.pdf

[27] Shuyan Huang, Zitao Liu, Xiangyu Zhao, Weiqi Luo, and Jian Weng. 2023. Towards Robust Knowledge Tracing Models via k-Sparse Attention. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (, Taipei, Taiwan,) *(SIGIR '23)*. Association for Computing Machinery, New York, NY, USA, 2441–2445. https://doi.org/10.1145/3539618.3592073

[28] Yoonjin Im, Eunseong Choi, Heejin Kook, and Jongwuk Lee. 2023. Forgetting-aware Linear Bias for Attentive Knowledge Tracing. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23)*. ACM. https://doi.org/10.1145/3583780.3615191

[29] Tumaini Kabudi, Ilias Pappas, and Dag Håkon Olsen. 2021. AI-enabled adaptive learning systems: A systematic mapping of the literature. *Computers and Education: Artificial Intelligence* 2 (2021), 100017. https://doi.org/10.1016/j.caeai.2021.100017

[30] Firuz Kamalov, David Santandreu Calonge, and Ikhlaas Gurrib. 2023. New Era of Artificial Intelligence in Education: Towards a Sustainable Multifaceted Revolution. *Sustainability* 15, 16 (2023). https://doi.org/10.3390/su151612451

[31] Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. arXiv:1412.6980 [cs.LG]

[32] K. Koedinger, R. Baker, Kyle Cunningham, Alida Skogsholm, Brett Leber, and John C. Stamper. 2010. A Data Repository for the EDM Community: The PSLC DataShop. https://api.semanticscholar.org/CorpusID:63729977

[33] Wonsung Lee, Jaeyoon Chun, Youngmin Lee, Kyoungsoo Park, and Sungrae Park. 2022. Contrastive Learning for Knowledge Tracing. In *Proceedings of the ACM Web Conference 2022* (, Virtual Event, Lyon, France,) *(WWW '22)*. Association for Computing Machinery, New York, NY, USA, 2330–2338. https://doi.org/10.1145/3485447.3512105

[34] Chen Lin and Min Chi. 2016. Intervention-bkt: incorporating instructional interventions into bayesian knowledge tracing. In *Intelligent Tutoring Systems: 13th International Conference, ITS 2016, Zagreb, Croatia, June 7-10, 2016. Proceedings 13*. Springer, 208–218.

[35] Robert V Lindsey, Mohammad Khajah, and Michael C Mozer. 2014. Automatic Discovery of Cognitive Skills to Improve the Prediction of Student Learning. In *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger (Eds.), Vol. 27. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2014/file/d840cc5d906c3e9c84374c8919d2074e-Paper.pdf

[36] Qi Liu, Yan Zhuang, Haoyang Bi, Zhenya Huang, Weizhe Huang, Jiatong Li, Junhao Yu, Zirui Liu, Zirui Hu, Yuting Hong, Zachary A. Pardos, Haiping Ma, Mengxiao Zhu, Shijin Wang, and Enhong Chen. 2024. Survey of Computerized Adaptive Testing: A Machine Learning Perspective. arXiv:2404.00712 [cs.LG]

[37] Zitao Liu, Qiongqiong Liu, Jiahao Chen, Shuyan Huang, and Weiqi Luo. 2023. simpleKT: A Simple But Tough-to-Beat Baseline for Knowledge Tracing. arXiv:2302.06881 [cs.LG]

[38] Zitao Liu, Qiongqiong Liu, Jiahao Chen, Shuyan Huang, Jiliang Tang, and Weiqi Luo. 2022. pyKT: A Python Library to Benchmark Deep Learning based Knowledge Tracing Models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

[39] Frederic M Lord. 2012. *Applications of item response theory to practical testing problems*. Routledge.

[40] Yu Lu, Deliang Wang, Penghe Chen, and Zhi Zhang. 2024. Design and Evaluation of Trustworthy Knowledge Tracing Model for Intelligent Tutoring System. *IEEE Transactions on Learning Technologies* (2024).

[41] Hiromi Nakagawa, Yusuke Iwasawa, and Yutaka Matsuo. 2019. Graph-based knowledge tracing: modeling student proficiency using graph neural network. In *IEEE/WIC/ACM International Conference on Web Intelligence*. 156–163.

[42] Shalini Pandey and George Karypis. 2019. A self-attentive model for knowledge tracing. *arXiv preprint arXiv:1907.06837* (2019).

[43] Jan Papoušek, Radek Pelánek, and Vít Stanislav. 2016. Adaptive Geography Practice Data Set. *Journal of Learning Analytics* 3, 2 (Sep. 2016), 317–321. https://doi.org/10.18608/jla.2016.32.17

[44] Zachary A. Pardos, Matthew Tang, Ioannis Anastasopoulos, Shreya K. Sheel, and Ethan Zhang. 2023. OATutor: An Open-source Adaptive Tutoring System and Curated Content Library for Learning Sciences Research. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (, Hamburg, Germany,) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 416, 17 pages. https://doi.org/10.1145/3544548.3581574

[45] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. arXiv:1912.01703 [cs.LG]

[46] Judea Pearl. 2001. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence* (Seattle, Washington) *(UAI'01)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 411–420.

[47] Judea Pearl. 2009. Causal inference in statistics: An overview. *Statistics Surveys* 3, none (2009), 96 – 146. https://doi.org/10.1214/09-SS057

[48] Judea Pearl. 2009. *Causality*. Cambridge university press.

[49] Judea Pearl. 2010. An introduction to causal inference. *The international journal of biostatistics* 6, 2 (2010).

[50] Pipatsarun Phobun and Jiracha Vicheanpanya. 2010. Adaptive intelligent tutoring systems for e-learning systems. *Procedia - Social and Behavioral Sciences* 2, 2 (2010), 4064–4069. https://doi.org/10.1016/j.sbspro.2010.03.641 Innovation and Creativity in Education.

[51] Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J Guibas, and Jascha Sohl-Dickstein. 2015. Deep Knowledge Tracing. In *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (Eds.), Vol. 28. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2015/file/bac9162b47c56fc8a4d2a519803d51b3-Paper.pdf

[52] Stephen Powell. 2018. The Book of Why: The New Science of Cause and Effect. Pearl, Judea, and Dana Mackenzie. 2018. Hachette UK. *Journal of MultiDisciplinary Evaluation* 14, 31 (Aug. 2018), 47–54. https://doi.org/10.56645/jmde.v14i31.507

[53] Liqing Qiu, Menglin Zhu, and Jingcheng Zhou. 2023. OPKT: Enhancing Knowledge Tracing with Optimized Pre-training Mechanisms in Intelligent Tutoring. *IEEE Transactions on Learning Technologies* (2023).

[54] Georg Rasch. 1993. *Probabilistic models for some intelligence and attainment tests.* ERIC.

[55] Steven P Reise and Niels G Waller. 2009. Item response theory and clinical measurement. *Annual review of clinical psychology* 5 (2009), 27–48.

[56] Paul Rosenbaum. 1984. Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika* 49, 3 (September 1984), 425–435. https://doi.org/10.1007/BF02306030

[57] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. 2016. Meta-learning with memory-augmented neural networks. In *International conference on machine learning*. PMLR, 1842–1850.

[58] Shuanghong Shen, Qi Liu, Enhong Chen, Zhenya Huang, Wei Huang, Yu Yin, Yu Su, and Shijin Wang. 2021. Learning process-consistent knowledge tracing. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*. 1452–1460.

[59] Xiangyu Song, Jianxin Li, Taotao Cai, Shuiqiao Yang, Tingting Yang, and Chengfei Liu. 2022. A survey on deep learning based knowledge tracing. *Knowledge-Based Systems* 258 (2022), 110036.

[60] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research* 15, 56 (2014), 1929–1958. http://jmlr.org/papers/v15/srivastava14a.html

[61] J Stamper, A Niculescu-Mizil, S Ritter, G Gordon, and K Koedinger. 2010. Algebra I 2005-2006 and Bridge to Algebra 2006-2007. Development data sets from KDD Cup 2010 Educational Data Mining Challenge.

[62] Jianwen Sun, Fenghua Yu, Qian Wan, Qing Li, Sannyuya Liu, and Xiaoxuan Shen. 2024. Interpretable Knowledge Tracing with Multiscale State Representation. In *Proceedings of the ACM on Web Conference 2024*. 3265–3276.

[63] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. 2020. Long-Tailed Classification by Keeping the Good and Removing the Bad Momentum Causal Effect. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1513–1524. https://proceedings.neurips.cc/paper_files/paper/2020/file/1091660f3dff84fd648efe31391c5524-Paper.pdf

[64] Qingyun Tang. 2023. A Case Study on The Application of Artificial Intelligence in Education Industry. In *2023 3rd International Conference on Modern Educational Technology and Social Sciences (ICMETSS 2023)*. Atlantis Press, 99–109.

[65] Tyler J VANDERWEELE. 2013. A Three-way Decomposition of a Total Effect into Direct, Indirect, and Interactive Effects. *Epidemiology (Cambridge, Mass.)* (2013).

[66] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[67] Wenjie Wang, Fuli Feng, Xiangnan He, Xiang Wang, and Tat-Seng Chua. 2021. Deconfounded recommendation for alleviating bias amplification. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 1717–1725.

[68] Wenjie Wang, Fuli Feng, Xiangnan He, Hanwang Zhang, and Tat-Seng Chua. 2021. Clicks can be Cheating: Counterfactual Recommendation for Mitigating Clickbait Issue. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (, Virtual Event, Canada,) *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 1288–1297. https://doi.org/10.1145/3404835.3462962

[69] Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. 2021. A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 15, 5 (2021), 1–46.

[70] Wendy M Yen and Anne R Fitzpatrick. 2006. Item response theory. *Educational measurement* 4 (2006), 111–153.

[71] Chun Kit Yeung. 2019. Deep-IRT: Make deep learning based knowledge tracing explainable using item response theory. In *Proceedings of the 12th International Conference on Educational Data Mining*. 683–686.

[72] Ramazan Yilmaz, Halil Yurdugül, Fatma Gizem Karaoğlan Yilmaz, Muhittin Şahin, Sema Sulak, Furkan Aydin, Mustafa Tepgeç, Cennet Terzi Müftüoğlu, and Ömer ORAL. 2022. Smart MOOC integrated with intelligent tutoring: A system architecture and framework model proposal. *Computers and Education: Artificial Intelligence* 3 (2022), 100092. https://doi.org/10.1016/j.caeai.2022.100092

[73] Yu Yin, Le Dai, Zhenya Huang, Shuanghong Shen, Fei Wang, Qi Liu, Enhong Chen, and Xin Li. 2023. Tracing Knowledge Instead of Patterns: Stable Knowledge Tracing with Diagnostic Transformer. In *Proceedings of the ACM Web Conference*

2023. 855–864.
[74] Li Yuan. 2024. Where does AI-driven Education, in the Chinese Context and Beyond, go next? *International Journal of Artificial Intelligence in Education* 34, 1 (2024), 31–41.
[75] Michael V Yudelson, Kenneth R Koedinger, and Geoffrey J Gordon. 2013. Individualized bayesian knowledge tracing models. In *Artificial Intelligence in Education: 16th International Conference, AIED 2013, Memphis, TN, USA, July 9-13, 2013. Proceedings 16*. Springer, 171–180.
[76] Jiani Zhang, Xingjian Shi, Irwin King, and Dit-Yan Yeung. 2017. Dynamic key-value memory networks for knowledge tracing. In *Proceedings of the 26th international conference on World Wide Web*. 765–774.
[77] Shengyu Zhang, Dong Yao, Zhou Zhao, Tat-Seng Chua, and Fei Wu. 2021. CauseRec: Counterfactual User Sequence Synthesis for Sequential Recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21)*. ACM. https://doi.org/10.1145/3404835.3462908
[78] Yan Zhuang, Qi Liu, Zhenya Huang, Zhi Li, Shuanghong Shen, and Haiping Ma. 2022. Fully adaptive framework: Neural computerized adaptive testing for online education. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 36. 4734–4742.

A Datasets

We evaluate the performance of DisKT on 11 public datasets:

- **assist09**[2] [16]: The assist09 dataset, composed of math exercises, is collected from the ASSISTment intelligent tutoring system in the school year 2009-2010 and is widely used as a standard benchmark in KT research.
- **algebra05, algebra06**[3] [61]: The algebra05 and algebra06 datasets come from the KDD Cup 2010 EDM Challenge, containing detailed step-level student responses to algebra questions.
- **statics**[4] [32]: The statics dataset is a collection of records from a college-level engineering statics course at Carnegie Mellon University during the Fall semester of 2011.
- **ednet**[5] [12]: The ednet dataset, collected by the multi-platform AI tutoring service Santa, stands as the largest publicly released interactive educational system dataset to date.
- **prob, comp, linux, database**[6] [26]: The prob, comp, linux, database datasets are collected from the Programming Teaching Assistant platform, specifically from course exercises in Probability and Statistics, Computational Thinking, Linux System, and Database Technology and Application.
- **spanish**[7] [35]: The spanish dataset is from middle-school students practicing Spanish exercises, including translations and applications of basic skills like verb conjugation, over a 15-week semester.
- **slepemapy**[8] [43]: The slepemapy dataset originates from slepemapy.cz, an online platform dedicated to the adaptive practice of geography facts.

Following the data preprocessing approach in [19], we exclude students with fewer than five interactions and all interactions involving nameless concepts. **Since a question may involve multiple concepts, we convert the unique combinations of concepts within a single question into a new concept.** The statistical information after processing is shown in Table 4. It's noted that we randomly sample 5000 students from three large datasets, ednet, comp and slepemapy.

Table 4: Statistics of 11 datasets. "#concepts*" denotes the total number of concepts after converting multiple concepts into a new concept.

Datasets	#students	#questions	#concepts	#concepts*	#interactions
assist09	3,644	17,727	123	150	281,890
algebra05	571	173,113	112	271	607,014
algebra06	1,138	129,263	493	550	1,817,450
statics	333	1,223	N/A	N/A	189,297
ednet	5,000	12,117	189	1,769	676,276
prob	512	1,054	247	247	42,869
comp	5,000	7,460	445	445	668,927
linux	4,375	2,672	281	281	365,027
database	5,488	3,388	291	291	990,468
spanish	182	409	221	221	578,726
slepemapy	5,000	2,723	1,391	1,391	625,523

[2]https://sites.google.com/site/assistmentsdata/home/2009-2010-assistment-data/skill-builder-data-2009-2010
[3]https://pslcdatashop.web.cmu.edu/KDDCup
[4]https://pslcdatashop.web.cmu.edu/DatasetInfo?datasetId=507
[5]https://github.com/riiid/ednet
[6]https://github.com/wahr0411/PTADisc
[7]https://github.com/robert-lindsey/WCRP
[8]https://www.fi.muni.cz/adaptivelearning/?a=data

B Baselines

We compare DisKT with 14 state-of-the-art models as follows:

- **DKT** [51]: DKT is the first model that employs Recurrent Neural Networks (RNNs) to solve the KT task. Over the past few years, it has been widely used as a standard baseline in KT research.
- **DKVMN** [76]: DKVMN leverages a dual-matrix approach to refine KT, using a static key matrix for mapping interconnections among concepts and a dynamic value matrix for real-time updates of a student's knowledge state.
- **SKVMN** [1]: SKVMN integrates the recurrent modeling capabilities of DKT with the advanced memory network structure of DKVMN, enhancing the representation and tracking of students' knowledge states over time.
- **Deep-IRT** [71]: Deep-IRT provides a detailed understanding of student learning trajectories and the difficulty of concepts, integrating the DKVMN for feature extraction with the psychometric insights of IRT.
- **GKT** [41]: GKT redefines the KT task by modeling the knowledge structure as a graph, transforming it into a node-level classification challenge.
- **SAKT** [42]: SAKT employs a self-attention mechanism within a Transformer architecture to dynamically weigh past learning interactions, capturing long-term dependencies and the relevance among concepts and historical interactions.
- **AKT** [20]: AKT accounts for the learner's tendency to forget over time within its monotonic attention framework by integrating embeddings inspired by the Rasch model.
- **ATKT** [22]: ATKT applies these perturbations to the original student interaction sequences, utilizing an attention-based LSTM framework.
- **CL4KT** [33]: CL4KT introduces a novel contrastive learning framework for KT, aiming to enhance representation learning by distinguishing between similar and dissimilar learning histories.
- **CoreKT** [14]: CORE framework enhances KT by addressing answer bias through a causality perspective. It differentiates between total and direct causal effects of questions on student responses to mitigate bias, improving the accuracy of tracing students' knowledge states. We introduce CORE with AKT, namely CoreKT.
- **DTransformer** [73]: DTransformer revolutionizes KT by diagnosing learner's knowledge states from question-level mastery using a novel architecture. It employs Temporal and Cumulative Attention (TCA) mechanisms for dynamic analysis and a contrastive learning-based algorithm for stable knowledge state tracking.
- **simpleKT** [37]: simpleKT introduces a simple but tough-to-beat baseline for KT, focusing on simplicity and robust performance across diverse KT datasets.
- **FoLiBiKT** [28]: FoLiBi, leveraging the forgetting-aware linear bias concept, innovatively addresses the challenge of modeling forgetting behavior in KT by introducing a linear bias mechanism. We introduce FoLiBi with AKT, namely FoLiBiKT.
- **sparseKT** [27]: sparseKT introduces a k-selection module designed to select items that achieve the highest attention scores, integrating two distinct sparsification strategies: soft-thresholding sparse attention and top-K sparse attention.
- **Mamba4KT** [8]: By leveraging Mamba, a state-space model supporting parallelized training and linear-time inference, Mamba4KT achieves efficient resource utilization, balancing time and space consumption.
- **MIKT** [62]: MIKT is a novel Knowledge Tracing model that combines coarse-grained and fine-grained representations to monitor students' domain and conceptual knowledge states, respectively. It leverages attention mechanisms and an IRT prediction module to enhance interpretability without sacrificing performance. The model extends the Rasch model to handle multi-concept questions and demonstrates superior performance and interpretability in experiments.

C Interpretable Prediction

To demonstrate the interpretable output of the variant IRT, we focus on the values of $d_{q_{1:t}}$ (question difficulty), X_{t+1} (student's overall ability), H^+_{t+1} (student's familiar abilities), and H^-_{t+1} (student's unfamiliar abilities) in equation 14. We randomly selected a student from the assist09 dataset and observed the prediction process of DisKT for the student's response to question 3127 and its related concept 35 at time 56. For intuitive analysis, we normalized them as follows:

$d_{q_{1:56}}$	X_{57}	H^+_{57}	H^-_{57}
0.3261	0.2158	0.7155	0.3214

From this, it can be seen that the question difficulty and the overall score weight assessed by the model for this question are close, and it cannot clearly predict whether the student has mastered the related concepts. However, the weight of the student's familiar abilities is much greater than that of the unfamiliar abilities. Therefore, the student is proficient in the question but may have made a mistake, leading to a lower overall score. Equation 14 mitigates the negative impact of mistaking, meaning that since the student has a good grasp of the related concepts, the predicted score should be higher. As a result, DisKT correctly predicts the probability of the student responding to the question correctly as 0.6739. Generally, we clarify the following template for assessing student state:

student	state	$d_{q_{1:t}}$	X_t	H^+_t	H^-_t
overperformer	ordinary	high	low	high	-
	mistaking	low	low	high	low
underperformer	ordinary	low	high	low	-
	guessing	high	high	low	high

Through the above template, the IRT variant distinguishes itself from traditional IRT by filtering out the contradictory psychology

of mistaking and guessing, and determining the student's true state through the student's ability and question difficulty.

D Evaluation Metric

In the experiments, we design a calibration metric E_{KL} to measure the amplification degree of cognitive bias within KT models, which is defined by the following equation.

$$
\begin{aligned}
E_{KL} &= \sum_{c} \frac{p_c}{\sum_c p_c} \frac{\log \frac{p_c}{\sum_c p_c}}{\frac{q_c}{\sum_c q_c}}, \\
p_c &= \frac{\sum_{i=1}^{t} \mathbb{1}(c_i = c, r_i = 1, (f_i < 0.5) = r_i)}{\sum_{i=1}^{t} \mathbb{1}(c_i = c, (f_i < 0.5) = r_i)}, \\
q_c &= \frac{\sum_{i=1}^{t} f_i \cdot \mathbb{1}(c_i = c, (f_i < 0.5) = r_i)}{\mathbb{1}(c_i = c, (f_i < 0.5) = r_i)}, \\
f_i &= f(q_{1:i-1}, c_{1:i-1}, r_{1:i-1}),
\end{aligned}
$$

where $\mathbb{1}(\cdot)$ is the indicator function. f is the model to be evaluated. c_i and r_i represent the i-th concept and its response, respectively. There are a total of t interactions.