

RESEARCH

Free and Open Access

Proficiency informed knowledge tracing: Leveraging large language models to enhance predictive methods in mathematics learning

Kannan Nataraj ^{1*}, Charles Y. C. Yeh ¹, Chih-Yueh Chou ² and Tak-Wai Chan ¹

*Correspondence:
kannannataraj94@gmail.com
Graduate Institute of Network
Learning Technology, National
Central University, Zhongli,
Taiwan.
Full list of author information is
available at the end of the article

Abstract

Predictive methods play a key role in learning systems to enable effective interventions. However, most existing approaches emphasize concept-level suggestions and pay limited attention to cognitive constructs aligned with instructional goals. For instance, in mathematics, mastery depends on the integrated and balanced development of multiple interwoven proficiency strands. When predictive outputs do not reflect instructional objectives, educators may struggle to trust or utilize the insights generated by these models. To address this, we propose a predictive framework that incorporates cognitive construct-aligned features into the modeling process. The research focuses on elementary mathematics, where proficiency strands represent fundamental cognitive dimensions of learning. Since manually labeling these strands is time-intensive for experts, large language models (LLMs) are used to classify semantic features. These features are integrated into the proposed Proficiency-informed Knowledge Tracing (PKT) model as auxiliary signals to enhance mastery estimation. PKT is evaluated on the XES3G5M dataset, which contains elementary mathematics problems, and is compared with established deep learning models. The results indicate that PKT achieves overall competitive performance and surpasses base models, demonstrating a capacity to capture latent mastery patterns. Further experiments reveal that integrating LLM-generated features improves model calibration, as measured by Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), compared to base models. PKT ranked first overall when predictive performance and calibration were considered together, followed by AKT and sparseKT. Additionally, the model generates interpretable diagnostic profiles at both cognitive and concept levels. These findings underscore the value of incorporating cognitively relevant constructs into knowledge tracing.

Keywords: mastery, predictive modeling, mathematics education, learning analytics, large language models (LLMs)



© The Author(s). 2026 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

Introduction

Technology-enhanced learning systems increasingly adopt predictive methods to support mastery-oriented learning, particularly when timely feedback and individualized support are central to instruction. In mathematics education, however, the value of such prediction depends not only on whether a model can estimate future performance, but also on whether it can represent the multidimensional nature of developing proficiency. Recent work has shown that predictive methods often remain limited in their ability to capture mastery at both content and cognitive levels, especially when they rely primarily on raw behavioral features and produce outputs that offer little guidance for personalized intervention (Simpson & Wang, 2023; Xing & Du, 2019; Zhang et al., 2023). Although knowledge tracing has long provided an important foundation for modeling learners' changing knowledge states over time (Corbett & Anderson, 1995; Pelánek, 2017; Piech et al., 2015), its instructional value in mathematics remains constrained when prediction is detached from meaningful learning constructs.

A central difficulty is that what constitutes mastery and how its thresholds are defined remain largely arbitrary across existing systems, even though the appropriate definition depends heavily on the cognitive demands of the subject. In computer science education, for instance, binary correct or incorrect responses fail to distinguish procedural coding steps from conceptual programming understanding (Liang et al., 2022). In language proficiency assessment, such as TOEIC, question difficulty and semantic context fundamentally shape what mastery actually means (Kim et al., 2023). In mathematics, mastery reflects qualitatively different forms of proficiency, including conceptual understanding, procedural fluency, and strategic competence, within a broader account of mathematical proficiency (Kilpatrick et al., 2001; Yeh et al., 2019). Yet many predictive systems continue to operationalize mastery through narrow thresholds such as percentage correct or consecutive correct responses, which may overlook the cognitive demand, transfer potential, and developmental character of learning (Guskey & Anderman, 2013; Kelly & Heffernan, 2016). This concern is compounded by the fact that a single concept may span multiple levels of difficulty and question formats, so that concept-level interpretation alone can lead to unnecessary repetition without meaningfully advancing learner understanding.

More broadly, contemporary predictive methods often privilege observable efficiency measures such as correctness, response time, and hint usage, while paying less attention to the deeper cognitive outcomes central to instructional judgment (Cen et al., 2025; Liang et al., 2022; Wang et al., 2023). As a result, prediction can become technically effective yet pedagogically thin. Three limitations follow from this pattern. First, predictive models emphasize accuracy over cognitive depth, limiting their usefulness for instructional decision-making. Second, cognitively meaningful features typically rely on teacher

judgment or manual annotation, making them difficult to scale. Third, even when such features are available, they are often misaligned with subject-specific curricular goals.

Recent advances in large language models create a potential pathway for addressing these gaps by enabling richer feature construction from educational text and assessment metadata. Because cognitively meaningful representations are difficult to annotate at scale, predictive models have often remained dependent on sparse behavioral logs and manually specified skill maps (Bier et al., 2014; Kim et al., 2023). Large language models offer new opportunities to extract latent semantic and contextual information from question content and curricular descriptions, potentially better aligning predictive systems with learning goals (Cen et al., 2025; Chen et al., 2025). Against this background, the present study develops a knowledge-tracing framework that incorporates proficiency-strand information to better represent learners' evolving knowledge states in elementary mathematics, thereby moving prediction beyond surface-level correctness toward a more cognitively grounded account of mastery. The main contributions of this paper can be summarized as:

- We develop a predictive framework that integrates behavioral and cognitive features grounded in mathematics-specific learning frameworks. At the core of the framework is a proficiency-informed mastery rate method that builds personalized knowledge state profiles by modeling student performance across distinct proficiency strands.
- We propose a prior-biased gated transformer for modeling knowledge acquisition across heterogeneous proficiency strands. The model integrates causal mastery tracking, temporal forgetting, and strand-aware gating, combined with structured attention biases that encode curricular priors.
- We conduct extensive experiments on elementary mathematics data to evaluate the proposed approach. The results show that the model performs competitively with existing knowledge tracing baselines while providing richer diagnostic information through strand-level mastery estimates.
- Because the model reasons at the strand level, it produces interpretable per-strand mastery estimates that offer diagnostic value for instructional decision making, extending prediction beyond a single aggregate correctness score.

The remainder of this paper is organized as follows. Section 2 reviews the related work. Section 3 presents the overall model architecture and methodological steps. Section 4 presents experimental results and analysis. Section 5 concludes the paper.

Literature review

The growing use of predictive models in educational technology has raised important questions about how such models can support mastery oriented instruction (Slater & Baker, 2019). Mastery learning is concerned not only with whether learners answer correctly, but also with how understanding develops over time across meaningful learning goals. In mathematics, this poses a particular challenge because mastery is inherently multidimensional rather than reducible to surface level performance. Recent work has therefore questioned whether commonly used predictive approaches adequately represent the cognitive and instructional complexity of mathematical learning (Boyd & Ash, 2018; Simpson & Wang, 2023). Although models such as logistic regression, Bayesian Knowledge Tracing, and Deep Knowledge Tracing have been useful for identifying response patterns, they often remain limited in their ability to capture deeper forms of mathematical proficiency, including conceptual reasoning and problem solving (Kilpatrick et al., 2001; Rittle-Johnson, 2017).

A central concern in this literature is that mastery is often operationalized through narrow indicators such as fixed accuracy thresholds or binary labels of mastered and not mastered. Such approaches may simplify evaluation, but they provide limited insight into the gradual and uneven development of understanding, particularly in domains where reasoning, abstraction, and transfer are central to competence (Guskey & Anderman, 2013; Kelly & Heffernan, 2016). In mathematics, this limitation is especially salient because mathematical proficiency includes interrelated strands such as conceptual understanding, procedural fluency, and strategic competence, which may not develop uniformly within the same learner (Yeh et al., 2019). Yet most available datasets privilege correctness, response time, and number of attempts, while offering little access to the qualitative traces of thinking that would make these distinctions more visible (Chen et al., 2025). As a result, predictive systems may achieve technical performance while remaining weakly aligned with the learning processes they are intended to represent.

Related research in knowledge tracing has also pointed to the limits of feature representation. Models that rely heavily on expert defined Q matrices or sparse behavioral logs often overlook the semantic and conceptual characteristics of exercises, thereby constraining the granularity with which learner knowledge can be modeled (Chen et al., 2025; Huang et al., 2021). Although incorporating exercise content has been shown to improve prediction, the field continues to face a tension between predictive performance and pedagogical transparency. Deep learning based models may achieve strong results, yet their black box character can reduce their usefulness for instructional interpretation (Liang et al., 2022). This has led to increasing interest in feature representations that are not only predictive, but also cognitively and pedagogically meaningful.

Within this context, large language models have emerged as a promising auxiliary tool for feature engineering. Recent studies suggest that LLMs can support the extraction of latent semantic and contextual information from educational text, making it possible to generate features that are more closely aligned with domain specific learning goals (Cen et al., 2025; Chen et al., 2025). Early work integrating LLMs into knowledge tracing frameworks indicates that they can assist in generating question descriptions, identifying knowledge concepts, and enriching representations of task demands, while still functioning as supportive rather than standalone components (Guo et al., 2024; Tan et al., 2024; Yu et al., 2024). Taken together, this literature suggests an important gap: predictive models in mathematics need representations that connect behavioral traces with cognitively meaningful dimensions of proficiency. Building on this gap, the present study investigates whether LLM generated proficiency related features can strengthen predictive modeling and improve the interpretability of mastery prediction in elementary mathematics.

Research question

RQ1. To what extent does incorporating proficiency-strand information improve the predictive performance of the models in mathematics learning?

RQ2. How does the inclusion of cognitive construct-aligned features influence model performance across different proficiency strands?

RQ3. How does incorporating proficiency-strand information enhance the interpretability and instructional usefulness?

Methodology

To support the development of mastery prediction models grounded in educational theory, we employed large language models (LLMs) to systematically label mathematics achievement items with proficiency-related features. Specifically, we focused on the three core strands of mathematical proficiency: Conceptual Understanding, which involves comprehension of mathematical concepts, structures, and relationships; Procedural Fluency, which requires the accurate and efficient application of standard procedures; and Strategic Competence, which engages students in formulating and executing solution strategies for non-routine problems, as defined in foundational educational frameworks for elementary-level mathematics learning (Kilpatrick et al., 2001).

Dataset description

Data were used from the learning datasets containing elementary level mathematics. The XES3G5M¹ corpus comprises 7,652 unique questions (6,142 fill-in-the-blank; 1,510 multiple-choice), each annotated with knowledge-concept (KC) routes that trace the sequence of concepts implicated in solving the item (Liu, Liu, Guo, et al., 2023). The item bank spans 865 KCs organized in a hierarchical concept tree, enabling the modelling of prerequisite relationships and dependencies. The corresponding interaction log includes 5,549,635 student-item records from 18,066 learners. The questions are designed for third-grade level mathematics. The XES3G5M datasets were selected because this study focuses on elementary-level mathematics. These datasets include detailed question metadata essential for assessing cognitive levels, which most publicly available datasets lack. Questions are fundamentally important for assessing students' mathematical knowledge, as they serve as the primary mechanism for cognitive diagnosis, identifying specific learning needs, and informing instructional decisions (Barton, 2018).

LLM-based feature generation

We applied the gpt-4o-mini model to label a corpus of 7,652 mathematics problems collected from a XES3G5M dataset. The model was selected due to its strong instruction-following capability and favorable cost performance trade-off for large-scale annotation tasks. The temperature parameter was set to 0.2 to reduce stochastic variation and encourage consistent outputs across similar inputs. For each problem, the model produced (1) a short explanation describing the reasoning required to solve the task and (2) a corresponding proficiency strand label (CU, PF, or SC). The generated explanation was retained to make the classification process interpretable and to allow subsequent human reviewers to assess whether the reasoning supported the assigned label (see Appendix A). The strands counts are CU (4438), PF (1016), and SC (2198).

Prompt construction

The prompt was designed to operationalize the theoretical definitions of conceptual understanding, procedural fluency, and strategic competence described by Kilpatrick et al. (2001). These definitions were embedded directly in the instructions given to the model so that the classification task remained grounded in established educational theory. To improve reasoning transparency, the prompt also incorporated elements of Chain of Thought prompting (Wei et al., 2022). The model was instructed to provide a short reasoning step before generating the final classification label. Prior work shows that

¹ <https://github.com/ai4ed/XES3G5M>.

prompting models to generate intermediate reasoning can improve performance and interpretability in reasoning tasks (Kojima et al., 2022; Wei et al., 2022). The full prompt specification used in the experiments is provided in Appendix A.

Prompt stability analysis

Because LLM outputs can vary depending on stochastic decoding parameters, a prompt stability analysis was conducted to evaluate the consistency of the labeling procedure. A validation subset of 200 problems was randomly selected from the dataset. The same prompt was applied to this subset under repeated runs using a slightly higher temperature setting of 0.35. Two additional runs were performed to examine whether the generated strand labels remained stable across repeated executions.

For each item, the strand labels generated across the three runs were compared and a consensus score was calculated. The consensus score represents the proportion of runs that produced the same strand label for a given item. The results showed the following distribution: Consensus = 1.0 for 150 items, Consensus = 0.67 for 48 items, Consensus = 0.33 for 2 items. These results indicate that 75 percent of the validation items received identical labels across all runs, while most of the remaining items exhibited only minor disagreement between two labels. Only two items showed complete disagreement across runs. This suggests that the prompt produces relatively stable classifications under moderate stochastic variation. The detailed implementation of the stability analysis procedure is provided in Appendix B.

Human validation of generated labels

Automated labeling using large language models has increasingly been adopted in educational and natural language processing research as a scalable approach when manual annotation is costly or impractical. Because such labels are generated automatically, their reliability should be examined empirically rather than assumed (Chen et al., 2025). Accordingly, a human validation procedure was conducted to evaluate the quality of the LLM-generated proficiency-strand labels. A total of 800 problems was sampled from the full dataset for review. For each problem, annotators were given the problem statement, the model-generated explanation, and the assigned proficiency strand. Two annotators with experience in elementary mathematics education independently judged whether the explanation and assigned label were appropriate for the item. Interrater reliability was assessed using Krippendorff's α^2 , which is appropriate for nominal judgments and agreement analysis (Krippendorff, 2018).

² <https://drive.google.com/drive/folders/1B28A02NxY217PBsvtXwki7X2zBwIXjEQ?hl=en> - supplement file link of Krippendorff's α report.

As shown in Table 1, the overall coefficient was .813. At the strand level, alpha was .819 for Conceptual Understanding, .728 for Procedural Fluency, and .803 for Strategic Competence, indicating substantial agreement overall (Krippendorff, 2018). Because no externally established ground truth was available for proficiency-strand classification, the purpose of the validation process was not to identify a single objectively correct label for each item. Instead, it was used to determine whether the automatically generated labels were sufficiently consistent with expert human judgment to serve as theory-guided reference annotations for subsequent analysis. When disagreements occurred, disputed cases were reviewed through discussion with reference to the strand definitions in order to clarify ambiguous items and refine the annotation criteria. These results provide empirical support for the reliability of the labeling procedure while acknowledging the interpretive nature of proficiency-strand classification (refer Table 1).

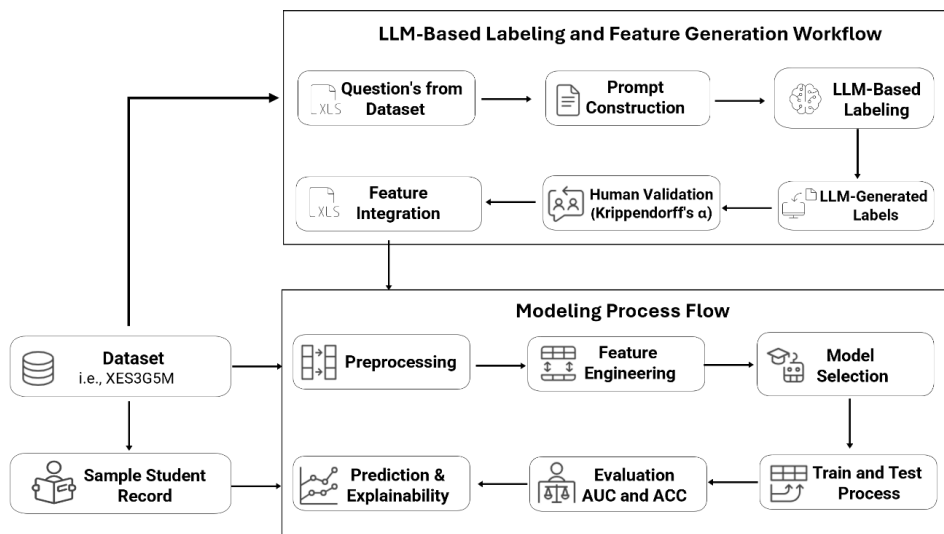
Table 1

Inter-rater reliability report

Strand Group	N Items	P1		P2		Agreements %	Krippendorff's α
		Agree (1)	Disagree (0)	Agree (1)	Disagree (0)		
All Items (Overall)	800	600	200	567	233	92.63%	0.8132
CU	478	337	141	320	158	92.26%	0.8199
PF	129	117	12	112	17	94.57%	0.7281
SC	193	146	47	135	58	92.23%	0.8038

Figure 1

LLM-based feature engineering and overall predictive modeling workflow



Model architecture

We propose the Proficiency-informed Knowledge Tracing (PKT) model, a prior-biased gated transformer architecture for knowledge tracing that captures long-term dependencies in student learning sequences. Unlike standard transformer-based trackers, PKT integrates three complementary mechanisms: (1) a context-sensitive strand gate that dynamically weights proficiency-strand information based on the current question and concept context; (2) a causal mastery module that accumulates per-concept and per-strand correctness histories with temporal decay via a learned forget gate; and (3) a prior-biased causal self-attention mechanism that encodes structural inductive priors, same-concept relevance, same-strand coherence, and recency as additive attention biases with learned weights. Together, these components enable PKT to model both short-term response patterns and long-term knowledge acquisition trajectories across heterogeneous curriculum strands. Given a student interaction sequence $\{(q_t, c_t, s_t, r_t, \Delta t)\}_{t=1}^T$, where q_t denotes the question identifier, c_t the associated knowledge concept, s_t the proficiency-strand, $r_t \in \{0,1\}$ the binary response, and Δt the elapsed time since the previous interaction, the model predicts the probability of a correct response to the next question.

Figure 2

The PKT model architecture.

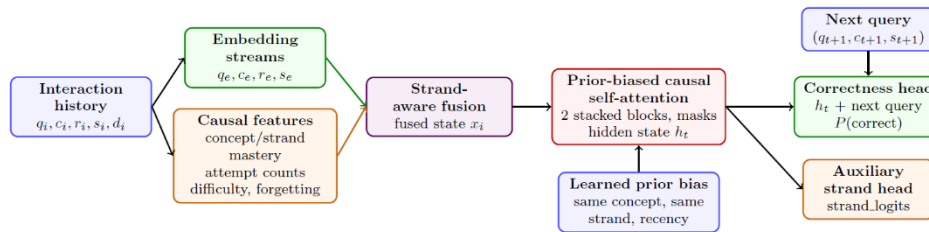
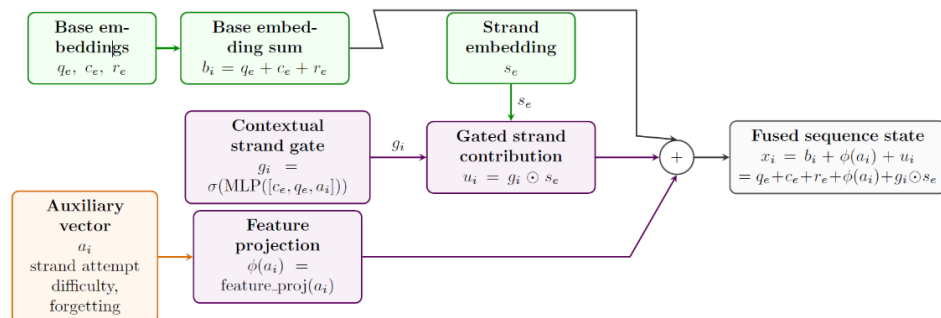


Figure 3

Strand-Aware fusion mechanism



Input representations

Each interaction at time step t is represented by four embedding components. The concept, question, and response embeddings are obtained as:

$$\mathbf{e}_t^c = E_c(c_t), \mathbf{e}_t^q = E_q(q_t), \mathbf{e}_t^r = E_r(r_t)$$

where $E_c \in \mathbb{R}^{(C+1) \times d}$, $E_q \in \mathbb{R}^{(Q+1) \times d}$, and $E_r \in \mathbb{R}^{2 \times d}$ are learnable embedding matrices, d is the model dimension, and padding index 0 is reserved. The strand embedding is projected into the model space and modulated by a context-sensitive gate:

$$\begin{aligned} \mathbf{e}_t^s &= g_t \cdot W_s E_s(s_t) \\ g_t &= \sigma \left(\text{MLP}_{\text{gate}}([\mathbf{e}_t^c; \mathbf{e}_t^q; \phi_t]) \right) \in \mathbb{R} \end{aligned}$$

where $E_s \in \mathbb{R}^{S \times d_s}$ is the strand embedding matrix, $W_s \in \mathbb{R}^{d \times d_s}$ is a linear projection, and ϕ_t denotes the causal mastery feature vector defined in Section 3.3.

Causal mastery features

To prevent data leakage, all mastery statistics are computed causally, that is, using only the responses observed prior to step t . Let n_t^c and k_t^c denote the number of attempts and correct responses for concept c_t before step t , and let n_t^{cs} and k_t^{cs} denote the corresponding counts for the joint concept-strand pair (c_t, s_t) . The raw mastery estimates are:

$$\hat{m}_t^c = \begin{cases} k_t^c/n_t^c & \text{if } n_t^c > 0 \\ 0.5 & \text{otherwise} \end{cases}, \hat{m}_t^{cs} = \begin{cases} k_t^{cs}/n_t^{cs} & \text{if } n_t^{cs} > 0 \\ 0.5 & \text{otherwise} \end{cases}$$

A learned forget gate models temporal decay of mastery using the log-transformed time interval:

$$f_t = \sigma, (w_f \log(1 + \Delta t) + b_f)$$

The adjusted mastery features are then:

$$m_t^c = f_t \cdot \hat{m}_t^c, m_t^{cs} = f_t \cdot \hat{m}_t^{cs}$$

The full causal mastery feature vector is:

$$\phi_t = [m_t^c, \log(1 + n_t^c), m_t^{cs}, \log(1 + n_t^{cs}), \delta_{q_t}, f_t] \in \mathbb{R}^6$$

where $\delta_{q_t} = \sigma(E_\delta(q_t)) \in \mathbb{R}$ is a learned per-question difficulty scalar. This vector is projected into the model dimension via a two-layer MLP with LayerNorm and GELU activation:

$$\mathbf{e}_t^{\text{aux}} = \text{LN} \left(\text{GELU}(W_\phi \phi_t + b_\phi) \right) \in \mathbb{R}^d$$

Input token construction

The final input representation for each time step is formed by summing all embedding components, followed by LayerNorm and dropout:

$$\mathbf{x}_t = \text{Dropout} \left(\text{LN}(\mathbf{e}_t^c + \mathbf{e}_t^q + \mathbf{e}_t^r + \mathbf{e}_t^{\text{aux}} + g_t \cdot W_s E_s(s_t)) \right)$$

Invalid (padding) positions are zeroed out via an element-wise validity mask $\mathbf{v} \in \{0,1\}^T$:

$$x_t \leftarrow v_t \cdot x_t$$

Prior-biased causal self-attention

The core sequence encoder consists of L stacked transformer blocks, each incorporating a prior-biased causal multi-head self-attention mechanism. Prior bias matrix. An additive attention bias $B \in \mathbb{R}^{T \times T}$ is constructed to inject structural inductive priors without supervision:

$$B_{ij} = w_c \cdot 1[c_i = c_j, c_i > 0] + w_s \cdot 1[s_i = s_j, s_i > 0] + w_r \cdot (-\log(1 + \max(i - j, 0)))$$

where $w_c = \text{softplus}(\alpha_c)$, $w_s = \text{softplus}(\alpha_s)$, and $w_r = \text{softplus}(\alpha_r)$ are learned scalar weights initialized at 0.50, 0.25, and 0.15, respectively. The bias rewards attention to same-concept interactions, same-strand interactions, and more recent history.

Attention computation. For each attention head h with projections $W_Q^h, W_K^h, W_V^h \in \mathbb{R}^{d \times d_h}$ where $d_h = d/H$:

$$\begin{aligned} Q^h &= XW_Q^h, K^h = XW_K^h, V^h = XW_V^h \\ A^h &= \text{softmax} \left(\frac{Q^h K^{h\top}}{\sqrt{d_h}} + B + M_{\text{causal}} + M_{\text{pad}} \right) \end{aligned}$$

where M_{causal} applies $-\infty$ to all future positions (upper triangular) and M_{pad} masks invalid key positions. The multi-head output is:

$$\text{MHA}(X) = \text{Concat}(A^1 V^1, \dots, A^H V^H) W_O$$

Block residuals. Each block applies pre-norm residual connections:

$$\begin{aligned} X' &= X + \text{MHA}(\text{LN}(X)) \\ H &= v_t \cdot (X' + \text{FFN}(\text{LN}(X'))) \end{aligned}$$

where FFN is a two-layer feed-forward network with inner dimension $4d$ and GELU activation. Padding positions are re-zeroed after each block.

Prediction heads

Next-query projection. To condition prediction on the forthcoming question, the next question, concept, and strand embeddings are concatenated and projected:

$$q_t^* = \text{GELU} \left(W_{nq} [E_q(q_{t+1}); E_c(c_{t+1}); W_s E_s(s_{t+1})] \right) \in \mathbb{R}^d$$

Main prediction head. The predicted probability of a correct response is:

$$\hat{p}_t = \sigma \left(\text{MLP}_{\text{pred}}([h_t; q_t^*]) \right) \in (0,1)$$

where h_t is the final encoder hidden state at position t after the last LayerNorm.

Auxiliary strand head. An auxiliary linear classifier operates on the hidden state to predict the strand label, providing an additional training signal:

$$\hat{s}_t = W_{\text{aux}} h_t \in \mathbb{R}^{|S|}$$

Settings and implementation details

The proposed PKT model was implemented in PyTorch and trained on the XES3G5M dataset using identical splits and preprocessing as baselines. The architecture uses 128-dimensional embeddings, 4 attention heads, 2 transformer blocks, and 0.10 dropout. Training used Adam (learning rate $1e-3$) for up to 100 epochs with early stopping (patience=5). Loss combined binary cross-entropy for response prediction with auxiliary cross-entropy for strand prediction (weight=0.10). Batch sizes: 64 (train) / 128 (eval), gradient clip norm: 1.0.

Baseline knowledge tracing models were implemented using the official pyKT toolkit (Liu et al., 2022) and trained under identical XES3G5M data splits and preprocessing procedures. Model configurations followed pyKT specifications, with comparable hyperparameters where applicable. Training optimized learner response prediction using masked binary cross-entropy, Adam optimization, validation-AUC-based early stopping, and gradient clipping. Predictive performance on held-out learning sequences was evaluated using AUC, ACC, MAE, and RMSE.

Base models

To evaluate the effectiveness of PKT, state of art baseline models were employed for comparison. The baseline models are described as follows.

Deep Knowledge Tracing (DKT)

DKT was the first model to introduce deep learning techniques into knowledge tracing (Piech et al., 2015). It is implemented using recurrent neural networks (RNNs), specifically Long Short-Term Memory (LSTM) networks. The model estimates students' knowledge states by modeling their learning processes over time and predicts their mastery of questions and related knowledge concepts.

Attentive Knowledge Tracing (AKT)

AKT utilizes an attention mechanism to characterize the temporal relationships between questions and students' historical interactions (Ghosh et al., 2020).

Dynamic Key-Value Memory Network (DKVMN)

DKVMN is a memory network-based knowledge tracing model (Zhang et al., 2017). It employs a static key matrix to represent latent knowledge concepts and a dynamic value matrix to store students' knowledge states. Student knowledge states are updated through read and write operations between these matrices.

DTransformer

DTransformer is a two-layer Transformer-based framework designed to enhance model stability through contrastive learning (Yin et al., 2023). The model introduces a novel architecture for capturing students' learning activity patterns and improves generalization by leveraging contrastive learning techniques.

Self-Attentive Knowledge Tracing (SAKT)

SAKT is the first knowledge tracing model that applies the Transformer architecture to the task (Pandey & Karypis, 2019). It utilizes a self-attention mechanism to capture long-term dependencies in students' learning histories, which improves the accuracy of knowledge state prediction.

Sparse Knowledge Tracing (sparseKT)

SparseKT is an extension of the SAKT model that introduces a k-sparse attention mechanism (Huang et al., 2023). It applies soft-thresholding and top-K sparsification strategies to select the most relevant historical interactions. Additionally, it incorporates question discrimination factors to capture variations among questions associated with the same knowledge concept.

simpleKT

SimpleKT is a lightweight yet powerful baseline model built upon a standard dot-product attention mechanism (Liu, Liu, Chen, et al., 2023). Inspired by the Rasch model from psychometrics, it explicitly models question-specific variations to capture differences among questions associated with the same knowledge concept.

This comparative implementation allows us to evaluate how the integration of proficiency-informed model performed with diverse architectures. In this study, we evaluated learner performance prediction using established learning analytics metrics (Li et al., 2024). ACC and AUC assessed binary correctness prediction, with 0.5 indicating chance-level performance, while MAE and RMSE quantified deviations between probabilistic learner-knowledge estimates and observed response outcomes.

Result

This section presents the evaluation outcomes of the deep learning-based models, focusing on three key comparative dimensions: (i) performance of the proposed Proficiency-informed Knowledge Tracing (PKT) model relative to baseline models, (ii) strand level analysis, (iii) calibration analysis, (iv) visualization analysis. The results are summarized in the following Tables.

Overall predictive performance

Table 2 summarizes the overall predictive performance of the proposed PKT model and the comparison knowledge tracing models on the validation and test sets. Across the validation set, PKT achieved an AUC of .823 and an ACC of .830. These values were close to those of the strongest-performing baseline models, particularly AKT (AUC = .825, ACC = .830), sparseKT (AUC = .825, ACC = .830), and simpleKT (AUC = .824, ACC = .830). This pattern suggests that PKT demonstrated stable predictive performance during model validation and performed comparably to several established knowledge tracing approaches.

Table 2

Performance comparison of baseline models with proposed model on the valid and test dataset

Model	Validation dataset		Test dataset			
	AUC	ACC	AUC	ACC	MAE	RMSE
AKT	0.825	0.830	0.827	0.831	0.234	0.347
PKT	0.823	0.830	0.825	0.830	0.233	0.348
sparseKT	0.825	0.830	0.821	0.828	0.243	0.349
simpleKT	0.824	0.830	0.821	0.828	0.234	0.350
DTransformer	0.820	0.828	0.820	0.828	0.236	0.350
SAKT	0.802	0.820	0.798	0.819	0.243	0.359
DKT	0.789	0.818	0.789	0.819	0.253	0.361
DKVMN	0.787	0.818	0.787	0.818	0.254	0.361

On the test dataset, PKT obtained an AUC of .825 and an ACC of .830. Its performance was slightly below that of AKT (AUC = .827, ACC = .831), but modestly higher than that of sparseKT and simpleKT (AUCs = .821, ACCs = .828), DTransformer (AUC = .820, ACC = .828), SAKT (AUC = .798, ACC = .819), DKT (AUC = .789, ACC = .819), and DKVMN (AUC = .787, ACC = .818). These results indicate that PKT remained among the better-performing models for predicting learners' future response correctness, although the differences between PKT and the strongest baseline were small. With respect to prediction

error, PKT produced the lowest MAE on the test set ($MAE = .233$) and an RMSE of .348, which was close to the best observed RMSE (AKT = .347). From a learning analytics perspective, this finding suggests that PKT provided comparatively accurate probability estimates of learner performance, even though its classification-based indices were not the highest in all cases.

Taken together, the results suggest that PKT achieved performance that was broadly comparable to the strongest baseline models. Although PKT did not outperform all comparison models on every metric, it showed a consistent and balanced pattern across discrimination, classification accuracy, and prediction error, with the most favorable MAE among the models examined.

Strand-level analysis

To examine whether model performance varied across dimensions of mathematical proficiency, strand-level results for Procedural Fluency (PF), Conceptual Understanding (CU), and Strategic Competence (SC) were analyzed (see Table 3). These strands represent related but distinct aspects of learners' mathematical development; therefore, strand-level analysis helps clarify how knowledge tracing models capture different forms of learning performance.

Across models, a consistent pattern emerged. PF generally showed the highest ACC and the lowest MAE and RMSE, suggesting that learner responses in procedural tasks were more predictable. CU showed intermediate performance, whereas SC tended to yield lower ACC and higher error values, indicating greater difficulty in modeling strategic performance. Because this pattern appeared across PKT and the comparison models, the observed variation likely reflects differences in task demands rather than model-specific behavior alone.

For PF, PKT achieved an AUC of .808 and an ACC of .852, with MAE of .212 and RMSE of .333. These results were close to AKT (AUC = .810, ACC = .851, MAE = .213, RMSE = .332), indicating comparable performance in tracking procedural fluency. For CU, PKT obtained an AUC of .825, ACC of .832, MAE of .232, and RMSE of .347, again closely matching AKT (AUC = .827, ACC = .833, MAE = .233, RMSE = .346), while showing a slightly lower MAE. For SC, PKT yielded an AUC of .828, ACC of .816, MAE of .248, and RMSE of .360. Although AKT showed slightly stronger discrimination and accuracy in this strand (AUC = .832, ACC = .819), PKT produced a marginally lower MAE (.248 vs. .249).

Overall, the strand-level results suggest that PKT remained broadly comparable to the strongest baseline across PF, CU, and SC. Differences between PKT and AKT were small across most metrics, whereas both models performed less strongly on SC than on PF and CU. From a learning science perspective, this pattern is theoretically plausible: procedural

fluency often involves more structured and repeated task formats, whereas strategic competence requires higher-order reasoning, flexible problem solving, and adaptive selection of solution strategies. Thus, learner performance associated with strategic competence may be less readily captured by current knowledge tracing approaches than performance associated with procedural fluency.

Table 3

Strand-based performance comparison of baseline models with proposed model on the test dataset

Model	PF				CU				SC			
	AUC	ACC	MAE	RMSE	AUC	ACC	MAE	RMSE	AUC	ACC	MAE	RMSE
AKT	0.810	0.851	0.213	0.332	0.827	0.833	0.233	0.346	0.832	0.819	0.249	0.358
PKT	0.808	0.852	0.212	0.333	0.825	0.832	0.232	0.347	0.828	0.816	0.248	0.360
sparseKT	0.805	0.850	0.220	0.334	0.822	0.830	0.242	0.348	0.825	0.814	0.259	0.361
simpleKT	0.805	0.850	0.213	0.334	0.821	0.830	0.233	0.349	0.825	0.814	0.250	0.362
DTransformer	0.806	0.850	0.214	0.334	0.820	0.830	0.236	0.349	0.824	0.815	0.252	0.362
SAKT	0.784	0.843	0.219	0.342	0.799	0.822	0.241	0.358	0.799	0.802	0.262	0.374
DKT	0.784	0.846	0.226	0.340	0.786	0.820	0.254	0.361	0.794	0.804	0.268	0.374
DKVMN	0.783	0.846	0.228	0.340	0.784	0.819	0.254	0.361	0.793	0.803	0.270	0.361

In terms of model comparison, AKT again demonstrates the most consistent performance across all strands, achieving the highest AUC and ACC in CU and SC, and maintaining strong results in PF. PKT remains highly competitive, particularly excelling in terms of MAE across strands, which indicates more precise estimation of student performance at a granular level. Importantly, the performance gap between models becomes more pronounced in the SC strand. Even the best-performing models show noticeable degradation in this domain, suggesting that modeling higher-order cognitive processes remains a significant challenge. This highlights the need for future work to better capture complex learning behaviors beyond procedural knowledge.

Calibration analysis

Beyond predictive accuracy, it is critical to examine how well model predictions align with actual student performance. To this end, we analyze calibration using the mean predicted probability (mean_prob), the mean observed correctness (mean_true), and their difference (Diff). A well-calibrated model should produce predictions that closely match the true

distribution, resulting in a gap close to zero. As shown in Table 4, calibration was examined and the gap between these values. A smaller absolute gap indicates closer correspondence between predicted and observed learner performance. In addition, Table 5 presents an overall descriptive ranking of the models based on predictive performance and calibration.

Table 4

Strand-based performance comparison of baseline models and calibration

Model	PF			CU			SC		
	M_prob	M_true	Diff	M_prob	M_true	Diff	M_prob	M_true	Diff
AKT	0.837	0.829	+0.0079	0.804	0.799	+0.0052	0.778	0.768	+0.0095
PKT	0.834	0.829	+0.0048	0.804	0.799	+0.0053	0.776	0.768	+0.0082
sparseKT	0.827	0.829	-0.0019	0.794	0.799	-0.0047	0.768	0.768	-0.0005
simpleKT	0.840	0.829	+0.0106	0.812	0.799	+0.0136	0.783	0.768	+0.0149
DTransformer	0.842	0.829	+0.0131	0.811	0.799	+0.0127	0.785	0.768	+0.0170
SAKT	0.841	0.829	+0.0117	0.809	0.799	+0.0099	0.781	0.768	+0.0131
DKT	0.834	0.829	+0.0049	0.806	0.799	+0.0073	0.787	0.768	+0.0186
DKVMN	0.831	0.829	+0.0015	0.804	0.799	+0.0053	0.784	0.768	+0.0153

Across models, Table 4 shows a general pattern of mild overestimation. For most models, the gap values were positive across PF, CU, and SC, indicating that predicted probabilities were slightly higher than the corresponding observed correctness rates. This tendency was most apparent for simpleKT and DTransformer, both of which showed comparatively larger positive gaps across all three strands. By contrast, Table 4 indicates that sparseKT produced gap values closest to zero, with slightly negative gaps across strands.

Table 4 also shows that PKT had relatively small and stable calibration gaps across PF, CU, and SC. In PF, PKT yielded $M_{\text{prob}} = .834$ and $M_{\text{true}} = .829$, resulting in a gap of $+0.0048$. In CU, the corresponding values were $.804$ and $.799$, with a gap of $+0.0053$. In SC, PKT produced $M_{\text{prob}} = .776$ and $M_{\text{true}} = .768$, with a gap of $+0.0082$. Relative to AKT, Table 4 suggests that PKT showed slightly smaller gaps in PF and SC, while the two models were nearly identical in CU.

At a descriptive level, Table 5 shows that PKT ranked first overall when predictive performance and calibration were considered together, followed by AKT and sparseKT. This ranking should be interpreted cautiously because rank-based summaries are descriptive rather than inferential. Still, Table 5 provides a compact summary of the broader pattern observed across the preceding analyses.

Table 5

Overall ranking of knowledge tracing models based on performance and calibration

Model	AUC Rank	ACC Rank	MAE Rank	RMSE Rank	Perf Avg Rank	Calib Rank	Final Avg Rank	Overall Rank
PKT	2	2	1	2	1.75	2	1.88	1
AKT	1	1	2	1	1.25	3	2.13	2
sparseKT	3	3	6	3	3.75	1	2.38	3
simpleKT	4	4	3	4	3.75	7	5.38	4
SAKT	6	6	5	6	5.75	6	5.88	5
DKT	7	7	7	7	7.00	5	6.00	6
DKVMN	8	8	8	8	8.00	4	6.00	7
DTransformer	5	4	4	4	4.25	8	6.13	8

Taken together, the results highlight several important insights. First, AKT achieves the best overall predictive performance, making it a strong choice when accuracy is the primary objective. Second, PKT provides the best balance between accuracy and calibration, combining low prediction error with stable and reliable probability estimates. Third, while sparseKT excels in calibration, its predictive performance is comparatively weaker, indicating a trade-off between these two aspects. More broadly, the findings emphasize that predictive accuracy alone is insufficient for educational applications. Models must also provide well-calibrated estimates to support reliable interpretation of student knowledge states. This is particularly critical for higher-order cognitive skills, where both prediction and calibration remain challenging. These results suggest that model selection should consider not only predictive performance but also explainability and deployment feasibility within educational systems.

PKT strand-feature ablation and permutation-control analysis

To examine whether proficiency-strand features provided calibration-relevant information, we compared three PKT conditions: no strand input, randomly shuffled strand labels, and true strand labels. The no-strand condition served as an ablation baseline, whereas the shuffled condition served as a permutation control. The shuffled condition preserved the overall distribution of strand labels but removed the true relationship between each item and its strand.

Table 6

Calibration and metrics by strand features and ablation condition

Strand	Condition	ECE M (SD)	Max gap M (SD)	Brier M (SD)	N
PF	None	.010 (.004)	.069 (.030)	.111 (.000)	144,168
	Shuffled	.012 (.003)	.070 (.004)	.111 (.000)	
	True	.011 (.004)	.079 (.009)	.111 (.000)	
CU	None	.010 (.002)	.049 (.028)	.121 (.000)	681,521
	Shuffled	.011 (.003)	.049 (.002)	.120 (.000)	
	True	.010 (.004)	.041 (.010)	.120 (.000)	
SC	None	.012 (.003)	.041 (.024)	.129 (.000)	273,508
	Shuffled	.014 (.003)	.039 (.006)	.129 (.000)	
	True	.010 (.006)	.032 (.012)	.129 (.000)	

Note. PF = Procedural Fluency; CU = Conceptual Understanding; SC = Strategic Competence. None = model without strand input; Shuffled = strand labels randomly permuted across items, preserving marginal frequencies; True = correct strand assignments. ECE = Expected Calibration Error with 10 equal-mass bins; Max Gap = maximum per-bin calibration gap. Lower values are better for ECE, Max Gap, and Brier. Cells report mean (standard deviation) across random seeds. Bolded values indicate the best-performing condition where the true-vs-shuffled difference exceeded one SD of the true condition.

As shown in Table 6, Brier scores were nearly unchanged across conditions, suggesting that strand information had little descriptive effect on overall prediction error. Calibration metrics, however, showed some strand-specific differences. Relative to the shuffled condition, the true-strand condition showed lower ECE for CU and SC, with reductions of approximately 10% and 25%, respectively. The maximum calibration gap was also lower for CU and SC, with reductions of approximately 16% and 19%, respectively. In contrast, PF showed no consistent calibration advantage: ECE was slightly lower under the true-strand condition, but the maximum calibration gap was higher.

These results suggest that proficiency-strand identity carries calibration-relevant structural information for conceptual and strategic items, but less so for procedural items. In other words, the true strand labels appear to help PKT produce more reliable probability estimates for mathematical knowledge, especially CU and SC. Following the evaluation of predictive performance, we further examined the role of proficiency-strand features through visualizations. Whether the visualizations provide deeper insight into how proficiency-level information contributes to the interpretability and effectiveness of the proposed PKT model (see fig 4, 5 and 6).

Visualization of the effectiveness of incorporating proficiency-strands

Integrating proficiency strand variables resulted in a notable shift in model behavior. Rather than relying solely on response sequences, the models began to condition their

predictions on the learner's historical performance within a specific strand (see figure 4, 5, and 6).

Figures 4 and 5 provide an illustrative learner-level comparison of AKT and PKT for UID 5187. In figure 4, the learner completed 444 attempts and answered 228 correctly (51.4%). Both models broadly followed the temporal pattern of performance, suggesting that each captured general variation in the learner's estimated proficiency over time. For this case, however, PKT was more closely aligned with the observed outcomes, with higher accuracy (70.9% vs. 64.0%) and lower mean absolute error (MAE = .384 vs. .417). The strand-level summaries in figure 4 further show lower prediction error for PKT in Procedural Fluency (PF: .400 vs. .425), Conceptual Understanding (CU: .362 vs. .408), and Strategic Competence (SC: .421 vs. .433). The largest reduction was observed in CU. Figure 4 also indicates that most attempts were concentrated in CU ($n = 251$), followed by SC ($n = 105$) and PF ($n = 88$). In the 30-attempt case window shown in figure 5 and 6, AKT and PKT produced broadly similar concept-level patterns, but PKT additionally provided an aggregated strand-level profile. Within this slice, the strand summary indicated lower observed performance in CU (33%) than in PF (50%) and SC (62%). These visualizations are best interpreted as descriptive evidence that PKT may support more interpretable strand-level profiling while remaining comparable to AKT at the concept level.

Figure 4

Longitudinal performance of student UID 5187

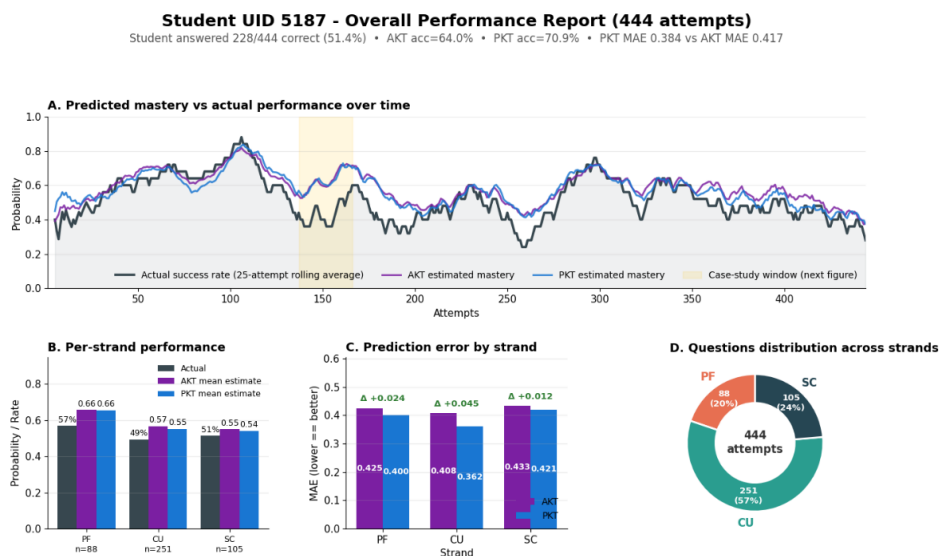


Figure 5
Concept and strand mastery states for 30 attempts

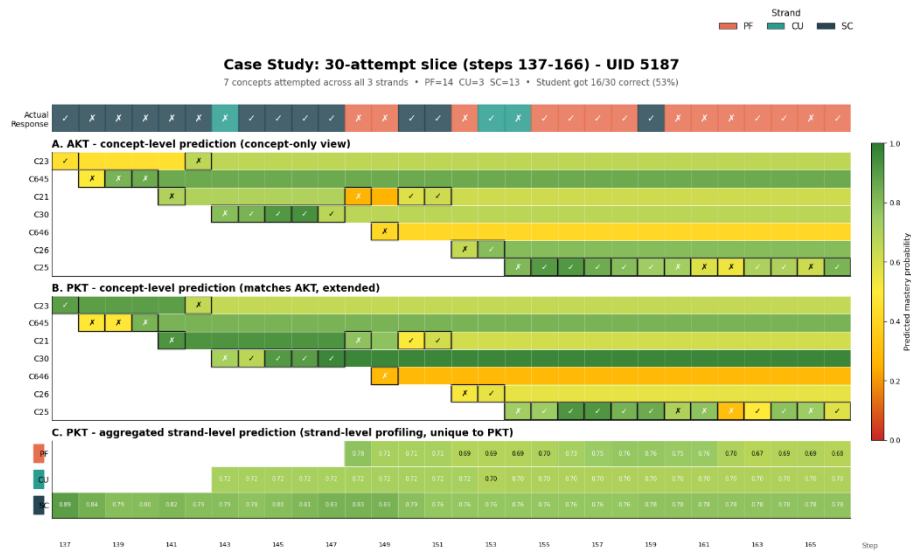
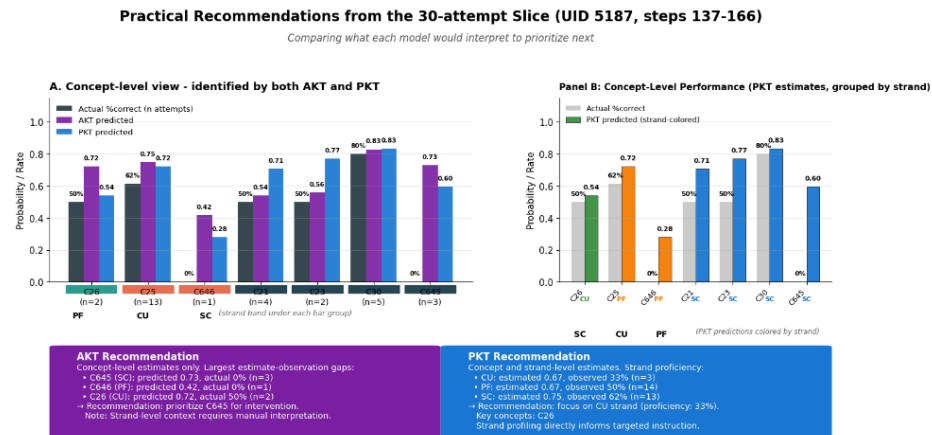


Figure 6
Actionable guidance comparison



Discussion

This study set out to explore how predictive models can be enhanced to support mastery-oriented instruction in mathematics learning. The findings suggest that such integration is not only effective in improving predictive accuracy but also advances the broader instructional value of prediction systems.

Addressing RQ1: Enhancing predictive models for mastery-oriented mathematics instruction

This study examined whether incorporating proficiency-strand information could strengthen predictive models for mastery-oriented mathematics instruction. Overall, the findings suggest that the inclusion of strand-level features was associated with a more educationally meaningful representation of learner performance, while maintaining competitive predictive performance relative to established knowledge tracing baselines (Huang & Liu, 2025; Ozyurt et al., 2025). In particular, the proposed PKT model did not uniformly outperform all comparison models on every metric, but it showed a consistent pattern of lower prediction error, especially in terms of MAE, and stable performance across overall, strand-level, and calibration analyses.

From a modeling perspective, the inclusion of proficiency-strand information enabled prediction to be linked more closely to pedagogically relevant dimensions of mathematics learning. (Liu et al., 2021), whereas strand-informed modeling introduces a broader representation of learner performance across related domains of proficiency. In the present study, this additional information appeared to support more precise probability estimation, as reflected in PKT's comparatively favorable MAE values and its relatively small calibration gaps across strands. These results suggest that strand-aligned features may help capture aspects of learning progression that are less visible in concept-only representations.

Importantly, the value of this approach should not be interpreted solely in terms of predictive improvement. (Li et al., 2025). The visual analyses in this study suggest that proficiency-strand information may help organize predictions into profiles that are more readily connected to instructional decision-making, such as identifying relative strengths and weaknesses across Procedural Fluency, Conceptual Understanding, and Strategic Competence. This does not by itself establish instructional effectiveness, but it does indicate that the model outputs may be more interpretable for formative use.

The findings point to a modest but meaningful shift in the role of predictive modeling in mathematics learning. Rather than treating prediction as an end in itself, the integration of proficiency-strand information supports a view of knowledge tracing as a tool for representing learner development in ways that are more consistent with mastery-oriented instruction. Accordingly, success in educational predictive modeling may need to be

considered not only in terms of discrimination and accuracy, but also in terms of calibration, interpretability, and alignment with meaningful learning constructs.

Addressing RQ2: The role of cognitive construct-aligned features across proficiency strands

In addressing RQ2, this study examined whether cognitive construct-aligned features contributed to variation in model performance across different proficiency strands. In the present work, these features were operationalized through proficiency-strand representations that were intended to reflect meaningful dimensions of mathematical learning, rather than relying only on response sequences or concept tags. The findings suggest that such features were associated with a more differentiated account of learner performance across strands, although the gains were not uniform across all metrics or all models.

A consistent pattern emerged across the strand-level analyses. PF was generally associated with higher accuracy and lower prediction error, whereas SC showed lower accuracy and larger error values. CU tended to fall between these two strands. This pattern was observed across models, suggesting that differences in predictive performance were shaped not only by model architecture, but also by the nature of the underlying mathematical proficiency being modeled. From a learning sciences perspective, this result is plausible, as procedural tasks are often more constrained and predictable, whereas tasks involving strategic competence may require more flexible reasoning and multi-step coordination.

Within this context, the inclusion of construct-aligned strand features appears to have supported more precise estimation of learner performance, particularly in error-based metrics. PKT did not outperform all comparison models on every strand-level index, but it remained competitive across PF, CU, and SC and showed consistently lower MAE than AKT in the strand-level analysis. This pattern suggests that construct-aligned features may be especially useful for refining probability estimates, even when improvements in AUC or ACC are modest. Such findings are consistent with prior arguments that educational prediction may benefit from features that better reflect the cognitive and curricular structure of learning tasks (Zhang et al., 2025).

At the same time, these results should be interpreted cautiously. The strand-level differences do not establish that construct-aligned features fully explain learner performance, nor do they imply that such features resolve the challenges of modeling more complex forms of mathematical thinking. Rather, the findings indicate that incorporating cognitively meaningful representations may improve how learner performance is organized and estimated across strands. In this sense, the contribution of construct-aligned features

appears to lie not only in predictive performance, but also in their potential to make model outputs more closely aligned with educationally relevant dimensions of learning.

Addressing RQ3: Interpretability and instructional usefulness of proficiency-strand information

In addressing RQ3, this study examined whether incorporating proficiency-strand information enhanced the interpretability and potential instructional usefulness of knowledge tracing outputs. The findings suggest that strand-informed modeling contributed to a more structured representation of learner performance by organizing predictions in relation to pedagogically meaningful dimensions of mathematics learning. Compared with concept-only representations, this additional layer did not simply indicate whether a learner was likely to answer correctly, but also situated that prediction within a broader proficiency profile.

This contribution was most evident in the strand-level analyses, calibration results, and visualizations. The strand-based results showed that model performance varied systematically across PF, CU, and SC, making visible differences in how learners' performance was distributed across dimensions of mathematical proficiency. The calibration analysis further suggested that PKT produced relatively stable probability estimates across strands, which is relevant when model outputs are interpreted as indicators of learner readiness or developing mastery. In the case-based visualizations, the inclusion of strand information made it easier to identify relative strengths and weaker areas within a learner's profile, especially when concept-level patterns alone were less immediately interpretable.

At the same time, the instructional usefulness of these outputs should be interpreted cautiously. The present study did not directly evaluate teachers' use of the model, classroom decision-making, or intervention outcomes. Accordingly, the findings do not demonstrate instructional effectiveness in a causal sense. Rather, they suggest that proficiency-strand information may improve the extent to which model outputs are aligned with curricular constructs and therefore more readily interpretable for formative purposes. In this sense, the main contribution of strand-aware modeling lies not only in prediction, but also in making learner performance more visible in ways that may better support mastery-oriented diagnosis and instructional reflection.

Conclusion

This study examines how predictive models can better support mastery-oriented mathematics instruction by using proficiency-strand information. Specifically, unlike models that rely mainly on concept-level or response-sequence data, which do not capture broader proficiency aspects, this work introduces strand-informed representations based on

mathematical proficiency, including conceptual understanding, procedural fluency, and strategic competence, as described by Kilpatrick et al. (2001). While these dimensions are central to decisions in mathematics learning and teaching, many knowledge-tracing approaches do not explicitly represent them (Park et al., 2026).

Large language models assisted with feature engineering by generating strand-aligned labels from assessment metadata. This process reduced manual coding, while human review of label checks maintained context and trustworthiness (Qian et al., 2026). Within this framework, the Proficiency-informed Knowledge Tracing (PKT) model combines sequence-based modeling with strand-level inputs to generate predictions aligned with meaningful learner proficiency dimensions. These findings show that adding proficiency-strand information strengthens the educational relevance of knowledge tracing. PKT provides not only better predictive performance but also profiles learner performance in a way that is more interpretable for mastery-oriented instruction. Instead of predicting correctness alone, strand-informed modeling lets instructors consider student performance across broader domains of mathematical proficiency. This supports framing predictive modeling as a means to more consistently align learner growth with instructional goals and the knowledge-skill gap discussed in educational research (e.g., Simpson & Wang, 2023).

Recommendations

Based on the findings, several recommendations are proposed for researchers, practitioners, and developers. Predictive models in education should align with cognitive theories that reflect domain-specific learning constructs, such as mathematical proficiency strands (Koedinger et al., 2012). While Kilpatrick et al. (2001) established a foundation for mathematics, similar frameworks exist in other areas, including scientific reasoning and reading comprehension. This theoretical alignment ensures that model outputs remain both accurate and instructionally meaningful.

Large language models can efficiently automate feature generation, but integrating human oversight within these systems ensures that predictions remain contextually grounded, reliable, and ethical. Researchers should also move beyond behavioral features derived solely from interaction logs. Features that represent cognitive processes and instructional goals, as demonstrated through proficiency strands, can enrich interpretation and improve learner support. Finally, predictive systems should be designed to assist rather than confuse instructors. Models that offer transparent, theory-driven explanations can enhance educator trust, promote adoption, and strengthen instructional decision-making.

Limitations and future work

Several limitations should be acknowledged. The study was conducted using a single elementary mathematics dataset, XES3G5M, with relatively well-structured item metadata,

which supported strand-based feature construction but may limit transferability to less richly documented settings. Although the model produced interpretable outputs, the study did not directly examine teacher use, classroom decision-making, or instructional outcomes. In addition, inter-rater agreement, while acceptable overall, varied across strands, suggesting that some proficiency categories may be more difficult to label consistently. Moreover, although PKT performed competitively relative to the selected baselines, formal statistical significance tests across models were not conducted. Therefore, the observed performance differences should be interpreted cautiously rather than as statistically validated effects. Future work should examine additional datasets, involve more human validators, apply appropriate paired significance tests across repeated runs, and evaluate strand-aware modeling in authentic instructional settings. Extending the framework to domains such as science or language arts would further clarify the generalizability of theory-aligned predictive modeling.

Appendix A

Full prompt specification and example input/output

This appendix provides the complete prompt used for the LLM-based labeling procedure together with an example of the input and output format.

Table A1. Full prompt specification used for LLM-based strand labeling	
Component	Content
Purpose	Classify mathematics problems into one of three proficiency strands (CU, PF, SC) based on Kilpatrick et al. (2001).
Model Role	Expert mathematics educator
Input Fields	Question ID ({qid}), Question text in Chinese ({question})
Output Format	Structured JSON with keys: "Question ID", "Reasoning (English)", "Strand"
Full Prompt Text	You are an expert mathematics educator. You will analyze math problems written in Chinese and classify them into one of three proficiency strands. 1. CU (Conceptual Understanding): Understanding mathematical concepts, properties, and relationships. 2. PF (Procedural Fluency): Carrying out mathematical procedures accurately and efficiently using known rules or algorithms. 3. SC (Strategic Competence): Formulating and solving non-routine or multi-step problems by selecting or creating effective strategies. For the question below: 1. Interpret the meaning of the mathematical task. 2. Provide a short reasoning in English explaining the type of thinking required. 3. Output the final strand label. Output format must be JSON with keys: Question ID, Reasoning (English), Strand Question ID: {qid} Question (Chinese): {question}
Note	The complete implementation code for label generation and parsing is available in the project repository.

Table A2. Example input and output of the LLM labeling procedure	
Field	Content
Question ID	1342
Question (Chinese)	学校有舞蹈、唱歌、围棋、绘画四种兴趣班。小宇、小明、小丽三个小朋友准备报名，每人只能报一个兴趣班。请问共有多少种不同的报名方式。
Model Output - Reasoning (English)	The problem requires applying a counting rule to determine the number of possible combinations when each student selects one activity. The solution mainly involves applying a standard combinational procedure.
Model Output - Strand Label	PF (Procedural Fluency)
Note	The reasoning explanation was retained for human validation to allow reviewers to assess consistency between the label and the model's justification.

Appendix B

Prompt stability analysis procedure

This appendix describes the procedure used to evaluate the stability of the prompt-based labeling process. The goal was to examine whether the strand label assigned to a problem remains consistent across repeated runs of the same prompt under moderate stochastic variation.

Table B1. Prompt stability analysis procedure	
Component	Description
Objective	LLM outputs may vary due to stochastic decoding parameters. The analysis assessed whether strand labels remain stable across repeated runs of the same prompt.
Validation dataset	200 problems randomly sampled from the full dataset of 7,652 items, selected to represent diverse mathematics topics and problem formats.
Repeated runs	The labeling procedure was executed three times using the same prompt. Run 1: temperature = 0.20; Run 2: temperature = 0.35; Run 3: temperature = 0.35.
Consensus score formula	Consensus score = majority count / number of runs. Let L1, L2, L3 represent strand labels from each run.
Possible consensus values	1.00 = all three runs identical; 0.67 = two runs identical, one different; 0.33 = all three runs different.

Table B2. Distribution of consensus scores across 200 validation items			
Consensus Score	Interpretation	Number of Items	Percentage (%)
1.00	All three runs produced the same label	150	75.0
0.67	Two runs identical; one run different	48	24.0
0.33	All three runs produced different labels	2	1.0
Total		200	100.0
Note. Minor disagreements occurred primarily for items involving both procedural and strategic reasoning, where the dominant proficiency strand is less clearly defined.			

Abbreviations

ACC: Accuracy; AKT: Attentive Knowledge Tracing; AUC: Area Under the Receiver Operating Characteristic Curve; CU: Conceptual Understanding; DKT: Deep Knowledge Tracing; DKVMN: Dynamic Key-Value Memory Network; ECE: Expected Calibration Error; FFN: Feed-Forward Network; GELU: Gaussian Error Linear Unit; KC: Knowledge Concept; LLM: Large Language Model; LSTM: Long Short-Term Memory; MAE: Mean Absolute Error; MHA: Multi-Head Attention; MLP: Multilayer Perceptron; PF: Procedural Fluency; PKT: Proficiency-informed Knowledge Tracing; RMSE: Root Mean Square Error; RNN: Recurrent Neural Network; SAKT: Self-Attentive Knowledge Tracing; SC: Strategic Competence; SD: Standard Deviation; UID: User Identifier.

Acknowledgements

The authors gratefully acknowledge the guidance and support provided by Tak-Wai Chan and Chih-Yueh Chou throughout this research. The authors also extend their sincere appreciation to all participants who contributed to this study. Finally, the authors acknowledge the Ministry of Science and Technology, Taiwan (114-2423-H-008-004 and 114-2410-H-008-014-MY3), and the Ministry of Education, Taiwan, for their financial support.

Author's contributions

Kannan Nataraj was the main author of the manuscript and was responsible for writing the various sections of the paper. Charles Y. C. Yeh and Chih-Yueh Chou contributed to data acquisition, data analysis, and revision of the manuscript. Tak-Wai Chan was responsible for funding acquisition and provided guidance throughout the research. All authors read and approved the final manuscript.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The authors would like to thank the Ministry of Science and Technology of the Republic of China, Taiwan, for financial support (114-2423-H-008-004 and 114-2410-H-008-014-MY3), and the Research Center for Science and Technology for Learning, National Central University, Taiwan.

Availability of data and materials

The datasets and supplementary materials used and/or generated during the current study are available from the corresponding author upon reasonable request.

Declarations

Competing interests

The authors declare that they have no competing interests

Author details

¹ Graduate Institute of Network Learning Technology, National Central University, Taoyuan, Taiwan.

² Department of Computer Science and Engineering, Yuan Ze University, Taoyuan, Taiwan.

Received: 24 November 2025 Accepted: 25 July 2026

Published online: 1 January 2027 (Online First: 7 October 2026)

References

- Barton, C. (2018). How I wish I'd taught maths: Lessons learned from research, conversations with experts, and 12 years of mistakes. John Catt Educational.
- Bier, N., Lip, S., Strader, R., Thille, C., & Zimmaro, D. (2014). *An approach to knowledge component/skill modeling in online courses*. Open Learning Initiative. <https://research.google/pubs/an-approach-to-knowledge-componentskill-modeling-in-online-courses/>
- Boyd, P., & Ash, A. (2018). Mastery mathematics: Changing teacher beliefs around in-class grouping and mindset. *Teaching and Teacher Education*, 75, 214–223. <https://doi.org/10.1016/j.tate.2018.06.016>
- Cen, J., Lin, J., Zhou, D., Zhong, W., Chen, J., Yang, A., & Zhou, Y. (2025). LLM-driven effective knowledge tracing by integrating dual-channel difficulty. In *2025 IEEE International Conference on Big Data (BigData)* (pp. 647–652). IEEE. <https://doi.org/10.1109/BIGDATA66926.2025.11402203>
- Chen, X., Zhang, J., Zhou, T., & Zhang, F. (2025). LLM-CDM: A large language model enhanced cognitive diagnosis for intelligent education. *IEEE Access*, 13, 47165–47180. <https://doi.org/10.1109/ACCESS.2025.3549309>
- Corbett, A. T., & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4(4), 253–278. <https://doi.org/10.1007/BF01099821>
- Ghosh, A., Heffernan, N. T., & Lan, A. S. (2020). Context-aware attentive knowledge tracing. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 2330–2339). Association for Computing Machinery. <https://doi.org/10.1145/3394486.3403282>
- Guo, Y., Shen, S., Liu, Q., Huang, Z., Zhu, L., Su, Y., & Chen, E. (2024). Mitigating cold-start problems in knowledge tracing with large language models: An attribute-aware approach. In *Proceedings of the 33rd ACM International*

- Conference on Information and Knowledge Management (pp. 727–736). Association for Computing Machinery. <https://doi.org/10.1145/3627673.3679664>
- Guskey, T. R., & Anderman, E. M. (2013). In search of a useful definition of mastery. *Educational Leadership*, 71(4), 18–23.
- Huang, S., Liu, Z., Zhao, X., Luo, W., & Weng, J. (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* (pp. 2441–2445). Association for Computing Machinery. <https://doi.org/10.1145/3539618.3592073>
- Huang, T., Liang, M., Yang, H., Li, Z., Yu, T., & Hu, S. (2021). Context-aware knowledge tracing integrated with the exercise representation and association in mathematics. In *Proceedings of the 14th International Conference on Educational Data Mining* (pp. 360–366). International Educational Data Mining Society.
- Huang, Z., & Liu, Z. (2025). HCGKT: Hierarchical contrastive graph knowledge tracing with multi-level feature learning. In A. I. Cristea, E. Walker, Y. Lu, O. C. Santos, & S. Isotani (Eds.), *Artificial intelligence in education: 26th International Conference, AIED 2025, proceedings, Part III* (Lecture Notes in Computer Science, Vol. 15879, pp. 274–288). Springer. https://doi.org/10.1007/978-3-031-98420-4_20
- Kelly, K. M., & Heffernan, N. T. (2016). Optimizing the amount of practice in an online learning platform. In *Proceedings of the 3rd 2016 ACM Conference on Learning at Scale* (pp. 145–148). Association for Computing Machinery. <https://doi.org/10.1145/2876034.2893393>
- Kilpatrick, J., Swafford, J., & Findell, B. (2001). *Adding it up: Helping children learn mathematics*. National Academy Press.
- Kim, J., Koo, S., & Lim, H. (2023). A multi-faceted exploration incorporating question difficulty in knowledge tracing for English proficiency assessment. *Electronics*, 12(19), Article 4171. <https://doi.org/10.3390/electronics12194171>
- Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2012). The Knowledge-Learning-Instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive Science*, 36(5), 757–798. <https://doi.org/10.1111/j.1551-6709.2012.01245.x>
- Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 35, 22199–22213.
- Krippendorff, K. (2018). *Content analysis: An introduction to its methodology* (4th ed.). SAGE Publications. <https://doi.org/10.4135/9781071878781>
- Li, C., Xing, W., & Leite, W. (2024). Using fair AI to predict students' math learning outcomes in an online platform. *Interactive Learning Environments*, 32(3), 1117–1136. <https://doi.org/10.1080/10494820.2022.2115076>
- Li, H., Yu, J., Ouyang, Y., Liu, Z., Rong, W., Liu, H., Li, J., & Xiong, Z. (2025). Explainable few-shot knowledge tracing. *Frontiers of Digital Education*, 2(4), Article 34. <https://doi.org/10.1007/s44366-025-0071-x>
- Liang, Y., Peng, T., Pu, Y., & Wu, W. (2022). HELP-DKT: An interpretable cognitive model of how students learn programming based on deep knowledge tracing. *Scientific Reports*, 12, Article 4012. <https://doi.org/10.1038/s41598-022-07956-0>
- Liu, Q., Huang, Z., Yin, Y., Chen, E., Xiong, H., Su, Y., & Hu, G. (2021). EKT: Exercise-aware knowledge tracing for student performance prediction. *IEEE Transactions on Knowledge and Data Engineering*, 33(1), 100–115. <https://doi.org/10.1109/TKDE.2019.2924374>
- Liu, Z., Liu, Q., Chen, J., Huang, S., & Luo, W. (2023). simpleKT: A simple but tough-to-beat baseline for knowledge tracing. In *The Eleventh International Conference on Learning Representations*. <https://openreview.net/forum?id=9HiGqC9C-KA>
- Liu, Z., Liu, Q., Guo, T., Chen, J., Huang, S., Zhao, X., Tang, J., Luo, W., & Weng, J. (2023). XES3G5M: A knowledge tracing benchmark dataset with auxiliary information. *Advances in Neural Information Processing Systems*, 36, 32958–32970. <https://doi.org/10.52202/075280-1429>
- Liu, Z., Liu, Q., Chen, J., Huang, S., Tang, J., & Luo, W. (2022). pyKT: A Python library to benchmark deep learning-based knowledge tracing models. *Advances in Neural Information Processing Systems*, 35, 18542–18555.
- Ozyurt, Y., Feuerriegel, S., & Sachan, M. (2025). *Automated knowledge concept annotation and question representation learning for knowledge tracing* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2410.01727>
- Pandey, S., & Karypis, G. (2019). A self-attentive model for knowledge tracing. In *Proceedings of the 12th International Conference on Educational Data Mining* (pp. 384–389). International Educational Data Mining Society.
- Park, J., Kang, S., Park, J., Kim, J. H., Shin, J., Park, S., & Yu, Y. (2026). Tracing mathematical proficiency through problem-solving processes. In *Findings of the Association for Computational Linguistics: ACL 2026* (pp. 19251–19269). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2026.findings-acl.961>
- Pelánek, R. (2017). Bayesian knowledge tracing, logistic models, and beyond: An overview of learner modelling techniques. *User Modeling and User-Adapted Interaction*, 27(3–5), 313–350. <https://doi.org/10.1007/s11257-017-9193-2>
- Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L. J., & Sohl-Dickstein, J. (2015). Deep knowledge tracing. *Advances in Neural Information Processing Systems*, 28, 505–513.
- Qian, M., Ji, H., & Li, L. (2026). Empirical validation of a generative AI framework for personalized education assessment. *Scientific Reports*, 16, Article 11538. <https://doi.org/10.1038/s41598-026-42169-9>
- Rittle-Johnson, B. (2017). Developing mathematics knowledge. *Child Development Perspectives*, 11(3), 184–190. <https://doi.org/10.1111/cdep.12229>

- Simpson, A., & Wang, Y. (2023). Making sense of ‘mastery’: Understandings of a policy term among a sample of teachers in England. *International Journal of Science and Mathematics Education*, 21(2), 581–600. <https://doi.org/10.1007/s10763-021-10178-x>
- Slater, S., & Baker, R. (2019). Forecasting future student mastery. *Distance Education*, 40(3), 380–394. <https://doi.org/10.1080/01587919.2019.1632169>
- Tan, M., Merrill, M. A., Gupta, V., Althoff, T., & Hartvigsen, T. (2024). Are language models actually useful for time series forecasting? *Advances in Neural Information Processing Systems*, 37, 60162–60191. <https://doi.org/10.52202/079017-1922>
- Wang, F., Huang, Z., Liu, Q., Chen, E., Yin, Y., Ma, J., & Wang, S. (2023). Dynamic cognitive diagnosis: An educational priors-enhanced deep knowledge tracing perspective. *IEEE Transactions on Learning Technologies*, 16(3), 306–323. <https://doi.org/10.1109/TLT.2023.3254544>
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35, 24824–24837.
- Xing, W., & Du, D. (2019). Dropout prediction in MOOCs: Using deep learning for personalized intervention. *Journal of Educational Computing Research*, 57(3), 547–570. <https://doi.org/10.1177/0735633118757015>
- Yeh, C. Y. C., Cheng, H. N. H., Chen, Z. H., Liao, C. C. Y., & Chan, T. W. (2019). Enhancing achievement and interest in mathematics learning through Math-Island. *Research and Practice in Technology Enhanced Learning*, 14(1), Article 5. <https://doi.org/10.1186/s41039-019-0100-9>
- Yin, Y., Dai, L., Huang, Z., Shen, S., Wang, F., Liu, Q., Chen, E., & Li, X. (2023). Tracing knowledge instead of patterns: Stable knowledge tracing with diagnostic transformer. In *Proceedings of the ACM Web Conference 2023* (pp. 855–864). Association for Computing Machinery. <https://doi.org/10.1145/3543507.3583255>
- Yu, Y., Zhou, Y., Zhu, Y., Ye, Y., Chen, L., & Chen, M. (2024). ECKT: Enhancing code knowledge tracing via large language models. In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Meeting of the Cognitive Science Society* (pp. 1638–1645). Cognitive Science Society.
- Zhang, J., Shi, X., King, I., & Yeung, D. Y. (2017). Dynamic key-value memory networks for knowledge tracing. In *Proceedings of the 26th International Conference on World Wide Web* (pp. 765–774). International World Wide Web Conferences Steering Committee. <https://doi.org/10.1145/3038912.3052580>
- Zhang, T., Mao, H., Liu, H., Liu, Y., Yu, M., Wu, W., Yu, G., Wei, B., & Guan, Y. (2023). Parallel prediction method of knowledge proficiency based on Bloom’s cognitive theory. *Mathematics*, 11(24), Article 5002. <https://doi.org/10.3390/math11245002>
- Zhang, Y., Lin, Z., Bu, C., Yu, K., Hu, X., & Wu, X. (2025). A cognitive diagnosis model with nonlinear dependence between students and exercises. *IEEE Transactions on Computational Social Systems*, 12(5), 2508–2519. <https://doi.org/10.1109/TCSS.2024.3525053>

Publisher’s Note

The Asia-Pacific Society for Computers in Education (APSCE) remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Research and Practice in Technology Enhanced Learning (RPTeL)
is an open-access journal and free of publication fee.