Revisiting Applicable and Comprehensive Knowledge Tracing in Large-Scale Data

Yiyun Zhou, Wenkang Han, and Jingyuan Chen (✉)

Zhejiang University {yiyunzhou, wenkangh, jingyuanchen}@zju.edu.cn

Abstract. Knowledge Tracing (KT) is a fundamental component of Intelligent Tutoring Systems (ITS), enabling the modeling of students' knowledge states to predict future performance. The introduction of Deep Knowledge Tracing (DKT), the first deep learning-based KT (DLKT) model, has brought significant advantages in terms of applicability and comprehensiveness. However, recent DLKT models, such as Attentive Knowledge Tracing (AKT), have often prioritized predictive performance at the expense of these benefits. While deep sequential models like DKT have shown potential, they face challenges related to parallel computing, storage decision modification, and limited storage capacity. To address these limitations, we propose DKT2, a novel KT model that leverages the recently developed xLSTM architecture. DKT2 enhances applicable input representation using the Rasch model and incorporates Item Response Theory (IRT) for output interpretability, allowing for the decomposition of learned knowledge into familiar and unfamiliar knowledge. By integrating this knowledge with predicted questions, DKT2 generates comprehensive knowledge states. Extensive experiments conducted across three large-scale datasets demonstrate that DKT2 consistently outperforms 18 baseline models in various prediction tasks, underscoring its potential for real-world educational applications. This work bridges the gap between theoretical advancements and practical implementation in KT. Our code, datasets and Appendix are fully available at https://github.com/zyy-2001/DKT2.

Keywords: Knowledge Tracing · Information Interaction.

1 Introduction

The rapid expansion of educational data within Intelligent Tutoring Systems (ITS) [26] (*e.g.*, AutoTutor [29]) has exposed significant limitations in traditional machine learning approaches [4]. In contrast, the advent of deep learning has introduced novel opportunities for addressing these challenges [18, 36]. A critical component of ITS is Knowledge Tracing (KT), which models students' knowledge states and predicts future performance by analyzing their interaction data. Deep learning, with its advanced feature learning paradigm, offers enhanced modeling power and predictive accuracy in this context.

Deep Knowledge Tracing (DKT) [31] represents the first significant application of deep learning to KT, employing Long Short-Term Memory (LSTM) networks [13] to capture the complexity of students' learning processes. As a pioneering deep learning-based KT (DLKT) model, DKT has demonstrated superior predictive performance compared to traditional machine learning-based KT models (*e.g.*, Bayesian Knowledge Tracing (BKT) [7]), offering notable advantages in applicability and comprehensiveness.

DKT encodes students' **historical** interactions to generate a comprehensive representation of their knowledge states (*i.e.*, proficiency scores[1] for **each concept** at each time step) and predicts future performance. **However, recent DLKT models, such as the Attentive Knowledge Tracing (AKT) [10], while excelling in predictive accuracy [23, 16, 15, 40, 42], present limitations in applicability and comprehensiveness (the related details are in Sec. 3.2).** Specifically, AKT requires both historical and future interactions as input, complicating its practical application since future responses are typically unavailable. Additionally, unlike DKT, AKT directly predicts scores on future questions without generating a comprehensive knowledge state, potentially weakening the correlations between different concepts and narrowing the definition of knowledge states in KT. Our review of 60 KT-model-related papers published in top AI/ML conferences and journals over the past decade (see Appendix A.1) reveals a trend where evaluation performance has been prioritized at the expense of practical applicability, risking a disconnect between theoretical advancements and real-world implementation.

Deep sequential models like DKT have intrinsic limitations that may prevent them from achieving optimal performance. LSTM networks, for instance, face challenges in dynamically updating stored information and exhibit limited storage capacity due to their scalar cell state design. Moreover, their inherent sequential processing nature hinders parallelization, limiting their scalability to large datasets. The recently proposed xLSTM [3], however, addresses these challenges by introducing two new variants: sLSTM, which improves LSTM's storage decision by incorporating an exponential activation function, and mLSTM, which replaces scalar cell states with matrix memory for increased storage capacity and improved retrieval efficiency, while achieving full parallelization by abandoning memory mixing. Building on the strengths of xLSTM, we introduce DKT2, an enhanced DLKT model designed for greater applicability and comprehensiveness. DKT2 integrates the Rasch model [32] from educational psychology to process historical interactions, using xLSTM for knowledge learning. DKT2 then incorporates Item Response Theory (IRT) [25, 37] to interpret the learned knowledge, differentiating between familiar and unfamiliar knowledge, and ultimately integrates this knowledge with predicted questions to generate comprehensive knowledge states.

Our primary contributions are as follows:

- We provide a systematic analysis of input and output settings in KT, proposing DLKT models optimized for real-world applicability and comprehensiveness.

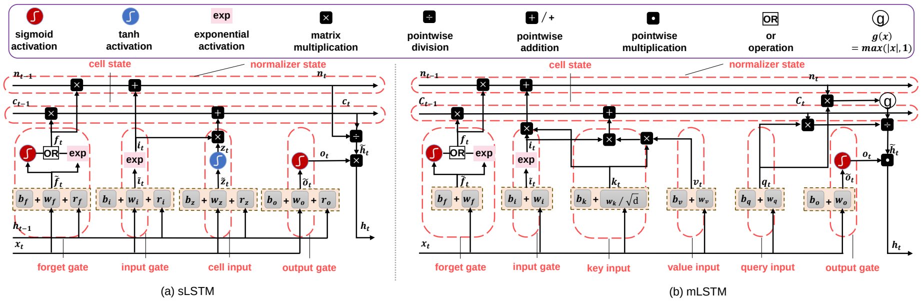
[1] Proficiency scores range from 0 to 1, with higher values indicating greater knowledge and skill level.

- We introduce DKT2, a model built on xLSTM, adhering to rigorous applicable input and comprehensive output settings, and incorporating both the Rasch model for input and an interpretable IRT-based output module.
- We conduct extensive experiments, including one-step prediction, multi-step prediction, and predictions with varying history lengths, across three large-scale datasets. Our findings demonstrate that DKT2 consistently outperforms 18 baseline models, with additional analysis on the impact of input settings and multi-concept output predictions on KT performance.

2 Related Work

Since DKT [31] first applied deep learning methods to the KT task a decade ago, deep learning techniques have flourished in KT. Current DLKT models can be categorized into the following 8 types:

- **Deep sequential models** use recurrent structures to encode students' chronologically ordered interactions, *e.g.*, DKT uses LSTM to model complex student cognitive processes. Two variants of DKT have emerged in subsequent research. DKT+ [39] introduces two regularization terms to improve the consistency of KT predictions, while DKT-F [27] enhances KT by considering forgetting behavior.
- **Attention-based models** capture long-term dependencies between interactions through attention mechanisms, *e.g.*, SAKT [30] is the first to use attention mechanisms to capture correlations between concepts and interactions. AKT [10] employs a novel monotonic attention to represent the time distance between questions and students' historical interactions. Due to AKT's outstanding predictive performance, numerous powerful KT models are subsequently derived, such as simpleKT [23], FoLiBiKT [16], sparseKT [15], DTransformer [40], and stableKT [20].
- **Mamba-based models** are strong competitors to Transformer models. The recently proposed Mamba4KT [5] is the first KT model to explore evaluation efficiency and resource utilization.
- **Graph-based models** use graph structures to characterize the relationships between questions, concepts, or interactions, *e.g.*, GKT [28] uses a graph to model the intrinsic relationships between concepts.
- **Memory-augmented models** capture latent relationships between concepts through memory networks, *e.g.*, DKVMN [41] uses a static key matrix to store relationships between different concepts and updates students' knowledge states through a dynamic value matrix. SKVMN [1], a variant of DKVMN, also integrates the advantages of LSTM in recurrent modeling.
- **Adversarial-based models** use adversarial techniques to enhance the model's generalization ability, *e.g.*, ATKT [11] mitigates overfitting and improves generalization by adding perturbations to student interactions during training.
- **Contrastive learning-based models** use contrastive learning to learn rich representations of student interactions, *e.g.*, CL4KT leverages contrastive



(a) sLSTM

(b) mLSTM

Fig. 1. Architecture of xLSTM.

learning to strengthen representation learning by distinguishing between similar and dissimilar learning histories.

- **Other representative models** include interpretable models and models with auxiliary tasks, *e.g.*, Deep-IRT [38] introduces item response theory [25] based on DKVMN to make deep learning-based KT explainable. AT-DKT [22] enhances KT by introducing two auxiliary learning tasks: question tagging prediction and individualized prior knowledge prediction.

Our proposed DKT2, by breaking the parallelization limitations of deep sequential models, can be classified as a new type of deep sequential models (**Deep sequential models***).

3 Methodology

3.1 Problem Statement

In the KT task, formally, let $\mathcal{S}$, $\mathcal{Q}$, and $\mathcal{C}$ represent the sets of students, questions, and concepts respectively. For each student $s \in \mathcal{S}$, there exists a sequence of k time steps $X_k = \{(q_1, c_1, r_1, t_1), (q_2, c_2, r_2, t_2), \ldots, (q_k, c_k, r_k, t_k)\}$, where $q_i \in \mathcal{Q}, c_i \subset \mathcal{C}, r_i \in \{0, 1\}$, and t_i represent the question attempted by the student, the concepts related to question q_i, whether the student responded correctly (0 for incorrect, 1 for correct), and the timestamp of the response, respectively. At time step $k+1$, DKT2 predicts $\hat{r}_{k+1}$ based on the student's interaction sequence X_k:

$$\hat{r}_{k+1} = \text{DKT2}(X_k, q_{k+1}, c_{k+1}, t_{k+1} \mid \theta), \tag{1}$$

where θ represents the parameters learned during training.

3.2 Preliminaries

LSTM and the Extended LSTM LSTM[2] is one of the earliest popular deep learning methods applied to NLP, but it has been overshadowed for a period by

[2] Refer to Appendix A.2 for details on LSTM.

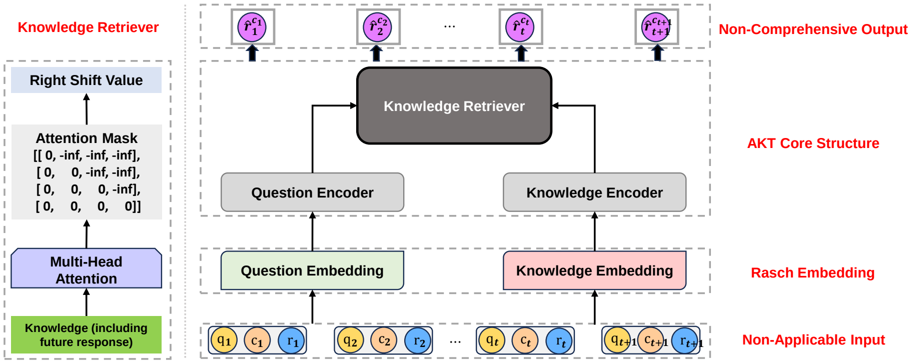


Fig. 2. Structural sketch of AKT.

the success of Transformers [34, 43]. However, the architecture is recently regaining attention and undergoing significant improvements. The improved LSTM is called extended Long Short-Term Memory (xLSTM) [3], which mainly addresses three limitations in traditional LSTM: (1) inability to revise storage decisions, (2) limited storage capacities, and (3) lack of parallelizability. xLSTM introduces two new members to the LSTM family to overcome these limitations: sLSTM and mLSTM, as described in Fig. 1. **Since our work does not focus on the architecture of xLSTM, we have placed the detailed introduction of xLSTM in A.3.**

Weakly Applicable Input and Comprehensive Output Settings in DLKT Models We use AKT [10] as an example to describe the common weakly applicable input and comprehensive output settings in DLKT models. Fig. 2 shows a structural sketch of AKT. Clearly, AKT takes both historical interactions and future interactions as input during training and inference, ignoring future information through attention masking while representing knowledge learned up to the current time step through offset (right-shifting values in attention), and directly predicts questions at each time step. From this, we can see that although AKT's setup is reasonable and does not lead to future information leakage, this input setting, while convenient, also **causes complications in engineering implementation** (engineering often requires cumbersome representation of future information as a padding value, and this common processing method does not seem suitable for KT, as KT tasks typically involve predicting future questions $2 \sim t+1$ based on historical interactions $1 \sim t$). Moreover, AKT only outputs the response for the current time step's question, without considering the student's proficiency in different dimensions, which **contradicts the multidimensional nature of real-world student knowledge and narrows the definition of KT.**

3.3 DKT2

Fig. 3 illustrates the architecture of our proposed DKT2, as described below.

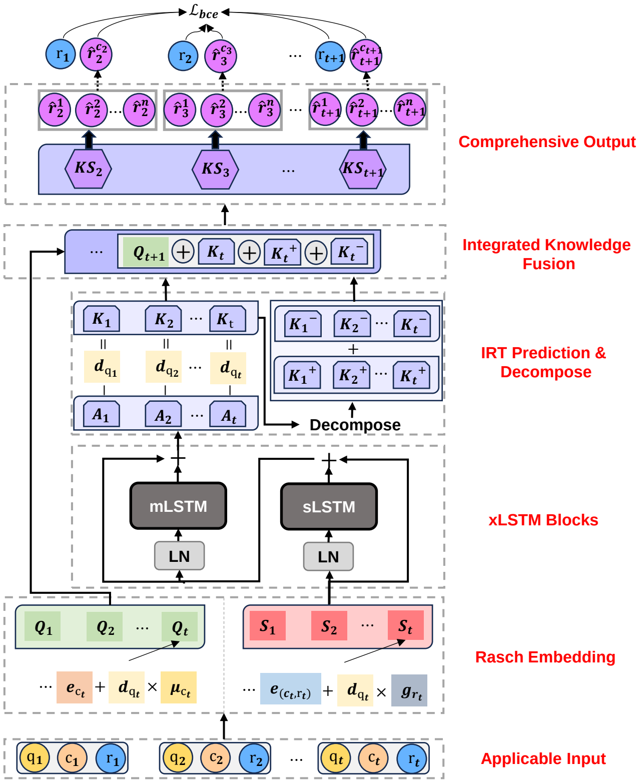

$\mathcal{L}_{bce}$
r_1
$\hat{r}_2^{c_2}$
r_2
$\hat{r}_3^{c_3}$
r_{t+1}
$\hat{r}_{t+1}^{c_{t+1}}$
$\hat{r}_2^1$ $\hat{r}_2^2$ ⋯ $\hat{r}_2^n$
$\hat{r}_3^1$ $\hat{r}_3^2$ ⋯ $\hat{r}_3^n$
$\hat{r}_{t+1}^1$ $\hat{r}_{t+1}^2$ ⋯ $\hat{r}_{t+1}^n$
KS_2
KS_3
KS_{t+1}
Comprehensive Output
⋯ $Q_{t+1} \oplus K_t \oplus K_t^+ \oplus K_t^-$
Integrated Knowledge Fusion
K_1 K_2 ⋯ K_t
d_{q_1} d_{q_2} ⋯ d_{q_t}
A_1 A_2 ⋯ A_t
K_1^- K_2^- ⋯ K_t^-
+
K_1^+ K_2^+ ⋯ K_t^+
Decompose
IRT Prediction & Decompose
mLSTM
sLSTM
LN
LN
xLSTM Blocks
Q_1 Q_2 ⋯ Q_t
S_1 S_2 ⋯ S_t
⋯ e_{c_t} + d_{q_t} × μ_{c_t}
⋯ $e_{(c_t,r_t)}$ + d_{q_t} × g_{r_t}
Rasch Embedding
q_1 c_1 r_1
q_2 c_2 r_2
⋯
q_t c_t r_t
Applicable Input

Fig. 3. Architecture of DKT2.

Rasch Embedding We use the classic Rasch model [32] from educational psychology to construct embeddings of questions and student skills. This model explicitly uses scalars to represent the degree of deviation between questions and the concepts they cover. Additionally, we choose to use question-specific difficulty vectors to capture differences among various questions within the same concept. DKT2 takes **applicable interactions (*i.e.*, inputs not involving the future response r_{t+1}, distinguishing it from models like AKT)** as input, denoted as $\{q_i, c_i, r_i\}_{i=1}^{t}$, and at time step t, the embeddings of questions and student skills, Q_t and S_t respectively, are represented as:

$$Q_t = e_{c_t} + d_{q_t} \cdot \mu_{c_t}, S_t = e_{(c_t,r_t)} + d_{q_t} \cdot g_{r_t}, e_{(c_t,r_t)} = e_{c_t} + e_{r_t}, \qquad (2)$$

where $e_{c_t} \in \mathbb{R}^d$ and $e_{r_t} \in \mathbb{R}^d$ are the embeddings of concept c_t and response r_t, respectively. $d_{q_t} \in \mathbb{R}$ is a difficulty scalar and $\mu_{c_t} \in \mathbb{R}^d$ summarizes the variation of questions containing concept c_t. $e_{(c_t,r_t)} \in \mathbb{R}^d$ is the interaction representation of the concept and student response, $g_{r_t} \in \mathbb{R}^d$ is the variant embedding of the response. d is the dimension of the embeddings.

xLSTM Blocks DKT2 further learns the student's ability representation $A_{1:t}$ at time step t through two xLSTM blocks (sLSTM and mLSTM) based on the original representation of student ability $S_{1:t}$:

$$A_{1:t} = \mathrm{Res}\left(\mathrm{LN}\left(\mathrm{mLSTM}\left(\mathrm{Res}\left(\mathrm{LN}\left(\mathrm{sLSTM}(S_{1:t})\right)\right)\right)\right)\right), \tag{3}$$

where LN and Res refer to layer normalization [2] and residual connection [12], respectively.

IRT Prediction & Decompose The core idea of IRT (Item Response Theory) lies in the interactive relationship between student ability and question difficulty [37]. Specifically, **if a student's ability is far above the question's difficulty, the probability of the student responding to the question correctly is very high, and vice versa.** This is also why IRT is often used for interpretable predictions in KT [38, 33] (**our work focuses not on the interpretability of KT models but on evaluating their applicability and comprehensive setup**). Therefore, the knowledge acquired by a student, denoted as $K_{1:t}$, can be represented as:

$$K_{1:t} = A_{1:t} - d_{q_{1:t}}, \tag{4}$$

where $d_{q_{1:t}}$ is the sequence representation of d_{q_t} from Eq. 2 up to time step t.

Further, DKT2 roughly distinguishes between the familiar and unfamiliar knowledge $K^{+}_{1:t}$ and $K^{-}_{1:t}$ based on correct and incorrect responses:

$$K^{+}_{1:t} = \exp(r_{1:t}, d) \circ K_{1:t}, K^{-}_{1:t} = \exp(\mathbf{one} - r_{1:t}, d) \circ K_{1:t}, \tag{5}$$

where $\exp(\cdot, d)$ denotes expanding the last dimension of the tensor to d dimensions. $\circ$ denotes element-wise multiplication. $\mathbf{one} \in \mathbb{R}^t$ is a vector of all ones.

Integrated Knowledge Fusion DKT2 estimates the student's knowledge $X_{2:t+1}$ based on the knowledge $K_{1:t}$ and the questions $Q_{2:t+1}$ that need to be predicted:

$$X_{2:t+1} = Q_{2:t+1} \oplus K_{1:t} \oplus K^{+}_{1:t} \oplus K^{-}_{1:t}, \tag{6}$$

where $\oplus$ denotes the concatenation operation. In addition to integrating the questions and the student's current knowledge, DKT2 also includes the student's familiar and unfamiliar knowledge $K^{+}_{1:t}$ and $K^{-}_{1:t}$. This is because, intuitively, **if the knowledge required to respond to a question is familiar to the student, the predicted score tends to be higher, and conversely, lower if unfamiliar.**

Finally, DKT2 predicts the student's comprehensive knowledge states $\mathrm{KS}_{2:t+1}$:

$$\mathrm{KS}_{2:t+1} = \sigma(\mathrm{ReLU}(X_{2:t+1}W_1 + b_1)W_2 + b_2), \tag{7}$$

where $W_1 \in \mathbb{R}^{4d\times 2d}, W_2 \in \mathbb{R}^{2d\times n}, b_1 \in \mathbb{R}^{2d}, b_2 \in \mathbb{R}^n$ are learnable parameters in the MLP. $\sigma(\cdot)$ is the Sigmoid function and $\mathrm{ReLU}(\cdot)$ is the activation function. n is the number of concepts (due to data sparsity, KT often predicts the concepts corresponding to questions).

Eq. 7 can be further represented as:

$$\mathrm{KS}_i = (\hat{r}^1_i, \hat{r}^2_i, \ldots, \hat{r}^n_i), 2 \leq i \leq t+1, \tag{8}$$

where $\hat{r}_i^j$ represents the prediction score of DKT2 for concept j at time step i. **The comprehensive output of DKT2 enables the prediction of multiple concepts at the same time step, whereas models like AKT can only predict $\hat{r}_i^{c_i}$ at time step i.** We will analyze the multi-concept prediction scenario in detail in Sec. 4.3, where some unexpected results have been discovered.

Model Training The loss of DKT2 is defined as the binary cross-entropy loss between the prediction $\hat{r}_t$ and the actual response r_t, calculated as follows:

$$\mathcal{L}_{\text{DKT2}} = -\sum_{i=2}^{t+1} r_i \log(\hat{r}_i) + (1 - r_i)\log(1 - \hat{r}_i). \tag{9}$$

Conversion of Input and Output Settings We attempt to convert the weakly applicable input and comprehensive output settings in DLKT models into strongly applicable input and comprehensive output settings. Similarly, using AKT as an example, like DKT2, as shown in Fig. 3, the transformed AKT only takes the historical interactions $\{(q_i, c_i, r_i)\}_{i=1}^t$ as input, with everything else remaining unchanged (note that the right-shift operation still needs to be retained because the attention does not mask the knowledge of the current time step). Before outputting the predicted score $\hat{r}_i^{c_i}$, it first concretizes knowledge into the knowledge of each concept (by converting the original dimensions into the number of concepts through an MLP) and then the comprehensive knowledge state is obtained through a Sigmoid function.

4 Experiments

Our goal is to answer the following research questions:

- **RQ1**: How does DKT2 perform compared to 18 baselines from 8 different categories under applicable input and comprehensive output settings?
- **RQ2**: How do different input settings for KT models with weak applicability and comprehensiveness and multi-concept prediction of various KT models affect their performance?
- **RQ3**: What are the impacts of the components (*e.g.*, the Rasch embedding and IRT prediction) on DKT2?

4.1 Experimental Setup

Datasets We conduct extensive experiments on three of the latest large-scale benchmark datasets from different platforms: Assist17 [9], EdNet [6], and Comp [14]. Details of the datasets are provided in Appendix A.4.

Baselines To comprehensively and systematically evaluate the performance of DKT2 and analyze the impact of input-output settings on KT models, we compare DKT2 with 18 DLKT baselines from 8 categories, as mentioned in Sec. 2. Detailed descriptions of the aforementioned DLKT baselines can be found in Appendix A.5.

Implementation Similar to CL4KT [19], we employ five-fold cross-validation, with folds divided by students. 10% of the training set is used for model evaluation and also for the early stopping strategy: if the AUC does not improve within 10 epochs during the 300 epochs, the training will be stopped. The averages across five test folds are reported. We focus on the most recent 100 interactions (history length) for each student, as this latest information is crucial for future predictions. During training, all models are trained using the Adam optimizer [17] with the following settings: batch size is fixed at 512, learning rate is 0.001, dropout rate is 0.05, and embedding dimension is 64. The seed is set to 12405 to reproduce experimental results. Similar to existing DLKT research, our evaluation metrics include two classification metrics, Area Under the ROC Curve (AUC) and Accuracy (ACC), and one regression metric, Root Mean Square Error (RMSE). Note that our experimental parameter configuration is consistent with CL4KT.

4.2 Applicable and Comprehensive Performance Comparison (RQ1)

Under applicable input and comprehensive output settings, we evaluate three common prediction tasks in KT [24]: 1) one-step prediction, 2) multi-step prediction, and 3) prediction with varying history lengths.

One-step Prediction KT's one-step prediction can provide immediate feedback for ITS and be used for short-term adjustments of personalized learning paths [7]. Table 1 shows the one-step prediction performance of DKT2 and 18 baselines from 8 different categories in three large-scale datasets. Overall, in this fair large-scale data competition, our DKT2 has emerged as the final winner by a narrow margin. We observe:

- Compared to previous research [10], under the input-output settings, attention-based models like AKT still generally outperform deep sequential models like DKT, **suggesting that attention-based models like AKT may be less affected by these settings.**
- The recently proposed Mamba4KT performs well on Assist17, but underperforms compared to DKT on larger-scale datasets like EdNet and Comp. This may be due to mamba's poorer performance in context learning in large-scale experiments, which is consistent with previous research findings [35].
- DLKT models based on graph, memory augmentation, adversarial, or contrastive learning do not show significant performance improvements. We believe this is because large-scale data contains more noise and diversity, making

Category	Model	Assist17			EdNet			Comp		
		AUC↑	ACC↑	RMSE↓	AUC↑	ACC↑	RMSE↓	AUC↑	ACC↑	RMSE↓
Deep sequential	DKT✓	0.6621	0.6370	0.4731	0.6834	0.6451	0.4687	0.7585	0.8129	0.3681
	DKT+✓	0.6668	0.6415	0.4711	0.6884	0.6483	0.4673	0.7593	0.8129	0.3679
	DKT-F✓	0.6633	0.6429	0.4724	0.6917	0.6503	0.4668	0.7615	0.8138	0.3672
Attention-based	SAKT†	0.6211	0.6108	0.4828	0.6773	0.6415	0.4708	0.7560	0.8123	0.3690
	AKT✗	0.6789	0.6464	0.4723	0.6855	0.6440	0.4686	0.7601	0.8119	0.3686
	simpleKT✗	0.6709	0.6441	0.4746	0.6865	0.6444	0.4686	0.7633	0.8135	0.3672
	FoLiBiKT✗	0.6771	0.6444	0.4750	0.6849	0.6432	0.4687	0.7599	0.8120	0.3685
	sparseKT✗	0.6674	0.6424	0.4740	0.6856	0.6430	0.4701	**0.7690**	**0.8178**	**0.3604**
	DTransformer✗	0.6480	0.6305	0.4770	0.6727	0.6355	0.4722	0.7551	0.8106	0.3699
	stableKT✗	0.6781	0.6455	0.4751	0.6841	0.6411	0.4695	0.7591	0.8126	0.3683
Mamba-based	Mamba4KT✓	0.7001	0.6555	0.4701	0.6667	0.6351	0.4764	0.7575	0.8121	0.3687
Graph-based	GKT✓	0.6408	0.6185	0.4802	0.6841	0.6361	0.4724	0.7390	0.8055	0.3766
Memory-augmented	DKVMN✗	0.6505	0.6308	0.4774	0.6778	0.6410	0.4705	0.7534	0.8113	0.3697
	SKVMN✗	0.6350	0.6184	0.4809	0.6800	0.6427	0.4696	0.7220	0.8040	0.3790
Adversarial-based	ATKT✓	0.6453	0.6313	0.4821	0.6780	0.6403	0.4714	0.7560	0.8123	0.3688
Contrastive learning-based	CL4KT✗	0.6540	0.6319	0.4783	-	-	-	0.7645	0.8146	0.3669
Other representative	Deep-IRT✗	0.6448	0.6268	0.4814	0.6661	0.6317	0.4769	0.7517	0.8108	0.3703
	AT-DKT✓	0.6720	0.6433	0.4708	0.6888	0.6494	0.4673	0.7655	0.8141	0.3663
Deep sequential*	**DKT2✓**	**0.7042**	**0.6594**	**0.4630**	**0.6929**	**0.6504**	**0.4660**	0.7679	0.8165	0.3652

Table 1. One-step prediction performance of DKT2 and 18 baselines from different categories. The **best result** is in bold, the second best is underlined. ✓ indicates strong applicability and comprehensiveness, ✗ indicates weak applicability and comprehensiveness, † indicates strong applicability but weak comprehensiveness. - indicates the model fails to be applied to such a large-scale dataset, resulting in a program crash.

it challenging for complex models (*e.g.*, graph-based and memory-augmented models) to effectively extract useful information during training. Moreover, large-scale data usually covers various student learning behaviors and knowledge states, meaning that basic models might already be sufficient for effective knowledge tracing, thus the advantages of adversarial-based and contrastive learning-based models are not pronounced.

- Our proposed DKT2 performs almost the best on all metrics across all datasets. This performance improvement can be attributed to the superiority of DKT2, which includes the exponential activation function in sLSTM that helps improve memory and forgetting processes, and the matrix memory introduced in mLSTM that gives DKT2 advantages in large-scale applications and long sequence processing.

Multi-step Prediction KT's accurate multi-step prediction not only provides valuable feedback for selecting and constructing personalized learning materials, but also assists ITS in flexibly adjusting future curriculum based on student needs [23]. Table 2 and Table 7 in Appendix A.6 show the multi-step (step=5, 10, 15, 20) prediction performance of DKT2 and several representative baselines from different categories. The main observations are as follows: (1) As the prediction steps increase, the performance of all models consistently decreases. This is due to error accumulation, meaning that small errors in one-step prediction can accumulate over multiple steps, leading to a decrease in multi-step predic-

Step	5			10			15			20		
Metric	AUC↑	ACC↑	RMSE↓	AUC↑	ACC↑	RMSE↓	AUC↑	ACC↑	RMSE↓	AUC↑	ACC↑	RMSE↓
DKT	0.6244	0.6104	0.4831	0.6048	0.5978	0.4868	0.5962	0.5960	0.4874	0.5902	0.5918	0.4890
SAKT	0.6103	0.6010	0.4860	0.6013	0.5966	0.4860	0.5989	0.5983	0.4860	0.5961	0.5960	0.4869
AKT	0.6486	0.6285	**0.4763**	0.6321	0.6213	**0.4798**	0.6231	0.6140	**0.4819**	0.6189	0.6134	**0.4827**
Mamba4KT	0.6222	0.6077	0.4869	0.5909	0.5938	0.4876	0.5875	0.5911	0.4880	0.5858	0.5907	0.4884
DKVMN	0.6205	0.6096	0.4851	0.6008	0.5958	0.4880	0.5905	0.5923	0.4879	0.5830	0.5856	0.4893
ATKT	0.6246	0.6186	0.4831	0.6176	0.6139	0.4847	0.6125	0.6118	0.4855	0.6094	0.6090	0.4865
CL4KT	0.6347	0.6186	0.4832	0.6128	0.6037	0.4882	0.6043	0.5987	0.4896	0.5991	0.5971	0.4890
Deep-IRT	0.6100	0.6020	0.4959	0.5867	0.5834	0.5022	0.5737	0.5713	0.5072	0.5652	0.5666	0.5049
AT-DKT	0.6424	0.6260	0.4782	0.6271	0.6154	0.4820	0.6206	0.6115	0.4832	0.6170	0.6082	0.4849
DKT2	**0.6496**	**0.6313**	**0.4763**	**0.6335**	**0.6221**	0.4802	**0.6246**	**0.6160**	0.4822	**0.6199**	**0.6148**	0.4828

Table 2. Multi-step prediction performance of DKT2 and several representative baselines on Assist17. The results for EdNet and Comp can be found in Appendix A.6.

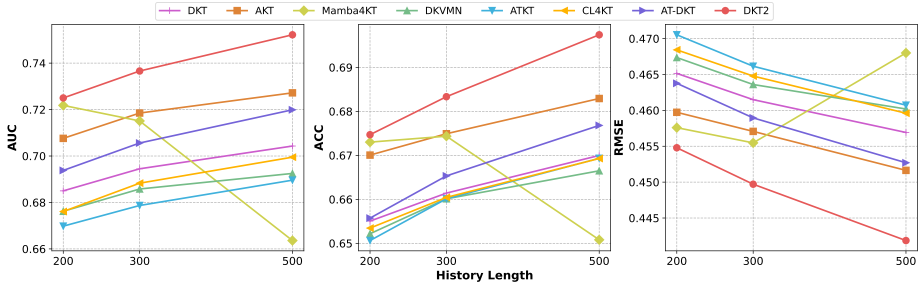


Fig. 4. The prediction performance of DKT2 and several representative baselines on Assist17 with different history lengths. The results for EdNet and Comp are in Appendix A.6.

tion performance. (2) Compared to one-step prediction, attention-based models perform well in multi-step prediction. This is because the attention mechanism can capture long-distance dependencies, making its advantages more apparent. In contrast, Mamba4KT performs poorly, as mamba-based models are highly dependent on context [21] and are more susceptible to error accumulation. (3) Our DKT2 generally outperforms all models in multi-step prediction. We can similarly attribute this to the exponential activation function introduced in sLSTM of DKT2, which can mitigate error accumulation by modifying storage decisions, as it allows the model to update its internal state at each step, while the matrix memory introduced in mLSTM provides support for large-capacity storage space.

Varying-history-length Prediction Analyzing the impact of different history lengths can help ITS better understand students' knowledge acquisition and forgetting processes, thereby improving teaching strategies. Fig. 4, Fig. 7 and Fig. 8 in Appendix A.6 show the prediction performance of DKT2 and several representative baselines with different history lengths. From these, we can observe: 1) As the history length increases, the prediction performance of almost

Setting	Metric	AKT	simpleKT	FoLiBiKT	sparseKT	DTransformer	stableKT	DKVMN	CL4KT	Deep-IRT
△	AUC↑	0.6554	0.6507	0.6545	0.6405	0.5995	0.6490	0.6228	0.5941	0.6234
	ACC↑	0.6154	0.6129	0.6117	0.6120	0.5755	0.6159	0.5899	0.5696	0.5915
	RMSE↓	0.4822	0.4840	0.4835	0.4865	0.5040	0.4837	0.4890	0.5090	0.4892
○	AUC↑	0.6505	0.6675	0.6471	0.6574	0.5989	0.6329	0.6203	0.6293	0.6182
	ACC↑	0.6202	0.6240	0.6226	0.6210	0.5625	0.6045	0.5862	0.6016	0.5884
	RMSE↓	0.4853	0.4866	0.4842	0.4836	0.5206	0.4908	0.5010	0.5037	0.5015
●	AUC↑	0.6320	0.6508	0.6192	0.6474	0.5994	0.6533	0.6087	0.6195	0.6001
	ACC↑	0.6066	0.6153	0.5980	0.6060	0.5770	0.6223	0.5787	0.5933	0.5728
	RMSE↓	0.4881	0.4999	0.4944	0.4944	0.4981	0.4833	0.5074	0.5083	0.5012

Table 3. The prediction performance of KT models with weak applicability and comprehensiveness in the last 5 steps on Assist17 under three different input settings. The △ setting represents masking all interaction information (including questions, concepts and responses) for the last 5 steps, the ○ setting represents masking the responses for the last 5 steps, without masking questions and concepts, and the ● setting represents no masking, *i.e.*, predicting the responses under the regular setting. The results for EdNet and Comp can be found in Appendix A.6.

all models generally improves, as longer sequences provide more historical information. Surprisingly, Mamba4KT's performance consistently decreases. A possible reason is that mamba-based models are better at capturing local temporal dependencies but may struggle to effectively capture long-distance dependencies within longer sequences. 2) Notably, DKT, using only one LSTM, can maintain a strong ranking position across different history lengths, further encouraging KT researchers to design simple yet effective models [23]. 3) Our DKT2 maintains optimal performance across different history lengths, with more significant performance improvements as the history length increases. This is not only due to the increased storage capacity of mLSTM but also related to sLSTM providing a broader output range as the sequence length increases.

4.3 In-Depth Analysis (RQ2 & RQ3)

Analysis of Different Input Settings We analyze three different input settings for the KT models with weak applicability and comprehensiveness. In Table 3 and Table 8 in Appendix A.6, we present the prediction performance of these models in the last 5 steps. From these, we have the following findings: (i) The performance differences among these three settings are more pronounced on EdNet and Comp, as larger-scale data can provide richer information for more accurate prediction. (ii) The models under guessing △ setting seems to perform well on Assist17, which may be because the models remember the answer bias [8] and make predictions directly, while the models under the ○ and ● settings achieve comparable performance, indicating that **the applicable ○ setting does not significantly reduce the model's performance.** This confirms the hypothesis proposed in the Sec. 4.2 (One-step Prediction).

Multi-concept Prediction Comprehensive KT can be used for multi-concept prediction. Multi-concept prediction can provide a more comprehensive learning

Dataset	Assist17			EdNet			Comp		
Metric	AUC↑	ACC↑	RMSE↓	AUC↑	ACC↑	RMSE↓	AUC↑	ACC↑	RMSE↓
DKT	0.5841	0.5787	0.4913	0.6600	0.6225	0.4775	0.7091	0.8045	0.3806
SAKT	0.5596	0.5534	0.5048	0.6546	0.6198	0.4788	0.6994	0.8037	0.3824
AKT	0.6185	0.6040	0.4862	0.6649	0.6241	**0.4765**	0.7054	0.8039	0.3815
Mamba4KT	0.5660	0.5660	0.4956	0.6531	0.6192	0.4796	0.7054	0.8043	0.3813
DKVMN	0.5730	0.5701	0.4964	0.6572	0.6205	0.4781	0.7050	0.8034	0.3817
ATKT	0.6205	0.6077	0.4836	0.6639	0.6229	0.4768	0.7111	0.8049	0.3802
CL4KT	0.5892	0.5911	0.4890	-	-	-	0.7044	0.8032	0.3820
Deep-IRT	**0.6445**	**0.6263**	**0.4808**	**0.6664**	**0.6321**	0.4767	**0.7515**	**0.8107**	**0.3704**
AT-DKT	0.6087	0.5962	0.4882	0.6644	0.6245	0.4766	0.7093	0.8046	0.3805
DKT2	0.6174	0.6041	0.4872	0.6646	0.6243	0.4768	0.7064	0.8047	0.3810

Table 4. Multi-concept prediction performance of DKT2 and several representative baselines.

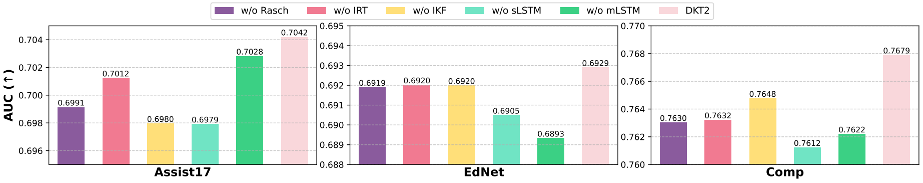


Fig. 5. Ablation study on AUC.

assessment, explore relationships between concepts, and create precise personalized learning plans for students. Due to the lack of datasets for multi-concept prediction (to our knowledge, existing datasets do not include students' proficiency scores for all concepts at different learning stages), our experiments are conducted under a weak assumption: the change in a student's knowledge state is a gradual process and is unlikely to experience sudden shifts over the long term. In our experiments, we use the knowledge state at the intermediate time step to predict subsequent questions. Table 4 shows the multi-concept prediction performance of DKT2 and several representative baselines. From this, we discover an unexpected phenomenon: Deep-IRT and ATKT, which are generally not advantageous in previous performance comparisons, achieve impressive results, while our DKT2 can only rank in the top four. These results might make us question the validity of the weak assumption, but the empirical evidence of the almost consistent performance rankings of Deep-IRT and ATKT across the three datasets dispels our doubts. This interesting phenomenon makes us ponder: is it necessary to excessively pursue prediction accuracy while neglecting the assessment of multiple concepts in practice? We will explore this important topic in depth in future KT research.

Ablation Study Fig. 5 and Fig. 9 in Appendix A.6 illustrate the impact of different components on DKT2. “w/o. Rasch" indicates the removal of Rasch embedding from DKT2 (setting d_{q_t} to 0 in Eq. 2), “w/o. IRT" represents the removal of the IRT module, “w/o. IKF" means DKT2 ignores the integrated knowledge fusion, while “w/o. sLSTM" and “w/o. mLSTM" denote the removal of sLSTM block and mLSTM block, respectively. The results show that DKT2 achieves the highest AUC scores across all datasets compared to other variants, demonstrating the importance of each component on DKT2. Notably, “w/o. mLSTM" generally outperforms DKT2 on ACC and RMSE scores on Assist17, which is due to mLSTM’s inability to demonstrate significant advantages in small-scale data, as evidenced by its poorer performance on larger datasets, EdNet and Comp.

5 Conclusion

This paper introduces DKT2, an applicable and comprehensive DLKT model that addresses key limitations of deep sequential models like DKT. By leveraging xLSTM, the Rasch model, and Item Response Theory (IRT), DKT2 effectively balances predictive performance with practical applicability. Our extensive experiments across three large-scale datasets demonstrate DKT2’s superiority over 18 baseline models in various prediction tasks, highlighting its robustness and potential for real-world educational applications.

6 Limitations

Our work represents an attempt to apply xLSTM in the KT domain on large-scale data with fair input and output settings. In our experiments, we observe that as the number of students increases, DKT2 gradually demonstrates a performance advantage that widens the gap with other DLKT models. Additionally, in our multi-concept prediction experiments, we find that Deep-IRT exhibits a leading, dataset-independent advantage, the reasons for which give us pause for reflection. Therefore, our future research directions include: 1) further exploration of deeper knowledge tracing methodologies based on xLSTM, particularly in the context of ultra-large-scale data, and 2) enhancing multi-concept predictive analysis by collecting and analyzing students’ proficiency scores across different concepts at various learning stages.

Acknowledgments

This research was supported by grants from the "Pioneer" and "Leading Goose" R&D Program of Zhejiang (2025C02022), National Natural Science Foundation of China (No.62307032), and Shanghai Rising-Star Program (23QA1409000).

References

1. Abdelrahman, G., Wang, Q.: Knowledge tracing with sequential key-value memory networks. In: Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval. pp. 175–184 (2019)
2. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization (2016)
3. Beck, M., Pöppel, K., Spanring, M., Auer, A., Prudnikova, O., Kopp, M., Klambauer, G., Brandstetter, J., Hochreiter, S.: xlstm: Extended long short-term memory. arXiv preprint arXiv:2405.04517 (2024)
4. Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence **35**(8), 1798–1828 (2013)
5. Cao, Y., Zhang, W.: Mamba4kt: An efficient and effective mamba-based knowledge tracing model. arXiv preprint arXiv:2405.16542 (2024)
6. Choi, Y., Lee, Y., Shin, D., Cho, J., Park, S., Lee, S., Baek, J., Bae, C., Kim, B., Heo, J.: 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. pp. 69–73. Springer (2020)
7. Corbett, A.T., Anderson, J.R.: Knowledge tracing: Modeling the acquisition of procedural knowledge. User modeling and user-adapted interaction **4**, 253–278 (1994)
8. Cui, C., Ma, H., Zhang, C., Zhang, C., Yao, Y., Chen, M., Ma, Y.: Do we fully understand students' knowledge states? identifying and mitigating answer bias in knowledge tracing. arXiv preprint arXiv:2308.07779 (2023)
9. Feng, M., Heffernan, N., Koedinger, K.: Addressing the assessment challenge with an online system that tutors as it assesses. User modeling and user-adapted interaction **19**, 243–266 (2009)
10. Ghosh, A., Heffernan, N., Lan, A.S.: Context-aware attentive knowledge tracing. In: Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining. pp. 2330–2339 (2020)
11. Guo, X., Huang, Z., Gao, J., Shang, M., Shu, M., Sun, J.: Enhancing knowledge tracing via adversarial training. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 367–375 (2021)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385 (2015)
13. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
14. Hu, L., Dong, Z., Chen, J., Wang, G., Wang, Z., Zhao, Z., Wu, F.: Ptadisc: A cross-course dataset supporting personalized learning in cold-start scenarios. In: Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2023)
15. Huang, S., Liu, Z., Zhao, X., Luo, W., Weng, J.: 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. pp. 2441–2445 (2023)
16. Im, Y., Choi, E., Kook, H., Lee, J.: Forgetting-aware linear bias for attentive knowledge tracing. In: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. pp. 3958–3962 (2023)
17. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization (2017)
18. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. nature **521**(7553), 436–444 (2015)

19. Lee, W., Chun, J., Lee, Y., Park, K., Park, S.: Contrastive learning for knowledge tracing. In: Proceedings of the ACM Web Conference 2022. pp. 2330–2338 (2022)
20. Li, X., Bai, Y., Guo, T., Liu, Z., Huang, Y., Zhao, X., Xia, F., Luo, W., Weng, J.: Enhancing length generalization for attention based knowledge tracing models with linear biases
21. Lieber, O., Lenz, B., Bata, H., Cohen, G., Osin, J., Dalmedigos, I., Safahi, E., Meirom, S., Belinkov, Y., Shalev-Shwartz, S., et al.: Jamba: A hybrid transformer-mamba language model. arXiv preprint arXiv:2403.19887 (2024)
22. Liu, Z., Liu, Q., Chen, J., Huang, S., Gao, B., Luo, W., Weng, J.: Enhancing deep knowledge tracing with auxiliary tasks. In: Proceedings of the ACM Web Conference 2023. pp. 4178–4187 (2023)
23. Liu, Z., Liu, Q., Chen, J., Huang, S., Luo, W.: simplekt: a simple but tough-to-beat baseline for knowledge tracing. arXiv preprint arXiv:2302.06881 (2023)
24. Liu, Z., Liu, Q., Chen, J., Huang, S., Tang, J., Luo, W.: 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 (2022)
25. Lord, F.: A theory of test scores. Psychometric monographs (1952)
26. Luckin, R., Holmes, W.: Intelligence unleashed: An argument for ai in education (2016)
27. Nagatani, K., Zhang, Q., Sato, M., Chen, Y.Y., Chen, F., Ohkuma, T.: Augmenting knowledge tracing by considering forgetting behavior. In: The world wide web conference. pp. 3101–3107 (2019)
28. Nakagawa, H., Iwasawa, Y., Matsuo, Y.: Graph-based knowledge tracing: modeling student proficiency using graph neural network. In: IEEE/WIC/ACM International Conference on Web Intelligence. pp. 156–163 (2019)
29. Nye, B.D., Graesser, A.C., Hu, X.: Autotutor and family: A review of 17 years of natural language tutoring. International Journal of Artificial Intelligence in Education **24**, 427–469 (2014)
30. Pandey, S., Karypis, G.: A self-attentive model for knowledge tracing. arXiv preprint arXiv:1907.06837 (2019)
31. Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L.J., Sohl-Dickstein, J.: Deep knowledge tracing. Advances in neural information processing systems **28** (2015)
32. Rasch, G.: Probabilistic models for some intelligence and attainment tests. ERIC (1993)
33. Sun, J., Yu, F., Wan, Q., Li, Q., Liu, S., Shen, X.: Interpretable knowledge tracing with multiscale state representation. In: Proceedings of the ACM on Web Conference 2024. pp. 3265–3276 (2024)
34. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
35. Waleffe, R., Byeon, W., Riach, D., Norick, B., Korthikanti, V., Dao, T., Gu, A., Hatamizadeh, A., Singh, S., Narayanan, D., et al.: An empirical study of mamba-based language models. arXiv preprint arXiv:2406.07887 (2024)
36. Wu, T., Chen, J., Lin, W., Li, M., Zhu, Y., Li, A., Kuang, K., Wu, F.: Embracing imperfection: Simulating students with diverse cognitive levels using llm-based agents (2025), https://arxiv.org/abs/2505.19997
37. Yen, W.M., Fitzpatrick, A.R.: Item response theory. Educational measurement **4**, 111–153 (2006)

38. Yeung, C.K.: Deep-irt: Make deep learning based knowledge tracing explainable using item response theory. arXiv preprint arXiv:1904.11738 (2019)
39. Yeung, C.K., Yeung, D.Y.: Addressing two problems in deep knowledge tracing via prediction-consistent regularization. In: Proceedings of the fifth annual ACM conference on learning at scale. pp. 1–10 (2018)
40. Yin, Y., Dai, L., Huang, Z., Shen, S., Wang, F., Liu, Q., Chen, E., Li, X.: Tracing knowledge instead of patterns: Stable knowledge tracing with diagnostic transformer. In: Proceedings of the ACM Web Conference 2023. pp. 855–864 (2023)
41. Zhang, J., Shi, X., King, I., Yeung, D.Y.: Dynamic key-value memory networks for knowledge tracing. In: Proceedings of the 26th international conference on World Wide Web. pp. 765–774 (2017)
42. Zhou, Y., Lv, Z., Zhang, S., Chen, J.: Cuff-kt: Tackling learners' real-time learning pattern adjustment via tuning-free knowledge state guided model updating (2025), https://arxiv.org/abs/2505.19543
43. Zhou, Y., Yao, C., Chen, J.: Cola: Collaborative low-rank adaptation. arXiv preprint arXiv:2505.15471 (2025)