An Efficient and Generic Method for Interpreting Deep Learning based Knowledge Tracing Models

Deliang WANG[a,c], **Yu Lu** [a,b*], **Zhi ZHANG**[a] **& Penghe CHEN** [a,b]
[a]*School of Educational Technology, Faculty of Education, Beijing Normal University, China*
[b]*Advanced Innovation Center for Future Education, Beijing Normal University, China*
[c]*Faculty of Education, The University of Hong Kong, Hong Kong, China*
*luyu@bnu.edu.cn

Abstract: Deep learning-based knowledge tracing (DLKT) models have been regarded as the promising solution to estimate learners' knowledge states and predict their future performance based on historical exercise records. However, the increasing complexity and diversity make DLKT models still difficult for users, typically including both learners and teachers, to understand models' estimation results, directly hindering the model's deployment and application. Previous studies have explored using methods from explainable artificial intelligence (xAI) to interpret DLKT models, but the methods have been limited in their generalizing capability and inefficient interpreting procedures. To address these limitations, we proposed a simple but efficient model-agnostic interpreting method, called Gradient*Input, to explain the predictions made by these models in two datasets. Comprehensive experiments have been conducted on the existing five DLKT models with representative neural network architectures. The experiment results showed that the method was effective in explaining the predictions of DLKT models. Further analysis of the interpreting results revealed that all five DLKT models share a similar rule in predicting learners' item responses, and the role of skill and temporal information was found and discussed. We also suggested potential avenues for investigating the interpretability of DLKT models.

Keywords: Knowledge tracing models, deep learning, explainable artificial intelligence

1. Introduction

The ability to automatically identify learners' knowledge states is crucial for personalized learning, and it plays a fundamental role in sustaining learning motivation (Pelánek et al., 2017) and improving academic performance (Koedinger & Aleven, 2016). To achieve this, researchers have developed a range of knowledge tracing (KT) models by leveraging on learners' historical exercise records to predict their future performance. With the advancement of artificial intelligence (AI), KT models that employ deep learning techniques are considered effective due to their strong capability to capture inherent information. However, the complex structures and large number of variables in the deep learning-based knowledge tracing (DLKT) models make them difficult for users (e.g., teachers, students, and education researchers) to understand the models' decisions (Tsai & Gasevic, 2017), which may reduce users' trust and accordingly hinder the DLKT models' deployment and application, as indicated in the case of automated recommendation systems (Dietvorst et al., 2015). Additionally, blindly trusting the incorrect decisions would cause a wrong diagnosis of knowledge status and accordingly reduce learning efficiency.

To address the interpretability issue in deep learning knowledge tracing (DLKT) models, researchers have started exploring solutions. One approach is the incorporation component of the model remains a black box. Another approach is using the explainable artificial intelligence (xAI) techniques to interpret DLKT models, such as the model-specific layer-wise relevance propagation (LRP) method (Lu et al., 2020; Lu et al., 2022), but it is

only applicable to the specific DLKT models and hard to be generalized to other DLKT models. The DeepSHAP method has been also proposed as a more generic method for interpreting DLKT models (Wang et al., 2022), but it heavily relies on reference samples, which are normally scarce in the real cases. In addition, it is still in lack of systematic study on interpreting the existing representative DLKT models, typically including the recurrent neural networks (RNNs), memory-augmented neural networks (MANNs), attention, graph neural networks (GNNs), and convolutional neural networks (CNNs).

We thus propose a cost-effective and generic method to interpret the diverse DLKT models, where the model-agnostic interpreting method is designed to explain the predictions made by the five representative DLKT models. The comprehensive experiments have validated the effectiveness of the proposed method. Further investigations show that despite their different configurations, the five DLKT models follow a similar rule for estimating learners' knowledge states and making predictions.

2. Related Work

2.1 Deep Learning based Knowledge Tracing (DLKT) Models

DKT (Piech et al., 2015) was the pioneer DLKT model, and it utilized RNNs to achieve superior performance compared to traditional Bayesian knowledge tracing (BKT) (Corbett & Anderson, 1994). Researchers subsequently improved upon the DKT model by incorporating additional input features such as question difficulty (Sonkar et al., 2020) and extending the model's structure such as adding another layer of RNN (Su et al., 2018). Other types of deep neural networks have also been adopted for use in DLKT models. For example, because DKT summarizes the states of all concepts in a single vector, memory-augmented neural networks (MANNs) were used to build the DKVMN (Zhang et al., 2017), SKVMN (Abdelrahman & Wang, 2019), and EKT (Liu et al., 2019) models, which store the status of each concept in a separate space. Due to the limited performance of RNN-based and MANN-based KT models on concepts with little data, the attention mechanism has been employed to extract the similarity between different concepts and questions, leading to the development of the SAKT (Pandey & Karypis, 2019) and AKT (Ghosh et al., 2020) models. To capture the interconnected relationship between concepts and questions, graph neural networks (GNNs) have been used to build relationship graphs, resulting in the creation of the GKT (Nakagawa et al., 2019) and GIKT (Y. Yang et al., 2020) models. Additionally, convolutional neural networks (CNNs) have been explored for use in DLKT models, resulting in the development of the CKT (Shen et al., 2020) and CAKT (S. Yang et al., 2020).

2.2 Interpreting Methods for Deep Learning Models

Researchers have developed multiple interpreting methods to understand the internal workings and individual decisions of non-transparent deep learning models. These interpreting methods can be classified into model-specific and model-agnostic approaches (Adadi & Berrada, 2018). The former ones are tailored to the models with specific structures, while the latter ones can be applied to a wider range of models. Model-agnostic interpreting techniques include visualization (Goldstein et al., 2015), knowledge extraction (Hinton et al., 2015), influence methods (Cortez & Embrechts, 2011), and example-based explanations (Wachter et al., 2017). Influence methods are commonly used to explain models by estimating the importance or contribution of a feature to the model's prediction. Popular methods in this category include perturbation-based and backpropagation-based approaches, such as Gradient (Simonyan et al., 2013) and DeepSHAP (Lundberg & Lee, 2017).

Although there are a large number of interpreting methods in the field of explainable AI (xAI), currently only two methods (i.e., LRP and DeepSHAP) have been utilized to interpret DLKT models (Lu et al., 2020; Wang et al., 2022). These methods are limited in their generalizing capability or their high computational complexity, leading to the need for simpler yet effective methods to provide explanations of the diverse and complex DLKT models.

3. Interpreting Method

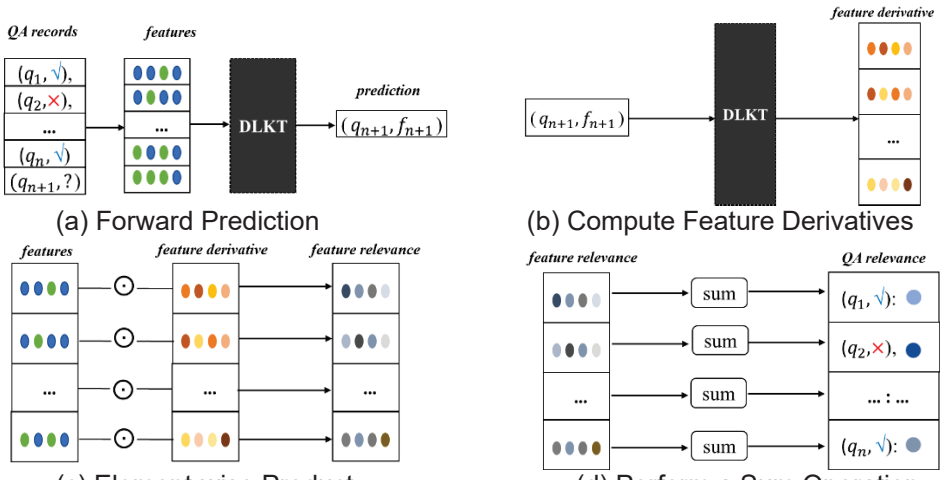
3.1 Method

We propose to utilize an existing xAI method, called Gradient*Input (Arras et al., 2019), for interpreting DLKT models. Compared to other methods previously investigated for DLKT models (i.e., LRP and DeepSHAP), Gradient*Input has two advantages. First, it can be applied to a wide range of models regardless of their internal structure. Second, it has relatively low computational complexity, requiring only one forward pass and one backward pass to obtain an explanation, and does not rely on auxiliary information, such as the reference samples used in DeepSHAP. And this method has been demonstrated to be effective in various tasks (Arras et al., 2019; Shrikumar et al., 2017).

The Gradient*Input method explains the model prediction by decomposing it into the contributions of the input features. Formally, given an input x with K features, the prediction of a deep learning model on class c is a highly non-linear function $f_c(x)$. Based on Taylor expansion (Li et al., 2016; Montavon et al., 2017), the non-linear prediction $f_c(x)$ is approximated by a linear function, as shown in Equation 1:

$$f_c(x) \approx \sum_{i=1}^{K} \frac{\partial f_c(x)}{\partial x_i} x_i + b \qquad (1)$$

where i represents the *i*-th feature in the input, $\sum_{i=1}^{K} \frac{\partial f_c(x)}{\partial x_i}$ represents the partial derivative of $f_c(x)$ with respect to x_i, and x_i reflects the contribution of feature x_i to the prediction.



(a) Forward Prediction (b) Compute Feature Derivatives

(c) Element-wise Product (d) Perform a Sum Operation

Figure 1. The interpreting procedure of Gradient*Input for DLKT models.

3.2 Explaining DLKT Models

Figure 1 illustrates the procedure of using Gradient*Input to interpret the predictions of DLKT models. Specifically, given a set of question-answer records from learners, denoted as $\{(q_1, a_1), (q_2, a_2), \ldots, (q_n, a_n)\}$, DLKT models can make predictions about their future performance, e.g., the probability $f(n+1)$ of correctly answering a future question q_{n+1}, as Figure 1(a) shows. To explain the prediction, Gradient*Input uses backpropagation to calculate the partial derivative of the prediction with respect to the input features, as depicted in Figure 1(b). Then, the element-wise product between the feature and its derivative is performed, which allows it to obtain the feature relevance (i.e., contributions) for each question-answer record, as Figure 1(c) indicates. Because each record is represented as a vector (e.g., an embedding or a one-hot vector with *m* dimensions), the sum of the feature relevance in a question-answer record can be used to determine the overall relevance of the record to the prediction (i.e., QA relevance), as shown in Figure 1(d).

4. Evaluation

4.1 Construction of DLKT Models

We select five typical and representative existing DLKT models for the experiment, including (1) DKT (Piech et al., 2015), the first DLKT model that adopts RNN as its inner structure; (2) DKVMN (Zhang et al., 2017), a DLKT model that adopts MANN to store concept status; (3) AKT (Ghosh et al., 2020), a DLKT model that employs attention to extract similarity between concepts and questions; (4) GKT (Nakagawa et al., 2019), A DLKT model that uses GNN to build skill relationship graphs; (5) CKT (Shen et al., 2020), A DLKT model that utilizes CNN to model individualization.

4.1.1 Dataset

We adopt two commonly used KT datasets, ASSISTment2009 and ASSISTment2015 (Feng et al., 2009). Specifically, the datasets were preprocessed to eliminate repetitive data and data without skill or question tags, and the length of the learner answer sequence was set to range from 10 to 200. After preprocessing, ASSISTment2009 contained 325,637 records on 110 skills and ASSISTment2015 contained 682,223 records on 100 skills. 80% of the data was randomly chosen for training, while the remaining 20% was used for testing.

4.1.2 Model Training

For all five models, we uniformly set the optimizer, dropout rate, mini-batch size, initial learning rate, and iteration epoch to Adam, 0.5, 64, 0.005, and 100, respectively. For the DKT model, the hidden dimension was set to 64. For the DKVMN model, the state dimension was set to 64 and the memory size was set to 110 for the ASSISTment2009 dataset and 100 for the ASSISTment2015 dataset. For the AKT model, the hidden dimension was set to 256 and the number of heads was set to 8. For the GKT model, the hidden dimension was set to 64 and the number of heads was set to 4. For the CKT model, the hidden dimension was set to 64. The performance (i.e., AUC and accuracy) of these five DLKT models in ASSISTment2009 and ASSISTment2015 can be seen in Table 1. Note that given the focus of this work is not on the model performance, we do not optimize the accuracy for each model.

Table 1. *The performance of five DLKT models in ASSISTment2009 and ASSISTment2015.*

Dataset	Metric	ASSISTment2009	ASSISTment2015
RNN-based DKT	AUC	0.74	0.72
	ACC	0.72	0.73
MANN-based DKVMN	AUC	0.76	0.72
	ACC	0.74	0.75
Attention-based AKT	AUC	0.75	0.73
	ACC	0.72	0.75
GNN-based GKT	AUC	0.74	0.72
	ACC	0.73	0.74
CNN-based CKT	AUC	0.75	0.72
	ACC	0.72	0.74

4.2 Method Validation

We first validate the capability of the proposed method on interpreting the decisions of DLKT models. We split the test data for both ASSISTment2009 and ASSISTment2015 into sequences of 15 question-answer records, resulting in 48,670 sequences for ASSISTment2009 and 97,637 sequences for ASSISTment2015. In each sequence, the first

14 records were used as input to predict the correctness of the last record, allowing us to identify correctly-predicted sequences. The number of positive and negative predictions made by all five DLKT models in each dataset is shown in Table 2. By applying the proposed method, we could calculate the relevance of each question-answer record to the prediction.

Table 2. *The number of positive and negative predictions in correctly-predicted sequences.*

DLKT Models	ASSISTment2009			ASSISTment2015		
	Positive	Negative	Sum	Positive	Negative	Sum
RNN-based DKT	26,258	8,777	35,035	60,322	10,865	71,187
MANN-based DKVMN	27,477	7,660	35,137	62,975	8,694	71,669
Attention-based AKT	27,271	7,628	34,899	61,787	9,314	71,101
GNN-based GKT	27,005	8,066	35,071	59,690	11,229	70,919
CNN-based CKT	26,629	8,679	35,308	63,101	8,992	72,093

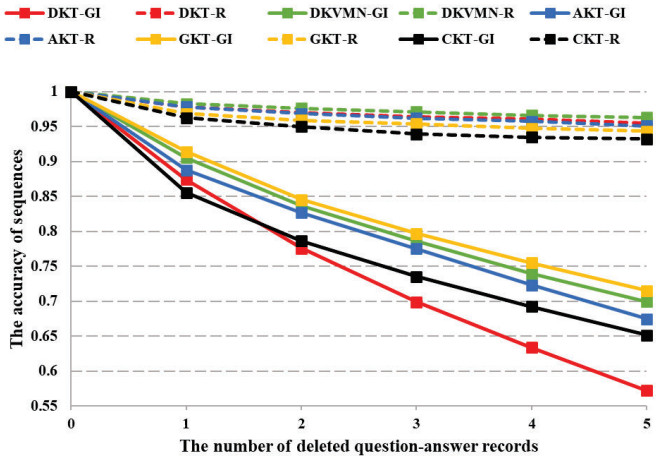


Figure 2. Question-answer record deletion results for DLKT models in ASSISTment2009.

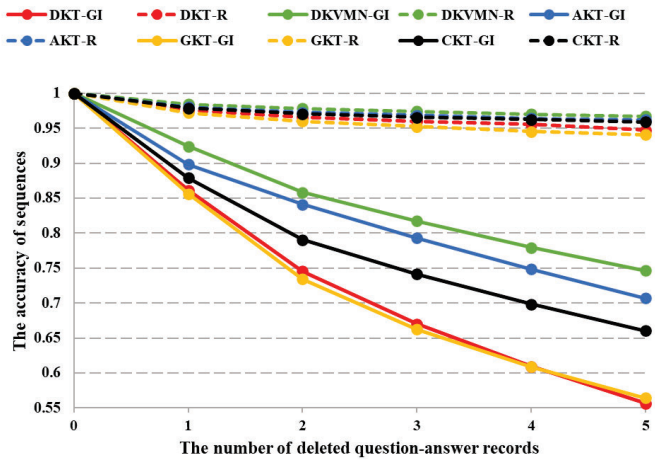


Figure 3. Question-answer record deletion results for DLKT models in ASSISTment2015.

To validate the effectiveness of the relevance of each question-answer record, we further conducted the experiment by removing question-answer records with high relevance from each sequence and observing the resulting change in accuracy. Specifically, we deleted question-answer records in the descending order of relevance for positive predictions (i.e., predictions of a correct response in the last question), and in the ascending order of relevance for negative predictions (i.e., predictions of a wrong response in the last question). By observing the change in accuracy, we could evaluate whether question-answer

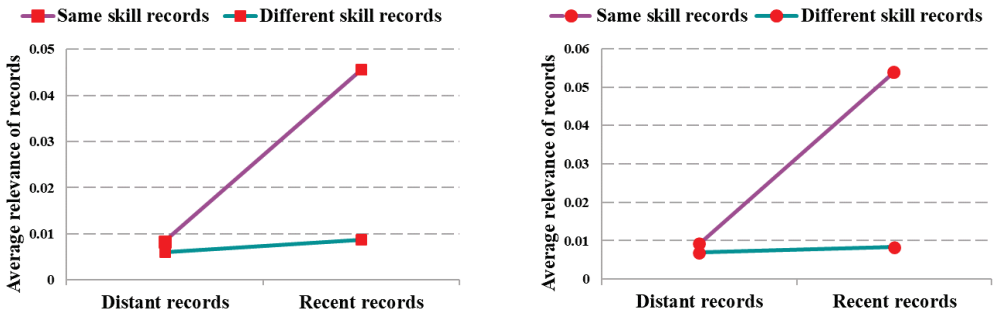
records with high relevance were more important for model predictions. We also performed random removal of records for comparison.

Figure 2 and 3 show the results of deleting question-answer records for all five DLKT models in the ASSISTment2009 and ASSISTment2015 datasets, where DLKT-GI represents the proposed method and DLKT-R means random deletion. As can be seen, compared to random deletion, the proposed method leads to a significant drop in accuracy. For example, in the case of the DKT model, removing 5 question-answer records based on relevance causes the prediction accuracy to drop from 100% to around 57%, while random deletion only slightly reduces the accuracy to around 95%. Similarly, for other DLKT models, deletions based on relevance cause the accuracy to drop to a range between 65% and 71%, while random deletions only reduce the accuracy to around 95%. Overall, the results suggest that the proposed method is effective in explaining the decisions of all five DLKT models.

4.3 Interpreting Model Rules

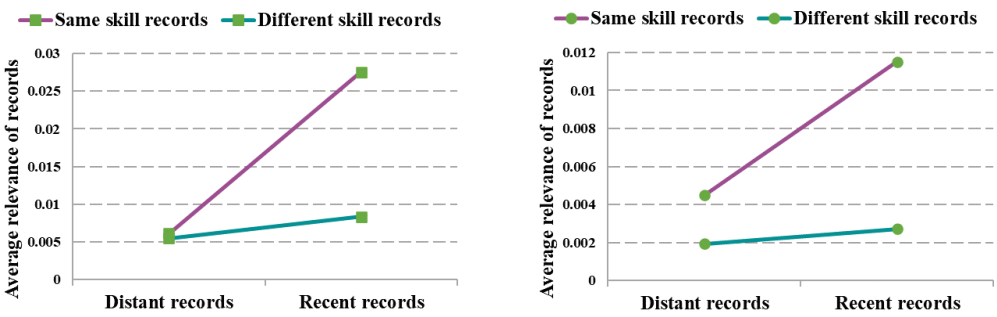
With the validated QA relevance, we further interpret how the DLKT models make predictions. We mainly analyze how the model input question-answer records influence the model output, i.e., the predictions on learners' future performance. In particular, we consider the effects of both the skill information (i.e., the specific skill being tested) and temporal information (i.e., when the exercise was completed) on the model's decision.

We selected correctly-predicted sequences and set the skill and position of the last record in each sequence as the target skill and prediction position. Then each question-answer record in the sequence was tagged as either *Same skill* or *Different skill* based on whether its skill matched the target skill, and *Recent* (i.e., the first half QA records) and *Distant* (i.e., the second half QA records) based on its distance from the prediction position. This resulted in four groups (i.e., 2 × 2): *Distant & Same skill, Distant & Different skill, Recent & Same skill*, and *Recent & Different skill*. For each group, the absolute relevance of each question-answer record was summed to compute the mean. We computed and compared the mean of four groups among all correctly-predicted sequences across the five DLKT models in the ASSISTment2009 and ASSISTment2015 datasets.



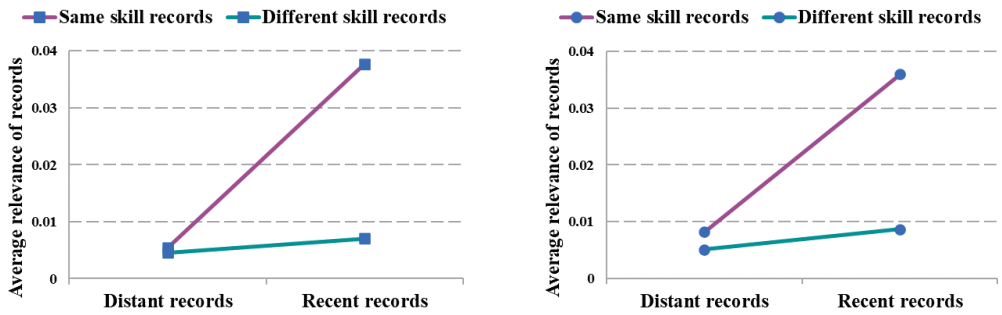
(a) ASSISTment2009 (b) ASSISTment2015

Figure 4. Skill-distance relevance comparison for RNN-based DKT.



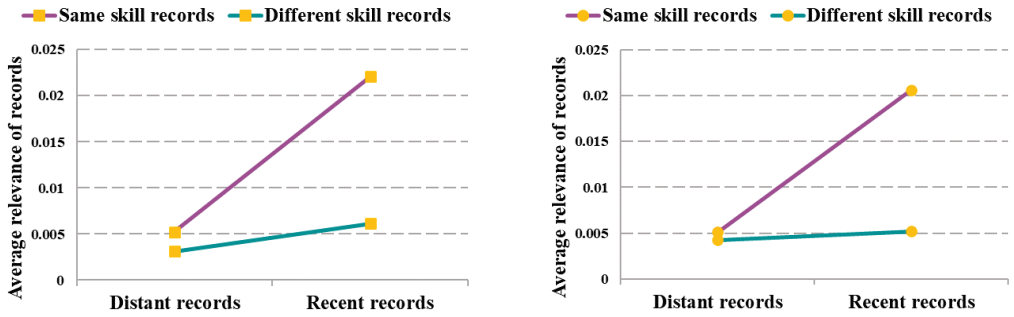
(a) ASSISTment2009 (b) ASSISTment2015

Figure 5. Skill-distance relevance comparison for MANN-based DKVMN.



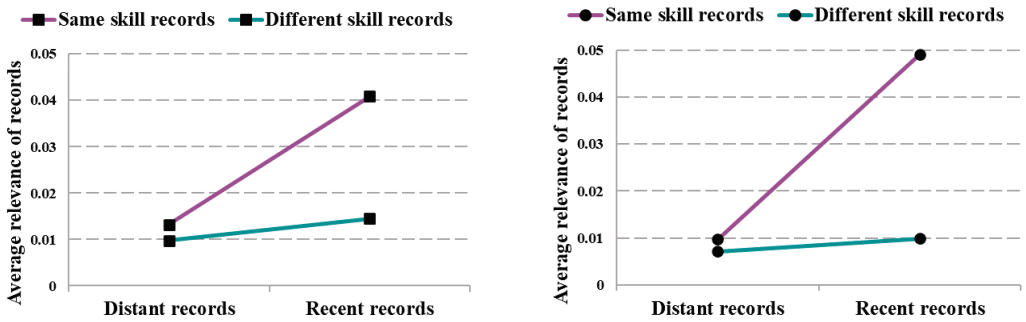
(a) ASSISTment2009 (b) ASSISTment2015

Figure 6. Skill-distance relevance comparison for Attention-based AKT.



(a) ASSISTment2009 (b) ASSISTment2015

Figure 7. Skill-distance relevance comparison for GNN-based GKT.



(a) ASSISTment2009 (b) ASSISTment2015

Figure 8. Skill-distance relevance comparison for CNN-based CKT.

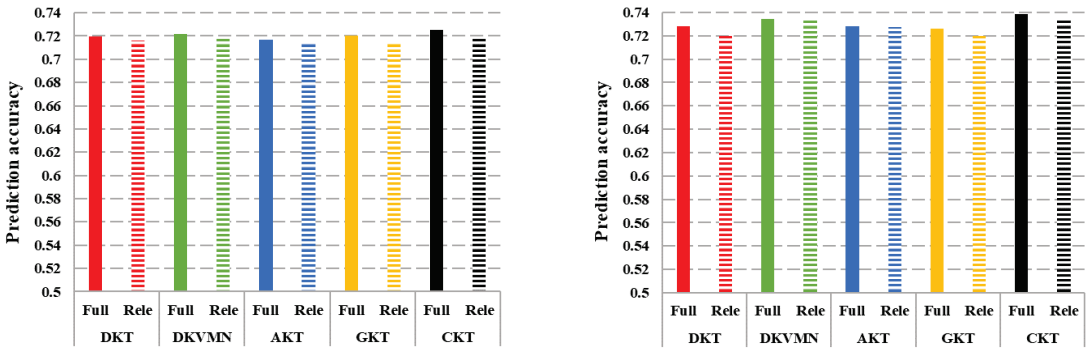
The results, shown in Figure 4, 5, 6, 7, and 8, indicate that for all five DLKT models, the recent-same-skill records (i.e., records that are close to the prediction position and on the target skill) have the highest relevance to the prediction, while those distant-different-skill records (i.e., records that are far from the prediction position and on non-target skills) have the lowest relevance. For example, in the case of AKT, the relevance of recent-same-skill records is approximately 0.036 in both the ASSISTment2009 and ASSISTment2015 datasets, while the relevance of distant-different-skill records is only about 0.005. The other four DLKT models show similar results for these two groups. Furthermore, it is difficult to differentiate between the relevance of recent-different-skill records or distant-same-skill records, as these groups show similar relevance in some cases (e.g., DKT and CKT) and differing relevance in others (e.g., DKVMN in ASSISTment2009 and ASSISTment2015).

We also find that the same skill records are more sensitive to distance (i.e., time) compared to different skill records. Specifically, for all five DLKT models in both ASSISTment2009 and ASSISTment2015, the average relevance difference between recent-same-skill records and distant-same-skill records is much larger than that between recent-

different skill records and distant-different-skill records. In contrast, distance seems to have a small effect on the relevance of different skill records for model prediction. For instance, the relevance of recent-same-skill records for the DKT model in ASSISTment2015 is about 0.055, 0.045 higher than distant-same-skill records, while the difference between recent-different-skill records and distant-different-skill records is close to 0.002. Other DLKT models show similar results.

Based on the findings discussed above, we conducted an additional experiment to further investigate the influence of skill and distance information on the decisions of DLKT models. Specifically, for each input sequence of length 15, we excluded the records that were distant from the prediction position and on non-target skills (i.e., distant-different-skill records) and thus only kept the records that were close to the prediction position or on the target skill (i.e., recent-same-skill records, recent-different-skill records, and distant-same-skill records). The leftover records made up 73% of the raw records in the ASSISTment2009 dataset and 63% in the ASSISTment2015 dataset. We then compared the model prediction performance with cases using all the raw records.

Figure 9 illustrates the experiment results: we see that for all five DLKT models, despite containing significantly fewer data points, the models with the leftover records have similar prediction accuracy to the models with raw full sequences. For example, DKVMN achieves 0.734 in accuracy with 1,464,555 records in ASSISTment2015. After excluding about 37% of data, the accuracy of DKVMN remains at 0.734 with 932,799 records. The experiment results partially validate the findings and the rules that how the DLKT models utilize the skill and distance information to make the decision.



(a) ASSISTment2009 (b) ASSISTment2015

Figure 9. The influence of skill and distance on the decisions of DLKT models.

5. Conclusion

In this work, we propose a simple and efficient xAI method to address the interpretability issue of increasingly complex and diverse DLKT models. The method only requires a lower complexity and can be used to explain a wide range of DLKT models. The experiment results on five DLKT models in two datasets validate the effectiveness of the proposed method. Further analysis of the explanations reveals that all five DLKT models use a similar rule when making the decision: question-answer records that are close to the predicted question and on the same skill as the predicted question are found to be the most relevant indicators, while records that are distant and on different skills are the least important. We also have found that records of the same skill are more sensitive to changes in distance compared to records of different skills. Additionally, it is also observed that using fewer but relevant question-answer records to make predictions can achieve similar accuracy as using full sequences, which supports the findings about the decision rules of DLKT models.

This work has significant impacts on practice. First, the rules obtained can be integrated into intelligent tutoring systems that utilize DLKT models. Automatically identifying learners' knowledge states and providing explanations can potentially increase their trust in the system and help them adjust their learning behavior when receiving incorrect diagnoses of their knowledge states. Second, the findings about the decisions of DLKT models

contribute to making these models more transparent and provide valuable insights for researchers to design more interpretable KT models. It can be a promising direction for future research to evaluate the impact of these explanations on education and consider the effect of skill and temporal information when designing DLKT models.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (no. 62077006, 62177009) and in part by the Fundamental Research Funds for the Central Universities.

References

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

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE access, 6,* 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Arras, L., Osman, A., Müller, K.-R., & Samek, W. (2019). Evaluating recurrent neural network explanations. *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 113–126.

Corbett, A.T., & Anderson, J.R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction, 4*(4), 253–278. https://doi.org/10.1007/BF01099821

Cortez, P., & Embrechts, M. J. (2011). Opening black box data mining models using sensitivity analysis. *2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, 341–348.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General, 144*(1), 114–126. https://doi.org/10.1037/xge0000033

Feng, M., Heffernan, N., & Koedinger, K. (2009). Addressing the assessment challenge with an online system that tutors as it assesses. *User Modeling and User-Adapted Interaction*, *19*(3), 243–266. https://doi.org/10.1007/s11257-009-9063-7

Ghosh, A., Heffernan, N., & Lan, A. S. (2020). Context-aware attentive knowledge tracing. *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, 2330–2339.

Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, *24*(1), 44–65. https://doi.org/10.1080/10618600.2014.907095

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. *stat*, *1050*, 9.

Koedinger, K. R., & Aleven, V. (2016). An interview reflection on “intelligent tutoring goes to school in the big city”. *International Journal of Artificial Intelligence in Education*, *26*(1), 13–24.

Li, J., Chen, X., Hovy, E., & Jurafsky, D. (2016). Visualizing and understanding neural models in nlp. *International Conference of the North American Chapter of the Association for Computational Linguistics*, 681–691.

Liu, Q., Huang, Z., Yin, Y., Chen, E., Xiong, H., Su, Y., & Hu, G. (2019). Ekt: Exerciseaware 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

Lu, Y., Wang, D., Chen, P., Meng, Q., & Yu, S. (2022). Interpreting deep learning models for knowledge tracing. *International Journal of Artificial Intelligence in Education*, 1–24. https://doi.org/10.1007/s40593-022-00297-z

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

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

Montavon, G., Lapuschkin, S., Binder, A., Samek, W., & Müller, K.-R. (2017). Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern recognition*, *65*, 211–222. https://doi.org/10.1016/j.patcog.2016.11.008

Nakagawa, H., Iwasawa, Y., & Matsuo, Y. (2019). Graph-based knowledge tracing: Modeling student proficiency using graph neural network. *2019 IEEE/WIC/ACM International Conference On Web Intelligence (WI)*, 156–163.

Pandey, S., & Karypis, G. (2019). A self-attentive model for knowledge tracing. *12th International Conference on Educational Data Mining, EDM 2019*, 384–389.

Pelánek, R., Papoušek, J., Řihák, J., Stanislav, V., & Nižnan, J. (2017). Elo-based learner modeling for the adaptive practice of facts. *User Modeling and User-Adapted Interaction*, *27*(1), 89–118. https://doi.org/10.1007/s11257-016-9185-7

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*.

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

Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. *International conference on machine learning*, 3145–3153.

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Sonkar, S., Lan, A. S., Waters, A. E., Grimaldi, P., & Baraniuk, R. G. (2020). Qdkt: Question-centric deep knowledge tracing. *Proceedings of the 13th International Conference on Educational Data Mining, EDM 2020, Fully virtual conference, July 1013, 2020*.

Su, Y., Liu, Q., Liu, Q., Huang, Z., Yin, Y., Chen, E., Ding, C., Wei, S., & Hu, G. (2018). Exercise-enhanced sequential modeling for student performance prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1).

Tsai, Y. S., & Gasevic, D. (2017). Learning analytics in higher education---challenges and policies: a review of eight learning analytics policies. *Proceedings of the seventh international learning analytics & knowledge conference,* 233-242.

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech., 31,* 841.

Wang, D., Lu, Y., Zhang, Z., & Chen, P. (2022). A generic interpreting method for knowledge tracing models. *International Conference on Artificial Intelligence in Education*, 573–580.

Yang, S., Zhu, M., & Lu, X. (2020). Deep knowledge tracing with learning curves. *arXiv preprint arXiv:2008.01169*.

Yang, Y., Shen, J., Qu, Y., Liu, Y., Wang, K., Zhu, Y., Zhang, W., & Yu, Y. (2020). Gikt: A graph-based interaction model for knowledge tracing. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 299–315.

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

Zhang, J., Shi, X., King, I., & Yeung, D.-Y. (2017). Dynamic key-value memory networks for knowledge tracing. *Proceedings of the 26th international conference on World Wide Web*, 765–774.