Improving Interpretability of Deep Sequential Knowledge Tracing Models with Question-centric Cognitive Representations

Jiahao Chen[1], Zitao Liu[2*], Shuyan Huang[1], Qiongqiong Liu[1], Weiqi Luo[2]

[1] TAL Education Group, Beijing, China
[2] Guangdong Institute of Smart Education, Jinan University, Guangzhou, China
chenjiahao@tal.com, liuzitao@jnu.edu.cn, huangshuyan@tal.com, liuqiongqiong1@tal.com, lwq@jnu.edu.cn

Abstract

Knowledge tracing (KT) is a crucial technique to predict students' future performance by observing their historical learning processes. Due to the powerful representation ability of deep neural networks, remarkable progress has been made by using deep learning techniques to solve the KT problem. The majority of existing approaches rely on the *homogeneous question* assumption that questions have equivalent contributions if they share the same set of knowledge components. Unfortunately, this assumption is inaccurate in real-world educational scenarios. Furthermore, it is very challenging to interpret the prediction results from the existing deep learning based KT models. Therefore, in this paper, we present QIKT, a question-centric interpretable KT model to address the above challenges. The proposed QIKT approach explicitly models students' knowledge state variations at a fine-grained level with question-sensitive cognitive representations that are jointly learned from a question-centric knowledge acquisition module and a question-centric problem solving module. Meanwhile, the QIKT utilizes an item response theory based prediction layer to generate interpretable prediction results. The proposed QIKT model is evaluated on three public real-world educational datasets. The results demonstrate that our approach is superior on the KT prediction task, and it outperforms a wide range of deep learning based KT models in terms of prediction accuracy with better model interpretability. To encourage reproducible results, we have provided all the datasets and code at https://pykt.org/.

Introduction

Knowledge tracing (KT) is the task of using students' historical learning interaction data (e.g., responses to a series of questions) to model their knowledge mastery over time so as to make predictions on their future performance (e.g., predicting correctly on next question) (Corbett and Anderson 1994). Figure 1 gives an illustrative example of the KT task. Such predictive capabilities can potentially help students learn better and faster when paired with high-quality learning materials and instructions and the KT models have been widely used to support intelligent tutoring systems and MOOC platforms (Käser et al. 2017; Cen, Koedinger, and Junker 2006; Lavoué et al. 2018; Liu et al. 2021a).

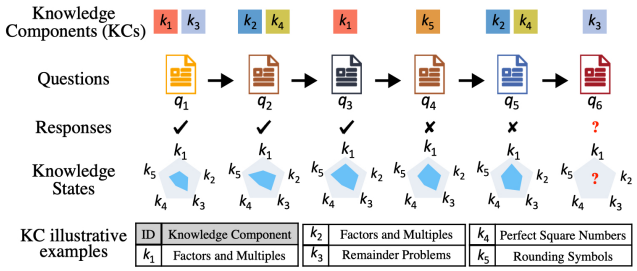

Figure 1: A graphical illustration of the KT problem.

Recently, remarkable progress has been made by applying deep learning to solve the KT problem (Piech et al. 2015; Abdelrahman and Wang 2019; Ghosh, Heffernan, and Lan 2020; Nakagawa, Iwasawa, and Matsuo 2019; Pandey and Karypis 2019; Pandey and Srivastava 2020; Shen et al. 2021, 2020; Yang et al. 2020; Zhang et al. 2017, 2021; Wang et al. 2019; Liu et al. 2023a,b). One of the representative approaches among them is the deep sequential KT modeling, which utilizes auto-regressive architectures, such as LSTM and GRU, to represent student's knowledge states (e.g., the mastery level of the concepts) as the hidden states of recurrent units (Piech et al. 2015; Chen et al. 2018; Guo et al. 2021; Lee and Yeung 2019; Liu et al. 2019). Due to the ability to learn sequential dependencies from student interaction data, deep sequential KT models draw attention from researchers from different communities and achieve great success in improving prediction accuracy (Minn et al. 2018; Nagatani et al. 2019; Su et al. 2018; Yeung and Yeung 2018).

In spite of the promising results demonstrated by previous methods, some important limitations still exist when applying deep sequential KT models on real-world educational data. First, most existing approaches rely on the *homogeneous assumption* that questions nested under a particular set of knowledge components (KCs) are equivalent (Zhang et al. 2017; Nagatani et al. 2019; Nakagawa, Iwasawa, and Matsuo 2019; Lee and Yeung 2019). The homogeneous assumption is inaccurate in two perspectives: (1) it assumes that students have the same knowledge increment after they give the same responses to homogeneous questions[1] during the knowledge acquisition learning processes; and (2) it as-

*The corresponding author: Zitao Liu

[1] In this paper, we refer to questions that have the same set of

sumes that students will give the same responses to different questions as long as these questions are homogeneous during the problem solving process. Such unrealistic assumption limits the KT performance of the previous works. While in some cases the problem may be alleviated by implicitly modeling the question difficulty or question discrimination (Zhang et al. 2021; Liu et al. 2021b; Ghosh, Heffernan, and Lan 2020), they suffer from the lack of ground truth labels or the exclusions of cognitive modeling (e.g., only used in pre-trained tasks), and jointly modeling the question-centric cognitive effects on knowledge states remains a big concern. Second, although deep learning based knowledge tracing (DLKT) models have shown advanced progress in terms of prediction accuracy compared with traditional cognitive models, it is difficult to extract psychologically meaningful explanations from their million-level parameters, that would relate to cognitive theory. The lack of nontransparent decision processes of DLKT models is unsatisfied for tutors and students who need to see a convincing diagnosis before they accept results generated from DLKT models.

In this paper, we address aforementioned challenges by proposing a novel KT model called *Q*uestion-centric *I*nterpretable *K*nowledge *T*racing, i.e., *QIKT*. More specifically, QIKT explicitly learns question-centric cognitive representations with a knowledge acquisition module and a problem solving module. The knowledge acquisition module aims to model the variations in students' knowledge states after receiving responses to specific questions. It estimates students' question-specific knowledge acquisition by a joint optimization including representations of students' current knowledge states, responses, questions and the corresponding KCs. The problem solving module estimates students' problem solving abilities on each specific question by projecting student knowledge states on the jointly learned representations of both questions and KCs. Furthermore, the QIKT incorporates an interpretable prediction layer to improve interpretability of prediction results. The interpretable prediction layer is built upon the Item Response Theory (IRT) in psychometrics, and integrates the parameters of an IRT model into the question-centric deep sequential KT model. This enables the QIKT model to generate explainable personalized parameters for each student at question level. We evaluate QIKT on three datasets by comparing it with 13 previous approaches under a rigorous KT evaluation protocol (Liu et al. 2022). Experimental results demonstrate that QIKT achieves superior prediction performance and the psychologically meaningful interpretability simultaneously.

The main contributions are summarized as follows:

- We introduce a knowledge acquisition module and a problem solving module to learn question-centric representations when students absorb knowledge after answering questions and apply knowledge to solve problems.
- We design a simple yet effective interpretable prediction layer based on the IRT theory and manage to seamlessly combine it with existing deep sequential KT models.
- We conduct comprehensive quantitative and qualitative experiments to validate the performance of QIKT on three public datasets with a wide range of baselines. The well-designed experiments illustrate the superiority of our approach in both prediction performance and model interpretability. To the best of our knowledge, **our QIKT model is able to achieve the best prediction performance in terms of AUC** on the publicly available reproducible KT experimental settings.

KCs as "*homogeneous questions*".

Background and Related Work

Deep Sequential Modeling for Knowledge Tracing

Deep sequential KT models utilize an auto-regressive architecture to capture the intrinsic dependencies among students' chronologically ordered interactions (Chen et al. 2018; Guo et al. 2021; Lee and Yeung 2019; Liu et al. 2019; Minn et al. 2018; Nagatani et al. 2019; Piech et al. 2015; Su et al. 2018; Yeung and Yeung 2018). Since the very first and successful research work of deep knowledge tracing (DKT) that applies recurrent neural networks to model students' dynamic learning behaviors by Piech et al. (2015), a large number of works have been done to improve DKT's performance (Yeung and Yeung 2018; Chen et al. 2018; Su et al. 2018; Nagatani et al. 2019; Lee and Yeung 2019; Liu et al. 2019; Guo et al. 2021). For example, Yeung and Yeung (2018) proposed to use two regularization terms to address the reconstruction and waviness issues in the DKT model. Chen et al. (2018) incorporated prerequisite relations between pedagogical concepts to enhance DKT model. Su et al. (2018) presented to aggregate textual representations to monitor student knowledge states. Nagatani et al. (2019) developed approaches to capture students' forgetting behaviors and Guo et al. (2021) leveraged adversarial training samples to enhance the deep sequential KT models' generalization.

Besides deep sequential KT models, other types of neural network based approaches are applied in the KT domain as well, such as memory augmented KT models that explicitly model latent relations between KCs with an external memory (Abdelrahman and Wang 2019; Shen et al. 2021; Zhang et al. 2017), graph based KT models that capture interaction relations with graph neural networks (Nakagawa, Iwasawa, and Matsuo 2019; Tong et al. 2020; Yang et al. 2020), and attention based KT models that use the attention mechanism and its variants to capture dependencies between interactions (Ghosh, Heffernan, and Lan 2020; Pandey and Srivastava 2020; Pu et al. 2020; Zhang et al. 2021).

Interpreting Deep Learning Based Knowledge Tracing Models

Recently, many interpretable approaches have been incorporated into DLKT models for both student modeling and prediction tasks. These techniques can be divided into the following three categories:

- **C1: Post-hoc local explanation**. Post-hoc local explanation techniques are used in the KT task aiming to examine each individual prediction result and figure out why the DLKT models make the decisions they make (Lu et al. 2020, 2022). For example, Lu et al. (2022) applied a layer-wise relevance propagation method to interpret a deep se-

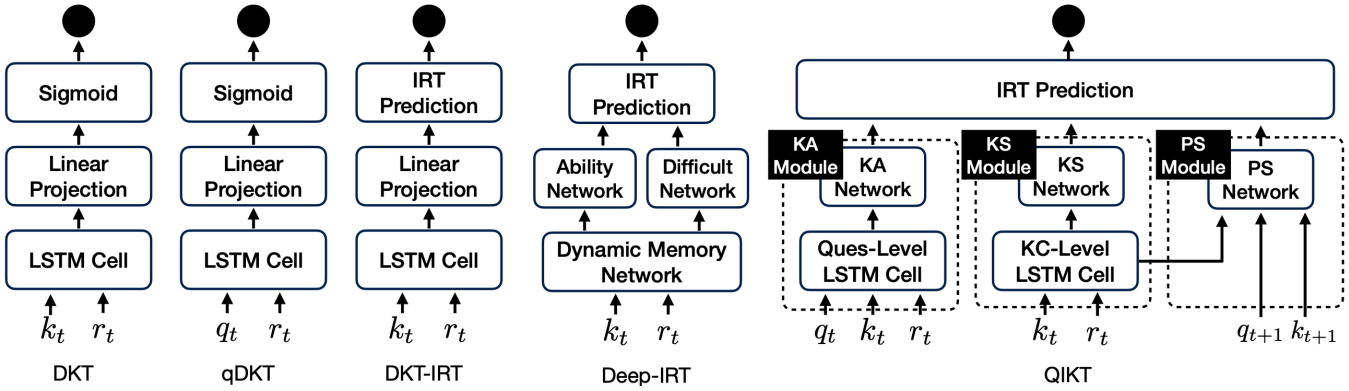

Figure 2: Graphical illustrations of our QIKT model along with some representative DLKT models including DKT (Piech et al. 2015), qDKT (Sonkar et al. 2020), DKT-IRT (Converse, Pu, and Oliveira 2021), and Deep-IRT (Yeung 2019).

quential KT model by back propagating relevance scores from the model's output layer to its input layer.

- **C2: Global interpretability with explainable structures**. Embed an interpretable cognitive module into existing DLKT architectures to better understand the knowledge state modeling process (Wang et al. 2020a; Zhao et al. 2020; Pu et al. 2022). For example, Wang et al. (2020a) designed an intermediate interaction layer based on multidimensional IRT and explicitly modeled both student factors and exercise factors. Pu et al. (2022) proposed to utilize an automatic temporal cognitive method to better capture the changes in students' knowledge states.

- **C3: Global interpretability with explainable parameters**. Directly use cognitively interpretable models to estimate the probability that a student will answer a question correctly. Explainable parameters in these models are obtained from outputs of the DLKT models (Converse, Pu, and Oliveira 2021; Yeung 2019). For example, Converse, Pu, and Oliveira (2021) linearly transformed the hidden states of the DKT model and then applied a hard thresholding operator to cast the parameters into the IRT-like form. Yeung (2019) proposed to explicitly learn levels of student abilities and KC difficulties with a dynamic key-value memory network for KT task and feed the learned results to an IRT layer for final prediction.

Our QIKT belongs to the C3 category since we utilize the IRT function for interpretable prediction. Different from existing approaches (Yeung 2019; Converse, Pu, and Oliveira 2021) that only optimize the model performance based on interpretable predicted outcomes, our QIKT approach directly incorporates the explainable parameter learning into the final model optimization objective, which improves the model interpretability and preserves the prediction performance as well. Compared with the methods developed based on memory networks such as Deep-IRT (Yeung 2019), our approach is based on the deep sequential architectures, which is more applicable and has better prediction accuracy (Liu et al. 2022).

Problem Statement

Our objective is given an arbitrary question q_* to predict the probability of whether a student will answer q_* correctly based on the student's historical interaction data. Specifically, for each student $\mathbf{S}$, we assume that we have observed a chronologically ordered collection of t past interactions i.e., $\mathbf{S} = \{\mathbf{s}_j\}_{j=1}^{t}$. Each interaction is represented as a 4-tuple $\mathbf{s}$, i.e., $\mathbf{s} =< q, \{k|k \in \mathcal{N}_q\}, r, s >$, where $q, \{k\}, r, s$ represent the specific question, the associated KC set, the binary valued student response[2], and student's response timestamp respectively. $\mathcal{N}_q$ is the set of KCs that are associated with question q. We would like to estimate the probability $\hat{r}_*$ of the student's future performance on arbitrary question q_*.

Interpretable KT Modeling with Question-centric Cognitive Representations

In this section, we discuss the five components QIKT model in detail: (1) the interaction encoder that assembles and encodes both question-level and KC-level information; (2) the question-centric knowledge acquisition (KA) module that examines students' knowledge acquisition after answering specific questions over time; (3) the question-agnostic knowledge state (KS) module that models the general knowledge state dynamics; (4) the question-centric problem solving (PS) module that estimates the capabilities of students to tackle a specific question with their current knowledge states; and (5) the interpretable prediction layer that aims to leverage the psychological theory of IRT to generate more interpretable results for both tutors and students.

Interaction Encoder

In real-world educational scenarios, the question bank is usually much bigger than the set of KCs, for example, the number of questions is more than 1500 times larger than the number of KCs in the well cited Algebra2005 dataset[3] (see Table 1). Therefore, most existing research works like DKT

[2]Response is a binary valued indicator variable where 1 represents the student correctly answered the question, and 0 otherwise.

[3]Details of Algebra2005 is described in the *Datasets* section.

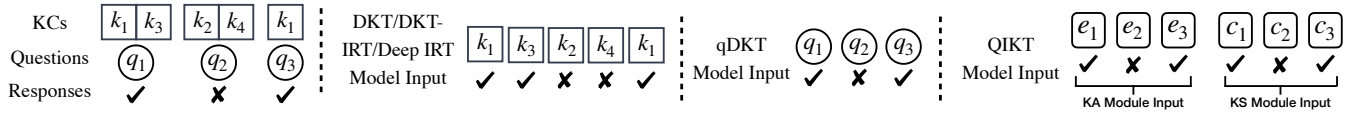

Figure 3: Illustrations of different interaction encoding approaches. $\mathbf{e}_i$s and $\mathbf{c}_i s$ are defined in eq.(1) and eq.(2) respectively.

(Piech et al. 2015), DKT-IRT (Converse, Pu, and Oliveira 2021), and Deep-IRT (Yeung 2019) alleviate the data sparsity issue by using KCs to index questions and all questions cover the same KC are treated as a single question. Although this conversion greatly relieves the sparsity problem (Ghosh, Heffernan, and Lan 2020; Liu et al. 2022), it enforces the DLKT models to follow the homogeneous assumption and hence omits different learning effects brought by questions of the same concepts. Please note that this conversion lets the DLKT models be learned from the extended KC-level sequences instead of the original question-level sequences, as illustrated in Figure 3. On the other hand, question-centric models like qDKT (Sonkar et al. 2020) completely ignore the relations between questions and KCs and purely uses the question sequence to track students' knowledge states.

In this work, we aim to improve the aforementioned DLKT models by capturing the intrinsic relations between questions and KCs at a more fine-grained level. More specifically, we have two different raw interaction encodings for the KA module and the KS module. For the KA module, similar to a recent work by Long et al. (2021), each question level interaction $\mathbf{e}_t$ is represented as a combination of question, response and the corresponding set of KCs, i.e.,

$$\mathbf{e}_t = \begin{cases} \mathbf{q}_t \oplus \bar{\mathbf{k}}_t \oplus \mathbf{0}, & r_t = 1 \\ \mathbf{0} \oplus \mathbf{q}_t \oplus \bar{\mathbf{k}}_t, & r_t = 0 \end{cases} \tag{1}$$

where $\mathbf{q}_t$ is the question embedding, $\mathbf{q}_t \in \mathbb{R}^{d\times 1}$ and $\bar{\mathbf{k}}_t$ is the average embeddings of all the KCs to the question, i.e.,

$$\bar{\mathbf{k}}_t = \frac{1}{|K_{q_t}|}\sum_{j=1}^{m} \mathbf{k}_j * \mathbb{I}(k_j \in \mathcal{N}_{q_t})$$

where $\mathbf{k}_j$ is the KC embedding, $\mathbf{k}_j \in \mathbb{R}^{d\times 1}$. m is the total number of KCs in the question bank. K_{q_t} is the size of $\mathcal{N}_{q_t}$. $\mathbb{I}(\cdot)$ is the indicator function and $\oplus$ is the concatenate operation. The response in each interaction is encoded as a $2d \times 1$ all-zero vector, $\mathbf{0}$. We use concatenation directions (left or right) to indicate different responses, i.e., correct or wrong.

We conduct a similar encoding mechanism for the KS module. Since the KS module only focuses on the general knowledge state changes regardless of question specific variations, the interaction embedding $\mathbf{c}_t$ of KS is

$$\mathbf{c}_t = \begin{cases} \bar{\mathbf{k}}_t \oplus \mathbf{0}, & r_t = 1 \\ \mathbf{0} \oplus \bar{\mathbf{k}}_t, & r_t = 0 \end{cases} \tag{2}$$

Question-centric Knowledge Acquisition Module

Students absorb knowledge as they interact with questions and their knowledge acquisition varies after solving the homogeneous questions. Hence, we propose to estimate students' question-specific knowledge acquisition with the joint representations $\mathbf{e}_t$s of the questions, concepts and responses. Similar to the standard DKT model, we use the LSTM cell to update the student's question-level knowledge state $\mathbf{a}_t$ after answering each question at timestamp t:

$$\begin{aligned}
\mathbf{i}_t &= \sigma\left(\mathbf{W}_1 \cdot \mathbf{e}_t + \mathbf{U}_1 \cdot \mathbf{a}_{t-1} + \mathbf{b}_1\right) \\
\mathbf{f}_t &= \sigma\left(\mathbf{W}_2 \cdot \mathbf{e}_t + \mathbf{U}_2 \cdot \mathbf{a}_{t-1} + \mathbf{b}_2\right) \\
\mathbf{o}_t &= \sigma\left(\mathbf{W}_3 \cdot \mathbf{e}_t + \mathbf{U}_3 \cdot \mathbf{a}_{t-1} + \mathbf{b}_3\right) \\
\tilde{\mathbf{c}}_t &= \sigma\left(\mathbf{W}_4 \cdot \mathbf{e}_t + \mathbf{U}_4 \cdot \mathbf{a}_{t-1} + \mathbf{b}_4\right) \\
\mathbf{c}_t &= \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tilde{\mathbf{c}}_t \\
\mathbf{a}_t &= \mathbf{o}_t \odot \tanh\left(\mathbf{c}_t\right)
\end{aligned}$$

where $\mathbf{W}_i$s, $\mathbf{U}_i$s, $\mathbf{b}_i$s are trainable parameters and $\mathbf{W}_i \in \mathbb{R}^{d\times 4d}$, $\mathbf{U}_i \in \mathbb{R}^{d\times d}$, $\mathbf{b}_i \in \mathbb{R}^{d\times 1}$ and $i = 1, 2, 3, 4$. σ, $\odot$, and $\tanh$ denote the sigmoid, element-wise multiplication and hyperbolic tangent functions.

Different from existing approaches that directly use the learned knowledge state ($\mathbf{a}_t$) to predict the student knowledge mastery, we apply a knowledge acquisition network to first extract the knowledge states with a fully connected neural layer and then project it into the question-centric space via non-linear transformation. The knowledge acquisition score α_t is computed as $\alpha_t = \text{S-Pool}\Big(\mathbf{w}^a \odot \text{ReLU}\big(\mathbf{W}_2^a \cdot \text{ReLU}(\mathbf{W}_1^a \cdot \mathbf{a}_t + \mathbf{b}_1^a) + \mathbf{b}_2^a\big)\Big)$ where $\mathbf{W}_1^a$, $\mathbf{W}_2^a$, $\mathbf{w}^a$, $\mathbf{b}_1^a$ and $\mathbf{b}_2^a$ are trainable parameters and $\mathbf{W}_1^a \in \mathbb{R}^{d\times d}$, $\mathbf{W}_2^a \in \mathbb{R}^{n\times d}$, $\mathbf{w}^a \in \mathbb{R}^{n\times 1}$, $\mathbf{b}_1^a \in \mathbb{R}^{d\times 1}$, $\mathbf{b}_2^a \in \mathbb{R}^{n\times 1}$, n is the total number of questions, and S-Pool means sum pooling.

Question-agnostic Knowledge State Module

In real educational contexts, students may frequently guess or slip when they interact with questions. This may cause the KA module overly sensitive to each interaction and hence lead prediction confusion during the inference stage. Therefore, as an important complement, we model the general question-agnostic changes of students' knowledge state in the KS module. Similar to knowledge state modeling in the KA module, we apply another LSTM cell to update the student's question-agnostic knowledge state ($\mathbf{g}_t$) after receiving each response. The LSTM cell in the KS module takes question-agnostic input $\mathbf{c}_t$ instead of $\mathbf{e}_t$. The iterative update equations of $\mathbf{g}_t$ are described in Appendix A.1 due to the space limit. Furthermore, we design a knowledge state extraction network to capsule the general knowledge states of a student by applying non-linear transformations to project the mastery level into the space of KCs and computing the knowledge mastery score β_t i.e., $\beta_t =$

$\text{S-Pool}\Big(\mathbf{w}^g \odot \text{ReLU}\big(\mathbf{W}_2^g \cdot \text{ReLU}(\mathbf{W}_1^g \cdot \mathbf{g}_t + \mathbf{b}_1^g) + \mathbf{b}_2^g\big)\Big)$ where $\mathbf{W}_1^g$, $\mathbf{W}_2^g$, $\mathbf{w}^g$, $\mathbf{b}_1^g$ and $\mathbf{b}_2^g$ are trainable parameters and $\mathbf{W}_1^g \in \mathbb{R}^{d\times d}$, $\mathbf{W}_2^g \in \mathbb{R}^{m\times d}$, $\mathbf{w}^g \in \mathbb{R}^{m\times 1}$, $\mathbf{b}_1^g \in \mathbb{R}^{d\times 1}$, $\mathbf{b}_2^g \in \mathbb{R}^{m\times 1}$.

Question-centric Problem Solving Module

Correctly answering a question not only depends on the students' knowledge mastery, but is highly relevant to the question itself such as its difficulty and discrimination. Therefore, we present a question-centric problem solving module to estimate students' knowledge application abilities to specific questions by projecting their knowledge mastery on questions. Specifically, we design a problem solving network to conduct such knowledge projection as follows:

$$\mathbf{p}_{t+1} = \mathbf{g}_t \oplus \mathbf{q}_{t+1} \oplus \bar{\mathbf{k}}_{t+1}$$
$$\zeta_{t+1} = \mathbf{w}^p \cdot \text{ReLU}\big(\mathbf{W}_2^p \cdot \text{ReLU}(\mathbf{W}_1^p \cdot \mathbf{p}_{t+1} + \mathbf{b}_1^p) + \mathbf{b}_2^p\big) + b^p$$

where ζ_{t+1} denotes the students' knowledge application score on question q_{t+1}. $\mathbf{p}_{t+1}$ contains both the student knowledge mastery at time t and all the available information about question q_{t+1} and $\mathbf{p}_{t+1} \in \mathbb{R}^{3d\times 1}$. $\mathbf{W}_i^p$s, $\mathbf{w}^p$, $\mathbf{b}_i^p$s and b^p are trainable parameters and $\mathbf{W}_i^p \in \mathbb{R}^{3d\times 3d}$, $\mathbf{b}_i^p \in \mathbb{R}^{3d\times 1}$,$\mathbf{w}^p \in \mathbb{R}^{1\times 3d}$, b^p is scalar and $i = 1, 2$.

Interpretable Prediction Layer

Generally, explaining DLKT models' parameters and decision-making is challenging. Thus, we design an IRT based prediction layer to enhance the prediction interpretability of the proposed QIKT model. Similar to previous work by Yeung (2019), we use the IRT function to calculate the probability of a correct answer. Furthermore, we strict the IRT function only takes the linear combined scores of question-centric knowledge acquisition score from the KA module, the knowledge mastery score from the KS module and the knowledge application score from the PS module i.e., $\hat{r}_{t+1} = \sigma(\alpha_t + \beta_t + \zeta_{t+1})$. And we explicitly choose not to include any learnable parameters inside the IRT based prediction function for better interpretability.

Optimization of QIKT

All learnable parameters of our QIKT model are optimized by minimizing the binary cross entropy loss between the ground-truth responses r_is and the estimated probabilities $\hat{r}_i$s from the IRT layer as the objective function i.e., $\mathcal{L}_{\text{IRT}} = -\sum_i \big(r_i \log \hat{r}_i + (1 - r_i)\log(1 - \hat{r}_i)\big)$. Furthermore, to directly improve the discriminative ability of the internal knowledge related scores from the KA, KS and PS modules, we explicitly cast these scores via sigmoid function and add the optimization terms about question-centric knowledge acquisition scores (α_ts), knowledge mastery scores (β_ts) and knowledge application scores (ζ_ts) into the overall model training process. Therefore, the final optimization function is $\mathcal{L} = \mathcal{L}_{\text{IRT}} + \lambda\big(\mathcal{L}_*(\boldsymbol{\alpha}) + \mathcal{L}_*(\boldsymbol{\beta}) + \mathcal{L}_*(\boldsymbol{\zeta})\big)$, where λ is tuning hyper-parameter. $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\zeta}$ denote the collections of the corresponding scores, i.e., $\boldsymbol{\alpha} = \{\alpha_i\}$, $\boldsymbol{\beta} = \{\beta_i\}$, and $\boldsymbol{\zeta} = \{\zeta_i\}$ and $\mathcal{L}_*(\mathbf{z})$ is defined as follows:

$$\mathcal{L}_*(\mathbf{z}) = -\sum_i \big(r_i \log \sigma(z_i) + (1 - r_i)\log(1 - \sigma(z_i))\big)$$

Experiment

In this section, we present details of experiment settings and the results. We conduct comprehensive analyses and investigations to illustrate the effectiveness of the QIKT model.

Datasets

We use three widely used publicly available datasets to evaluate the performance of QIKT:

- **ASSISTments2009**[4] **(ASSIST2009)**: is collected from ASSISTment online tutoring platform in the school year 2012-2013 that students are assigned to answer similar exercises from the skill builder problem sets.
- **Algebra 2005-2006**[5] **(Algebra2005)**: is provided by the KDD Cup 2010 EDM Challenge where students need to complete steps to achieve the mastery of the related KCs.
- **NeurIPS2020 Education Challenge**[6] **(NeurIPS34)**: is released in Task 3 and Task 4 of NeurIPS2020 Education Challenge, it includes students' answers to multiple-choice math diagnostic questions (Wang et al. 2020b).

To conduct reproducible experiments, we rigorously follow the data pre-processing steps suggested in (Liu et al. 2022). We remove student sequences shorter than 3 attempts. Data statistics are summarized in Table 1.

Baselines

We compare our QIKT with the following state-of-the-art DLKT models to evaluate the effectiveness of our approach:

- **DKT**: leverages an LSTM layer to encode the student knowledge state to predict the students' performances (Piech et al. 2015).
- **DKT+**: an improved version of DKT to solve the reconstruction and non-consistent prediction problems (Yeung and Yeung 2018).
- **KQN**: uses student knowledge state encoder and skill encoder to predict the student response performance via the dot product (Lee and Yeung 2019).
- **qDKT**: predicts the future performance of student knowledge state at the question level (Sonkar et al. 2020).
- **DKT-IRT**: incorporates IRT to improve the interpretability of DKT (Converse, Pu, and Oliveira 2021).
- **IEKT**: models student knowledge state via the student cognition and knowledge acquisition estimation (Long et al. 2021).
- **DKVMN**: designs a static key matrix to store the relations between the different KCs and a dynamic value matrix to update the students' knowledge state (Zhang et al. 2017).

[4]https://sites.google.com/site/assistmentsdata/home/2009-2010-assistment-data/skill-builder-data-2009-2010

[5]https://pslcdatashop.web.cmu.edu/KDDCup/

[6]https://eedi.com/projects/neurips-education-challenge

Dataset	# of Ss	# of Is	# of Qs	# of KCs	Avg. KCs	Avg. Qs
ASSIST2009	3,852	282,605	17,737	123	1.197	144.2
Algebra2005	574	607,013	173,109	112	1.364	1545.6
NeurIPS34	4,918	1,382,661	948	57	1.015	16.6

Table 1: Data statistics of three datasets. # of Ss/Is/Qs denote the number of students, interactions and questions. *Avg. KCs* and *Avg. Qs* denotes the number of KCs per question and the number of questions per KC.

Method	Model Type	Usage of Questions	Usage of KCs	Is interpretable	AUC		
					ASSIST2009	Algebra2005	NeurIPS34
DKT	Sequential	No	Yes	No	0.7541±0.0011*	0.8149±0.0011*	0.7689±0.0002*
DKT+	Sequential	No	Yes	No	0.7547±0.0017*	0.8156±0.0011*	0.7696±0.0002*
KQN	Sequential	No	Yes	No	0.7477±0.0011*	0.8027±0.0015*	0.7684±0.0003*
qDKT	Sequential	Yes	No	No	0.7016±0.0049*	0.7485±0.0017*	0.7995±0.0008*
DKT-IRT	Sequential	No	Yes	Yes	0.7591±0.0007*	0.8290±0.0004*	0.7695±0.0004*
IEKT	Sequential	Yes	Yes	No	0.7861±0.0027*	**0.8416±0.0014•**	**0.8045±0.0002•**
DeepIRT	Memory	No	Yes	Yes	0.7465±0.0006*	0.8040±0.0013*	0.7672±0.0006*
DKVMN	Memory	No	Yes	No	0.7473±0.0006*	0.8054±0.0011*	0.7673±0.0004*
ATKT	Adversarial	No	Yes	No	0.7470±0.0008*	0.7995±0.0023*	0.7665±0.0001*
GKT	Graph	No	Yes	No	0.7424±0.0021*	0.8110±0.0009*	0.7689±0.0024*
SAKT	Attention	No	Yes	No	0.7246±0.0017*	0.7880±0.0063*	0.7517±0.0005*
SAINT	Attention	Yes	Yes	No	0.6958±0.0023*	0.7775±0.0017*	0.7873±0.0007*
AKT	Attention	Yes	Yes	No	0.7853±0.0017*	0.8306±0.0019*	0.8033±0.0003*
QIKT	Sequential	Yes	Yes	Yes	**0.7878±0.0024**	0.8408±0.0007	0.8044±0.0005

Table 2: The overall prediction performance of all the baseline models and our QIKT. We highlight the highest results with bold. Marker $*$, $\circ$ and $\bullet$ indicates whether the proposed model is statistically superior/equal/inferior to the compared method (using paired t-test at 0.01 significance level).

- **DeepIRT**: a combination of the IRT and DKVMN to enhance the interpretability of memory augmented models (Yeung 2019).
- **ATKT**: performs adversarial perturbations into student interaction sequence to improve model's generalization ability (Guo et al. 2021).
- **GKT**: utilizes the graph structure to predict the students' performance (Nakagawa, Iwasawa, and Matsuo 2019).
- **SAKT**: uses self-attention to capture relations between exercises and student responses (Pandey and Karypis 2019).
- **SAINT**: uses the Transformer-based layer to capture students' exercise and response sequences (Choi et al. 2020).
- **AKT**: leverages an attention mechanism to characterize the time distance between questions and the past interaction of students (Ghosh, Heffernan, and Lan 2020).

Experimental Setup

We set the maximum length of model input sequence to 200 and perform 5-fold cross-validation for every combination of models and datasets. We use 80% of student sequences for training and validation, and use the rest 20% of student sequences for model evaluation. We adopt ADAM optimizer to train all the models (Kingma and Ba 2015). The number of training epochs is set to 200. We choose to use early stopping strategy that stops optimization when the AUC score is failed to get the improvement on the validation set in the latest 10 epochs. The hyper-parameter λ, the learning rate and the embedding size d are searching from [0,0.5,1,1.5,2], [1e-3, 1e-4, 1e-5], [64, 256] respectively. All the models are implemented in PyTorch and are trained on a cluster of Linux servers with the NVIDIA RTX A5000 GPU device. Following all existing DLKT research, we use the Area Under the Curve (AUC) as the main evaluation metric. We also choose to use Accuracy as the secondary evaluation metric.

Results

Overall Performance. Due to the space limit, results in terms of accuracy and the details of statistical tests are provided in Appendix A.2 and Appendix A.3. The overall model performance is reported in Table 2. From Table 2, we make the following observations: (1) Our proposed model QIKT significantly outperforms 13 baselines on all three datasets (except we have two loss with IEKT on ASSIST2009 and NeurIPS34 datasets). More importantly, as a representative of the deep sequential KT models, compared with DKT, our proposed model improves the AUC by 3.30%, 2.60% and 3.60% on three datasets. That show our proposed modules can significantly improve the performance. (2) When comparing performance on ASSIST2009, Algebra2005 to NeurIPS34, DLKT models behave quite differently. For example, DKT significantly outperforms qDKT on the ASSIST2009 and Algebra2005 datasets by 5.30% and 6.60% but is beaten by qDKT by 3.10% on the NeurIPS34 dataset. Meanwhile, SAINT performs terrible in ASSIST2009 and Algebra2005 datasets, but is pretty good on NeurIPS34 data. We believe this is because the ASSIST2009 and Algebra2005 datasets are much sparser than NeurIPS34 dataset. As we can see from Table 1, the average number of questions per KC is 16.6 in NeurIPS34 dataset,

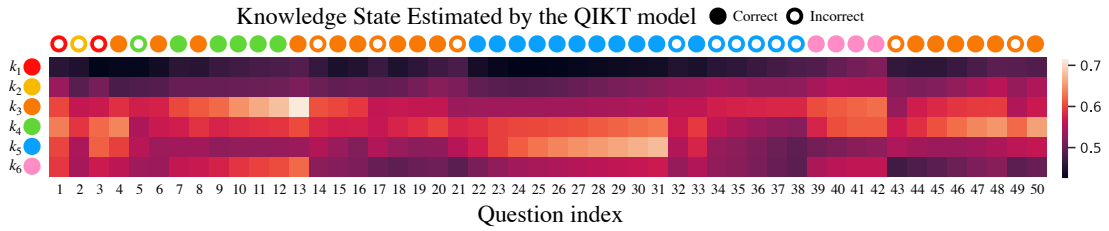

Figure 4: An example of a student knowledge states of 6 concepts as student's solve 50 questions of NeurIPS34.

which is much smaller compared to the numbers in ASSIST2009 and Algebra2005 (144.2 and 1545.6) datasets. (3) Results between DeepIRT and DKVMN are very close on three datasets, which empirically shows that the IRT function won't sacrifice the model prediction ability too much. (4) AKT and IEKT are very strong baselines. Both of them use both question and KC related information, which further empirically verifies the importance of considering question-centric representations when building the DLKT models.

Qualitative Question-centric Effects. We qualitatively show the question-centric effects of our QIKT model. Figure 5 shows the progressive knowledge state estimations of one student with and without question-centric modules, i.e., $\hat{\mathbf{r}}$ v.s. $\sigma(\boldsymbol{\beta})$. As we can see, predictions ($\sigma(\boldsymbol{\beta})$) without the question-centric information from the KS module are relatively smooth compared with results from QIKT. We believe this is because the KS module mainly focuses on the knowledge states at KC level, which is insensitive with question variations. When considering both the question-centric knowledge acquisition information and the question-centric problem solving information, the model outputs distinct predictive results even for the homogeneous questions. Due to the space limit, more illustrative and fine-grained results of $\hat{\mathbf{r}}$, $\sigma(\boldsymbol{\alpha})$, $\sigma(\boldsymbol{\beta})$ and $\sigma(\boldsymbol{\zeta})$ are provided in Appendix A.4.

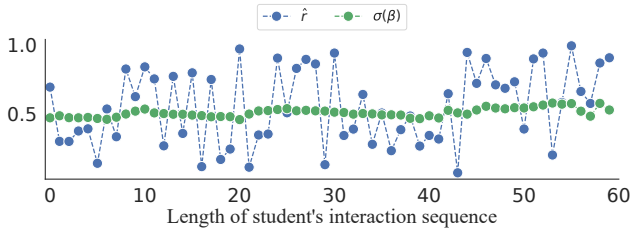

Figure 5: The outputs of QIKT ($\hat{r}$) and module KS ($\sigma(\beta)$) at each steps in student's interaction sequence.

Interpretable Student Diagnosis. To verify the interpretable and accurate estimations of students' knowledge states by the QIKT model, we randomly select a student sequence from NeurIPS34 and observe the knowledge state variations of the student in 50 questions with 6 KCs. From Figure 4, we observe that: (1) since the student always gives wrong responses to k_3 after answering the question 20 (e.g. question 21,43), the knowledge state of k_3 is constantly decline. On the other hand, the student always gives right answers to the questions (question 22-31) which are related to the k_5 hence the knowledge acquisition of k_5 is constantly increasing. (2) The knowledge states of little-attempted KCs are slightly lower than those of the diligent-attempted KCs. For example, the knowledge state of k_6 is relatively lower than others until the student attempts question 39.

Ablation Study. We examine the effect of key components by constructing four model variants in Table 3. "w/o" means excludes such module from QIKT. From Table 3, we can observe that (1) comparing QIKT and QIKT w/o IRT, we can see that our IRT based interpretable prediction layer is able to make a good enough trade-off between prediction performance and results interpretability. The AUC score of QIKT decreases 0.09% on the ASSIST2009 dataset and increases 0.81% and 0.07% on Algebra2005 and NeurIPS34 datasets. (2) compared to other variants (e.g., QIKT w/o KS, QIKT w/o PS, and QIKT w/o KS & PS) that have the IRT prediction layer, QIKT obtains the highest AUC score in all cases except QIKT w/o KS in the NeurIPS34 dataset. This suggests that prediction performance degrades when ignoring any type of question-centric information. Thus, it is important to incorporate question information in DLKT models.

Method	ASSIST2009	Algebra2005	NeurIPS34
QIKT	0.7878±0.0024	**0.8408±0.0007**	0.8044±0.0005
w/o IRT	**0.7887±0.0017**•	0.8327±0.0005*	0.8037±0.0004*
w/o KS	0.7813±0.0019*	0.8365±0.0008*	**0.8048±0.0002**•
w/o PS	0.7822±0.0022*	0.8345±0.0005*	0.8037±0.0002*
w/o KS & PS	0.7442±0.0043*	0.7487±0.0008*	0.8032±0.0002*

Table 3: The performance of different variants in QIKT. Marker $*$, $\circ$ and $\bullet$ indicates whether our proposed model is statistically superior/equal/inferior to the compared method (using paired t-test at 0.01 significance level).

Conclusions

In this paper, we propose an interpretable deep sequential KT model learning framework with question-centric cognitive representations. Comparing with existing DLKT models, our QIKT model is able to estimate students' knowledge acquisition and measure the student problem solving ability for each specific question. Furthermore, we design an IRT based interpretable layer to make the QIKT's prediction results more explainable. Quantitative and qualitative experiments on three real-world datasets show that QIKT outperforms other state-of-the-art DLKT models in AUC and generates explainable predictions for tutors and students.

Acknowledgments

This work was supported in part by National Key R&D Program of China, under Grant No. 2020AAA0104500; in part by Beijing Nova Program (Z201100006820068) from Beijing Municipal Science & Technology Commission; in part by NFSC under Grant No. 61877029 and in part by Key Laboratory of Smart Education of Guangdong Higher Education Institutes, Jinan University (2022LSYS003).

References

Abdelrahman, G.; and Wang, Q. 2019. Knowledge tracing with sequential key-value memory networks. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 175–184.

Cen, H.; Koedinger, K.; and Junker, B. 2006. Learning factors analysis–a general method for cognitive model evaluation and improvement. In *International Conference on Intelligent Tutoring Systems*, 164–175. Springer.

Chen, P.; Lu, Y.; Zheng, V. W.; and Pian, Y. 2018. Prerequisite-driven deep knowledge tracing. In *2018 IEEE International Conference on Data Mining*, 39–48. IEEE.

Choi, Y.; Lee, Y.; Cho, J.; Baek, J.; Kim, B.; Cha, Y.; Shin, D.; Bae, C.; and Heo, J. 2020. Towards an appropriate query, key, and value computation for knowledge tracing. In *Proceedings of the Seventh ACM Conference on Learning@Scale*, 341–344.

Converse, G.; Pu, S.; and Oliveira, S. 2021. Incorporating item response theory into knowledge tracing. In *International Conference on Artificial Intelligence in Education*, 114–118. Springer.

Corbett, A. T.; and Anderson, J. R. 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-adapted Interaction*, 4(4): 253–278.

Ghosh, A.; Heffernan, N.; and Lan, A. S. 2020. Context-Aware Attentive Knowledge Tracing. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Guo, X.; Huang, Z.; Gao, J.; Shang, M.; Shu, M.; and Sun, J. 2021. Enhancing Knowledge Tracing via Adversarial Training. In *Proceedings of the 29th ACM International Conference on Multimedia*, 367–375.

Käser, T.; Klingler, S.; Schwing, A. G.; and Gross, M. 2017. Dynamic Bayesian networks for student modeling. *IEEE Transactions on Learning Technologies*, 10(4): 450–462.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*.

Lavoué, E.; Monterrat, B.; Desmarais, M.; and George, S. 2018. Adaptive gamification for learning environments. *IEEE Transactions on Learning Technologies*, 12(1): 16–28.

Lee, J.; and Yeung, D.-Y. 2019. Knowledge query network for knowledge tracing: How knowledge interacts with skills. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, 491–500.

Liu, Q.; Huang, Z.; Yin, Y.; Chen, E.; Xiong, H.; Su, Y.; and Hu, G. 2019. EKT: Exercise-aware knowledge tracing for student performance prediction. *IEEE Transactions on Knowledge and Data Engineering*, 33(1): 100–115.

Liu, T.; Fang, Q.; Ding, W.; Li, H.; Wu, Z.; and Liu, Z. 2021a. Mathematical Word Problem Generation from Commonsense Knowledge Graph and Equations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 4225–4240.

Liu, Y.; Yang, Y.; Chen, X.; Shen, J.; Zhang, H.; and Yu, Y. 2021b. Improving knowledge tracing via pre-training question embeddings. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, 1556–1562.

Liu, Z.; Liu, Q.; Chen, J.; Huang, S.; Gao, B.; Luo, W.; and Weng, J. 2023a. Enhancing Deep Knowledge Tracing with Auxiliary Tasks. In *Proceedings of the 2023 World Wide Web Conference*.

Liu, Z.; Liu, Q.; Chen, J.; Huang, S.; and Luo, W. 2023b. simpleKT: A Simple But Tough-to-Beat Baseline for Knowledge Tracing. In *International Conference on Learning Representations*.

Liu, Z.; Liu, Q.; Chen, J.; Huang, S.; Tang, J.; and Luo, W. 2022. pyKT: A Python Library to Benchmark Deep Learning based Knowledge Tracing Models. In *Thirty-sixth Conference on Neural Information Processing Systems*.

Long, T.; Liu, Y.; Shen, J.; Zhang, W.; and Yu, Y. 2021. Tracing Knowledge State with Individual Cognition and Acquisition Estimation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 173–182.

Lu, Y.; Wang, D.; Chen, P.; Meng, Q.; and Yu, S. 2022. Interpreting Deep Learning Models for Knowledge Tracing. *International Journal of Artificial Intelligence in Education*, 1–24.

Lu, Y.; Wang, D.; Meng, Q.; and Chen, P. 2020. Towards interpretable deep learning models for knowledge tracing. In *International Conference on Artificial Intelligence in Education*, 185–190. Springer.

Minn, S.; Yu, Y.; Desmarais, M. C.; Zhu, F.; and Vie, J.-J. 2018. Deep knowledge tracing and dynamic student classification for knowledge tracing. In *2018 IEEE International Conference on Data Mining*, 1182–1187. IEEE.

Nagatani, K.; Zhang, Q.; Sato, M.; Chen, Y.-Y.; Chen, F.; and Ohkuma, T. 2019. Augmenting knowledge tracing by considering forgetting behavior. In *The World Wide Web Conference*, 3101–3107.

Nakagawa, H.; Iwasawa, Y.; and Matsuo, Y. 2019. Graph-based knowledge tracing: modeling student proficiency using graph neural network. In *2019 IEEE/WIC/ACM International Conference on Web Intelligence*, 156–163. IEEE.

Pandey, S.; and Karypis, G. 2019. A self-attentive model for knowledge tracing. In *12th International Conference on Educational Data Mining*, 384–389. International Educational Data Mining Society.

Pandey, S.; and Srivastava, J. 2020. RKT: relation-aware self-attention for knowledge tracing. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 1205–1214.

Piech, C.; Bassen, J.; Huang, J.; Ganguli, S.; Sahami, M.; Guibas, L. J.; and Sohl-Dickstein, J. 2015. Deep knowledge tracing. *Advances in neural information processing systems*, 28.

Pu, S.; Yudelson, M.; Ou, L.; and Huang, Y. 2020. Deep knowledge tracing with transformers. In *International Conference on Artificial Intelligence in Education*, 252–256. Springer.

Pu, Y.; Wu, W.; Peng, T.; Liu, F.; Liang, Y.; Yu, X.; Chen, R.; and Feng, P. 2022. EAKT: Embedding Cognitive Framework with Attention for Interpretable Knowledge Tracing. *Scientific Reports*.

Shen, S.; Liu, Q.; Chen, E.; Huang, Z.; Huang, W.; Yin, Y.; Su, Y.; and Wang, S. 2021. Learning Process-consistent Knowledge Tracing. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 1452–1460.

Shen, S.; Liu, Q.; Chen, E.; Wu, H.; Huang, Z.; Zhao, W.; Su, Y.; Ma, H.; and Wang, S. 2020. Convolutional knowledge tracing: Modeling individualization in student learning process. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1857–1860.

Sonkar, S.; Waters, A. E.; Lan, A. S.; Grimaldi, P. J.; and Baraniuk, R. G. 2020. qDKT: Question-centric deep knowledge tracing. In *Proceedings of The 13th International Conference on Educational Data Mining*, 677–681.

Su, Y.; Liu, Q.; Liu, Q.; Huang, Z.; Yin, Y.; Chen, E.; Ding, C.; Wei, S.; and Hu, G. 2018. Exercise-enhanced sequential modeling for student performance prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Tong, S.; Liu, Q.; Huang, W.; Huang, Z.; Chen, E.; Liu, C.; Ma, H.; and Wang, S. 2020. Structure-based Knowledge Tracing: An Influence Propagation View. In *2020 IEEE International Conference on Data Mining*, 541–550. IEEE.

Wang, F.; Liu, Q.; Chen, E.; Huang, Z.; Chen, Y.; Yin, Y.; Huang, Z.; and Wang, S. 2020a. Neural cognitive diagnosis for intelligent education systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 6153–6161.

Wang, Z.; Feng, X.; Tang, J.; Huang, G. Y.; and Liu, Z. 2019. Deep knowledge tracing with side information. In *International Conference on Artificial Intelligence in Education*, 303–308. Springer.

Wang, Z.; Lamb, A.; Saveliev, E.; Cameron, P.; Zaykov, Y.; Hernández-Lobato, J. M.; Turner, R. E.; Baraniuk, R. G.; Barton, C.; Jones, S. P.; et al. 2020b. Instructions and Guide for Diagnostic Questions: The NeurIPS 2020 Education Challenge. *ArXiv preprint*, abs/2007.12061.

Yang, Y.; Shen, J.; Qu, Y.; Liu, Y.; Wang, K.; Zhu, Y.; Zhang, W.; and Yu, Y. 2020. GIKT: a graph-based interaction model for knowledge tracing. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 299–315. Springer.

Yeung, C.-K. 2019. Deep-IRT: Make deep learning based knowledge tracing explainable using item response theory. *arXiv preprint arXiv:1904.11738*.

Yeung, C.-K.; and Yeung, D.-Y. 2018. Addressing two problems in deep knowledge tracing via prediction-consistent regularization. In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*, 1–10.

Zhang, J.; Shi, X.; King, I.; and Yeung, D. Y. 2017. Dynamic Key-Value Memory Networks for Knowledge Tracing. In *Proceedings of the 26th International Conference on World Wide Web*, 765.

Zhang, M.; Zhu, X.; Zhang, C.; Ji, Y.; Pan, F.; and Yin, C. 2021. Multi-Factors Aware Dual-Attentional Knowledge Tracing. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 2588–2597.

Zhao, J.; Bhatt, S.; Thille, C.; Zimmaro, D.; and Gattani, N. 2020. Interpretable personalized knowledge tracing and next learning activity recommendation. In *Proceedings of the Seventh ACM Conference on Learning@Scale*, 325–328.