

MDKT: Robust Multi-Dimensional Knowledge Tracing with Adaptive Dual-Phase State Evolution

Muhui Lin

Fujian Normal University

Jiangsong Xu

Fujian Normal University

Ziqi Wang

Fujian Normal University

Mingwei Lin

lmwfjnu@fjnu.edu.cn

Fujian Normal University

Shenbao Yu

Fujian Normal University

Jun Shen

University of Wollongong

Research Article

Keywords: Knowledge tracing, Multi-dimensional modeling, Adversarial training, Educational data mining, Deep learning

Posted Date: September 23rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10972645/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

MDKT: Robust Multi-Dimensional Knowledge Tracing with Adaptive Dual-Phase State Evolution

Abstract

Knowledge Tracing (KT) is a fundamental task in intelligent educational systems, aiming to dynamically model students’ evolving knowledge states from their learning interaction sequences. Despite the success of deep learning-based KT models, existing approaches still suffer from two major limitations: (1) reliance on single-dimensional knowledge representations—typically modeling either at the question level or the concept level only—which fail to capture the multifaceted and dynamic nature of students’ cognitive processes; and (2) insufficient robustness under sparse and noisy interaction data that are prevalent in real-world educational scenarios. To address these challenges, we propose a robust model, named MDKT (Robust Multi-Dimensional Knowledge Tracing with Adaptive Dual-Phase State Evolution). MDKT jointly models a student’s knowledge state at three complementary granularities: question-level, concept-level, and overall-level. This design provides a more comprehensive and expressive representation of learning dynamics. To better capture the temporal evolution of knowledge, we further introduce an adaptive dual-phase learning state encoder that explicitly models both short-term learning boosts and long-term forgetting effects. Finally, to enhance robustness against noisy and sparse data, we incorporate an adversarial smoothing mechanism during training. Extensive experiments on three real-world benchmark datasets demonstrate that MDKT consistently outperforms state-of-the-art KT models in terms of prediction accuracy and robustness, particularly in long-sequence and data-scarce scenarios. These results highlight the effectiveness of multi-dimensional knowledge modeling and adaptive temporal encoding for building reliable knowledge tracing systems in real-world intelligent educational applications.

Keywords: Knowledge tracing, Multi-dimensional modeling, Adversarial training, Educational data mining, Deep learning

1 Introduction

Knowledge Tracing (KT) is a core task in intelligent educational systems, aiming to dynamically model and predict students’ evolving knowledge states from their historical learning interactions (He et al. 2023; Yu et al. 2026). With the rapid growth of online learning platforms and educational data mining, vast amounts of student-question interaction logs are becoming available, enabling more effective real-time and

personalized interventions. A reliable KT model can accurately predict future performance, identify knowledge gaps, and support targeted learning strategies, thereby improving both learning efficiency and teaching quality (Huang et al. 2024; Pu et al. 2024; Shen et al. 2023). For instance, Figure 1 illustrates the KT task using a six-question interaction sequence. The responses to $\{Q_1, \dots, Q_5\}$, together with their associated concepts, constitute the observed interaction history, and the model predicts the student’s response to Q_6 while dynamically updating the latent knowledge-state estimates. This allows teaching efforts to focus on genuine weaknesses rather than redundantly reviewing mastered content (Liu et al. 2024; Wang et al. 2026a; Zhang et al. 2026).

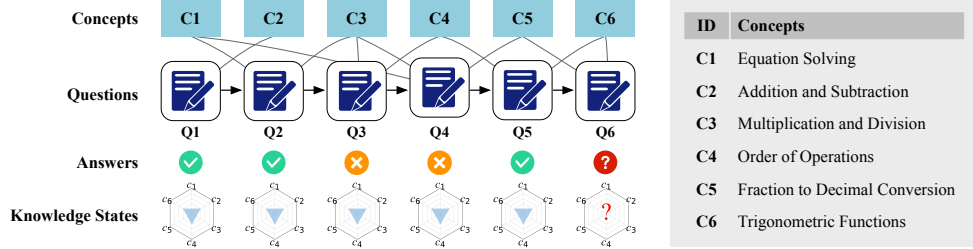


Fig. 1 Example of knowledge tracing with question, concept, response, and knowledge-state sequences

Earlier existing KT models primarily relied on Bayesian methods (Corbett and Anderson 1994; Šarić Grgić et al. 2024), using statistical inference to estimate students’ knowledge states. With the development of deep learning, current KT models have increasingly relied on complicated neural networks to model students’ learning sequences, enabling more accurate capture of the dynamic changes in their knowledge states (Huang et al. 2024; Shen et al. 2023). As illustrated in Figure 2, these deep learning-based KT models can be broadly categorized into two mainstream paradigms (Shen et al. 2024): **concept-based KT** (Piech et al. 2015; Zhang et al. 2017), which focuses on modeling the mastery of underlying concepts, and **question-based KT** (Ghosh et al. 2020; Liu et al. 2023), which emphasizes learning trajectories at the questions level. Both approaches aim to estimate the probability that a student will correctly answer future questions based on their historical interactions.

Despite these advances, existing KT approaches still suffer from two major limitations. First, they typically rely on single-dimensional representations of knowledge states. Most current models model students’ knowledge from either a question perspective or a concept perspective alone, failing to capture the multifaceted dynamics of students’ cognitive processes and thus providing an incomplete picture of real learning. Second, they are prone to overfitting due to severe data sparsity. In real-world scenarios, many questions have low interaction frequency with students, leading to highly sparse datasets. As a result, deep learning models often overfit to noise in the training data rather than learning meaningful knowledge patterns. This issue becomes particularly evident when the model encounters new questions or unseen learning scenarios,

where performance tends to degrade significantly and limits its practical applicability in actual educational settings.

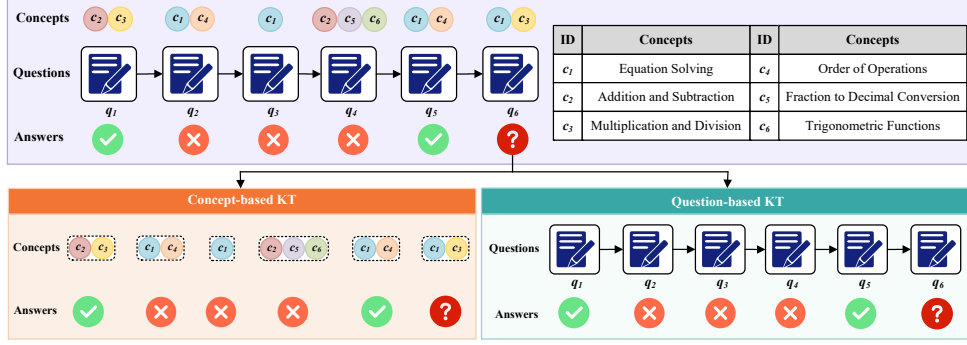


Fig. 2 Concept-based and question-based knowledge tracing paradigms

Considering the above issues, we have envisioned a new KT model that can not only characterize the students' true knowledge status from multiple cognitive dimensions, but also effectively improve the robustness of the model. However, achieving this goal is not easy, and it mainly faces the following two major challenges: (1) How to unify the modeling of students' knowledge states across different cognitive levels. Inspired by cognitive theory (Shen et al. 2024), we draw on the multi-dimensional evaluation strategies commonly used by teachers to assess student learning progress. Specifically, when evaluating a student's ability to solve a given question, teachers often consider: (i) the student's prior performance on the same question, (ii) their performance on similar questions to gauge understanding of related concepts, and (iii) their overall learning history when neither direct nor similar experience exists. Motivated by this, we model three types of knowledge states: question-level (specific mastery of a question), concept-level (understanding of shared concepts across similar questions), and overall-level (overall subject proficiency). The challenge still lies in effectively integrating these dimensions to comprehensively represent a student's evolving knowledge state. (2) How to improve the model's robustness and generalization ability. This could be enhanced through techniques such as adversarial training (AT). However, since the raw data in KT consists of discrete student interactions, this discrete nature makes it challenging to directly apply AT methods (Guo et al. 2021). To address this, it is necessary to design specialized embedding representations tailored to the KT context.

To address the aforementioned challenges, this paper proposes a robust knowledge tracing approach named MDKT (Robust Multi-Dimensional Knowledge Tracing with Adaptive Dual-Phase State Evolution), which will achieve two primary objectives: (1) capturing students' knowledge states through multi-dimensional modeling, and (2) enhancing the robustness of the model by introducing AT. To fulfill the objective (1), we trace students' knowledge states from three different perspectives: (i) question-level knowledge state, derived from the interaction history limited to specific questions, indicating the student's mastery of individual questions; (ii) concept-level knowledge

state, derived from the interaction history restricted to specific concepts, indicating the student’s mastery of questions that share these concepts; (iii) overall-level knowledge state, derived from the student’s complete interaction history, reflecting their overall learning performance. These three perspectives are integrated to comprehensively represent the student’s knowledge state. Additionally, each knowledge state considers both knowledge acquisition and forgetting. Notably, the student’s learning state also influences their knowledge mastery. To capture this dimension, we design an adaptive dual-phase learning state encoder to model the temporal dynamics of learning behavior. To achieve the objective (2), we first propose a new response embedding mechanism and construct a dedicated embedding space to address the challenge of representing discrete interaction data in a continuous form. Then, we introduce an AT framework to further enhance the model’s robustness. We conduct experiments on three real-world datasets, and the results demonstrate that the proposed method achieves competitive and consistently strong performance in predicting students’ future performance across multiple evaluation metrics. MDKT is designed to serve as a core component of expert systems in intelligent educational platforms. It can be directly integrated into adaptive learning systems to provide real-time personalized learning recommendations and intelligent decision support for educators. By accurately modeling students’ evolving knowledge states across multiple dimensions, MDKT enables targeted interventions and helps educators focus on genuine student weaknesses, thereby improving overall learning efficiency and teaching quality. The main contributions are summarized as follows:

- 1) We propose MDKT, a novel multi-dimensional knowledge tracing model that captures students’ knowledge states from multiple cognitive granularities. By integrating an adaptive dual-phase learning state encoder and an adversarial smoothing mechanism, MDKT effectively models both multi-level knowledge dynamics and temporal evolution while maintaining strong robustness under noisy and sparse data conditions.
- 2) To obtain a more comprehensive representation of students’ knowledge, MDKT jointly models the knowledge state at three distinct levels: question-level, concept-level, and overall-level. In addition, we design an adaptive dual-phase learning state encoder to explicitly capture the short-term and long-term temporal dynamics inherent in learning behaviors.
- 3) We evaluate MDKT against several state-of-the-art deep learning-based KT models on three real-world benchmark datasets. Experimental results demonstrate that MDKT consistently achieves superior prediction accuracy and exhibits greater robustness, particularly in long-sequence and data-scarce scenarios.

The remainder of this paper is organized as follows: Section 2 presents a review of related work. Section 3 provides a detailed explanation of the proposed model, followed by Section 4, which presents extensive experiments and detailed reports of the results. Section 5 concludes the paper.

2 Related Work

We first review the related work from two main perspectives in this section. First, we reexamine the representative knowledge tracing models based on deep learning, focusing on two types of deep tracing models: concept-based and question-based. Then, we investigate the recent adversarial training techniques and their applications in enhancing the robustness of machine learning models, especially in the field of education.

2.1 Deep Learning Knowledge Tracing

With the advent of deep learning, numerous deep learning-based knowledge tracing models have been proposed to capture students’ evolving knowledge states from their historical learning interactions. Recent studies have further explored heterogeneous graph-based representation, multi-source feature fusion, and cross-scale temporal modeling (Liu et al. 2026; Ma et al. 2026; Wang et al. 2026b). These models can be broadly categorized based on their modeling granularity and architecture.

Concept-based KT. This line of work focuses on estimating students’ understanding of abstract concepts or skills that underlie the learning content. One of the earliest deep models in this category is Deep Knowledge Tracing (DKT) (Piech et al. 2015), which uses LSTM to model sequential learning behavior. Subsequent methods, such as DKVMN (Zhang et al. 2017), further advance this direction by employing memory-augmented networks to store and dynamically update concept-level knowledge representations. Attention-based methods like SAKT (Pandey and Karypis 2019) emphasized the alignment between student responses and concept embeddings. Graph-based approaches (e.g., GKT (Nakagawa et al. 2019)) further exploit structured concept dependencies by modeling knowledge diffusion over concept graphs. Recent advances, including SKT (Tong et al. 2020) and SKVMN (Abdelrahman and Wang 2019), refine concept modeling through structure-aware mechanisms or enhanced memory operations. Moreover, robust learning strategies such as adversarial training (e.g., ATKT (Guo et al. 2021)) and contrastive representation learning (e.g., CL4KT (Lee et al. 2022)) have been incorporated to improve model generalization and feature expressiveness under limited or noisy data.

Question-based KT. In contrast to concept-based models that focus solely on abstract skills, question-based KT approaches incorporate question-level features to capture more granular learning dynamics. These methods aim to trace how students interact with specific assessment questions, often leveraging additional metadata such as questions difficulty or response context. For example, AKT (Ghosh et al. 2020) adopts a Transformer-based architecture with temporal decay functions to account for forgetting effects during sequential modeling. SAINT (Choi et al. 2020) utilizes a dual-embedding scheme within a Transformer to jointly model both questions and associated concepts. GIKT (Yang et al. 2021) introduces graph convolutional layers to represent the structured relationships between questions and concepts, improving both prediction accuracy and interpretability. Other variants such as LPKT (Shen et al. 2021) focus on aligning students’ dynamic knowledge states with their actual learning process, while DTransformer (Yin et al. 2023) employs contrastive objectives

to stabilize knowledge representations. Lightweight designs like SimpleKT (Liu et al. 2023) demonstrate that even simplified attention mechanisms can maintain strong performance across diverse datasets.

On the other hand, the model we proposed, MDKT, breaks through the traditional one-dimensional modeling and comprehensively considers students’ knowledge from three cognitive levels: question, concept, and overall, so as to be able to characterize students’ learning status more comprehensively and meticulously. At the same time, different from most previous studies that focus on the concept or question level, MDKT designs an adaptive two-phase encoder that can capture the dynamics of students’ learning behavior and has significant advantages in tracking students’ knowledge status.

2.2 Adversarial Training in Educational Modeling

Originally proposed to enhance the robustness of deep neural networks in computer vision tasks (Goodfellow et al. 2015), AT introduces small, deliberately generated perturbations to model inputs during training, thereby improving generalization and resistance to overfitting. Although AT was initially designed for continuous input domains such as images, its effectiveness has led to its adaptation in natural language processing (NLP) tasks. To overcome the limitation of discrete text inputs, (Miyato et al. 2021) pioneered the idea of applying adversarial noise to continuous word embedding spaces for text classification. Following this direction, AT has since been utilized in diverse NLP applications, including cross-lingual classification (Dong et al. 2020), relation extraction (Wu et al. 2017), and neural machine translation (Sato et al. 2019).

While AT has shown promise in improving the robustness of neural models, its integration into KT remains relatively limited. One of the few attempts, ATKKT (Guo et al. 2021), applies adversarial perturbations to student interaction embeddings to regularize the learned knowledge representations. However, ATKKT adopts a relatively shallow representation, focusing solely on question-level interactions without explicitly modeling the underlying structure of students’ knowledge states or their learning dynamics. In contrast, our proposed model, MDKT, extends adversarial regularization into a more expressive KT framework that models student knowledge from a multi-dimensional perspective. By embedding AT within this richer structure, MDKT not only inherits the robustness benefits of adversarial training, but also improves knowledge representation and prediction under conditions of data sparsity, behavioral noise, and learning variability—challenges that are not adequately addressed by existing adversarial KT methods.

3 Methodology

Figure 3 illustrates the overall framework of MDKT. To facilitate understanding, this paper will first introduce the basic question definition of the KT task, and then provide a detailed explanation of each key module of the MDKT model.

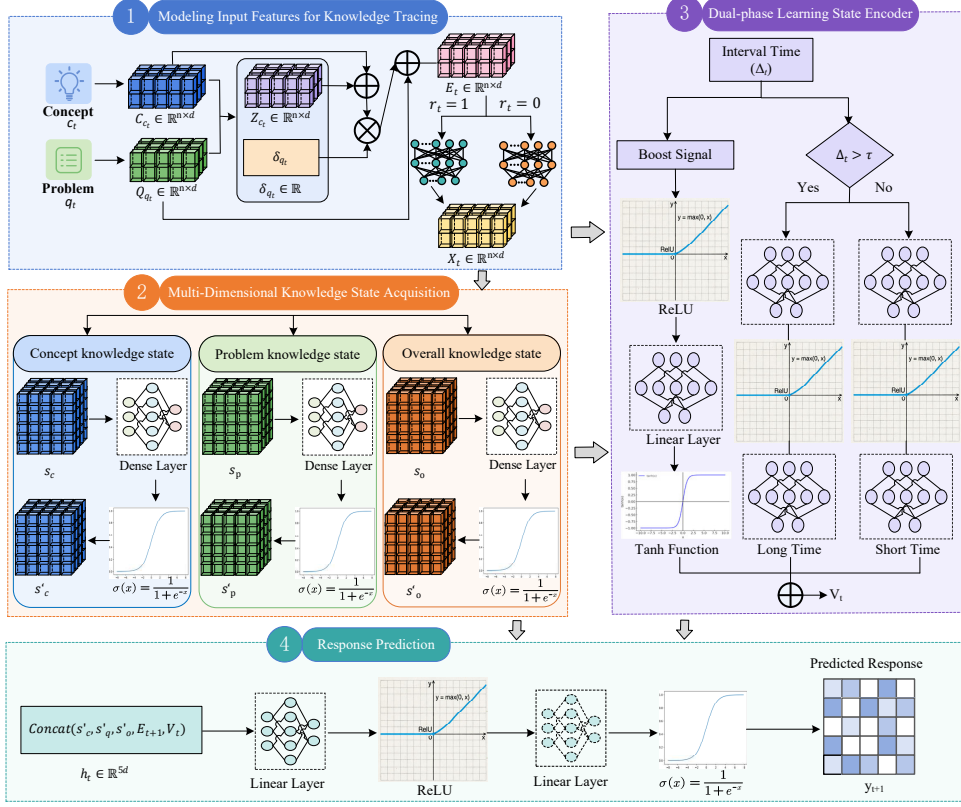


Fig. 3 Overall architecture and forward-computation stages of MDKT

3.1 Problem Definition

In online learning systems, a learner's behavior is primarily represented as a sequence of practice interactions. At each time step t , the learner attempts a question $q_t \in \mathcal{Q}$ associated with a specific knowledge concept $c_t \in \mathcal{C}$, producing a binary response $r_t \in \{0, 1\}$, where $r_t = 1$ indicates a correct answer and $r_t = 0$ indicates an incorrect one.

Formally, the learning trajectory of any learner can be represented as an ordered sequence of quadruples:

$$\mathcal{X} = ((q_t, c_t, r_t, \Delta t_t))_{t=1}^n, \quad (1)$$

$$(q_t, c_t, r_t, \Delta t_t) \in \mathcal{Q} \times \mathcal{C} \times \{0, 1\} \times \mathbb{R}_{\geq 0}.$$

where n is the sequence length, $\mathcal{Q}$ and $\mathcal{C}$ denote the question bank and the set of knowledge concepts, respectively, and Δt_t represents the time interval between consecutive practices. These sequences objectively record the learner's practice history, while the underlying knowledge state remains latent and unobservable.

A key observation is that the external manifestation of knowledge mastery is the learner’s ability to correctly answer subsequent questions. Consequently, the knowledge tracing task can be formally defined as follows: given the historical interaction sequence up to time step $t - 1$, denoted as $\mathcal{X}_{1:t-1} = \{(q_1, c_1, r_1, \Delta t_1), \dots, (q_{t-1}, c_{t-1}, r_{t-1}, \Delta t_{t-1})\}$, the model must infer the learner’s current knowledge state at time t and predict the probability that the learner will correctly answer the current question q_t . This objective is mathematically expressed as the conditional probability:

$$P(r_t = 1 \mid \mathcal{X}_{1:t-1}, q_t) \quad (2)$$

This probability quantifies the likelihood that a learner will correctly solve question q_t based on all prior learning experiences, providing a crucial foundation for subsequent personalized instructional interventions.

3.2 MDKT Framework

To address the limitations of insufficient representational capacity inherent in traditional knowledge tracing methods, this section proposes a knowledge tracing model—MDKT—based on multi-dimensional knowledge state modeling. Concurrently, an adversarial smoothing mechanism is introduced to enable the model to maintain robust stability while simultaneously enhancing its predictive performance. Overall, MDKT takes a student’s historical interaction sequence as input and employs the collaborative efforts of multiple modules to dynamically model the student’s knowledge state, which is then utilized to predict future performance. The model primarily comprises five modules: Input Feature Modeling, Multi-dimensional Knowledge State Modeling, Dual-phase Learning State Encoder, Response Prediction, and Adversarial Smoothing Module. Among these, the first four modules constitute the foundational MDKT model, responsible for tracing and predicting knowledge states; the Adversarial Smoothing Module functions primarily during the training phase, introducing specific constraints to enhance the model’s robustness in the presence of noisy and sparse data.

To facilitate an understanding of the model’s overall architecture and information flow, this section presents a schematic diagram of the model’s framework, as illustrated in [Figure 3](#). As depicted in this figure, the overall workflow encompasses several key stages: Modeling Input Features for Knowledge Tracing, Multi-Dimensional Knowledge State Acquisition, Dual-phase Learning State Encoder, and Response Prediction. It is worth noting that the adversarial smoothing mechanism does not alter the model’s forward computation structure; rather, it imposes constraints on the representation space during the training process and is therefore not explicitly depicted as a separate entity in the diagram.

3.2.1 Modeling Input Features for Knowledge Tracing

Effectively encoding questions, concepts, and student responses is crucial for knowledge tracing, as their combined semantics collectively determine the evolving knowledge state of a student. In MDKT, we adopt a unified embedding-based framework

to jointly model these three components and their interactions within the learning sequence.

Let $Q \in \mathbb{R}^{n \times d}$, $C \in \mathbb{R}^{m \times d}$, and $R \in \mathbb{R}^{2 \times d}$ denote the embedding matrices for questions, knowledge concepts, and responses, respectively, where n , m , and d represent the number of questions, concepts, and the embedding dimension. Inspired by the Rasch model from Item Response Theory (IRT) (Ghosh et al. 2020), we introduce a learnable scalar difficulty parameter δ_{q_t} for each question to capture its inherent difficulty.

Given a question q_t and its associated concept c_t , we compute their composite representation as follows:

$$\mathbf{E}_t = Q_{q_t} + C_{c_t} + \delta_{q_t} \cdot Z_{c_t} \quad (3)$$

where $Q_{q_t}, C_{c_t} \in \mathbb{R}^{1 \times d}$ are the embedding vectors of the question and concept, $\delta_{q_t} \in \mathbb{R}$ is a learnable scalar indicating question difficulty, and $Z_{c_t} \in \mathbb{R}^{1 \times d}$ is a variation vector that models the differences among questions sharing the same concept.

For each step in the interaction sequence, we define an interaction tuple $L_t = (q_t, c_t, r_t)$, where $r_t \in \{0, 1\}$ denotes the correctness of the response ($r_t = 1$ for correct, $r_t = 0$ for incorrect). To capture the nuanced behavioral differences between correct and incorrect responses, we propose an asymmetric interaction mechanism. Specifically, the input representation $\mathbf{X}_t \in \mathbb{R}^{1 \times d}$ at time step t is computed as:

$$\mathbf{X}_t = \begin{cases} \mathbf{E}_t \oplus R_{r_t}, & \text{if } r_t = 1 \\ R_{r_t} \oplus \mathbf{E}_t, & \text{if } r_t = 0 \end{cases} \quad (4)$$

where $\oplus$ denotes vector concatenation. Unlike element-wise addition, vector concatenation preserves the ordering of the input embeddings. Therefore, the representations $\mathbf{E}_t \oplus R_{r_t}$ and $R_{r_t} \oplus \mathbf{E}_t$ are mathematically different and encode different interaction semantics. This asymmetric design allows the model to distinguish subtle behavioral patterns associated with correct and incorrect answers.

3.2.2 Multi-Dimensional Knowledge State Acquisition

Effectively modeling a student’s evolving knowledge state lies at the core of knowledge tracing. Unlike previous methods that typically adopt a single perspective, MDKT explicitly represents the student’s knowledge from three complementary granularities: concept-level, question-level, and overall-level.

Given the historical interaction sequence $\mathcal{L} = \{(q_1, c_1, r_1), \dots, (q_t, c_t, r_t)\}$, at the current time step t where the student is about to answer question q_t associated with concept c_t , we construct the three knowledge states as follows:

For concept knowledge state modeling, we gather all prior interactions involving the same concept c_t :

$$\mathcal{L}_t^C = \bigcup_{j < t, c_j = c_t} (q_j, c_j, r_j) \quad (5)$$

To model question knowledge state, we extract the set of historical interactions involving the same question:

$$\mathcal{L}_t^Q = \bigcup_{j < t, q_j = q_t} (q_j, c_j, r_j) \quad (6)$$

To construct overall knowledge state, we consider the complete history of interactions prior to the current step:

$$\mathcal{L}_t^O = \{X_1, X_2, \dots, X_{t-1}\} \quad (7)$$

To dynamically maintain and update multiple knowledge states, we adopt a unified two-stage mechanism consisting of a forgetting phase followed by a learning phase. This design is motivated by the observation from educational psychology that knowledge retention gradually changes with the elapsed time between learning activities. Instead of assuming a predefined forgetting function, we employ a learnable neural gating mechanism that adaptively estimates the forgetting intensity according to both the current knowledge state and the elapsed time. This enables the model to capture diverse temporal forgetting patterns across different students and learning contexts.

For each type of knowledge state s (i.e., concept-level s_c , question-level s_q , or overall-level s_o), we first retrieve its most recent value from the latest relevant interaction. To account for temporal decay, we introduce a time-aware forgetting gate:

$$f_t = \sigma([s \oplus I_\alpha]W_1 + b_1) \quad (8)$$

where I_α denotes the embedding of the elapsed time interval since the last relevant interaction, $\sigma(\cdot)$ is the Sigmoid activation function, and W_1 and b_1 are learnable parameters. The interval embedding projects the raw time interval into a continuous latent representation, allowing the model to capture nonlinear temporal patterns. The Sigmoid gate further generates an adaptive forgetting coefficient within $(0, 1)$, enabling the model to dynamically regulate the proportion of retained historical knowledge while maintaining stable optimization. The decayed state is then computed via element-wise modulation:

$$s'_t = f_t \odot s \quad (9)$$

After observing the response to question q_t , the model incorporates the new learning signal to update the state:

$$s_{t+1} = s'_t + \text{SeLU}([s'_t \oplus X_t]W_2 + b_2) \quad (10)$$

where X_t denotes the embedding of the current interaction, $\oplus$ represents vector concatenation, and W_2 and b_2 are learnable parameters. The residual update structure incrementally incorporates newly acquired knowledge while preserving the retained historical knowledge, leading to stable knowledge evolution and effective gradient propagation. Meanwhile, the SeLU activation enhances the nonlinear representation capability of the updated knowledge state.

By applying this forgetting-then-learning procedure to s_c , s_q , and s_o independently, MDKT maintains a multi-level and dynamically evolving representation of student knowledge, enabling both fine-grained modeling at the concept and question levels and a holistic characterization of overall mastery.

3.2.3 Dual-phase Learning State Encoder

Temporal information is crucial in knowledge tracing, as students' knowledge states evolve over time due to both natural forgetting and reinforcement from recent practice. To explicitly model these temporal dynamics, we propose a dual-phase learning state encoder, which captures short-term forgetting and long-term retention patterns while incorporating the reinforcing effect of recent learning activities.

Given the time interval $\Delta t \in \mathbb{R}^+$ between two consecutive interactions, we design two independent transformation branches to characterize different temporal regimes. Specifically, the short-term and long-term temporal representations are computed as:

$$\mathbf{s}_t = \text{ReLU}(W_s \cdot \Delta t + \mathbf{b}_s) \quad (11)$$

$$\mathbf{l}_t = \text{ReLU}(W_l \cdot \Delta t + \mathbf{b}_l) \quad (12)$$

where $W_s, W_l \in \mathbb{R}^{d \times 1}$ and $\mathbf{b}_s, \mathbf{b}_l \in \mathbb{R}^d$ are learnable projection weights and biases, respectively. The two branches provide flexible nonlinear representations for different temporal forgetting patterns without imposing a predefined decay function.

To adaptively balance the contributions of short-term forgetting and long-term retention, we introduce a learnable temporal transition threshold τ . Instead of using a hard binary selection, which leads to a non-differentiable optimization problem, we employ a differentiable sigmoid-based gating mechanism:

$$g_t = \sigma(\gamma(\Delta t - \tau)) \quad (13)$$

where $\sigma(\cdot)$ denotes the sigmoid function and γ is a fixed scaling factor controlling the sharpness of the transition, which is set to 0.001 in our implementation. The threshold τ is treated as a trainable parameter and optimized jointly with other model parameters through backpropagation.

The final temporal representation is obtained by adaptively combining the two temporal branches:

$$\mathbf{T}_t = W_o \cdot ((1 - g_t)\mathbf{s}_t + g_t\mathbf{l}_t) \quad (14)$$

where $W_o \in \mathbb{R}^{d \times d}$ is a learnable fusion matrix. When Δt is smaller than the learned transition boundary τ , g_t approaches zero and the encoder emphasizes short-term forgetting dynamics. Conversely, when Δt becomes larger than τ , g_t gradually approaches one and the long-term retention representation receives more attention. This differentiable transition allows the model to automatically discover an appropriate temporal boundary from the training data.

In our implementation, τ is initialized to 10000 before training. This value only provides a reasonable temporal scale for the initial optimization stage and does not constrain the final transition boundary. During training, τ is continuously updated through gradient-based optimization, and its converged value is determined by the

temporal characteristics of the dataset. Therefore, no additional cross-validation is required to manually select τ .

In addition to modeling forgetting effects, we explicitly incorporate the reinforcement effect caused by recent learning activities. Intuitively, recently practiced knowledge is more likely to be reinforced and maintained. We model this practice-induced boost signal using an exponentially decaying transformation:

$$\mathbf{B}_t = \tanh \left(W_b \cdot \exp \left(-\frac{\Delta t}{\lambda} \right) \right) \quad (15)$$

where $W_b \in \mathbb{R}^{d \times 1}$ is a learnable projection vector and $\lambda > 0$ is a learnable decay scaling factor. The hyperbolic tangent activation constrains the magnitude of the boost signal and improves optimization stability.

Finally, the complete time-aware learning state representation $\mathbf{V}_t$ is obtained by integrating the temporal encoding and the practice-induced reinforcement signal:

$$\mathbf{V}_t = \mathbf{T}_t + \mathbf{B}_t \quad (16)$$

This dual-phase formulation enables MDKT to model the nonlinear temporal evolution of human learning, including rapid forgetting over short intervals, gradual retention over long intervals, and reinforcement from recent practice. By learning the temporal transition boundary directly from data, the proposed encoder provides a flexible and adaptive estimation of student knowledge states under diverse learning scenarios.

3.2.4 Response Prediction

At each time step t , we derive three independent knowledge states to represent the student’s cognitive mastery: the concept knowledge state $\mathbf{s}'_c$, the question knowledge state $\mathbf{s}'_q$, and the overall knowledge state $\mathbf{s}'_o$. These latent representations are dynamically updated based on recent interactions within their respective granularities.

To incorporate future content information, we compute the question–concept joint embedding $\mathbf{E}_{t+1}$, as defined in Eq. (3), which captures the semantics and difficulty of the upcoming question. In parallel, the dual-phase learning state encoder produces a temporal representation $\mathbf{V}_t$, which models both short-term forgetting and practice-induced reinforcement based on the interval time since the last relevant interaction. These five components are concatenated to form a joint predictive feature:

$$\mathbf{h}_t = \text{Concat} [\mathbf{s}'_c, \mathbf{s}'_q, \mathbf{s}'_o, \mathbf{E}_{t+1}, \mathbf{V}_t] \in \mathbb{R}^{5d} \quad (17)$$

The probability $y_{t+1} \in [0, 1]$ that the student will correctly answer the next question q_{t+1} is predicted as:

$$y_{t+1} = \sigma (W_2 \cdot \text{ReLU}(W_1 \cdot \mathbf{h}_t + \mathbf{b}_1) + b_2) \quad (18)$$

where $W_1 \in \mathbb{R}^{d \times 5d}$, $W_2 \in \mathbb{R}^{1 \times d}$, and $\mathbf{b}_1 \in \mathbb{R}^d$, $b_2 \in \mathbb{R}$ are learnable parameters, and $\sigma(\cdot)$ denotes the sigmoid activation function. Binary cross-entropy loss is used to

measure the difference between the predicted response y_{t+1} and the actual response r_{t+1} . Specifically, the loss at time step $t + 1$ is calculated as:

$$l(y_{t+1}, r_{t+1}) = -(r_{t+1} \log y_{t+1} + (1 - r_{t+1}) \log(1 - y_{t+1})) \quad (19)$$

3.2.5 Adversarial Smoothing

Although MDKT effectively models the multi-dimensional knowledge states and temporal dynamics, real-world educational data are often sparse, irregular, and noisy, which may introduce instability into the learned knowledge representations and increase the risk of overfitting. To improve the robustness and smoothness of the learned representation space, we introduce an adversarial smoothing mechanism during training.

As illustrated in Figure 3, the proposed adversarial smoothing module serves as an auxiliary regularization strategy and does not participate in the forward inference process. Specifically, we generate controlled perturbations on the high-level semantic representation $\mathbf{h}_t$, which integrates multi-dimensional knowledge states and temporal information, and encourage the model to maintain stable predictions under small perturbations in the latent representation space.

Unlike Virtual Adversarial Training (VAT), which typically constructs perturbations by maximizing the divergence between the original and perturbed prediction distributions without using ground-truth labels, our method adopts a supervised gradient-based adversarial smoothing strategy. Specifically, the perturbation is generated according to the gradient of the supervised prediction loss, and the perturbed representation is optimized to preserve the correctness of the original response prediction.

Formally, let $\mathbf{h}_t$ denote the joint representation used for predicting the probability of correctly answering the next question. We first perform a standard forward pass to obtain the original prediction:

$$y_{t+1} = \sigma(\text{MLP}(\mathbf{h}_t)) \quad (20)$$

where $\text{MLP}(\cdot)$ denotes the prediction head and $\sigma(\cdot)$ is the sigmoid activation function that maps the output to a probability in $(0, 1)$. The binary cross-entropy loss is computed as:

$$\mathcal{L}_{\text{BCE}} = -(r_{t+1} \log y_{t+1} + (1 - r_{t+1}) \log(1 - y_{t+1})) \quad (21)$$

where $r_{t+1} \in \{0, 1\}$ is the ground-truth response label.

Next, the gradient of the supervised loss with respect to the latent representation $\mathbf{h}_t$ is calculated:

$$\mathbf{g} = \nabla_{\mathbf{h}_t} \mathcal{L}_{\text{BCE}} \quad (22)$$

which indicates the direction in the representation space that increases the prediction loss most rapidly. Based on this gradient, we construct the normalized adversarial perturbation:

$$\hat{\delta} = \frac{\mathbf{g}}{\|\mathbf{g}\|_2 + \eta} \quad (23)$$

where $\|\cdot\|_2$ denotes the ℓ_2 norm and $\eta = 10^{-16}$ is a small constant introduced to avoid numerical instability.

The perturbed representation is generated as:

$$\tilde{\mathbf{h}}_t = \mathbf{h}_t + \epsilon \cdot \hat{\delta} \quad (24)$$

where $\epsilon > 0$ controls the perturbation magnitude. A small perturbation scale is adopted to ensure that the generated representation remains within a local neighborhood of the original representation, thereby encouraging smoothness rather than introducing excessive noise.

The perturbed representation is then fed into the same prediction head:

$$\tilde{y}_{t+1} = \sigma(\text{MLP}(\tilde{\mathbf{h}}_t)) \quad (25)$$

The corresponding supervised adversarial loss is defined as:

$$\mathcal{L}_{\text{adv}} = -(r_{t+1} \log \tilde{y}_{t+1} + (1 - r_{t+1}) \log(1 - \tilde{y}_{t+1})) \quad (26)$$

Finally, the overall training objective combines the original prediction loss and the adversarial regularization term:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{BCE}} + \beta \cdot \mathcal{L}_{\text{adv}} \quad (27)$$

where $\beta \in [0, 1]$ controls the contribution of adversarial smoothing. In our experiments, ϵ and β are selected according to validation performance to balance prediction accuracy and representation robustness.

It is worth noting that the proposed adversarial smoothing mechanism does not introduce additional parameters or modify the inference architecture. Since the perturbation is generated from the first-order gradient of the supervised loss with respect to the continuous representation $\mathbf{h}_t$, and no discrete operation is involved, the optimization process remains differentiable without introducing gradient masking issues.

Importantly, the adversarial smoothing module is designed as a complementary regularization component rather than the primary source of performance improvement. The main improvements of MDKT are attributed to the proposed multi-dimensional knowledge state modeling and dual-phase temporal evolution mechanism, while adversarial smoothing further enhances representation stability by encouraging local smoothness under small perturbations. Therefore, although the performance gains introduced by this module are relatively modest, the consistent improvements across different datasets demonstrate its effectiveness in improving robustness under sparse and noisy educational scenarios.

The perturbation is applied only to the high-level semantic representation $\mathbf{h}_t$ rather than the raw discrete inputs. This design encourages smoothness in the fused knowledge representation space while preserving the original inference procedure.

4 Experiment Results

We first present the real-world datasets and baseline models used in our experiments. Then, we describe the experimental setup and implementation details. Following that, we will detail the experimental results and conduct in-depth analysis to address the following research questions (RQs):

RQ1: How does MDKT perform compared with state-of-the-art methods in student performance prediction?

RQ2: What is the impact of each proposed component on prediction performance?

RQ3: How robust is MDKT to variations in sequence length and training data proportions?

RQ4: Does MDKT enable interpretable visualization of students' knowledge state evolution?

4.1 Experimental Setup

4.1.1 Datasets

To comprehensively evaluate the performance and generalizability of the proposed MDKT model, we utilize three widely adopted real-world benchmark datasets in knowledge tracing: ASSISTments2009¹, Bridge2006², and Junyi³. These datasets differ significantly in knowledge structure complexity, student behavior patterns, and temporal characteristics.

- **ASSISTments2009 (AS2009):** This dataset comes from the ASSISTments online learning platform and contains student interaction logs with math exercises from the 2009–2010 school year. It has been widely used as a benchmark for knowledge tracing research due to its scale and detailed skill tagging, making it suitable for evaluating different KT models (Lee et al. 2022).
- **Bridge2006 (BD2006):** This dataset is part of the KDD Cup 2010 Educational Data Mining Challenge. It includes step-by-step solution logs from students solving math questions between November 2006 and June 2007 (Stamper et al. 2010). The data offers fine-grained annotations of each step and related skills.
- **Junyi:** This dataset was collected from the Junyi Academy, an e-learning platform similar to Khan Academy. It provides a large number of randomly generated math exercises. The dataset records student learning behaviors between October 2012 and January 2015, including their attempts, answers, timestamps, and associated skills (Ma et al. 2024). It is often used to study sequential learning patterns in real educational settings.

Our data preprocessing procedure consists of several key steps. First, following the common preprocessing protocol adopted in existing knowledge tracing studies (Wang et al. 2021), questions associated with multiple knowledge concepts are decomposed into multiple question-concept interaction records. Specifically, each question-concept pair is treated as an independent interaction while preserving the original response

¹<https://sites.google.com/site/assistmentsdata/home/2009-2010-assistment-data/skill-builder-data-2009-2010>.

²<https://pslcdatashop.web.cmu.edu/KDDCup>.

³<https://pslcdatashop.web.cmu.edu/DatasetInfo?datasetId=1198>.

label and timestamp, so that every processed interaction corresponds to a single target concept.

Next, following the approach in (Wang et al. 2021), we remove interaction records lacking skill annotations and exclude students with fewer than five total interactions to mitigate overfitting caused by sparse data. We then normalize all timestamps and order each student’s learning records chronologically to preserve temporal integrity. To effectively segment learning sessions, we set a maximum allowable interval of 86,400 seconds (24 hours) between consecutive interactions. This threshold is motivated by two considerations: (1) cognitive continuity—intervals exceeding 24 hours are regarded as breaks in knowledge retention, reflecting possible forgetting effects; and (2) behavioral validity—this cutoff helps filter out irregular or sporadic logins from inactive accounts that do not represent meaningful learning activity, thereby reducing noise in model training. Table 1 provides detailed information on the original and processed data for these datasets.

Table 1 Details of the original and preprocessed data for the three real-world datasets.

Stage	Details	AS2009	BD2006	Junyi
Original	Interactions	525,534	3,679,199	359,804
	Students	4,217	1,146	2,063
	Questions	26,688	19,355	1,880
	KCs	123	564	631
Preprocessed	Interactions	445,734	1,809,644	289,605
	Students	3,678	1,138	1,520
	Questions	16,868	19,001	1,872
	KCs	111	558	630

4.1.2 Baseline

To thoroughly assess the effectiveness of the proposed MDKT model, we conduct comparisons against 19 robust baseline methods. The evaluation is performed on the three real-world datasets introduced earlier. The baseline models are summarized as follows:

- **DKT** (Piech et al. 2015): As one of the earliest deep learning-based knowledge tracing models, DKT employs recurrent neural networks (RNNs) to capture the sequential patterns of students’ skill acquisition and track the evolution of their knowledge states over time.
- **DKT+** (Yeung and Yeung 2018): This variant enhances the original DKT by incorporating additional regularization terms aimed at improving reconstruction and reducing prediction fluctuations, thereby increasing the consistency of the model’s output without sacrificing accuracy.
- **DKT-F** (Nagatani et al. 2019): By integrating a forgetting mechanism, DKT-F models the natural decay of knowledge over time, providing a more realistic simulation of the learning and forgetting processes.

- **DKVMN** (Zhang et al. 2017): Utilizing a key-value memory network, DKVMN dynamically updates students’ knowledge states and efficiently retrieves relevant information during learning, enabling fine-grained tracking of knowledge components.
- **Deep-IRT** (Yeung 2019): Building upon the DKVMN framework, Deep-IRT combines neural networks with Item Response Theory (IRT) to improve interpretability and provide probabilistic insights into student responses.
- **GKT** (Nakagawa et al. 2019): This approach applies graph neural networks (GNNs) to model the complex relationships between students and knowledge concepts, capturing rich structural information within the learning data.
- **SAKT** (Pandey and Karypis 2019): SAKT pioneers the use of self-attention mechanisms in knowledge tracing, effectively modeling long-range dependencies among knowledge components.
- **AKT** (Ghosh et al. 2020): AKT enhances the attention mechanism by incorporating temporal features to model forgetting behavior during the interaction between past exercises and current questions.
- **SimpleKT** (Liu et al. 2023): Employing a straightforward attention mechanism, SimpleKT achieves competitive performance and serves as a strong benchmark within attention-based KT models.
- **SparseKT** (Huang et al. 2023): This model introduces sparse attention techniques to mitigate overfitting in attention-based KT models, particularly on smaller datasets.
- **FolibiKT** (Im et al. 2023): FolibiKT offers a simple yet effective attention-based method that explicitly models students’ forgetting behavior.
- **SAINT** (Choi et al. 2020): SAINT adopts a transformer architecture, encoding exercises and student responses separately through an encoder-decoder framework to capture rich contextual information.
- **SAINT+** (Liu et al. 2022): An improved version of SAINT, this model incorporates novel positional embeddings that enhance prediction accuracy.
- **KTM** (Vie and Kashima 2019): KTM applies factorization machines to model feature interactions, emphasizing temporal factors in knowledge acquisition.
- **HawkesKT** (Wang et al. 2021): Modeling learning as a temporal point process, HawkesKT is the first to include skill cross-effects as time-varying features in knowledge tracing.
- **KQN** (Lee and Yeung 2019): This model encodes student activities into knowledge states and skill vectors using neural networks, then models their interactions via dot product operations.
- **StableKT** (Li et al. 2024b): StableKT improves the length generalization capabilities of attention-based knowledge tracing models by effectively capturing intricate interactions between questions and their associated concepts, utilizing a hybrid approach that combines dot-product attention with hyperbolic attention mechanisms.
- **ExtraKT** (Li et al. 2024a): ExtraKT is a powerful model designed specifically to tackle the challenge of length extrapolation in knowledge tracing tasks, improving model performance on longer sequences.

Table 2 Performance comparison on the AS2009 dataset.

Model	ACC	AUC	MAE	RMSE
DKT	0.7221±0.0030	0.7616±0.0058	0.3690±0.0044	0.4301±0.0016
DKT+	0.6925±0.0070	0.7560±0.0076	<u>0.3413±0.0050</u>	0.4519±0.0042
DKT-F	0.7292±0.0074	0.7711±0.0060	0.3604±0.0060	0.4250±0.0032
DKVMN	0.7143±0.0069	0.7495±0.0060	0.3762±0.0066	0.4345±0.0033
Deep-IRT	0.7086±0.0062	0.7401±0.0060	0.3816±0.0042	0.4384±0.0031
GKT	0.7156±0.0110	0.7522±0.0130	0.3678±0.0106	0.4413±0.0121
HawkesKT	0.7311±0.0066	0.7697±0.0049	0.3706±0.0049	0.4271±0.0026
KQN	0.7150±0.0065	0.7405±0.0062	0.3858±0.0040	0.4372±0.0031
SAKT	0.6981±0.0064	0.7227±0.0065	0.3910±0.0075	0.4448±0.0030
AKT	0.7241±0.0028	0.7600±0.0069	0.3491±0.0120	0.4383±0.0053
SimpleKT	0.7195±0.0058	0.7588±0.0091	0.3488±0.0152	0.4363±0.0059
SparseKT	0.7196±0.0054	0.7463±0.0091	0.3501±0.0071	0.4448±0.0093
FolibiKT	0.7232±0.0047	0.7588±0.0101	0.3401±0.0140	0.4387±0.0079
StableKT	0.7117±0.0076	0.7608±0.0068	0.3861±0.0053	0.4382±0.0030
ExtraKT	0.7152±0.0067	0.7602±0.0060	0.3609±0.0069	0.4342±0.0025
SAINT	0.7250±0.0073	0.7725±0.0064	0.3591±0.0133	0.4273±0.0035
SAINT+	0.7036±0.0080	0.7239±0.0065	0.3811±0.0050	0.4485±0.0041
KTM	0.7224±0.0065	0.7533±0.0038	0.3743±0.0080	0.4359±0.0032
ATKT	<u>0.7424±0.0058</u>	<u>0.7837±0.0048</u>	0.3333±0.0036	<u>0.4186±0.0027</u>
MDKT	0.7520±0.0060*	0.8126±0.0045*	0.3170±0.0057	0.4064±0.0018*

The best results are shown in bold and the second-best results are underlined. * indicates that the improvement over the best baseline is statistically significant under a paired t -test ($p < 0.05$).

- **ATKT** (Guo et al. 2021): ATKT is an adversarial training-based knowledge tracing framework that introduces adversarial perturbations into the learning process to improve the robustness of knowledge state estimation.

4.1.3 Implementation Details

To establish a standardized evaluation framework, we first sort all students' learning records chronologically according to their response timestamps. Following widely adopted evaluation protocols in previous knowledge tracing studies (Sun et al. 2024; Wang et al. 2021), we use the first 50 interaction records per student as the default evaluation setting. This protocol has been commonly adopted to ensure a fair comparison among different models by controlling the sequence length variations across students. Moreover, it provides a consistent benchmark for evaluating models under limited historical interaction scenarios. To further investigate the effectiveness of MDKT in long-term learning scenarios, we additionally evaluate model performance with longer interaction sequences (i.e., 100, 150, 200, and 250 interactions). Following the established methodology, we adopt five-fold cross-validation, where 10% of the training

Table 3 Performance comparison on the BD2006 dataset.

Model	ACC	AUC	MAE	RMSE
DKT	0.8055±0.0059	0.7045±0.0123	0.2887±0.0092	0.3798±0.0061
DKT+	0.8033±0.0055	0.7268±0.0035	0.2716±0.0045	0.3751±0.0041
DKT-F	0.8081±0.0061	0.7186±0.0078	0.2812±0.0063	0.3749±0.0044
DKVMN	<u>0.8107±0.0053</u>	<u>0.7452±0.0042</u>	0.2691±0.0041	<u>0.3689±0.0043</u>
Deep-IRT	0.8102±0.0054	0.7414±0.0072	0.2706±0.0046	0.3698±0.0044
GKT	0.7954±0.0402	0.7051±0.0735	0.2702±0.0370	0.3943±0.0479
HawkesKT	0.8086±0.0041	0.7288±0.0151	0.2874±0.0043	0.3755±0.0031
KQN	0.8075±0.0048	0.6344±0.0053	0.3844±0.0348	0.4136±0.0169
SAKT	0.8079±0.0043	0.7058±0.0048	0.2789±0.0039	0.3771±0.0040
AKT	0.8073±0.0049	0.7446±0.0087	0.2686±0.0058	0.3701±0.0045
SimpleKT	0.8075±0.0048	0.7161±0.0055	0.2734±0.0066	0.37518±0.0038
SparseKT	0.8075±0.0048	0.7176±0.0050	0.2692±0.0054	0.3747±0.0042
FolibiKT	0.8075±0.0048	0.7144±0.0060	0.2726±0.0052	0.3753±0.0036
StableKT	0.8062±0.0060	0.7451±0.0060	0.2598±0.0069	0.3694±0.0043
ExtraKT	0.8065±0.0058	0.7441±0.0081	<u>0.2659±0.0098</u>	0.3695±0.0046
SAINT	0.8086±0.0038	0.7063±0.0129	0.2828±0.0131	0.3776±0.0048
SAINT+	0.8046±0.0043	0.6843±0.0071	0.2816±0.0060	0.3819±0.0042
KTM	0.8062±0.0052	0.7141±0.0196	0.2870±0.0162	0.3788±0.0066
ATKT	0.8072±0.0075	0.7356±0.0052	0.2820±0.0087	0.3774±0.0050
MDKT	0.8124±0.0063	0.7476±0.0103	0.2739±0.0024	0.3681±0.0048

The best results are shown in bold and the second-best results are underlined. * indicates that the improvement over the best baseline is statistically significant under a paired t -test ($p < 0.05$).

sequences are further reserved as the validation set for hyperparameter selection and early stopping.

The proposed MDKT is implemented in PyTorch and trained on an NVIDIA RTX 3090 GPU. Unless otherwise specified, the Adam optimizer is adopted with learning rates selected from $\{1 \times 10^{-3}, 1 \times 10^{-4}, 1 \times 10^{-5}\}$. The embedding dimension and hidden dimension are both set to 64, the dropout rate is fixed at 0.2, the batch size is set to 64, and the random seed is fixed at 3407 to ensure reproducibility. The model is trained for at most 200 epochs, and early stopping with a patience of 15 epochs is employed based on the validation AUC. The model achieving the best validation performance is used for final testing. All hyperparameters were selected according to the validation performance and kept unchanged across all datasets unless otherwise specified.

4.2 Experimental Results and Analysis

4.2.1 Overall Performance (RQ1)

KT aims to predict students' future learning outcomes. The core evaluation method involves comparing models' performance on next-step prediction tasks. As shown

Table 4 Performance comparison on the Junyi dataset.

Model	ACC	AUC	MAE	RMSE
DKT	0.7769±0.0157	0.8276±0.0158	0.3036±0.0123	0.3943±0.0109
DKT+	<u>0.8002±0.0096</u>	0.8476±0.0116	<u>0.2486±0.0133</u>	0.3860±0.0095
DKT-F	0.7771±0.0169	0.8253±0.0183	0.2839±0.0168	0.3956±0.0137
DKVMN	0.7579±0.0159	0.7996±0.0220	0.3012±0.0190	0.4111±0.0131
Deep-IRT	0.7390±0.0139	0.7709±0.0125	0.3314±0.0136	0.4226±0.0073
GKT	0.7840±0.0186	0.8008±0.0918	0.2731±0.0062	0.4027±0.0359
HawkesKT	0.7687±0.0152	0.8087±0.0126	0.3314±0.0184	0.4068±0.0110
KQN	0.6649±0.0104	0.6635±0.0202	0.4997±0.0001	0.4997±0.0001
SAKT	0.7039±0.0190	0.7212±0.0221	0.3720±0.0139	0.4457±0.0106
AKT	0.7810±0.0085	0.8034±0.0108	0.2860±0.0081	0.4060±0.0061
SimpleKT	0.7898±0.0091	0.8366±0.0123	0.2738±0.0062	0.3876±0.0092
SparseKT	0.7768±0.0108	0.8236±0.0117	0.2789±0.0072	0.3982±0.0115
FolibiKT	0.7905±0.0131	0.8373±0.0120	0.2746±0.0083	0.3882±0.0101
StableKT	0.7890±0.0100	0.8362±0.0137	0.2749±0.0065	0.3884±0.0099
ExtraKT	0.7870±0.0148	0.8372±0.0140	0.2738±0.0109	0.3900±0.0127
SAINT	0.7985±0.0101	<u>0.8503±0.0117</u>	0.2663±0.0158	0.3833±0.0085
SAINT+	0.7783±0.0133	0.8208±0.0125	0.2990±0.0138	0.3995±0.0089
KTM	0.7425±0.0117	0.7525±0.0238	0.3169±0.0140	0.4332±0.0113
ATKT	0.7972±0.0132	0.8486±0.0170	0.2763±0.0110	<u>0.3804±0.0121</u>
MDKT	0.8482±0.0057*	0.9091±0.0071*	0.2080±0.0079*	0.3305±0.0076*

The best results are shown in bold and the second-best results are underlined. * indicates that the improvement over the best baseline is statistically significant under a paired t -test ($p < 0.05$).

in Table 2, Table 3 and Table 4, we systematically compare our model against all baseline approaches across three benchmark datasets using four evaluation metrics: accuracy (ACC), area under the ROC curve (AUC), mean absolute error (MAE), and root mean square error (RMSE). The selection of these four metrics ensures a comprehensive evaluation of model performance (Yin et al. 2023), with particular emphasis on AUC as the primary evaluation criterion, consistent with prior research in sequential learning (Li et al. 2024a,b).

Table 2, Table 3 and Table 4 present the performance comparison between MDKT and multiple baseline models across three real-world datasets. From the results, the following observations can be drawn: (1) MDKT achieves consistently superior performance across the three benchmark datasets. Specifically, it obtains the highest AUC values of 0.8126, 0.7476, and 0.9091 on AS2009, BD2006, and Junyi, respectively, demonstrating its effectiveness in modeling students’ evolving knowledge states. Moreover, MDKT achieves the best ACC and RMSE results across all datasets. Regarding MAE, MDKT obtains the lowest prediction error on AS2009 and Junyi, while StableKT achieves a slightly lower MAE on BD2006. These results demonstrate the effectiveness and generalization capability of MDKT across diverse evaluation scenarios. (2) Different baseline models exhibit varying performance characteristics across

datasets. For example, DKT+ achieves competitive performance on the Junyi dataset, obtaining the second-best ACC among all compared baselines, while its performance on AS2009 and BD2006 is relatively weaker. Deep-IRT shows competitive results on BD2006 but performs less effectively on AS2009 and Junyi, indicating that different modeling strategies may have different adaptation capabilities under diverse learning environments. (3) Attention-based models, including SAKT, SimpleKT, AKT, and FolibiKT, achieve competitive overall performance. Notably, models incorporating forgetting mechanisms, such as AKT and FolibiKT, demonstrate advantages on several evaluation metrics, suggesting that explicitly modeling forgetting behaviors is beneficial for knowledge tracing. Transformer-based models, such as SAINT and SAINT+, also exhibit strong performance, with SAINT achieving the second-best AUC of 0.8503 on the Junyi dataset, highlighting the effectiveness of capturing long-range dependencies in learning sequences. (4) Memory network-based models, such as DKVMN, provide relatively stable performance but generally lag behind MDKT, particularly on AS2009 and Junyi. Meanwhile, KQN shows relatively weaker performance across the evaluated datasets, which may be attributed to its limited capacity to capture complex temporal knowledge evolution. (5) Overall, no single baseline consistently dominates across all datasets and evaluation metrics, reflecting the different advantages and limitations of existing KT approaches. In comparison, MDKT achieves comprehensive improvements by integrating multi-granularity knowledge state modeling with adaptive dual-phase temporal encoding, demonstrating stronger generalization capability for modeling dynamic learning processes.

4.2.2 Ablation Study (RQ2)

In this section, we conduct comprehensive ablation experiments to evaluate the contribution of each key component in the proposed MDKT model. In particular, we design eight ablated variants by selectively removing specific modules, including the three-level knowledge state encodings — question-level (Q), concept-level (C), and overall-level (O) — as well as the dual-time learning dynamic encoder (T) and the adversarial training mechanism (AT). The details of these variants are described as follows:

- MDKT w/o C,O: This variant retains only the question-level knowledge state, removing both concept-level and overall-level representations. It is used to assess the effect of relying solely on question-based modeling for knowledge tracing.
- MDKT w/o Q,O: This variant retains only the concept-level knowledge state, excluding both Q and O. It is used to assess the effect of relying solely on concept-based modeling for knowledge tracing.
- MDKT w/o Q,C: This variant retains only the overall-level knowledge state, excluding both Q and C. It is used to assess the effect of relying solely on overall knowledge trends for knowledge tracing.
- MDKT w/o O: This variant retains both Q and C, while removing the overall-level knowledge state. It is used to assess the effect of jointly modeling question- and concept-level knowledge.

- MDKT w/o C: This variant retains Q and O, while removing the concept-level knowledge state. It is used to assess the effect of jointly modeling question- and overall-level knowledge.
- MDKT w/o Q: This variant retains C and O, while removing the question-level knowledge state. It is used to assess the effect of jointly modeling concept- and concept-level knowledge.
- MDKT w/o T: This variant removes the dual-time learning dynamic encoder, while retaining Q, C, and O. It is used to assess the impact of modeling temporal learning patterns.
- MDKT w/o AT: This variant removes the adversarial training mechanism, while retaining the rest of the model. It is used to examine the effect of adversarial regularization.

Table 5, Table 6, and Table 7 present the ablation results of MDKT on the AS2009, Junyi, and BD2006 datasets, respectively. The following observations can be drawn: (1) When only one type of knowledge state is modeled, MDKT w/o Q,O achieves the best performance among the three single-state variants on most datasets. This suggests that concept-level representations provide informative knowledge structures for knowledge tracing. In contrast, MDKT w/o Q,C generally obtains inferior performance, indicating that relying only on overall-level knowledge states is insufficient to capture fine-grained learning dynamics. (2) Incorporating multiple types of knowledge states generally improves performance compared with single-state variants. Among the two-state variants, MDKT w/o O achieves competitive performance across datasets. For example, compared with MDKT w/o Q,C, it improves AUC by 4.10% on AS2009 and 2.40% on Junyi. These results demonstrate that combining different knowledge granularities provides complementary information for modeling students’ learning behaviors. (3) The complete MDKT model achieves the best overall performance across all three datasets, demonstrating that jointly modeling multi-granularity knowledge states, temporal learning dynamics, and adversarial regularization provides complementary benefits. These results indicate that the integration of different components contributes to a more comprehensive representation of students’ evolving knowledge states. (4) The temporal dynamic encoder and adversarial training mechanism contribute differently to the overall performance of MDKT. Removing the temporal encoder (MDKT w/o T) results in noticeable performance degradation, particularly on the Junyi dataset, where the AUC decreases from 0.9091 to 0.8897. This demonstrates the importance of capturing temporal learning patterns, including forgetting and retention effects. In comparison, removing the adversarial training mechanism (MDKT w/o AT) leads to relatively smaller performance changes. For example, the AUC decreases from 0.8126 to 0.8101 on AS2009 and from 0.9091 to 0.9078 on Junyi. Although the improvement introduced by adversarial training is modest, it provides additional regularization and improves the robustness of learned representations against potential perturbations in student interaction data.

Table 5 Ablation study on the MDKT model on the AS2009 dataset. Best results are in bold.

Model	Q	C	O	ACC	AUC	MAE	RMSE
MDKT w/o C,O	✓			0.7240	0.7681	0.3476	0.4295
MDKT w/o Q,O		✓		0.7216	0.7732	0.3444	0.4250
MDKT w/o Q,C			✓	0.7259	0.7679	0.3436	0.4296
MDKT w/o O	✓	✓		0.7479	0.8089	0.3196	0.4090
MDKT w/o C	✓		✓	0.7462	0.8064	0.3172	0.4106
MDKT w/o Q		✓	✓	0.7317	0.7766	0.3439	0.4250
MDKT w/o T	✓	✓	✓	0.7509	0.8099	0.3189	0.4075
MDKT w/o AT	✓	✓	✓	0.7512	0.8101	0.3171	0.4066
MDKT	✓	✓	✓	0.7520	0.8126	0.3169	0.4063

Table 6 Ablation study on the MDKT model on the Junyi dataset. Best results are in bold.

Model	Q	C	O	ACC	AUC	MAE	RMSE
MDKT w/o C,O	✓			0.8383	0.9001	0.2240	0.3405
MDKT w/o Q,O		✓		0.8411	0.9027	0.2178	0.3372
MDKT w/o Q,C			✓	0.8247	0.8837	0.2376	0.3554
MDKT w/o O	✓	✓		0.8474	0.9077	0.2132	0.3317
MDKT w/o C	✓		✓	0.8473	0.9076	0.2099	0.3318
MDKT w/o Q		✓	✓	0.8423	0.9021	0.2164	0.3379
MDKT w/o T	✓	✓	✓	0.8333	0.8897	0.2306	0.3472
MDKT w/o AT	✓	✓	✓	0.8475	0.9078	0.2115	0.3320
MDKT	✓	✓	✓	0.8482	0.9091	0.2080	0.3305

4.2.3 Robustness Analysis (RQ3)

To further evaluate the robustness of MDKT, we conduct extensive experiments under two realistic and challenging settings: (1) varying lengths of student interaction sequences and (2) different proportions of training data. These settings closely simulate real-world educational scenarios where students may have short or long learning histories and training data may be limited.

1) Robustness under varying sequence lengths.

We evaluate the performance of MDKT and several strong baselines under different interaction sequence lengths (50, 100, 150, 200, and 250) on the AS2009, BD2006, and Junyi datasets. The results are shown in [Figure 4](#). From the figure, we can draw the following observations: (1) MDKT consistently achieves the highest ACC and AUC across all sequence lengths on the three datasets, demonstrating superior predictive performance and strong stability. Notably, its performance remains robust even with very short sequences (length = 50). (2) On the AS2009 and BD2006 datasets, MDKT shows a clear upward trend in both ACC and AUC as the sequence length increases, indicating that the model can effectively leverage longer interaction histories to improve prediction accuracy. On the Junyi dataset, MDKT maintains exceptionally

Table 7 Ablation study on the MDKT model on the BD2006 dataset. Best results are in bold.

Model	Q	C	O	ACC	AUC	MAE	RMSE
MDKT w/o C,O	✓			0.8116	0.7393	0.2682	0.3701
MDKT w/o Q,O		✓		0.8070	0.7233	0.2861	0.3749
MDKT w/o Q,C			✓	0.8094	0.7308	0.2785	0.3724
MDKT w/o O	✓	✓		0.8117	0.7465	0.2650	0.3688
MDKT w/o C	✓		✓	0.8065	0.7320	0.2762	0.3725
MDKT w/o Q		✓	✓	0.8116	0.7427	0.2693	0.3694
MDKT w/o T	✓	✓	✓	0.8104	0.7409	0.2725	0.3702
MDKT w/o AT	✓	✓	✓	0.8122	0.7462	0.2708	0.3683
MDKT	✓	✓	✓	0.8124	0.7476	0.2739	0.3681

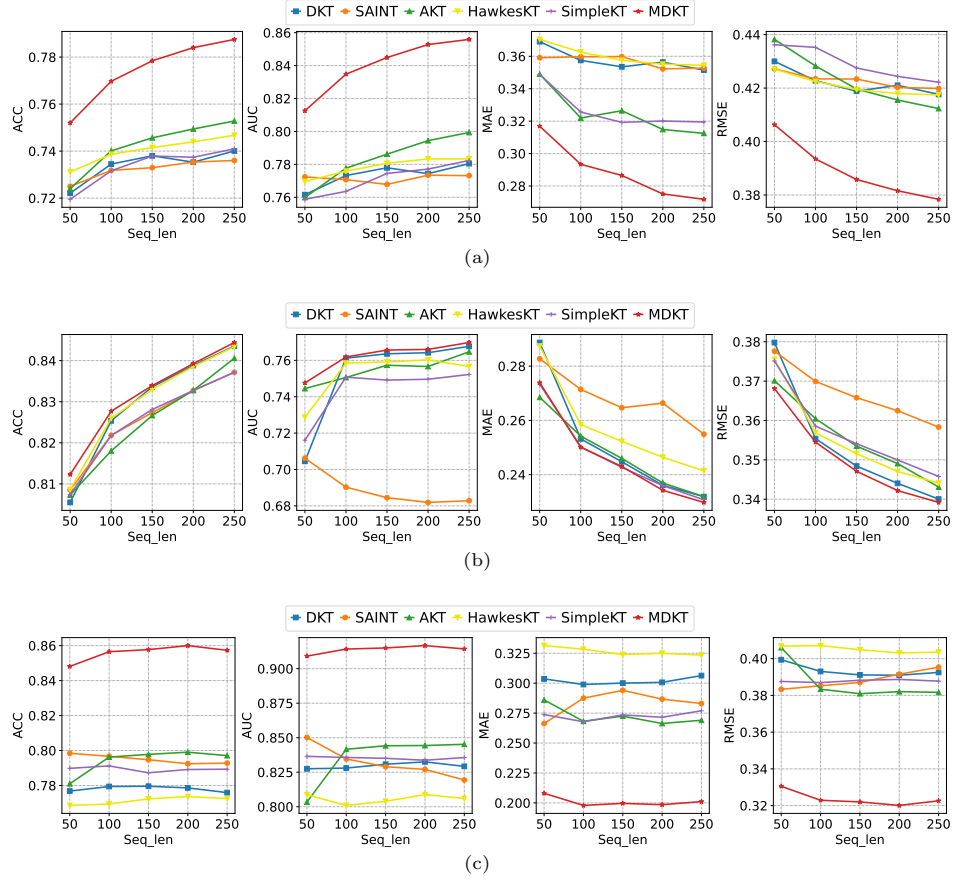


Fig. 4 Performance under varying interaction sequence lengths on (a) AS2009, (b) BD2006, and (c) Junyi

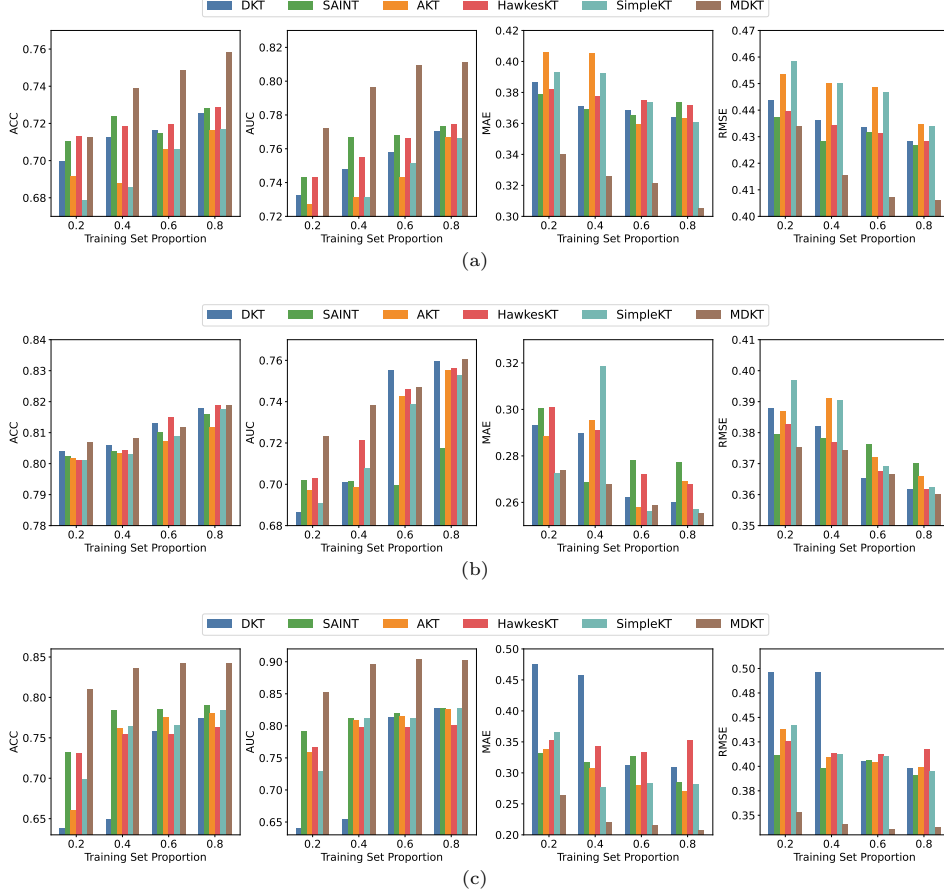


Fig. 5 Performance under varying training-set proportions on (a) AS2009, (b) BD2006, and (c) Junyi

high and stable performance across all lengths. (3) In contrast, several baseline models exhibit performance saturation or fluctuations. For example, DKT and SimpleKT show limited improvement or even slight declines as the sequence becomes longer, while HawkesKT and SAINT also display more modest gains compared to MDKT. MDKT maintains a consistent and substantial lead over all baselines, particularly in longer sequences. These results highlight MDKT’s excellent robustness to varying sequence lengths, making it well-suited for real-world scenarios where students’ interaction histories differ significantly in length.

2) Robustness under varying training set proportions. In real educational systems, data sparsity is a common issue, especially for new or inactive students. Therefore, a robust KT model should generalize well even with limited training data. We evaluated all models on AS2009, BD2006, and Junyi with training set proportions ranging from 0.2 to 0.8. The results are visualized in Figure 5. We summarize the following observations: (1) As expected, all models benefit from larger training sets,

with performance generally improving as the training proportion increases. (2) On the Junyi dataset, DKT performs poorly at low training proportions (0.2 and 0.4) but shows notable improvement at 0.6 and beyond. This suggests DKT struggles in sparse-data settings and is more sensitive to the data availability. (3) MDKT outperforms all baselines across all training proportions and datasets. This demonstrates its strong ability to infer student knowledge states even when data is limited. (4) Comparing performance gaps between MDKT and other models at different training sizes reveals its superior generalization. For instance, on AS2009 with only 20% training data, MDKT achieves an AUC more than 10% higher than SimpleKT, while the gap narrows to around 3% at 80%. This highlights MDKT’s advantage in data-scarce scenarios.

In summary, the experimental results clearly demonstrate that MDKT possesses strong robustness under both varying sequence lengths and limited training data conditions. Across all datasets and settings, MDKT consistently maintains or improves its predictive performance without exhibiting instability, degradation, or overfitting. This indicates that our model is highly resilient to variations in student interaction history and can effectively generalize from sparse data.

4.2.4 Visual Analysis (RQ4)

Visualization provides an intuitive way to understand how knowledge tracing models capture students’ learning dynamics. To better illustrate the advantages of MDKT, we randomly selected four representative students (S1 to S4) from the Junyi dataset and compared the knowledge state tracking results of AKT and MDKT. As shown in Figure 6, the upper part of each subfigure displays the models’ predicted mastery probabilities for different knowledge concepts using a color matrix, where darker colors indicate higher predicted probabilities. The lower part presents the student’s actual interaction records, including answers, accuracy, involved concepts, questions, answer times, and interval times.

The visualization clearly reveals MDKT’s superior ability in modeling the dynamic evolution of students’ knowledge states. For student S1, who answered multiple questions related to concept c5, AKT’s predicted probabilities remained stagnant at around 0.31, showing almost no sign of learning progress. In contrast, MDKT’s predictions steadily increased from 0.28 to 0.96, effectively capturing the student’s gradual mastery of the concept through repeated practice. The advantage becomes even more evident during concept transitions. Taking student S2 as an example, when the student switched from concept c5 to c1 and c0, AKT exhibited sharp fluctuations in predicted probabilities (e.g., dropping to only 0.53 for c0). MDKT, however, adapted more smoothly to the new concepts and quickly raised its prediction confidence above 0.90 in subsequent interactions, demonstrating stronger adaptability and generalization. Furthermore, in student S3, when facing the more difficult concept c329, the predicted probability of AKT fluctuated dramatically, whereas MDKT’s predictions rose steadily from 0.12 to 0.93 as the student’s correct rate improved. For student S4, AKT displayed significant instability across different concepts and time steps: its predictions started high at 0.96, then dropped sharply to 0.49 and even 0.36 in the middle of the sequence, before rising again to 0.95. In comparison, MDKT maintained more

consistent predictions throughout, reaching stable high values (0.94–0.96) in the later steps, particularly when the student answered correctly on concepts c2 and c3.

Overall, these representative examples suggest that MDKT, through its multi-dimensional knowledge state modeling and adaptive temporal encoding, is able to capture students’ knowledge evolution more smoothly than the baseline model while remaining consistent with the observed learning process. Such visualizations provide intuitive insights into how the proposed model models concept-level mastery evolution and may assist teachers in understanding students’ learning progress and identifying potential learning difficulties.

It should be noted that the analysis in Figure 6 is qualitative in nature. Since existing public knowledge tracing benchmarks do not provide expert annotations of students’ latent concept mastery, quantitative evaluation of interpretability (e.g., alignment between model predictions and expert judgments) cannot be directly conducted within the current benchmark setting. Developing standardized quantitative interpretability evaluation protocols for knowledge tracing remains an important direction for future research.

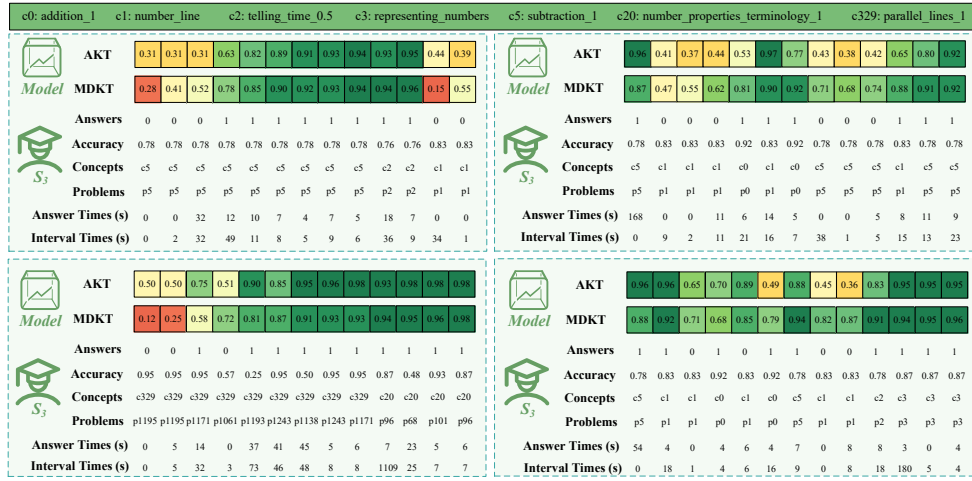


Fig. 6 AKT and MDKT prediction trajectories and interaction records for four students in the Junyi dataset

5 Conclusion

In this paper, we presented MDKT, a novel knowledge tracing model that addresses two key limitations in existing approaches: single-dimensional knowledge representation and insufficient robustness under real-world conditions. MDKT advances the field by modeling students’ knowledge states at three complementary levels, including question-level, concept-level, and overall-level representations, thereby capturing a more comprehensive view of learning dynamics. To better account for temporal

evolution, we designed an adaptive dual-phase learning state encoder that effectively models both short-term learning reinforcement and long-term forgetting effects. Furthermore, we introduced an adversarial smoothing mechanism to improve the stability and generalization capability of the learned knowledge representations, particularly under sparse and noisy interaction scenarios. We extensively evaluated MDKT on three large-scale real-world datasets. The experimental results demonstrate that MDKT consistently outperforms state-of-the-art baselines in terms of prediction accuracy (ACC and AUC) and error metrics (MAE and RMSE). Moreover, additional analyses under varying sequence lengths and different training data proportions confirm that MDKT maintains strong robustness and generalization ability under challenging learning scenarios.

Despite these promising results, several limitations remain. First, although MDKT achieves competitive performance on the benchmark datasets used in this study, its scalability to extremely large-scale educational platforms involving millions of students and massive interaction records requires further investigation. The current architecture may introduce additional computational overhead due to the joint modeling of multi-dimensional knowledge states and temporal dynamics, and future work will explore more efficient architectures for large-scale deployment. Second, similar to most existing knowledge tracing approaches, MDKT relies on historical interaction records to infer student knowledge states. Therefore, its performance may be limited in cold-start scenarios, such as newly enrolled students with insufficient learning histories or newly introduced questions without adequate response records. Incorporating auxiliary information, such as learner profiles, curriculum structures, or question metadata, may help alleviate this limitation. Third, although MDKT can handle questions associated with multiple knowledge concepts through a preprocessing strategy that decomposes a multi-concept question into multiple concept-specific interaction records, the current framework does not explicitly model the inherent correlations among multiple concepts within a single learning interaction. Developing an end-to-end multi-concept knowledge tracing mechanism remains an important direction for future research. Finally, while MDKT has been extensively validated on public benchmark datasets, further investigation is required to evaluate its effectiveness, efficiency, and interpretability in real-world intelligent tutoring systems and adaptive learning platforms.

Looking ahead, we identify several promising directions for future research. On one hand, incorporating higher-level curriculum structures, domain-specific educational priors, and explicit multi-concept relationship modeling may further enhance the interpretability and pedagogical value of knowledge tracing models. On the other hand, extending MDKT to multimodal learning environments by integrating interaction logs with additional educational signals, such as eye-tracking, video recordings, or physiological measurements, provides an opportunity to construct more comprehensive student models. Furthermore, improving computational efficiency and conducting large-scale online deployment studies will be essential steps toward applying MDKT in practical educational systems.

References

- Abdelrahman G, Wang Q (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. Association for Computing Machinery, New York, NY, USA, SIGIR'19, p 175–184, <https://doi.org/10.1145/3331184.3331195>
- Choi Y, Lee Y, Cho J, et al (2020) Towards an appropriate query, key, and value computation for knowledge tracing. In: Proceedings of the Seventh ACM Conference on Learning @ Scale. Association for Computing Machinery, New York, NY, USA, L@S '20, p 341–344, <https://doi.org/10.1145/3386527.3405945>
- Corbett AT, Anderson JR (1994) Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction* 4(4):253–278. <https://doi.org/10.1007/BF01099821>
- Dong X, Zhu Y, Zhang Y, et al (2020) Leveraging adversarial training in self-learning for cross-lingual text classification. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. Association for Computing Machinery, New York, NY, USA, SIGIR '20, p 1541–1544, <https://doi.org/10.1145/3397271.3401209>
- Ghosh A, Heffernan N, Lan AS (2020) Context-aware attentive knowledge tracing. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Association for Computing Machinery, New York, NY, USA, KDD '20, p 2330–2339, <https://doi.org/10.1145/3394486.3403282>
- Goodfellow IJ, Shlens J, Szegedy C (2015) Explaining and harnessing adversarial examples. URL <https://arxiv.org/abs/1412.6572>, [arXiv:1412.6572](https://arxiv.org/abs/1412.6572)
- Šarić Grgić I, Grubišić A, Gašpar A (2024) Twenty-five years of bayesian knowledge tracing: a systematic review. *User Modeling and User-Adapted Interaction* 34(4):1127–1173. <https://doi.org/10.1007/s11257-023-09389-4>
- Guo X, Huang Z, Gao J, et al (2021) Enhancing knowledge tracing via adversarial training. In: Proceedings of the 29th ACM International Conference on Multimedia. Association for Computing Machinery, New York, NY, USA, MM '21, p 367–375, <https://doi.org/10.1145/3474085.3475554>
- He L, Li X, Wang P, et al (2023) Man: Memory-augmented attentive networks for deep learning-based knowledge tracing. *ACM Trans Inf Syst* 42(1). <https://doi.org/10.1145/3589340>
- Huang CQ, Huang QH, Huang X, et al (2024) Xkt: Toward explainable knowledge tracing model with cognitive learning theories for questions of multiple knowledge concepts. *IEEE Trans on Knowl and Data Eng* 36(11):7308–7325. <https://doi.org/>

- Huang S, Liu Z, Zhao X, et al (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. Association for Computing Machinery, New York, NY, USA, SIGIR '23, p 2441–2445, <https://doi.org/10.1145/3539618.3592073>
- Im Y, Choi E, Kook H, et al (2023) Forgetting-aware linear bias for attentive knowledge tracing. In: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. Association for Computing Machinery, New York, NY, USA, CIKM '23, p 3958–3962, <https://doi.org/10.1145/3583780.3615191>
- Lee J, Yeung DY (2019) Knowledge query network for knowledge tracing: How knowledge interacts with skills. In: Proceedings of the 9th International Conference on Learning Analytics & Knowledge. Association for Computing Machinery, New York, NY, USA, LAK19, p 491–500, <https://doi.org/10.1145/3303772.3303786>
- Lee W, Chun J, Lee Y, et al (2022) Contrastive learning for knowledge tracing. In: Proceedings of the ACM Web Conference 2022. Association for Computing Machinery, New York, NY, USA, WWW '22, p 2330–2338, <https://doi.org/10.1145/3485447.3512105>
- Li R, Xu J, Cao Z, et al (2024a) Extending context window in large language models with segmented base adjustment for rotary position embeddings. *Applied Sciences* 14(7):3076
- Li X, Bai Y, Guo T, et al (2024b) Enhancing length generalization for attention based knowledge tracing models with linear biases. In: Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI '24, <https://doi.org/10.24963/ijcai.2024/654>
- Liu F, Bu C, Zhang H, et al (2024) Fdkt: Towards an interpretable deep knowledge tracing via fuzzy reasoning. *ACM Trans Inf Syst* 42(5). <https://doi.org/10.1145/3656167>
- Liu Y, Ji W, Zhou G (2026) Cross-scale knowledge tracking: A joint time-frequency modeling approach for global cycles and local fluctuations. *Journal of King Saud University Computer and Information Sciences* 38:179. <https://doi.org/10.1007/s44443-026-00600-1>
- Liu Z, Liu Q, Chen J, et al (2022) pykt: a python library to benchmark deep learning based knowledge tracing models. *Advances in Neural Information Processing Systems* 35:18542–18555
- Liu Z, Liu Q, Chen J, et al (2023) simplekt: A simple but tough-to-beat baseline for knowledge tracing. URL <https://arxiv.org/abs/2302.06881>, [arXiv:2302.06881](https://arxiv.org/abs/2302.06881)

- Ma F, Jiang Y, Li H, et al (2026) MSBR-KT: knowledge tracing via multifaceted structured-bias routing. *Journal of King Saud University Computer and Information Sciences* 38:631. <https://doi.org/10.1007/s44443-026-01029-2>
- Ma H, Yang Y, Qin C, et al (2024) Hd-kt: Advancing robust knowledge tracing via anomalous learning interaction detection. In: *Proceedings of the ACM Web Conference 2024*. Association for Computing Machinery, New York, NY, USA, WWW '24, p 4479–4488, <https://doi.org/10.1145/3589334.3645718>
- Miyato T, Dai AM, Goodfellow I (2021) Adversarial training methods for semi-supervised text classification. URL <https://arxiv.org/abs/1605.07725>, [arXiv:1605.07725](https://arxiv.org/abs/1605.07725)
- Nagatani K, Zhang Q, Sato M, et al (2019) Augmenting knowledge tracing by considering forgetting behavior. In: *The World Wide Web Conference*. Association for Computing Machinery, New York, NY, USA, WWW '19, p 3101–3107, <https://doi.org/10.1145/3308558.3313565>
- Nakagawa H, Iwasawa Y, Matsuo Y (2019) Graph-based knowledge tracing: Modeling student proficiency using graph neural network. In: *IEEE/WIC/ACM International Conference on Web Intelligence*. Association for Computing Machinery, New York, NY, USA, WI '19, p 156–163, <https://doi.org/10.1145/3350546.3352513>
- Pandey S, Karypis G (2019) A self-attentive model for knowledge tracing. URL <https://arxiv.org/abs/1907.06837>, [arXiv:1907.06837](https://arxiv.org/abs/1907.06837)
- Piech C, Bassen J, Huang J, et al (2015) Deep knowledge tracing. In: *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 1*. MIT Press, Cambridge, MA, USA, NIPS'15, p 505–513
- Pu Y, Liu F, Shi R, et al (2024) Elakt: Enhancing locality for attentive knowledge tracing. *ACM Trans Inf Syst* 42(4). <https://doi.org/10.1145/3652601>
- Sato M, Suzuki J, Kiyono S (2019) Effective adversarial regularization for neural machine translation. In: Korhonen A, Traum D, Màrquez L (eds) *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, pp 204–210, <https://doi.org/10.18653/v1/P19-1020>
- Shen S, Liu Q, Chen E, et al (2021) Learning process-consistent knowledge tracing. In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. Association for Computing Machinery, New York, NY, USA, KDD '21, p 1452–1460, <https://doi.org/10.1145/3447548.3467237>
- Shen S, Chen E, Liu Q, et al (2023) Monitoring student progress for learning process-consistent knowledge tracing. *IEEE Transactions on Knowledge and Data Engineering* 35(8):8213–8227. <https://doi.org/10.1109/TKDE.2022.3221985>

- Shen X, Yu F, Liu Y, et al (2024) Revisiting knowledge tracing: A simple and powerful model. In: Proceedings of the 32nd ACM International Conference on Multimedia. Association for Computing Machinery, New York, NY, USA, MM '24, p 263–272, <https://doi.org/10.1145/3664647.3681205>
- Stamper J, Niculescu-Mizil A, Ritter S, et al (2010) Algebra i 2005-2006 and bridge to algebra 2006-2007. Development Data Sets from KDD Cup
- Sun J, Yu F, Wan Q, et al (2024) Interpretable knowledge tracing with multiscale state representation. In: Proceedings of the ACM Web Conference 2024. Association for Computing Machinery, New York, NY, USA, WWW '24, p 3265–3276, <https://doi.org/10.1145/3589334.3645373>
- Tong S, Liu Q, Huang W, et al (2020) Structure-based knowledge tracing: An influence propagation view. In: 2020 IEEE International Conference on Data Mining (ICDM), pp 541–550, <https://doi.org/10.1109/ICDM50108.2020.00063>
- Vie JJ, Kashima H (2019) Knowledge tracing machines: factorization machines for knowledge tracing. In: Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence. AAAI Press, AAAI'19/IAAI'19/EAAI'19, <https://doi.org/10.1609/aaai.v33i01.3301750>
- Wang C, Ma W, Zhang M, et al (2021) Temporal cross-effects in knowledge tracing. In: Proceedings of the 14th ACM International Conference on Web Search and Data Mining. Association for Computing Machinery, New York, NY, USA, WSDM '21, p 517–525, <https://doi.org/10.1145/3437963.3441802>
- Wang J, Lu J, Su Z (2026a) Swdakt: Knowledge tracing using sliding window-based dynamic ability perception. Expert Systems with Applications 303:130735. <https://doi.org/10.1016/j.eswa.2025.130735>
- Wang T, Zhang Z, Qin X (2026b) A knowledge tracing model based on heterogeneous exercise graphs and multi-source behavioral feature fusion in blended learning environments. Journal of King Saud University Computer and Information Sciences 38:653. <https://doi.org/10.1007/s44443-026-01078-7>
- Wu Y, Bamman D, Russell S (2017) Adversarial training for relation extraction. In: Palmer M, Hwa R, Riedel S (eds) Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Copenhagen, Denmark, pp 1778–1783, <https://doi.org/10.18653/v1/D17-1187>
- Yang Y, Shen J, Qu Y, et al (2021) Gikt: a graph-based interaction model for knowledge tracing. In: Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18,

- 2020, Proceedings, Part I, Springer. Springer, pp 299–315, https://doi.org/10.1007/978-3-030-67658-2_18
- Yeung CK (2019) Deep-irt: Make deep learning based knowledge tracing explainable using item response theory. URL <https://arxiv.org/abs/1904.11738>, [arXiv:1904.11738](https://arxiv.org/abs/1904.11738)
- Yeung CK, Yeung DY (2018) Addressing two problems in deep knowledge tracing via prediction-consistent regularization. In: Proceedings of the Fifth Annual ACM Conference on Learning at Scale. Association for Computing Machinery, New York, NY, USA, L@S '18, <https://doi.org/10.1145/3231644.3231647>
- Yin Y, Dai L, Huang Z, et al (2023) Tracing knowledge instead of patterns: Stable knowledge tracing with diagnostic transformer. In: Proceedings of the ACM Web Conference 2023. Association for Computing Machinery, New York, NY, USA, WWW '23, p 855–864, <https://doi.org/10.1145/3543507.3583255>
- Yu G, Zhao Z, Zhou G, et al (2026) Attributes enhanced representation learning with large language model for cold-start knowledge tracing. Expert Systems with Applications 319:131985. <https://doi.org/10.1016/j.eswa.2026.131985>
- Zhang J, Shi X, King I, et al (2017) Dynamic key-value memory networks for knowledge tracing. In: Proceedings of the 26th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, WWW '17, p 765–774, <https://doi.org/10.1145/3038912.3052580>
- Zhang Z, Li J, Liu H, et al (2026) Drkt: Learning differential relationships for efficient knowledge tracing with learner’s knowledge internalization representation. Expert Systems with Applications 310:131319. <https://doi.org/10.1016/j.eswa.2026.131319>

Statements and Declarations

Data availability

The datasets analyzed in this study are publicly available from the following repositories: the ASSISTments 2009 dataset (<https://sites.google.com/site/assistmentsdata/home/2009-2010-assistment-data/skill-builder-data-2009-2010>), the Bridge to Algebra 2006–2007 dataset from the KDD Cup 2010 (<https://pslccdashop.web.cmu.edu/KDDCup>), and the Junyi Academy dataset (<https://pslccdashop.web.cmu.edu/DatasetInfo?datasetId=1198>). No new datasets were generated during the current study.