

Article

Robust English Knowledge Tracing via Profile-Driven Forgetting and Masked Consistency

Xibo Chen ¹, Ziqi Zhang ^{2,*}, Haize Hu ², Jie Jin ³ , Fei Yu ⁴  and Lv Zhao ⁵¹ College of Foreign Languages, Hunan Institute of Engineering, Xiangtan 411104, China² School of Computer Science and Technology, Guangxi Normal University, Guilin 541001, China³ School of Information and Electrical Engineering, Hunan University of Science and Technology, Xiangtan 411201, China⁴ School of Physics and Electronic Science, Changsha University of Science and Technology, Changsha 410114, China⁵ School of Electrical and Information Engineering, Hunan Institute of Technology, Hengyang 421002, China

* Correspondence: ziqizhang@stu.gxnu.edu.cn

Abstract

Knowledge Tracing (KT) plays a pivotal role in Intelligent Tutoring Systems (ITS) by dynamically assessing learners' evolving knowledge states. However, tracking the acquisition of English presents unique challenges. Existing KT models typically employ homogeneous, predefined forgetting mechanisms that fail to capture the highly individualized nature of linguistic memory retention. Furthermore, language assessment data is notoriously noisy, which leads models to overfit superficial performance rather than capturing true underlying linguistic competence. To address these issues, we propose a novel framework to robustly trace English language competence. First, we introduce a Learning-Profile-Driven Adaptive Forgetting mechanism. Unlike methods with shared forgetting rates, our approach constructs a dynamic and strictly causal profile from historical interactions to generate personalized cognitive parameters (e.g., individualized forgetting rates). These parameters synchronously modulate the decay of multi-level knowledge states, enabling the model to accurately capture the heterogeneous memory retention patterns of different learners. Second, we design a Masked Consistency Regularization training paradigm. By applying stochastic masking to historical responses and enforcing predictive consistency, we prevent the model from exploiting localized noise and “shortcut” learning, compelling it to mine robust and invariant language representations. Extensive experiments on real-world educational datasets demonstrate that our proposed framework significantly outperforms state-of-the-art baselines in both prediction accuracy and noise resistance, offering a robust and interpretable solution for personalized language learning.

Keywords: knowledge tracing; adaptive forgetting; robustness; language learning

Academic Editor: Christos Bouras

Received: 25 May 2026

Revised: 21 June 2026

Accepted: 23 June 2026

Published: 26 June 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

With the rapid development of science and technology, researchers have conducted extensive studies in areas such as artificial intelligence [1–3], deep learning [4–6], data center networks [7–9], neural networks [10–12], and intelligent tutoring systems (ITS) [13–15]. Among them, ITSs are a computer-assisted tool that uses artificial intelligence technology to imitate the teaching behaviors of human teachers. Its purpose is to provide personalized guidance and feedback by analyzing students' learning data [16]. ITSs increasingly rely on Knowledge Tracing (KT) to estimate the evolving knowledge states of learners from

sequential learning interactions [17]. Early neural KT models, such as Deep Knowledge Tracing (DKT), demonstrated the effectiveness of recurrent neural networks for modeling latent student knowledge from interaction sequences [18], while subsequent memory- and attention-based models further improved the representation of concept-level mastery and long-range dependencies [19–21]. While KT has been widely studied in general educational settings, tracking the process of English learning remains particularly challenging. In English vocabulary and grammar learning, the observable performance of a learner is shaped not only by accumulated knowledge but also by long-term memory retention, linguistic background, transfer ability, and noisy response behaviors. Prior studies in language learning further suggest that morphological awareness and vocabulary knowledge are closely associated across monolingual and bilingual learners, indicating that English competence is influenced by learner-specific linguistic foundations rather than isolated item responses alone [22,23]. Therefore, accurately tracing English proficiency requires models that go beyond isolated right-or-wrong interactions and capture more stable, individualized learning dynamics.

Despite recent progress, existing KT models still face two primary limitations when applied to the tracing of English competence. First, forgetting behavior is often modeled implicitly or through the assumption of shared temporal decay. Although forgetting-aware KT methods have introduced temporal factors into student modeling [24], recent studies show that attention-based KT models may still struggle to capture diverse and continuous forgetting behaviors in long interaction sequences [25]. However, second-language acquisition exhibits strong heterogeneity among learners. For example, students with stronger morphological awareness or richer prior linguistic knowledge may retain newly learned words or grammatical patterns more effectively than students with weaker foundations [22]. A homogeneous forgetting mechanism is therefore insufficient to capture individualized long-term retention patterns in English learning. Second, English assessment data are typically noisy. Multiple-choice questions may contain random guessing, while spelling or grammar exercises may include careless mistakes or typographical errors. Recent robust KT studies have explicitly recognized that error-prone educational data may contain random factors that deviate from learners' true cognitive patterns [26]. Such discrepancies between observable performance and underlying language competence can cause KT models to overfit local response fluctuations, leading to unstable estimations of knowledge states and reduced generalization.

To address these challenges, we propose a robust framework for English knowledge tracing that integrates profile-driven adaptive forgetting with masked consistency regularization. Specifically, the model first constructs a strictly causal dynamic profile of the learner from historical English learning interactions. This profile is utilized as a contextual prior to generate personalized forgetting parameters, including a learner-specific forgetting rate and a minimum retention bound. Crucially, these personalized physiological parameters are synchronously applied to modulate the decay of multi-level knowledge states based on their distinct exposure intervals. By doing so, the model can adaptively characterize heterogeneous patterns of multi-grained memory decay across different learners and learning histories. In addition, we introduce an attribute-masked consistency regularization strategy to improve robustness against noisy signals and item-level overfitting. Consistency regularization has been widely used to encourage stable predictions under input perturbations or data augmentation [27,28]. During training, specific historical problem identifiers are stochastically masked while their corresponding correctness labels are strictly retained to construct an augmented view of the interaction sequence. The model is then encouraged to produce consistent predictions on the unmasked future steps between the original and augmented views. This consistency constraint reduces the model's over-reliance on item-

specific co-occurrence patterns and promotes more stable representations of latent English proficiency derived from broader correctness trajectories.

To overcome the limitation of homogeneous forgetting assumptions, we propose a novel Learning-Profile-Driven Adaptive Forgetting mechanism. Unlike prior methods that apply a shared decay rate to all learners, our approach constructs a dynamic and strictly causal profile for each student from their historical interactions. This profile serves as a contextual prior to generate personalized cognitive parameters, such as an individualized forgetting rate and a minimum retention bound. These parameters are then used to synchronously modulate the decay of knowledge states at global, skill, and problem levels, allowing the model to accurately capture the heterogeneous memory retention patterns unique to each English learner.

Furthermore, to ensure stable and effective optimization during the training phase, we introduce a novel Masked Consistency Normalization training paradigm. Unlike standard approaches that may suffer from distribution shifts between masked and visible regions, our paradigm enforces feature alignment and prediction consistency across different masking ratios. This unified normalization mechanism provides robust self-supervision signals, effectively bridging the gap between partial observations and complete representations. The effectiveness of this strategy in preventing shortcut learning is empirically validated in our ablation studies.

The proposed framework combines individualized cognitive modeling and noise-resistant representation learning within a unified KT architecture. The adaptive forgetting module enables the model to better capture long-term English learning trajectories, while the masked consistency objective regularizes the prediction process under incomplete and noisy performance observations. Extensive experiments on benchmark educational datasets demonstrate that the proposed method achieves consistent improvements over competitive baselines, especially in terms of predictive accuracy and robustness. These results suggest that explicitly modeling personalized forgetting and reducing sensitivity to noisy responses are both crucial for reliable English knowledge tracing.

Before detailing our specific contributions, it is important to clarify the scope of this study to ensure the alignment between our theoretical framework and empirical validation. While the primary motivation of this study stems from the challenges in English Knowledge Tracing (KT)—specifically the heterogeneity in vocabulary retention and grammar acquisition—the proposed model is designed as a general cognitive framework applicable to broader educational contexts. The core mechanisms we introduce, such as [mention your specific module, e.g., adaptive forgetting or noise regularization], address fundamental cognitive bottlenecks that are domain-agnostic. Therefore, to rigorously validate the effectiveness of these underlying mechanisms, we utilize widely recognized general educational benchmarks (ASSISTments and Algebra2005) in this study, serving as a robust foundation for future domain-specific applications.

The main contributions of this work are summarized as follows:

(1) **Profile-Driven Multi-Level Adaptive Forgetting.** We propose a learner-profile-driven forgetting mechanism that dynamically generates individualized memory decay parameters from strictly causal historical interactions. These parameters synchronously govern the temporal decay of multi-level knowledge states, enabling more accurate modeling of heterogeneous retention patterns in English learning.

(2) **Masked Consistency Regularization.** We introduce a consistency regularization strategy that performs stochastic attribute masking. By aligning predictions on unmasked steps between the original and augmented views, this strategy improves the model's robustness against item-level overfitting and noisy assessment behaviors.

(3) Robust English Knowledge Tracing Framework. We integrate adaptive forgetting and masked consistency learning into a unified KT framework for English competence tracing, achieving more stable and generalizable proficiency estimation than conventional behavior-fitting approaches.

Extensive experiments on benchmark datasets demonstrate that our framework significantly outperforms state-of-the-art baselines in prediction accuracy and robustness. Detailed results and analyses are presented in Section 4.

2. Related Work

2.1. Deep Knowledge Tracing

The goal of Knowledge Tracing (KT) is to estimate the latent knowledge states of learners from their sequential interaction histories. Early approaches, such as Bayesian Knowledge Tracing (BKT) [29], used Hidden Markov Models to track the binary mastery states of individual skills. The field was revolutionized by Deep Knowledge Tracing (DKT) [18], which applied Recurrent Neural Networks (RNNs) to capture non-linear representations of student learning trajectories across multiple skills simultaneously.

Since DKT, various neural architectures have been proposed to improve representational capacity. Dynamic Key-Value Memory Networks (DKVMN) [19] partition the memory matrix to independently track underlying concepts and evolving student mastery. Attention- and Transformer-based models have further advanced KT by capturing long-range dependencies in interaction sequences. For example, AKT [20] introduced a Rasch-model-inspired embedding strategy to account for question difficulty, while DTransformer [30] uses a diagnostic Transformer architecture to trace knowledge states rather than superficial interaction patterns. More recently, simpleKT [31] demonstrated that a lightweight attention-based model with question-specific variation can compete with more complex KT architectures. Additionally, stableKT [32] addressed the length generalization problem in attention-based KT models, improving long-sequence prediction through linear attention biases.

While these deep learning-based approaches have significantly improved prediction performance by capturing complex non-linear relationships, they often treat the learning process as a purely data-driven sequence modeling task, largely overlooking the cognitive mechanism of human memory. Specifically, standard RNNs or attention mechanisms do not explicitly model the “forgetting” behavior inherent in human learning, assuming implicitly that knowledge states persist unless updated. This contradicts our thesis that accurate knowledge tracing in language learning must account for the temporal decay of memory. Therefore, simply stacking deep layers is insufficient; a mechanism that integrates cognitive forgetting laws into the deep learning framework is essential for modeling long-term dependencies in student learning behaviors.

2.2. Modeling Forgetting in Knowledge Tracing

Forgetting is a fundamental aspect of learning, particularly in language acquisition. Early attempts to model forgetting in educational data mining incorporated discrete time-gap features or Ebbinghaus curves into regression models, such as the Half-Life Regression (HLR) used for vocabulary retention [33]. In the deep learning era, researchers have integrated temporal contexts into neural KT architectures. HawkesKT [34] used Hawkes processes to model the mutual excitation and continuous temporal decay of historical interactions. Learning Process-consistent KT (LPKT) [21] explicitly models learning gains and forgetting effects, while FoLiBiKT [35] introduces a forgetting-aware linear bias into attention-based KT models.

Recent studies have highlighted the limitations of fixed-window and homogeneous temporal modeling. stableKT [32] and related methods show that attention-based KT models suffer degraded generalization when test sequences exceed the training context length. LefoKT [25] rethinks learning and forgetting processes in attention-based KT, proposing relative forgetting attention to decouple forgetting patterns from problem relevance. Furthermore, the Cognitive Fluctuations Enhanced Attention Network (CFAKT) [36] emphasizes that learners' cognitive states fluctuate dynamically rather than evolving monotonically, which is especially relevant for noisy educational sequences.

Despite these advances, existing temporal KT models typically rely on globally shared decay parameters, fixed temporal biases, or predefined mathematical formulations [35,37,38]. This homogeneous assumption fails to account for the intrinsic heterogeneity of second-language learners, who have diverse prior knowledge, linguistic backgrounds, and retention patterns. To address this limitation, our framework introduces a profile-driven multi-level adaptive forgetting mechanism. Instead of applying a universal decay rate, we dynamically generate individualized physiological parameters from causal profiles and synchronously apply them across global, skill-level, and problem-level knowledge states to simulate personalized memory decay.

While these studies have made significant progress in incorporating temporal dynamics, they predominantly rely on homogeneous assumptions, such as globally shared decay parameters or fixed mathematical formulations (e.g., Ebbinghaus curves). This approach fails to capture the intrinsic heterogeneity of second-language learners, whose retention patterns are heavily influenced by individual linguistic backgrounds and cognitive profiles. In contrast, our work posits that a one-size-fits-all forgetting mechanism is insufficient for English learning. Our proposed profile-driven adaptive forgetting mechanism directly addresses this gap by dynamically generating personalized forgetting parameters from a learner's unique interaction history, thereby offering a more nuanced and accurate model of individualized knowledge decay.

2.3. Robustness and Regularization in KT

To improve model robustness and sequence stability, recent methods have adopted adversarial training, sparse attention, contrastive learning, ideal-state alignment, uncertainty modeling, and consistency regularization. ATKKT [39] improves generalization by injecting adversarial perturbations into interaction embeddings. SparseKT [40] enhances the robustness of attention-based models by selecting only highly relevant historical interactions through k-sparse attention. CL4KT [41] introduces contrastive learning to align semantically similar learning histories and separate dissimilar ones. AlignKT [42] models knowledge states through ideal state alignment, enhancing interpretability and robustness by encouraging learned states to conform to structured cognitive trajectories.

More recent work has shifted from generic robustness toward explicitly modeling uncertainty and error-prone educational data. UKT [43] represents uncertainty in student interactions through stochastic distribution embeddings and incorporates uncertainty-aware contrastive learning. RobustKT [26] decouples cognitive patterns from error-prone response data, addressing the mismatch between observed responses and latent knowledge mastery. These studies align with our motivation, as English learning data often contain random guessing, careless mistakes, typographical errors, and item-specific shortcuts.

Consistency regularization provides another approach to robustness. General-purpose methods such as R-Drop [44] encourage consistent output distributions under stochastic dropout perturbations. In KT, sequence augmentation and representation alignment have been used to stabilize learning-state representations [41,43]. However, conventional augmentation and masking strategies may perturb correctness labels or disrupt causal

state-transition signals. This issue is problematic in KT, where correctness is the primary observable evidence of knowledge-state changes. To address this, our framework employs an attribute-masked consistency regularization mechanism. By masking specific historical problem identifiers while preserving their correctness labels, the model is discouraged from relying on superficial item co-occurrence patterns and is instead encouraged to trace underlying learner competence.

These robust KT methods have effectively highlighted the issue of noisy data and proposed various regularization techniques. However, many existing strategies, such as adversarial training or contrastive learning, primarily focus on perturbing input embeddings or aligning sequence-level representations globally. Our thesis argues that a more direct approach is needed to prevent the model from overfitting to specific item identifiers (ID-overfitting), a common issue in sparse educational datasets. Merely adding noise to embeddings does not guarantee that the model captures the underlying concept mastery. To this end, our work introduces attribute-masked consistency regularization. By forcing the model to maintain predictive consistency even when problem identifiers are masked, we compel the model to learn from the trajectory of correctness and exercise attributes rather than memorizing item-specific patterns, thus building a more robust and generalizable representation of learner competence.

3. Our Proposed Framework

3.1. Problem Formulation and Input Representation

English knowledge tracing aims to predict future response performance based on historical sequential interactions [17,18,29]. Given an interaction sequence $\mathcal{X} = \{x_1, x_2, \dots, x_T\}$ for a learner, each interaction at time step t is represented as a tuple $x_t = (q_t, s_t, r_t)$, where q_t denotes the specific exercise, s_t denotes the corresponding linguistic skill, and $r_t \in \{0, 1\}$ denotes the response correctness. The objective is to estimate the probability that the learner will correctly answer the next exercise. Formally, given the historical interactions before time step $t + 1$, the model predicts:

$$\hat{r}_{t+1} = P(r_{t+1} = 1 \mid x_1, x_2, \dots, x_t, q_{t+1}, s_{t+1}). \quad (1)$$

To support sequential neural modeling, discrete interactions must be projected into a continuous latent space [18–20], as illustrated in Figure 1. A naive representation strategy typically sums or concatenates the exercise and skill embeddings. However, exercises associated with the same skill often exhibit varying degrees of inherent difficulty. Overlooking this intra-skill difficulty variation restricts the representational capacity of the model.

To capture this variation, we adopt a difficulty-aware representation mechanism inspired by the Rasch model in psychometrics and its use in attention-based KT models [20,45]. Specifically, the representation for a given exercise $e_t \in R^d$ is formulated as:

$$e_t = p_{q_t} + s_{s_t} + \mu_{q_t} v_{s_t}, \quad (2)$$

where $p_{q_t} \in R^d$ and $s_{s_t} \in R^d$ are the base embeddings for exercise q_t and skill s_t , respectively. The scalar $\mu_{q_t} \in R$ models the inherent difficulty bias of the exercise, and $v_{s_t} \in R^d$ serves as a skill-specific variation vector. The interaction term $\mu_{q_t} v_{s_t}$ adjusts the semantic representation of the skill based on the difficulty of the exercise.

The interaction representation $x_t \in R^d$ is then obtained by incorporating the response correctness:

$$x_t = e_t + a_{r_t}, \quad (3)$$

where $ar_t \in R^d$ denotes the embedding of the response label r_t . The sequence of continuous interaction embeddings $x_1, x_2, \dots, x_T$ is then used as the input for tracking the multi-level knowledge states of the learner.

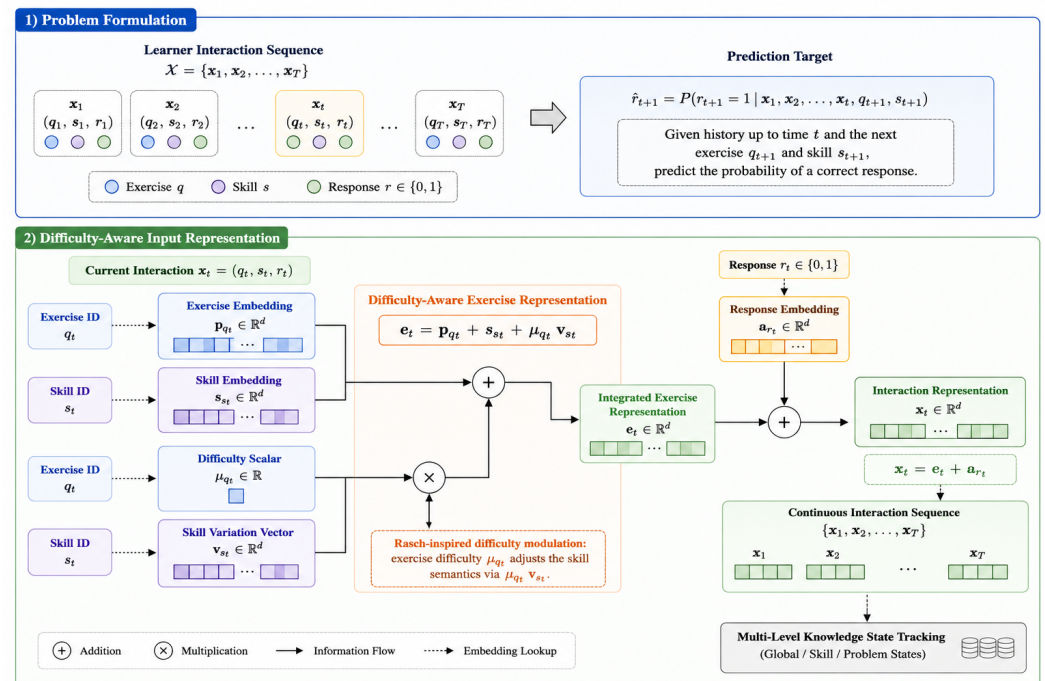


Figure 1. Illustration of the problem formulation and the difficulty-aware input representation. The framework maps discrete sequential interactions into a continuous latent space, where exercise difficulties adaptively modulate the skill semantics.

3.2. Profile-Driven Multi-Level State Tracking and Forgetting

To capture the cognitive dynamics of learners, our framework tracks knowledge states at three distinct granularities:

Global Ability State ($\mathbf{h}_t^g$): Represents the learner's generalized latent proficiency and overall learning pace across the entire domain.

Skill Mastery State ($\mathbf{h}_{t,s}^s$): Measures the explicit mastery level of specific abstract linguistic concepts or skills.

Problem Memory State ($\mathbf{h}_{t,q}^p$): Captures item-specific short-term familiarity, rote memorization, or localized impression towards exact exercises.

As illustrated in Figure 2, at each time step t , the framework retrieves the previous states, applies profile-driven cognitive decay, uses them for the current prediction, and finally updates them with the new interaction.

3.2.1. Profile-Driven Cognitive Decay

Let $\tau^s(t)$ and $\tau^p(t)$ denote the time steps of the most recent prior interactions for the current skill s_t and problem q_t , respectively. At step t , the framework retrieves the corresponding historical states: the previous global state $\mathbf{h}_{t-1}^g$, the historical skill state $\mathbf{h}_{\tau^s(t)}^s$, and the historical problem state $\mathbf{h}_{\tau^p(t)}^p$. For notational simplicity, we denote the retrieved historical state at level $k \in \{g, s, p\}$ as $\mathbf{h}_{\text{prev}}^k$.

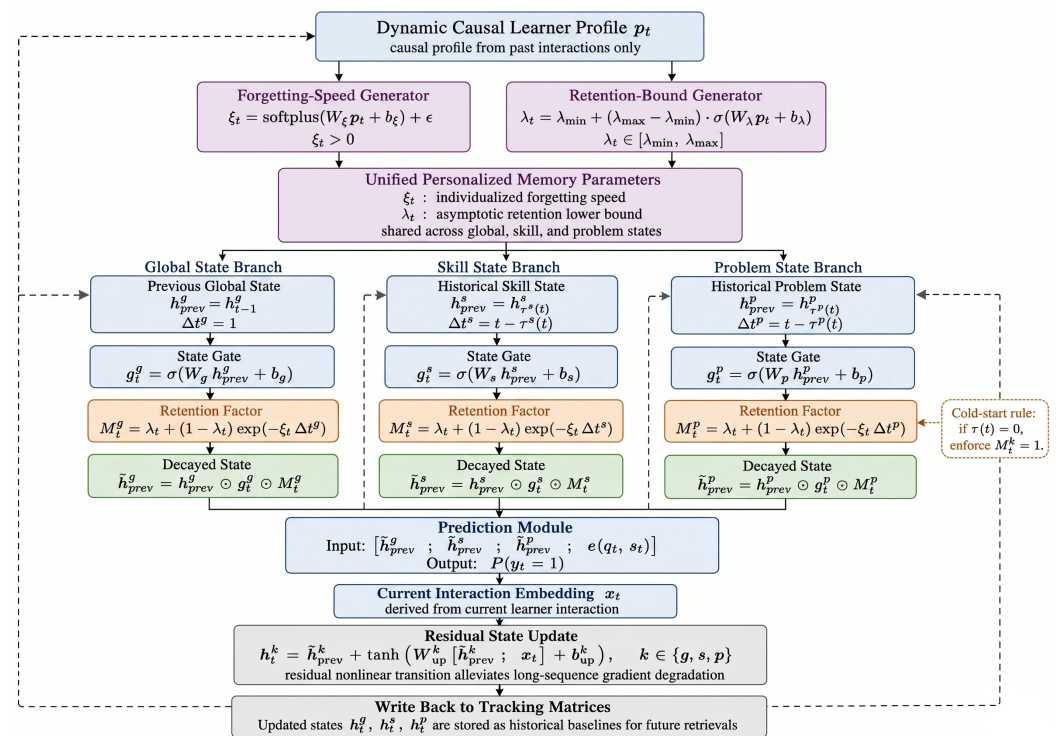


Figure 2. The overall architecture of the Profile-Driven Multi-Level State Tracking and Forgetting mechanism. A unified set of physiological parameters dynamically generated from the causal learner profile synchronously controls the temporal decay across the global ability, skill mastery, and problem memory states.

Because these multi-level states represent the knowledge of a single learner, their decay patterns are governed by a synchronized cognitive process, consistent with forgetting-aware KT studies that incorporate temporal decay into learner-state modeling [35,37]. We use the dynamic causal learner profile p_t to generate a unified set of memory parameters via two networks:

$$\xi_t = \text{softplus}(W_\xi p_t + b_\xi) + \epsilon, \quad (4)$$

$$\lambda_t = \lambda_{\min} + (\lambda_{\max} - \lambda_{\min}) \cdot \sigma(W_\lambda p_t + b_\lambda), \quad (5)$$

where $\xi_t > 0$ represents the individualized forgetting rate and $\lambda_t \in [\lambda_{\min}, \lambda_{\max}]$ denotes the asymptotic lower bound of retention.

In the English learning process, learners exhibit highly heterogeneous forgetting behaviors due to differences in cognitive abilities, prior knowledge, and learning habits. To model this heterogeneity, our mechanism abandons the traditional assumption of a uniform decay rate. Instead, we map each learner's unique causal profile to personalized cognitive parameters. Specifically, the individualized forgetting rate ξ_t and the minimum retention bound λ_t are dynamically generated for each student. Consequently, the state transition function applies distinct decay trajectories to different learners: students with strong foundational profiles experience slower cognitive decay (i.e., smaller ξ_t), while those with weaker profiles exhibit faster forgetting. By injecting these personalized parameters into the multi-level state tracking, the model effectively captures and distinguishes the diverse memory retention patterns across the learner population, thereby achieving a truly personalized modeling of heterogeneous forgetting.

The decay of each state differs based on its exposure interval. Let Δt^k denote the time gap for level k :

$$\Delta t^g = 1, \quad \Delta t^s = t - \tau^s(t), \quad \Delta t^p = t - \tau^p(t). \quad (6)$$

The profile-adaptive retention factor $\mathcal{M}_t^k$ is defined as:

$$\mathcal{M}_t^k = \lambda_t + (1 - \lambda_t) \cdot \exp(-\xi_t \Delta t^k). \quad (7)$$

To handle newly encountered skills or exercises (i.e., $\tau(t) = 0$), we set $\mathcal{M}_t^k = 1$.

Additionally, a state-level gate vector $\mathbf{g}_t^k$ is learned directly from the latent state representation, independent of explicit temporal features:

$$\mathbf{g}_t^k = \sigma(\mathbf{W}_k \mathbf{h}_{\text{prev}}^k + \mathbf{b}_k). \quad (8)$$

Combining the learned gate and the retention factor, the decayed hidden states $\tilde{\mathbf{h}}_{\text{prev}}^k$ used for the current prediction are computed via element-wise multiplication:

$$\tilde{\mathbf{h}}_{\text{prev}}^k = \mathbf{h}_{\text{prev}}^k \odot \mathbf{g}_t^k \odot \mathcal{M}_t^k, \quad k \in \{g, s, p\}. \quad (9)$$

3.2.2. Residual State Update

Following the prediction, the interaction embedding $\mathbf{x}_t$ is used to update the decayed knowledge states. To mitigate gradient degradation over long interaction sequences, we apply a residual update with a hyperbolic tangent non-linearity:

$$\mathbf{h}_t^k = \tilde{\mathbf{h}}_{\text{prev}}^k + \tanh\left(\mathbf{W}_{\text{up}}^k [\tilde{\mathbf{h}}_{\text{prev}}^k; \mathbf{x}_t] + \mathbf{b}_{\text{up}}^k\right), \quad k \in \{g, s, p\}, \quad (10)$$

where $[\cdot; \cdot]$ denotes vector concatenation, and $\mathbf{W}_{\text{up}}^k$ and $\mathbf{b}_{\text{up}}^k$ are learnable projection parameters for level k . These updated representations $\mathbf{h}_t^k$ are then written into their respective memory matrices, serving as the historical states for future retrievals.

3.3. Multi-Scale State Fusion and Prediction

Predicting the response of a learner to a specific English exercise requires evaluating the cognitive state of the learner. A learner with high global ability might answer a question incorrectly if the memory of a specific grammatical rule has decayed or if the target exercise is inherently difficult. Therefore, relying on a single granularity of knowledge states is insufficient for accurate competence tracing.

To address this, we propose a multi-scale state fusion strategy. At time step t , before the learner attempts the target exercise q_t , the model aggregates the multi-grained historical cognitive context and the target exercise representation. Specifically, the decayed global ability state $\tilde{\mathbf{h}}_{\text{prev}}^g$ (macro-level), the decayed skill mastery state $\tilde{\mathbf{h}}_{\text{prev}}^s$ (meso-level), and the decayed problem memory state $\tilde{\mathbf{h}}_{\text{prev}}^p$ (micro-level) are concatenated with the difficulty-aware exercise embedding $\mathbf{e}_t$. This operation forms a multi-scale feature vector $\mathbf{f}_t \in \mathbb{R}^{4d}$:

$$\mathbf{f}_t = [\tilde{\mathbf{h}}_{\text{prev}}^g; \tilde{\mathbf{h}}_{\text{prev}}^s; \tilde{\mathbf{h}}_{\text{prev}}^p; \mathbf{e}_t]. \quad (11)$$

By projecting these three tracking levels and the target exercise into a shared feature space, the model learns to dynamically weigh the importance of global proficiency, concept-specific mastery, and item-specific familiarity.

Finally, the fused state vector $\mathbf{f}_t$ is fed into a feed-forward neural network to estimate the probability of a correct response. The prediction network is formulated as a two-layer Multi-Layer Perceptron (MLP) with a non-linear activation:

$$\hat{r}_t = \sigma\left(\mathbf{W}_{\text{out}}^{(2)} \text{ReLU}\left(\mathbf{W}_{\text{out}}^{(1)} \mathbf{f}_t + \mathbf{b}_{\text{out}}^{(1)}\right) + \mathbf{b}_{\text{out}}^{(2)}\right), \quad (12)$$

where $\mathbf{W}_{\text{out}}^{(1)} \in \mathbb{R}^{d \times 4d}$, $\mathbf{W}_{\text{out}}^{(2)} \in \mathbb{R}^{1 \times d}$, and $\mathbf{b}_{\text{out}}^{(1)}, \mathbf{b}_{\text{out}}^{(2)}$ are learnable parameters. To prevent overfitting on sparse response trajectories, dropout regularization is applied to the hidden representations during training. The final output $\hat{r}_t \in (0, 1)$ obtained via the sigmoid function $\sigma(\cdot)$ represents the predicted probability that the learner will correctly answer the target exercise at step t .

This prediction score $\hat{r}_t$ serves as the basis for the supervised binary cross-entropy optimization and acts as the original probability distribution $\hat{r}_t^{\text{ori}}$ used in the subsequent masked consistency regularization.

3.4. Noise-Resistant Masked Consistency Regularization

English assessment data frequently contain noisy observations of learner competence. For tasks involving vocabulary discrimination or multiple-choice grammar questions, learners may answer correctly through random guessing. Conversely, in spelling or cloze exercises, an incorrect response might arise from a typographical error rather than a lack of knowledge. Consequently, observed response labels are not fully reliable indicators of underlying proficiency.

The gap between observable performance and latent competence can mislead knowledge tracing models. If a model relies strictly on local response labels, it may interpret a random guess as an improvement in mastery, or a careless error as a degradation of knowledge. This sensitivity to isolated response fluctuations is problematic in English learning, where assessment records are often sparse and noisy.

To improve robustness against noise, we introduce an attribute-masked consistency regularization strategy. The intuition is that a reliable model should not alter the assessment of future performance when a small portion of historical *problem identifiers* is hidden, provided that the actual correctness of those responses remains known. By masking specific exercises while retaining the response outcomes, we force the model to infer the cognitive state from the broader correctness trajectory rather than overfitting to item-specific co-occurrence patterns, as illustrated in Figure 3.

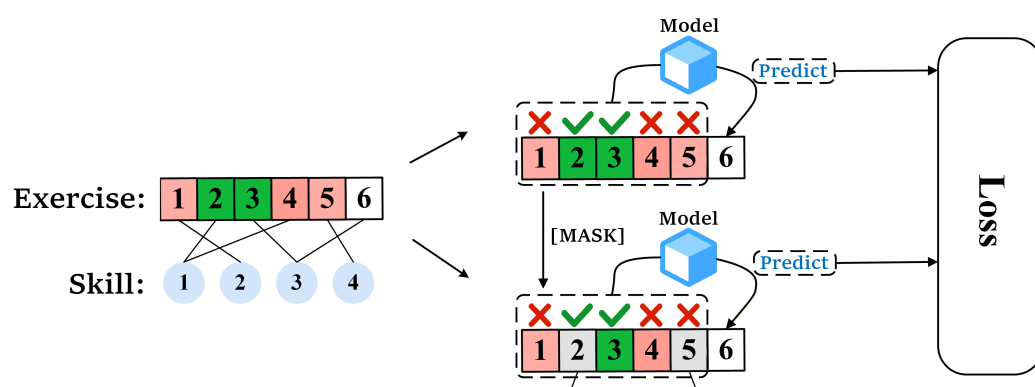


Figure 3. Illustration of the attribute-masked consistency regularization strategy. A subset of historical problem identifiers is replaced with a special mask token while their correctness labels remain strictly unchanged. The model is forced to yield consistent predictions on the unmasked future steps across both the original and masked views.

Given an original sequence of interactions $\mathcal{X} = \{(q_t, s_t, r_t)\}_{t=1}^T$, a subset of valid positions is randomly selected, and the corresponding problem identifiers are replaced

with a special mask token, following the general masked modeling paradigm widely used in representation learning [46]. The masked problem sequence is defined as:

$$\tilde{q}_t = \begin{cases} q_{\text{mask}}, & t \in \mathcal{M}, \\ q_t, & t \notin \mathcal{M}, \end{cases} \quad (13)$$

where $\mathcal{M}$ denotes the set of masked positions and q_{mask} denotes the special token indicating that the exercise identity is unavailable. The observed response r_t remains unchanged. This operation produces an augmented sequence of interactions:

$$\tilde{\mathcal{X}} = \{(\tilde{q}_t, s_t, r_t)\}_{t=1}^T. \quad (14)$$

The model processes both the original sequence and the masked sequence. Formally, the two predictive views are formulated as:

$$\hat{r}_t^{\text{ori}} = f_{\theta}(\mathcal{X}), \quad (15)$$

$$\hat{r}_t^{\text{mask}} = f_{\theta}(\tilde{\mathcal{X}}). \quad (16)$$

Although the masked view omits specific problem identifiers, it retains the correctness sequence and causal state updates. Therefore, predictions derived from both views for the remaining unmasked future steps should remain consistent. This consistency requirement prevents the model from overfitting to individual response outcomes and forces it to estimate competence from broader contextual evidence. By enforcing prediction stability under problem masking, the proposed strategy mitigates the impact of noisy assessment signals and improves the tracing of language competence.

3.5. Joint Optimization and Overall Objective

The proposed framework is optimized under a joint objective that balances two complementary requirements. First, the model must fit the observed assessment outcomes, as response correctness provides direct supervision for knowledge tracing. Second, the model should avoid overfitting to noisy response fluctuations, such as random guesses, careless mistakes, or typographical errors. Therefore, we combine a standard supervised prediction objective with a masked consistency regularization term.

For the supervised branch, the model takes the original interaction sequence as input and predicts the probability of a correct response at each valid time step. Given the original-view prediction $\hat{r}_t^{\text{ori}}$ and the observed response label r_t , we define the binary cross-entropy loss as:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{|\Omega|} \sum_{t \in \Omega} [r_t \log \hat{r}_t^{\text{ori}} + (1 - r_t) \log(1 - \hat{r}_t^{\text{ori}})], \quad (17)$$

where Ω denotes the set of valid non-padding positions. This term ensures that the learned knowledge state remains predictive of actual response performance.

For the consistency branch, the model processes the masked interaction sequence introduced and produces the masked-view prediction $\hat{r}_t^{\text{mask}}$. Because the masked view removes the problem context at step $t \in \mathcal{M}$, the prediction for that specific step will naturally deviate. However, subsequent estimates of the knowledge state for the remaining unmasked exercises should remain consistent with the original-view predictions. We therefore define the masked consistency objective strictly over the valid, non-padding positions that are not masked:

$$\mathcal{L}_{\text{cons}} = \frac{1}{|\Omega \setminus \mathcal{M}|} \sum_{t \in \Omega \setminus \mathcal{M}} \left(\hat{r}_t^{\text{mask}} - \text{sg}(\hat{r}_t^{\text{ori}}) \right)^2, \quad (18)$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operation. Using the Mean Squared Error (MSE) encourages the continuous probability distributions of the two views to align. By applying this constraint to the unmasked future trajectory, the model mitigates the impact of noisy historical interactions without penalizing the necessary uncertainty at the masked steps themselves.

The final optimization objective is formulated as:

$$\mathcal{L} = \mathcal{L}_{BCE} + \alpha \mathcal{L}_{cons}, \quad (19)$$

where α controls the strength of the consistency regularization.

Through this joint optimization, the model learns from observed assessment records while reducing sensitivity to unreliable response cues. The supervised objective anchors predictions to actual performance, whereas the consistency objective encourages stable predictions under partial attribute masking. Consequently, the framework learns competence-oriented representations from sparse and noisy learning data.

4. Experiments and Results Analysis

4.1. Experimental Preparation

4.1.1. Datasets

To ensure comparability with state-of-the-art baselines and to adhere to the standard evaluation protocols in the field of Knowledge Tracing, we utilize ASSISTments2009 and Algebra2005. These established benchmarks provide a rigorous testing ground for our model's cognitive mechanisms.

ASSISTments2009 (AS09).

The ASSISTments2009 dataset was collected from the ASSISTments online tutoring platform during the 2009–2010 school year. Serving as a standard benchmark for knowledge tracing research, it provides a medium-scale online tutoring scenario suitable for evaluating the capacity of the model to track the evolving knowledge states of learners from sequential interactions.

Algebra2005 (AL05). The Algebra2005 dataset originates from the KDD Cup 2010 Educational Data Mining Challenge and records step-level responses to algebra problem-solving tasks. Compared with AS09, AL05 contains longer learning sequences and a sparser concept-to-exercise structure, making it useful for examining whether a model can maintain stable estimations of learner states over extended interaction histories.

Although our research is motivated by English language learning, we utilize the ASSISTments2009 and Algebra2005 datasets for empirical validation for two key reasons:

- (1) **Benchmarking Cognitive Mechanisms:** These datasets are the most established benchmarks for evaluating KT models' ability to handle cognitive heterogeneity and noisy labels. Since our model aims to solve these exact theoretical problems, validating on these rigorous datasets ensures the robustness of our proposed mechanism.
- (2) **Generalizability:** Second-language acquisition shares strong statistical similarities with other learning processes (e.g., power-law forgetting curves and sparse interaction patterns). Therefore, demonstrating effectiveness on these standard mathematical datasets serves as a rigorous "proof of concept" for the model's underlying cognitive logic, which can be seamlessly transferred to English datasets in future work.

4.1.2. Baselines

To evaluate the proposed model, we compare it with eight representative knowledge tracing baselines. These methods cover a broad spectrum of architectures, including

recurrent neural networks, memory-augmented networks, attention mechanisms, and Transformer variants. Furthermore, they encompass recent advancements that emphasize forgetting awareness, length stability, and explicit state alignment.

DKT [18]. Deep Knowledge Tracing (DKT) is a classic KT model based on recurrent neural networks. It uses the hidden state of a recurrent network to summarize historical interactions and predict future responses.

DKVMN [19]. Dynamic Key-Value Memory Network (DKVMN) introduces an external key-value memory structure for knowledge tracing. The key matrix represents knowledge concepts, whereas the value matrix is updated dynamically to reflect the mastery states of learners.

AKT [20]. Context-Aware Attentive Knowledge Tracing (AKT) incorporates embeddings based on the Rasch model alongside attention modules to capture contextual dependencies among learning interactions.

DTransformer [30]. DTransformer is a Transformer-based KT model designed to obtain stable representations of knowledge states. It applies diagnostic sequence modeling to improve tracing beyond surface-level interaction fitting.

FolibiKT [35]. FolibiKT extends attentive knowledge tracing by introducing a linear bias that models forgetting behaviors. This mechanism explicitly incorporates temporal forgetting information into the prediction process.

simpleKT [31]. simpleKT serves as a streamlined yet competitive KT baseline. It reduces architectural complexity while maintaining strong predictive accuracy.

stableKT [32]. stableKT improves the length generalization of KT models. It introduces a linear bias mechanism to enhance stability when tracking long interaction sequences.

AlignKT [42]. AlignKT explicitly models the knowledge states of learners through an ideal-state alignment framework. It further employs contrastive learning to improve the quality of state representations.

4.1.3. Evaluation Metrics

Knowledge tracing is commonly formulated as a binary prediction task, where a model estimates the probability that a learner will answer the next exercise correctly. Following standard practice in KT research, we evaluate all models using two metrics: the Area Under the Receiver Operating Characteristic Curve (AUC) and Prediction Accuracy (ACC).

Area Under the Curve (AUC). AUC serves as the primary evaluation metric in our experiments. It measures the ranking quality of predicted probabilities by evaluating whether correct responses receive higher scores than incorrect responses. Formally, AUC is calculated as

$$\text{AUC} = \frac{\sum_{i \in \mathcal{P}} \sum_{j \in \mathcal{N}} \mathbb{I}(\hat{r}_i > \hat{r}_j) + \frac{1}{2} \mathbb{I}(\hat{r}_i = \hat{r}_j)}{|\mathcal{P}| |\mathcal{N}|}, \quad (20)$$

where $\mathcal{P}$ and $\mathcal{N}$ denote the sets of positive and negative response instances, respectively, and $\mathbb{I}(\cdot)$ represents the indicator function. This metric represents the probability that a randomly selected correct response is ranked higher than a randomly selected incorrect response. Because educational datasets frequently exhibit imbalanced distributions of correct and incorrect responses, AUC remains unaffected by classification thresholds, thereby providing a stable measure of discriminative capacity.

Prediction Accuracy (ACC). ACC measures the overall proportion of correctly classified responses. To compute ACC, the predicted probability $\hat{r}_t$ is converted into a binary prediction using a threshold of 0.5:

$$\hat{y}_t = \begin{cases} 1, & \hat{r}_t \geq 0.5, \\ 0, & \hat{r}_t < 0.5. \end{cases} \quad (21)$$

Subsequently, ACC is computed as:

$$\text{ACC} = \frac{1}{|\Omega|} \sum_{t \in \Omega} \mathbb{I}(\hat{y}_t = r_t), \quad (22)$$

where Ω denotes the set of valid test instances. Compared with AUC, ACC provides an intuitive interpretation but is more sensitive to label distribution and threshold selection. Consequently, we report both AUC and ACC to ensure a thorough evaluation.

4.1.4. Experimental Environment and Implementation Details

The proposed framework is implemented using PyTorch 2.11.0 and trained on a single NVIDIA GeForce RTX 5090 GPU. To ensure fair comparisons, all datasets undergo an identical preprocessing protocol, where interaction sequences are either truncated or padded to a maximum length of 200. Sequences containing fewer than three interactions are discarded to guarantee that each sample provides sufficient historical context for sequential modeling.

For the model configuration, the embedding dimension is set to 128. Optimization is performed using the Adam optimizer with a learning rate of 0.002 and a weight decay of 1×10^{-5} . The batch size is fixed at 128. To enhance numerical stability during optimization, gradient clipping is applied with a maximum norm of 15.0.

For the masked consistency regularization, the masking ratio is set to 0.3, meaning that 30% of valid response positions are randomly replaced by a mask token during training. The consistency coefficient is set to 0.1 to balance the supervised prediction objective and the consistency regularization term. During evaluation, no masking mechanism is applied, and the model predicts future responses based entirely on the original interaction sequences.

4.2. Experimental Analysis

To comprehensively evaluate the proposed framework, we conduct extensive experiments to answer the following research questions. The results demonstrate our model's superiority in prediction accuracy and its robustness against noisy data, as detailed below.

RQ1: How does the proposed framework perform against state-of-the-art baselines on real-world knowledge tracing datasets?

Table 1 presents the predictive performance of all evaluated models on the ASSISTments2009 and Algebra2005 datasets. The experimental results show that the proposed framework achieves the best AUC on both datasets, demonstrating its strong diagnostic capability in ranking learners' future response correctness. Since AUC measures the quality of continuous prediction scores independently of a fixed decision threshold, it is used as the primary evaluation metric in our analysis.

Table 1. Predictive performance on ASSISTments2009 and Algebra2005. Best in **bold**, second-best underlined.

Model	ASSISTments2009		Algebra2005	
	AUC	ACC	AUC	ACC
DKT	0.7552	0.7258	0.8149	0.8097
DKVMN	0.7479	0.7199	0.8065	0.8027
AKT	0.7805	0.7380	0.8139	0.8050
DTransformer	0.7763	0.7315	0.8188	0.8043
FolibiKT	0.7770	0.7325	0.8159	0.8035
simpleKT	0.7724	0.7293	0.8212	0.8069
stableKT	0.7766	0.7318	0.8275	<u>0.8102</u>
AlignKT	<u>0.7857</u>	<u>0.7382</u>	<u>0.8323</u>	0.8125
Ours	0.7929	0.7468	0.8495	0.8095

On the ASSISTments2009 dataset, the proposed model obtains the highest AUC of 0.7929 and ACC of 0.7468. Compared with the strongest baseline, AlignKT, the proposed approach yields absolute improvements of 0.72 percentage points in AUC and 0.86 percentage points in ACC. Furthermore, compared with FolibiKT, which incorporates a global forgetting mechanism, the superior performance of the proposed framework suggests the necessity of adaptive cognitive modeling. In language acquisition scenarios such as English learning, memory decay may vary substantially across individuals due to differences in linguistic intuition, prior vocabulary knowledge, and learning histories. By dynamically generating forgetting parameters conditioned on individual learning profiles, the proposed model can better capture heterogeneous cognitive dynamics than methods relying on uniformly shared decay patterns.

On the Algebra2005 dataset, the proposed model achieves the highest AUC of 0.8495, outperforming AlignKT and stableKT by 1.72 and 2.20 percentage points, respectively. Although the proposed framework obtains a competitive ACC of 0.8095, it does not achieve the best ACC on this dataset. This discrepancy is reasonable because ACC depends on a fixed binary classification threshold and may therefore be more sensitive to threshold calibration, response noise, and the distributional characteristics of long interaction sequences. In contrast, AUC evaluates the overall ranking quality of prediction probabilities. The substantial AUC improvement indicates that the proposed framework more effectively estimates learners' latent knowledge states, even when the final threshold-based classification accuracy does not achieve the highest value.

RQ2: How does each proposed component contribute to the overall performance of the framework?

To answer this question, we conduct an ablation study to evaluate the contribution of individual core modules. We compare the complete framework against two ablated variants. The first variant, denoted as *w/o* Profile-Driven Forgetting, completely removes the personalized memory decay mechanism. In this setting, the multi-level knowledge states are updated sequentially without any explicit temporal decay applied to historical tracking. The second variant, denoted as *w/o* Attribute-Masked Consistency, disables the stochastic problem masking during training and removes the corresponding consistency regularization objective, relying strictly on the standard supervised binary cross-entropy loss. The ablation results on both datasets are summarized in Table 2, and the absolute performance drops are visually compared in Figure 4.

Removing the profile-driven forgetting mechanism leads to a substantial degradation in predictive performance, indicating that it is the most critical component of the proposed framework. As detailed in Table 2 and Figure 4, the *"w/o* Profile-Driven Forgetting" variant experiences a severe drop on ASSISTments2009, with the AUC decreasing by 0.0137 to 0.7792 and the ACC falling by 0.0112 to 0.7356. A similar decline is observed on Algebra2005, where the AUC drops by 0.0114 to 0.8381 and the ACC decreases by 0.0098 to 0.7997. This performance drop empirically validates our core motivation: assuming a static or uniform memory process is highly insufficient. The dynamic, profile-driven multi-level decay captures the heterogeneous forgetting trajectories of different learners, providing the temporal dynamics needed for accurate long-term competence tracing.

Table 2. Ablation study of the proposed framework on ASSISTments2009 and Algebra2005.

Model	ASSISTments2009		Algebra2005	
	AUC	ACC	AUC	ACC
Full Model (Ours)	0.7929	0.7468	0.8495	0.8095
<i>w/o</i> Profile-Driven Forgetting	0.7792	0.7356	0.8381	0.7997
<i>w/o</i> Attribute-Masked Consistency	0.7918	0.7454	0.8484	0.8082

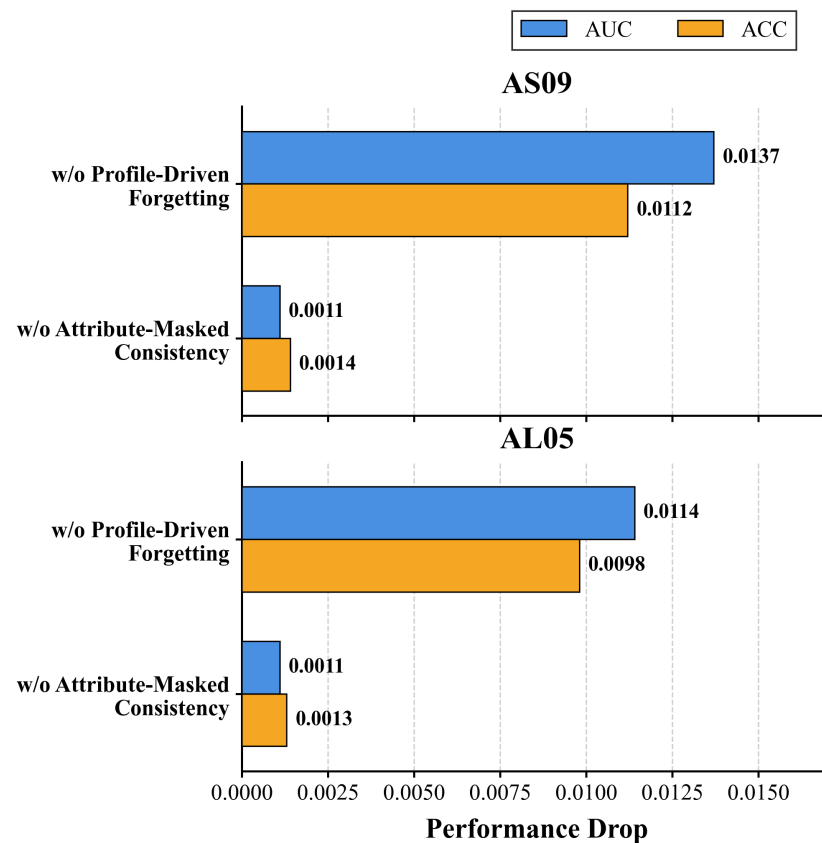


Figure 4. Absolute performance drop of the ablation variants compared to the full model on ASSISTments2009 (AS09) and Algebra2005 (AL05).

While the performance drop is smaller, removing the masking module also results in a consistent decline across all metrics on both datasets. For instance, without the consistency regularization, the AUC drops by 0.0011 to 0.7918 on ASSISTments2009 and to 0.8484 on Algebra2005. The ACC exhibits a similar decrease of 0.0014 and 0.0013 on the respective datasets. This degradation highlights the importance of the attribute-masked consistency strategy. By forcing the model to yield stable competence estimations even when specific problem identifiers are obscured, the regularization prevents the network from overfitting to item-specific co-occurrence patterns and noisy responses. Collectively, the ablation study demonstrates that both individualized cognitive modeling and noise-resistant regularization are necessary for achieving optimal performance.

RQ3: How sensitive is the proposed framework to the attribute masking ratio, and what is the optimal masking strategy?

To answer this question, we investigate the impact of the masking ratio, defined as the proportion of problem identifiers replaced by the mask token during the consistency regularization phase. We vary the mask ratio p from 0.15 to 0.60 with a step size of 0.15. Figure 5 illustrates the performance drop from the optimal setting for both AUC and ACC.

The results show an inverted U-shaped trend for absolute performance across both datasets, corresponding to the V-shaped drop curves in Figure 5. Peak performance occurs when the mask ratio is set to 0.30. Specifically, on the ASSISTments2009 dataset, the AUC peaks at 0.7929 and the ACC reaches 0.7468. Similarly, on Algebra2005, performance maximizes at an AUC of 0.8495 and an ACC of 0.8095. Deviating from this optimal value leads to a consistent degradation in predictive accuracy.

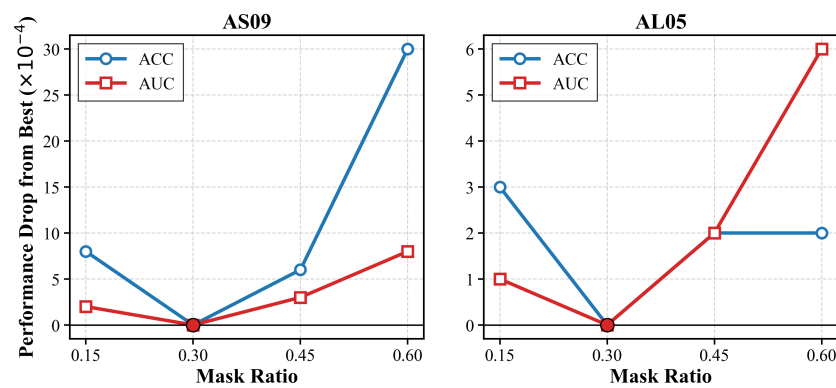


Figure 5. Performance drop from the optimal results under varying mask ratios on ASSISTments2009 (AS09) and Algebra2005 (AL05). The lowest performance drop is consistently achieved at a mask ratio of 0.30.

This phenomenon reflects a trade-off between noise-resistant regularization and the preservation of sequential learning context. For a small mask ratio ($p = 0.15$), the augmented view remains largely identical to the original interaction sequence. The consistency constraint is therefore too weak to reduce reliance on item-specific memorization, leaving the model vulnerable to noisy assessment behaviors such as random guessing or careless slips. Consequently, the model slightly underperforms.

Conversely, larger mask ratios ($p \geq 0.45$) cause more severe performance deterioration. At a mask ratio of 0.60, the AUC drops to 0.7921 on ASSISTments2009 and 0.8489 on Algebra2005, alongside a sharp decrease in accuracy. This decline occurs because masking too many problem identifiers removes necessary micro-level contextual anchors. Although the overarching correctness trajectory is preserved, the model loses track of the specific exercises the learner attempts, making it difficult to accurately update the localized problem memory state $\mathbf{h}^p$. In extreme cases, the augmented sequence becomes structurally fragmented, forcing the model to align predictions with distorted states and impairing tracking precision.

Therefore, a moderate mask ratio of 0.30 provides an optimal balance. It introduces sufficient perturbation to prevent item-level overfitting while preserving enough exercise-specific context for accurate, multi-grained knowledge state tracking.

RQ4: How does the learning rate affect the predictive performance and optimization stability of the proposed framework?

To answer this question, we evaluate the framework under three learning rates: 0.0002, 0.002, and 0.02. Figure 6 shows the predictive performance in terms of AUC and ACC on the ASSISTments2009 and Algebra2005 datasets.

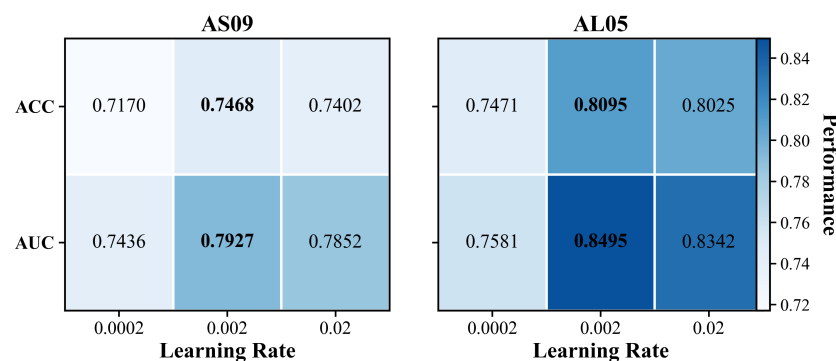


Figure 6. Predictive performance heatmaps of the proposed framework under varying learning rates on ASSISTments2009 (AS09) and Algebra2005 (AL05). The optimal performance is consistently achieved at a learning rate of 0.002.

The results show that the framework achieves optimal performance at a learning rate of 0.002 across both datasets. At this rate, the model obtains an AUC of 0.7927 and an ACC of 0.7468 on ASSISTments2009, and an AUC of 0.8495 and an ACC of 0.8095 on Algebra2005. Deviating from this value degrades predictive performance.

When the learning rate is reduced to 0.0002, performance drops substantially. The AUC decreases to 0.7436 on ASSISTments2009 and 0.7581 on Algebra2005. This decline indicates that a small learning rate causes the optimization process to converge too slowly. Consequently, the model becomes trapped in local minima and fails to learn the complex temporal dynamics of the multi-level knowledge states within the designated training epochs.

Conversely, increasing the learning rate to 0.02 also decreases performance, resulting in an AUC of 0.7852 on ASSISTments2009 and 0.8342 on Algebra2005. A large learning rate induces instability during gradient descent, causing parameter updates to oscillate and overshoot the optimal solution. This instability prevents the parameter-generation networks from finely tuning the personalized forgetting rates and minimum retention bounds.

Therefore, a learning rate of 0.002 provides an optimal balance between convergence speed and optimization stability, enabling accurate competence tracing.

RQ5: How does the hidden state dimensionality impact the representational capacity and prediction accuracy of the model?

To answer this question, we investigate the sensitivity of the proposed framework to the hidden embedding dimension d . We evaluate the model by setting $d \in \{64, 128, 256\}$. Figure 7 visualizes the predictive performance across the ASSISTments2009 and Algebra2005 datasets.

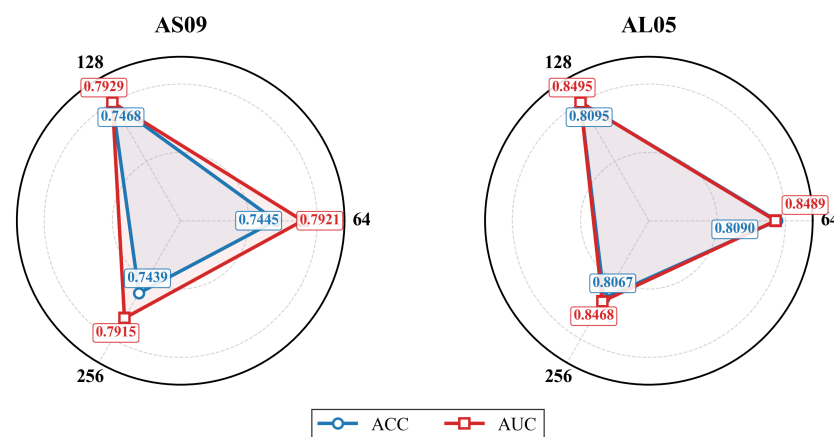


Figure 7. Predictive performance radar charts of the proposed framework under varying hidden dimensions ($d \in \{64, 128, 256\}$) on ASSISTments2009 (AS09) and Algebra2005 (AL05). The optimal balance is consistently achieved at $d = 128$.

Results indicate that the framework achieves peak performance at a hidden dimension of 128. At this dimensionality, the model yields an AUC of 0.7929 and an ACC of 0.7468 on ASSISTments2009, and an AUC of 0.8495 and an ACC of 0.8095 on Algebra2005. Deviating from this dimension decreases predictive accuracy.

Reducing the hidden dimension to 64 slightly decreases performance, with the AUC dropping to 0.7921 on ASSISTments2009 and 0.8489 on Algebra2005. This degradation suggests that a compact latent space lacks sufficient representational capacity. It restricts the model from capturing semantic interactions between exercises and skills, limiting the expressiveness required to maintain individualized multi-level knowledge states.

Conversely, increasing the hidden dimension to 256 causes a more pronounced performance drop. The AUC decreases to 0.7915 on ASSISTments2009 and 0.8468 on Algebra2005. This drop is primarily attributed to overfitting and parameter redundancy. For sparse

educational interaction sequences, high-dimensional embeddings introduce noise. Furthermore, the large parameter space makes it difficult for the parameter-generation networks to output stable memory decay parameters.

Therefore, a hidden dimension of 128 provides the optimal trade-off between representational expressiveness and model generalization, ensuring the accurate tracking of learners' language competence.

5. Conclusions and Discussion

5.1. Synthesis of Key Findings

In this study, we proposed a novel framework for robust English knowledge tracing. Through extensive empirical analysis on two benchmark datasets, our results demonstrate that integrating personalized cognitive mechanisms with noise-resilient training significantly outperforms state-of-the-art baselines. Specifically, our analysis reveals that the performance gain is primarily driven by the profile-driven adaptive forgetting mechanism. Unlike traditional models that treat forgetting as a global constant, our findings confirm that learners exhibit highly heterogeneous memory decay patterns. By dynamically generating individualized forgetting parameters, the model successfully captures these subtle cognitive differences, leading to more accurate predictions of future performance.

5.2. Discussion on Robustness and Prior Literature

Furthermore, the attribute-masked consistency regularization proved essential in mitigating the impact of noisy evaluation data. Prior studies often relied on adversarial training to improve robustness, which can sometimes introduce artificial perturbations that deviate from real learning behaviors. Our analysis suggests that enforcing predictive consistency under masked conditions is a more cognitively plausible approach. It forces the model to rely on the underlying trajectory of concept mastery rather than memorizing specific item identifiers. This finding not only validates our central thesis but also provides empirical evidence that addressing data sparsity and noise requires structural constraints rather than mere data augmentation.

Ultimately, this research leads us to the conclusion that reliable knowledge tracing in language learning cannot be achieved solely through deeper neural architectures; it requires an explicit integration of cognitive science principles. Beyond theoretical contributions, our proposed framework offers concrete implications for the current higher education scene. For instance, in large-enrollment English courses, instructors often struggle to identify individual learning bottlenecks. Our model can be integrated into Intelligent Tutoring Systems (ITS) to dynamically track students' heterogeneous forgetting patterns, thereby triggering personalized review interventions rather than relying on uniform assignments. Furthermore, by making the knowledge tracing process transparent, our approach supports the student-centered learning paradigm widely advocated in modern higher education, ultimately reducing teachers' workload while improving learning efficiency.

Looking forward, while our current experiments have rigorously validated the model's generalizability on standard benchmarks, we recognize the importance of domain-specific validation. Future work will focus on extending this framework to dedicated language learning corpora, such as the Duolingo SLAM datasets or EAnatole. This will allow us to further investigate how the model captures unique linguistic features (e.g., semantic relationships) and to fine-tune its scalability in real-world second-language acquisition platforms.

Author Contributions: X.C.: Conceptualization, Methodology, Software, Validation, Investigation, Data Curation, Writing—Original Draft Preparation, Writing—Review and Editing, Visualization; Z.Z.: Conceptualization, Methodology, Software, Validation, Resources, Data Curation, Writing—Original Draft Preparation, Writing—Review and Editing, Project Administration, Funding Acqui-

sition; H.H.: Formal Analysis; J.J.: Conceptualization, Validation; F.Y.: Formal Analysis, Funding Acquisition; L.Z.: Validation. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Hunan Social Science Foundation grant number 21WLH25. The APC was funded the Hunan Social Science Foundation.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Li, X.; Sun, Y.; Lin, J.; Li, L.; Feng, T.; Yin, S. The Synergy of Seeing and Saying: Revolutionary Advances in Multi-modality Medical Vision-Language Large Models. *Artif. Intell. Sci. Eng.* **2025**, *1*, 79–97.
2. Varghese, C.; Harrison, E.M.; O’Grady, G.; Topol, E.J. Artificial intelligence in surgery. *Nat. Med.* **2024**, *30*, 1257–1268.
3. Bouguettaya, S.; Zedadra, O.; Pupo, F.; Fortino, G. Toward Collaborative and Adaptive Learning: A Survey of Multi-Agent Reinforcement Learning in Education. *Artif. Intell. Sci. Eng.* **2026**, *2*, 1–19.
4. Shan, H.; Wei, C.; Ramos, N.; Yang, X.; Guo, C.; Li, H.; Chen, Y. Neuromorphic Computing in the Era of Large Models. *Artif. Intell. Sci. Eng.* **2025**, *1*, 17–30.
5. Zhang, J.; Li, D.; Zeng, Z.; Zhang, R.; Wang, J. Dual-branch crack segmentation network with multi-shape kernel based on convolutional neural network and Mamba. *Eng. Appl. Artif. Intell.* **2025**, *150*, 110536.
6. Khan, A.H.; Cao, X.; Luo, C.; Zhang, S.; Guo, W.; Katsikis, V.N.; Li, S. Spiking Neural Networks: A Comprehensive Survey of Training Methodologies, Hardware Implementations and Applications. *Artif. Intell. Sci. Eng.* **2025**, *1*, 175–207.
7. Hu, J.; Zeng, C.; Wang, Z.; Zhang, J.; Guo, K.; Xu, H.; Huang, J.; Chen, K. Load balancing with multi-level signals for lossless datacenter networks. *IEEE/ACM Trans. Netw.* **2024**, *32*, 2736–2748.
8. Hu, J.; Shen, H.; Huang, J.; Sherratt, R.S.; Wang, J. A high-performance Sketch with Dynamic Memory Allocation for priority-oriented data stream processing. *IEEE Trans. Comput.* **2025**, *75*, 720–733.
9. Hu, J.; He, Y.; Luo, W.; Huang, J.; Wang, J. Enhancing load balancing with in-network recirculation to prevent packet reordering in lossless data centers. *IEEE/ACM Trans. Netw.* **2024**, *32*, 4114–4127.
10. Yu, F.; Guo, R.; Zheng, M.; Yao, W.; Zhang, D.; Cai, S. Novel approach to time series forecasting based on memristive Hopfield Neural Network with hidden heterogeneous and homogeneous extreme multistability. *Chaos Solitons Fractals* **2026**, *210*, 118626.
11. He, T.; Yu, F.; Lin, Y.; He, S.; Yao, W.; Cai, S.; Jin, J. Multi-scroll hopfield neural network excited by memristive self-synapses and its application in image encryption. *Chin. Phys. B* **2025**, *34*, 120506.
12. Yu, F.; Wang, X.; Xiao, Y.; He, Y.; Yao, W.; Cai, S.; Jin, J. Extreme multistability in discrete memristive neuron maps and implications for dual-field applications. *Integration* **2026**, *109*, 102688.
13. Létourneau, A.; Deslandes Martineau, M.; Charland, P.; Karran, J.A.; Boasen, J.; Léger, P.M. A systematic review of AI-driven intelligent tutoring systems (ITS) in K-12 education. *npj Sci. Learn.* **2025**, *10*, 29.
14. Niño-Rojas, F.; Lancheros-Cuesta, D.; Jiménez-Valderrama, M.T.P.; Mestre, G.; Gómez, S. Systematic Review: Trends in Intelligent Tutoring Systems in Mathematics Teaching and Learning. *Int. J. Educ. Math. Sci. Technol.* **2024**, *12*, 203–229.
15. Ateş, H. Integrating augmented reality into intelligent tutoring systems to enhance science education outcomes. *Educ. Inf. Technol.* **2025**, *30*, 4435–4470.
16. Yang, A.C.; Lin, J.Y.; Lin, C.Y.; Ogata, H. Enhancing python learning with PyTutor: Efficacy of a ChatGPT-Based intelligent tutoring system in programming education. *Comput. Educ. Artif. Intell.* **2024**, *7*, 100309.
17. Abdelrahman, G.; Wang, Q.; Nunes, B.P. Knowledge Tracing: A Survey. *ACM Comput. Surv.* **2023**, *55*, 224. <https://doi.org/10.1145/3569576>.
18. Piech, C.; Bassen, J.; Huang, J.; Ganguli, S.; Sahami, M.; Guibas, L.J.; Sohl-Dickstein, J. Deep knowledge tracing. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Montreal, QC, Canada, 7–12 December 2015; Volume 28.
19. 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 (WWW), Perth, Australia, 3–7 April 2017; pp. 765–774.
20. 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, Virtual, 6–10 July 2020; pp. 2330–2339.

21. Shen, S.; Liu, Q.; Chen, E. Learning process-consistent knowledge tracing. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, Virtual, 14–18 August 2021; pp. 1452–1460.
22. Cheng, X.; Yin, L.; Zhang, H. Morphological Awareness and Vocabulary Knowledge in Monolingual and Bilingual Learners: A Meta-Analysis. *Int. J. Biling.* **2026**, *30*, 222–248. <https://doi.org/10.1177/13670069241311029>.
23. Liu, Y.; Groen, M.A.; Cain, K. The Association between Morphological Awareness and Reading Comprehension in Children: A Systematic Review and Meta-Analysis. *Educ. Res. Rev.* **2024**, *42*, 100571. <https://doi.org/10.1016/j.edurev.2023.100571>.
24. Nagatani, K.; Zhang, Q.; Sato, M.; Chen, Y.Y.; Chen, F.; Ohkuma, T. Augmenting Knowledge Tracing by Considering Forgetting Behavior. In Proceedings of the World Wide Web Conference, New York, NY, USA, 13–17 May 2019; pp. 3101–3107. <https://doi.org/10.1145/3308558.3313565>.
25. Bai, Y.; Li, X.; Liu, Z.; Huang, Y.; Tian, M.; Luo, W. Rethinking and Improving Student Learning and Forgetting Processes for Attention Based Knowledge Tracing Models. *Proc. AAAI Conf. Artif. Intell.* **2025**, *39*, 27822–27830. <https://doi.org/10.1609/aaai.v39i27.34998>.
26. Guo, T.; Qin, Y.; Xia, Y.; Hou, M.; Liu, Z.; Xia, F.; Luo, W. Enhancing Knowledge Tracing through Decoupling Cognitive Pattern from Error-Prone Data. In Proceedings of the ACM on Web Conference 2025, Sydney, Australia, 28 April–2 May 2025; pp. 5108–5116. <https://doi.org/10.1145/3696410.3714486>.
27. Xie, Q.; Dai, Z.; Hovy, E.; Luong, M.T.; Le, Q.V. Unsupervised Data Augmentation for Consistency Training. In Proceedings of the Advances in Neural Information Processing Systems, Virtual, 6–12 December 2020; Volume 33, pp. 6256–6268.
28. Sohn, K.; Berthelot, D.; Li, C.L.; Zhang, Z.; Carlini, N.; Cubuk, E.D.; Kurakin, A.; Zhang, H.; Raffel, C. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. In Proceedings of the Advances in Neural Information Processing Systems, Virtual, 6–12 December 2020; Volume 33, pp. 596–608.
29. Corbett, A.T.; Anderson, J.R. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Model. User-Adapt. Interact.* **1994**, *4*, 253–278.
30. 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; ACM: Sydney, Australia, 2023; pp. 855–864. <https://doi.org/10.1145/3543507.3583255>.
31. Liu, Z.; Liu, Q.; Chen, J.; Huang, S.; Luo, W. simpleKT: A Simple But Tough-to-Beat Baseline for Knowledge Tracing. In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Long Beach, CA, USA, 6–10 August 2023; pp. 4552–4562.
32. 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. In Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, Jeju, Republic of Korea, 3–9 August 2024. <https://doi.org/10.24963/ijcai.2024/654>.
33. Settles, B.; Meeder, B. A trainable spaced repetition model for language learning. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL), Berlin, Germany, 7–12 August 2016; pp. 1848–1858.
34. Wang, C.; Ma, W.; Zhang, M.; Lv, C.; Wan, F.; Lin, H.; Tang, T.; Liu, Y.; Ma, S. Temporal cross-effects in knowledge tracing. In Proceedings of the 14th ACM International Conference on Web Search and Data Mining (WSDM), Virtual, 8–12 March 2021; pp. 517–525.
35. 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; ACM: Sydney, Australia, 2023. <https://doi.org/10.1145/358378.0.3615191>.
36. Hou, M.; Li, X.; Guo, T.; Liu, Z.; Tian, M.; Luo, R.; Luo, W. Cognitive Fluctuations Enhanced Attention Network for Knowledge Tracing. *Proc. AAAI Conf. Artif. Intell.* **2025**, *39*, 14265–14273. <https://doi.org/10.1609/aaai.v39i13.33562>.
37. Cui, C.; Yao, Y.; Zhang, C.; Ma, H.; Ma, Y.; Ren, Z.; Zhang, C.; Ko, J. DGEKT: A Dual Graph Ensemble Learning Method for Knowledge Tracing. *ACM Trans. Inf. Syst.* **2024**, *42*, 1–24. <https://doi.org/10.1145/3638350>.
38. Su, Y.; Liu, Q.; Liu, Y. An attention-based knowledge tracing model with forgetting factor. *Expert Syst. Appl.* **2022**, *201*, 117180.
39. 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; ACM: Sydney, Australia, 2021; pp. 367–375. <https://doi.org/10.1145/347408.5.3475554>.
40. 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; ACM: Sydney, Australia, 2023; pp. 2441–2445. <https://doi.org/10.1145/3539618.3592073>.
41. Lee, W.; Chun, J.; Lee, Y.; Park, K.; Park, S. Contrastive Learning for Knowledge Tracing. In Proceedings of the ACM Web Conference 2022; ACM: Sydney, Australia, 2022; pp. 2330–2338. <https://doi.org/10.1145/3485447.3512105>.
42. Xiao, J.; You, C.; Chen, Z. AlignKT: Explicitly Modeling Knowledge State for Knowledge Tracing with Ideal State Alignment. In Proceedings of the 2025 IEEE International Conference on Multimedia and Expo (ICME), Nantes, France, 30 June–4 July 2025; pp. 1–6. <https://doi.org/10.1109/ICME59968.2025.11209278>.

43. Cheng, W.; Du, H.; Li, C.; Ni, E.; Tan, L.; Xu, T.; Ni, Y. Uncertainty-Aware Knowledge Tracing. *Proc. AAAI Conf. Artif. Intell.* **2025**, *39*, 27905–27913. <https://doi.org/10.1609/aaai.v39i27.35007>.
44. Liang, X.; Wu, L.; Li, J.; Wang, Y.; Meng, Q.; Qin, T.; Chen, W.; Zhang, M.; Liu, T.Y. R-Drop: Regularized Dropout for Neural Networks. In Proceedings of the Advances in Neural Information Processing Systems, Virtual, 6–14 December 2021; Volume 34, pp. 10890–10905.
45. Shen, X.; Yu, F.; Liu, Y.; Liang, R.; Wan, Q.; Yang, K.; Sun, J. Revisiting Knowledge Tracing: A Simple and Powerful Model. In Proceedings of the 32nd ACM International Conference on Multimedia, New York, NY, USA, 28 October–1 November 2024; pp. 263–272. <https://doi.org/10.1145/3664647.3681205>.
46. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186. <https://doi.org/10.18653/v1/N19-1423>.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.