Towards Robust Knowledge Tracing Models via *k*-Sparse Attention

Shuyan Huang
Think Academy International Education
TAL Education Group
Beijing, China
huangshuyan@tal.com

Zitao Liu*
Guangdong Institute of Smart Education
Jinan University
Guangzhou, China
liuzitao@jnu.edu.cn

Xiangyu Zhao
Applied Machine Learning Lab
School of Data Science
City University of Hong Kong
Hong Kong, China
xianzhao@cityu.edu.hk

Weiqi Luo
Guangdong Institute of Smart Education
Jinan University
Guangzhou, China
lwq@jnu.edu.cn

Jian Weng
College of Cyber Security
Jinan University
Guangzhou, China
cryptjweng@gmail.com

ABSTRACT

Knowledge tracing (KT) is the problem of predicting students' future performance based on their historical interaction sequences. With the advanced capability of capturing contextual long-term dependency, attention mechanism becomes one of the essential components in many deep learning based KT (DLKT) models. In spite of the impressive performance achieved by these attentional DLKT models, many of them are often vulnerable to run the risk of overfitting, especially on small-scale educational datasets. Therefore, in this paper, we propose sparseKT, a simple yet effective framework to improve the robustness and generalization of the attention based DLKT approaches. Specifically, we incorporate a k-selection module to only pick items with the highest attention scores. We propose two sparsification heuristics : (1) soft-thresholding sparse attention and (2) top-K sparse attention. We show that our sparseKT is able to help attentional KT models get rid of irrelevant student interactions and improve the predictive performance when compared to 11 state-of-the-art KT models on three publicly available real-world educational datasets. To encourage reproducible research, we make our data and code publicly available at https://github.com/pykt-team/pykt-toolkit[1].

CCS CONCEPTS

• **Social and professional topics** → **Student assessment**; • **Applied computing** → **Learning management systems**.

*The corresponding author: Zitao Liu.

[1]We merged our model to the pyKT benchmark at https://pykt.org/.

SIGIR '23, July 23–27, 2023, Taipei, Taiwan

ACM ISBN 978-1-4503-9408-6/23/07…$15.00
https://doi.org/10.1145/3539618.3592073

KEYWORDS

knowledge tracing, student modeling, AI in education, sparse attention, deep learning

ACM Reference Format:
Shuyan Huang, Zitao Liu, Xiangyu Zhao, Weiqi Luo, and Jian Weng. 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 (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3539618.3592073

1 INTRODUCTION

The process of knowledge tracing involves utilizing a student's past learning interactions to construct a model to estimate his/her knowledge mastery to predict his/her future performance over a period of time (as shown in Figure 1). Such predictive abilities have the potential to improve students' learning outcomes and accelerate their progress when combined with high-quality learning materials and instructions. Recently, with the remarkable capability of attention mechanisms in natural language processing (NLP) or computer vision (CV) tasks, many DLKT models achieve accurate students' knowledge state estimations by utilizing attention networks to capture the intrinsic relevance among past interactions [6, 10, 15].

Although the attentional DLKT approaches have achieved impressive results, they may run the risk of overfitting in real-world educational scenarios. Educational data is usually limited compared to large-scale language or image data. In the KT datasets, the question bank is usually bigger than the set of knowledge components (KCs) and a student has a very small number of question responses. Furthermore, since questions may be associated with limited relevant KCs, not all the past question responses contribute equally to the KT prediction task [1]. For instance, during predicting the student's performance for q_6 in Figure 1, the KT models need easily look back to important historical information of the student's response in q_1 due to both of them are associated by c_3. Besides the question q_1, other questions likely have limited correlations to q_6. However, due to the smooth characteristic of

the softmax function, irrelevant questions such as q_1 to q_5 may still get moderate attention scores in previous attention-based KT methods to predict the student's performance on q_6, which hinder the accurate inference of the student's knowledge state.

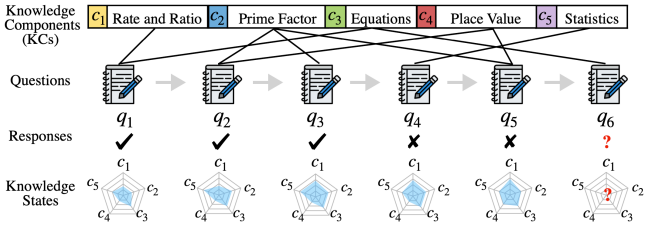

Figure 1: An illustration of the KT problem. A KC is a generality of everyday terms like concept, principle, or skill.

Therefore, in this work, to improve the robustness of attentional DLKT models meanwhile preserving the generalization performance under the assumption that it enables models to focus on the influential question inputs, we propose sparseKT that utilize sparse attention techniques to allow knowledge state estimations to be mapped to a limited number of pivotal interactions [2, 13]. Specifically, our sparseKT approach focuses on the refinements of a popular attention based KT model: the Self Attentive Knowledge Tracing (SAKT) [15]. SAKT is a classical and widely used model for KT due to its relative simplicity, mathematically predictable behavior, and the fact that it handles the data sparsity problem based on relatively few past interactions. However, the pure attention mechanism in SAKT gives weights to each historical interaction of students, which may bring noise to the model and interfere with the accurate knowledge state estimations. Hence, we aim to develop a sparse attention framework with two simple but effective sparse attention heuristics to extract the relevant information from students' past learning sequences to perform better predictions when excluding interferences from other historical interactions. Our sparse attention methods refine the original dot-product attention by selecting k most influential weights. We evaluate sparseKT on three benchmark datasets by comparing it with 11 previous approaches under a rigorous KT evaluation protocol [3, 9, 11]. Experimental results demonstrate that our sparseKT approach achieves superior prediction performance.

2 PRELIMINARY

2.1 Self Attentive Knowledge Tracing

The self-attentive knowledge tracing (SAKT) model is the first method to use attention mechanisms in the context of KT [15]. The standard encoder in the Transformer model is employed in the basic setup of SAKT to extract context-aware interaction information through historical query and key-value pairs for the KT scenario [19]. The definitions of the query and key-value pairs are as follows:

$$\begin{aligned} &\mathbf{h}_{t+1} = \text{SelfAttention}(Q, K, V) \\ &Q = \mathbf{z}_{t+1}; \quad K = \{\mathbf{y}_1, \cdots, \mathbf{y}_t\}; \quad V = \{\mathbf{y}_1, \cdots, \mathbf{y}_t\} \\ &\mathbf{y}_t = \mathbf{z}_t \oplus \mathbf{r}_t; \quad \mathbf{z}_t = \mathbf{W}^c \cdot \mathbf{e}_t^c; \quad \mathbf{r}_t = \mathbf{W}^q \cdot \mathbf{e}_t^q \end{aligned} \tag{1}$$

where $\mathbf{h}_{t+1}$ is the learned context-aware latent representation that summarizes all available information for prediction at the $(t+1)$th timestamp. $\mathbf{z}_t$ and $\mathbf{y}_t$ denote the latent representations of the related KCs and interaction at timestamp t. $\mathbf{r}_t$ denotes the representation of student response. $\mathbf{e}_t^c$ is the n-dimensional one-hot vector indicating the corresponding KC and $\mathbf{e}_t^q$ is the 2-dimensional one-hot vector indicating whether the question is answered correctly. $\mathbf{z}_t$ and $\mathbf{r}_t$ are d-dimensional learnable real-valued vectors. $\mathbf{W}^c \in \mathbb{R}^{d \times n}$ and $\mathbf{W}^q \in \mathbb{R}^{d \times 2}$ are learnable linear transformation operations. $\oplus$ is the element-wise addition operator and $\cdot$ is the standard matrix/vector multiplication. n is the total number of KCs.

2.2 Related Work

2.2.1 Attention based Knowledge Tracing. Attention based KT models utilize attention mechanisms to capture the intrinsic dependencies among students' chronologically ordered interactions. SAKT is the first research work that adopted a self-attention network to predict students' future performance [15]. Since then, many KT methods use attention networks to capture the potential relations between questions and responses [5, 6, 16]. Choi et al. applied deep self-attentive layers in a pure Transformer architecture to extract the question and response representations [5]. Ghosh et al. proposed a monotonic attention mechanism that computes attention weights with exponential time-related decay [6].

2.2.2 Sparse Attention. Sparse attention improves the ordinary attention mechanism by only computing a limited selection of similarity scores from a sequence rather than all possible pairs [4, 13, 17, 24], which has shown promising performance in NLP and CV domains [13, 24]. For example, Martins and Astudillo proposed a new activation function called sparsemax that is able to output sparse probabilities rather than traditional softmax [13]. Child et al. introduced several sparse factorizations of the attention matrix without sacrificing performance [4]. Zhao et al. designed an explicit sparse Transformer by selecting k most relevant components [24].

3 THE SPARSEKT APPROACH

We propose that to improve the performance of attention-based knowledge tracing models, it is necessary to further enhance the generalization of the model. One way to accomplish this is by incorporating recent advancements in attention sparsification techniques. Therefore, in this paper, we propose sparseKT, a simple yet effective framework to facilitate the robustness of the attention based KT approach. Briefly, our sparseKT approach incorporates an additional k-sparse selection module after the standard self-attention function to only select the top K interactions with highest attention scores. Only the selected K interactions are used to make future predictions. The idea of sparseKT is illustrated in Figure 2.

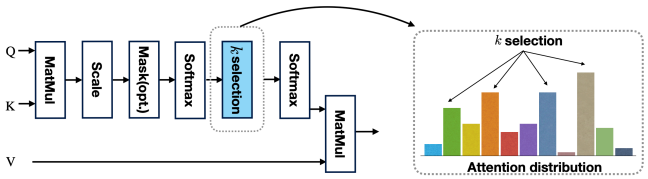

Figure 2: The sparseKT illustration.

3.1 Embedding

Inspired by the classic and simple Rasch model in psychometrics that explicitly uses a scalar to characterize the latent factor

of question discrimination, we improve the SAKT's interaction representation (shown in eq.(1)) by utilizing a question-specific discrimination factor to capture the individual differences among questions on the same KC. Specifically, let $\mathbf{x}_t$ be the enhanced representations that contain question-centric information, i.e.,

$$\mathbf{x}_t = \mathbf{m}^q \odot \mathbf{v}^c \oplus \mathbf{z}_t$$

where $\mathbf{m}^q$ denotes the question-specific discrimination factor of question q_t and $\mathbf{v}^c$ represents the variation of q_t covering this KC set $\mathbf{c}$. Both $\mathbf{q}_t$ and $\mathbf{v}^c$ are d-dimensional learnable real-valued vectors. $\odot$ is the element-wise product operator.

3.2 k-Sparse Attention

In this section, we leverage sparsification techniques to enhance the generalization of KT models for better performance. In our work, historical interactions that have limited correlations to the current question will not be assigned to the attention scores. More specifically, we enhance the SAKT's scaled dot-product attention mechanism by using sparse attention to allow the model to focus on only a few pieces of historical information through explicit selection. Let $\mathcal{I}$ be the attention distribution computed from our question enhanced representations $\mathbf{x}_{t+1}$, i.e.,

$$\mathcal{I} = \text{softmax}(\frac{QK^T}{\sqrt{d}}); Q = \mathbf{x}_{t+1}; K = \{\mathbf{x}_1, \cdots, \mathbf{x}_t\}$$

Let $\mathcal{M}(\cdot)$ be the mask operation that selects the k most informative historical interactions, $\mathcal{I}_i$ denotes the attention score of $\mathbf{x}_i$. There are two implementations of $\mathcal{M}(\cdot)$ including soft-thresholding and top-K sparse attention, described as follows:

- **soft-thresholding sparse attention**: we order all the attention scores $\mathcal{I}_i$ of attention distribution $\mathcal{I}$ from largest to smallest. And gradually pick up $\mathcal{I}_i$ into a weighting set $\mathcal{N} = \mathcal{N} \cup \{\mathcal{I}_i\}$ until the cumulative sum of $\mathcal{I}_i$s in $\mathcal{N}$ is larger than the predefined soft-threshold k. Hence, we treat all the historical interactions with the attention scores $\mathcal{I}_i$ as the most contributive ones to predict a student's future performance. Other interactions are likely to be irrelevant to the prediction which will not be assigned attention scores by:

$$\mathcal{M}_{\text{soft}}(\mathcal{I}_i) = \begin{cases} \mathcal{I}_i & \text{if } \mathcal{I}_i \in \mathcal{N} \\ -\infty & \text{otherwise} \end{cases}$$

- **top-K sparse attention**: let s be the k-th largest value in $\mathcal{I}$. We select the top k largest scores of in $\mathcal{I}$ as the most influential components. Practically, if $\mathcal{I}_i$ is larger than s, we will select $\mathcal{I}_i$ and vice versa, i.e.,

$$\mathcal{M}_{\text{topK}}(\mathcal{I}_i) = \begin{cases} \mathcal{I}_i & \text{if } \mathcal{I}_i \geq s \\ -\infty & \text{otherwise} \end{cases}$$

We then re-normalized the attention score distribution I, and the normalized scores of $\mathcal{I}_i$ in negative infinity are approximately 0. Therefore, we get a sparse attention distribution that explicitly chooses the highest attention scores that may influence the KT model decision. Finally, we obtain the knowledge state representation of q_{t+1} as:

$$\mathbf{h}_{t+1} = \text{softmax}(\mathcal{M}(\mathcal{I}))V; V = \{\mathbf{y}_1, \cdots, \mathbf{y}_t\}$$

3.3 Prediction Layer

We use a two-layer fully connected network to refine the knowledge state and the overall optimization function is as follows:

$$\hat{r}_{t+1} = \sigma(\mathbf{w}^\top \cdot \text{ReLU}\big(\mathbf{W}_2 \cdot \text{ReLU}(\mathbf{W}_1 \cdot [\mathbf{h}_{t+1}; \mathbf{x}_{t+1}] + \mathbf{b}_1) + \mathbf{b}_2\big) + b)$$
$$\mathcal{L} = -\sum_t \big(r_{t+1} \log \hat{r}_{t+1} \ + (1 - r_t) \log(1 - \hat{r}_{t+1})\big)$$

where $\mathbf{W}_1$, $\mathbf{W}_2$, $\mathbf{w}$, $\mathbf{b}_1$, $\mathbf{b}_2$ and b are trainable parameters and $\mathbf{W}_1 \in \mathbb{R}^{d \times 2d}$, $\mathbf{W}_2 \in \mathbb{R}^{d \times d}$, $\mathbf{w}, \mathbf{b}_1, \mathbf{b}_2 \in \mathbb{R}^{d \times 1}$, b is scalar. $\sigma(\cdot)$ is the sigmoid function.

4 EXPERIMENTS

We use three publicly educational datasets to evaluate the effectiveness of our model. We remove student sequences with fewer than 3 attempts and set the maximum length to 200 following the data preprocessing by [11]. The datasets are described as follows:

- ASSISTments2015[2] (AS2015): the dataset comprises mathematical exercises from the ASSISTments platform during the 2015 academic year. It ends up with 682,789 interactions, 19,292 sequences and 100 KCs after pre-processing.
- NeurIPS2020 Education Challenge[3] (NIPS34): the dataset is provided by NeurIPS 2020 Education Challenge. We use the dataset of Task 3 & Task 4 to evaluate our models [21]. There are 1,399,470 interactions, 9,401 sequences, 948 questions, 57 KCs.
- Peking Online Judge[4] (POJ): the dataset contains programming exercises on the Peking coding platform and is scraped by [16]. It has 987,593 interactions, 20,114 sequences and 2,748 questions.

We compare the two instances of our k-sparse self-attention framework, i.e., sparseKT-soft and sparseKT-topK to the following 11 KT models to evaluate the effectiveness of our approach:

- *DKT* [18]: uses an LSTM layer to encode the students' knowledge state for predicting their response performances.
- *DKT+* [22]: a variation of DKT that tackles the problems of reconstruction and non-consistent prediction.
- *KQN* [8]: calculates the relevance of student knowledge state encoder and skill encoder via the dot product.
- *DKVMN* [23]: exploits two memory networks to extract the relations between different KCs and students' knowledge states.
- *ATKT* [7]: exploits adversarial perturbations to the interaction embeddings to enhance the models' generalization capability.
- *GKT* [14]: utilizes graph neural networks to model the relation between KCs to predict the student's future performance on KCs.
- *SAKT* [15]: leverages a self-attention mechanism to capture relevance between exercises and responses.
- *SAINT* [5]: uses Transformer architecture to represent students' exercise and response sequences via encoder and decoder.
- *AKT* [6]: introduces monotonic attention to enhance self-attention by considering the students' forgetting behavior.
- *HawkesKT* [20]: utilizes the Hawkes process to model temporal cross-effects in student historical interactions.

[2] https://sites.google.com/site/assistmentsdata/datasets/2015-assistments-skill-builder-data.
[3] https://eedi.com/projects/neurips-education-challenge.
[4] https://drive.google.com/drive/folders/1LRljqWfODwTYRMPw6wEJ_mMt1KZ4xBDk.

- *IEKT* [12]: estimates students' knowledge states by modeling student cognition and knowledge acquisition behaviors.

4.1 Results

4.1.1 Overall Performance. Table 1 shows the overall performance of all models. From Table 1, we have the following observations: (1) our two sparseKT models all outperform 11 baselines in AS2015 and POJ datasets and have comparable performance with the IEKT in NIPS34 dataset; (2) compared to all attention based KT models, our two sparseKT approaches have the best performance on all three datasets. More importantly, our two sparseKT approaches are extensions of SAKT only by using sparse attention to replace scaled dot-product attention, they all have remarkable improvements of AUC scores by 3.26%, 5.21% and 2.67% on three datasets on average compared to SAKT. This indicates our sparse attention mechanism allows KT models to pay attention to limited influential historical information that improves the predictive performance; and (3) comparing sparseKT-topK and sparseKT-soft, we can see, sparseKT-topK performs slightly better than sparseKT-soft. We believe that the student performance on questions only depends on a very limited number of past interactions and the soft version of sparse attention may still bring some irrelevant information when estimating students' knowledge states compared to sparseKT-topK.

Table 1: AUC and accuracy results on AS2015, NIPS34 and POJ datasets. HAWKES and IEKT require both question IDs and KC IDs which are not available in AS2015 and POJ.

Model	AUC			Accuracy		
	AS2015	NIPS34	POJ	AS2015	NIPS34	POJ
DKT	0.7271±0.0005	0.7689±0.0002	0.6089±0.0009	0.7503±0.0003	0.7032±0.0004	0.6328±0.0020
DKT+	0.7285±0.0006	0.7696±0.0002	0.6173±0.0007	0.7510±0.0004	0.7039±0.0004	0.6482±0.0021
KQN	0.7254±0.0004	0.7684±0.0003	0.6080±0.0015	0.7500±0.0003	0.7028±0.0001	0.6435±0.0017
DKVMN	0.7227±0.0004	0.7673±0.0004	0.6056±0.0022	0.7508±0.0006	0.7016±0.0005	0.6393±0.0015
ATKT	0.7245±0.0007	0.7665±0.0001	0.6075±0.0012	0.7494±0.0002	0.7013±0.0002	0.6332±0.0023
GKT	0.7258±0.0012	0.7689±0.0024	0.6070±0.0036	0.7504±0.0010	0.7014±0.0028	0.6117±0.0147
SAKT	0.7114±0.0003	0.7517±0.0005	0.6095±0.0013	0.7474±0.0002	0.6879±0.0004	0.6407±0.0035
SAINT	0.7026±0.0011	0.7873±0.0007	0.5563±0.0012	0.7438±0.0010	0.7180±0.0006	0.6476±0.0003
AKT	0.7281±0.0004	0.8033±0.0003	0.6281±0.0013	0.7521±0.0005	0.7323±0.0005	0.6492±0.0010
HAWKES	-	0.7767±0.0010	-	-	0.7110±0.0007	-
IEKT	-	**0.8045±0.0002**	-	-	**0.7330±0.0002**	-
sparseKT-soft	0.7379±0.0018	0.8033±0.0007	0.6323±0.0034	0.7548±0.0011	0.7322±0.0012	0.6549±0.0019
sparseKT-topK	**0.7501±0.0004**	0.8043±0.0004	**0.6401±0.0013**	**0.7597±0.0008**	0.7325±0.0013	**0.6565±0.0024**

4.1.2 Impact of the Sparsity Level. We further explore the impacts of the sparse selection k on the model performance. We conduct experiments on the AS2015 to evaluate our two sparse attention approaches with different values of k on the validation set. The results are illustrated in Figure 3. We limit the range of k to $[0.1, 1]$ and $[1, 10]$ in soft-threshold and top-K sparse attention respectively. We observe that neither soft-threshold nor top-K sparse attention, the AUC scores increase at first and then decline with the increasing values of k. We suppose that if the value of k is small, e.g., k=1/2 in top-K sparse attention, the limited historical interactions are selected that the KT model can not obtain enough past learning information to predict students' future performance and hence get low AUC scores. With the increasing values of k, sparseKT-soft and sparseKT-topK obtain better performance gradually and perform best AUC scores when $k = 0.9/8$. After that, larger values of k may contain more noise that decreases the model's robustness and limit its generalization yield to get a decreasing AUC score.

4.1.3 Visualization of KC Relations via k-Sparse Attention. Figure 4 shows the KC relation visualization via our proposed k-sparse attention. We compute the cumulative sum of attention weights among all the KCs during the training of our sparseKT-topK. To better observe the relations, we compute the min-max normalization of the cumulative sum results. Due to the space limitation, we visualize the results between the top-6 KCs with the highest frequency on AS2015. We can see that, since pre-post sequence relations among KCs, the attention weights are different in the same KC pairs. For example, for a KC pair $<c_2,c_3>$, c_2 has a high influence (0.95 weight) on $c3$ when c_2 is the pre-interaction of $c3$. On the contrary, c_3 has a relatively small impact (0.43 weight) on $c2$ when c_3 is the pre-interaction of c_2. Furthermore, there is a limited correlation to $<c_2, c_8>$, so the attention weights between them are less than 0.2 regardless of which KC is the pre-interaction.

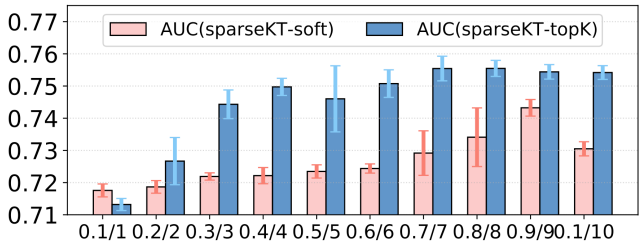

Figure 3: AUC performance of different values of k with our sparseKT-soft and sparseKT-topK on AS2015.

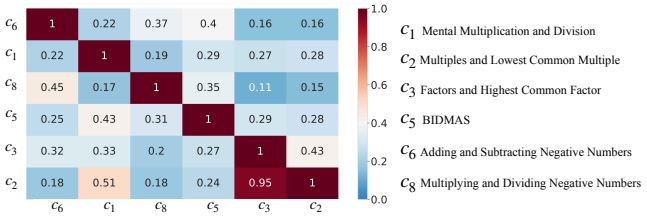

Figure 4: Attention weights visualization of sparseKT-topK. The y-axis is the pre-interaction KCs, and the x-axis is the post-interaction KCs.

5 CONCLUSION

In this paper, we propose sparseKT which enhances the classical scaled dot-product attention by extracting influential historical interactions to estimate students' mastery of knowledge. Extensive experimental results on three publicly educational datasets show the effectiveness and superior prediction outcomes and robustness of sparseKT. In the future, we would like to explore more sparse attention approaches like dynamic k selection or self-adaptive select k attention weights without the hyperparameter tuning.

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; in part by Key Laboratory of Smart Education of Guangdong Higher Education Institutes, Jinan University (2022LSY-S003) and in part by National Joint Engineering Research Center of Network Security Detection and Protection Technology.

REFERENCES

[1] Ghodai Abdelrahman and Qing Wang. 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.

[2] Sajjad Amini and Shahrokh Ghaemmaghami. 2022. Towards Robust Visual Transformer Networks via K-Sparse Attention. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 4053–4057.

[3] Jiahao Chen, Zitao Liu, Shuyan Huang, Qiongqiong Liu, and Weiqi Luo. 2023. Improving Interpretability of Deep Sequential Knowledge Tracing Models with Question-centric Cognitive Representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[4] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509* (2019).

[5] Youngduck Choi, Youngnam Lee, Junghyun Cho, Jineon Baek, Byungsoo Kim, Yeongmin Cha, Dongmin Shin, Chan Bae, and Jaewe Heo. 2020. Towards an appropriate query, key, and value computation for knowledge tracing. In *Proceedings of the Seventh ACM Conference on Learning@ Scale*. 341–344.

[6] Aritra Ghosh, Neil Heffernan, and Andrew S Lan. 2020. Context-aware attentive knowledge tracing. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2330–2339.

[7] Xiaopeng Guo, Zhijie Huang, Jie Gao, Mingyu Shang, Maojing Shu, and Jun Sun. 2021. Enhancing Knowledge Tracing via Adversarial Training. In *Proceedings of the 29th ACM International Conference on Multimedia*. 367–375.

[8] Jinseok Lee and Dit-Yan Yeung. 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.

[9] Zitao Liu, Qiongqiong Liu, Jiahao Chen, Shuyan Huang, Boyu Gao, Weiqi Luo, and Jian Weng. 2023. Enhancing Deep Knowledge Tracing with Auxiliary Tasks. In *Proceedings of the ACM Web Conference 2023*.

[10] Zitao Liu, Qiongqiong Liu, Jiahao Chen, Shuyan Huang, and Weiqi Luo. 2023. simpleKT: A Simple But Tough-to-Beat Baseline for Knowledge Tracing. In *The Eleventh International Conference on Learning Representations*.

[11] Zitao Liu, Qiongqiong Liu, Jiahao Chen, Shuyan Huang, Jiliang Tang, and Weiqi Luo. 2022. pyKT: A Python Library to Benchmark Deep Learning based Knowledge Tracing Models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

[12] Ting Long, Yunfei Liu, Jian Shen, Weinan Zhang, and Yong Yu. 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.

[13] Andre Martins and Ramon Astudillo. 2016. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *International Conference on Machine Learning*. PMLR, 1614–1623.

[14] Hiromi Nakagawa, Yusuke Iwasawa, and Yutaka Matsuo. 2019. Graph-based knowledge tracing: modeling student proficiency using graph neural network. In *2019 IEEE/WIC/ACM International Conference on Web Intelligence*. IEEE, 156–163.

[15] Shalini Pandey and George Karypis. 2019. A self-attentive model for knowledge tracing. In *12th International Conference on Educational Data Mining*. International Educational Data Mining Society, 384–389.

[16] Shalini Pandey and Jaideep Srivastava. 2020. RKT: relation-aware self-attention for knowledge tracing. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 1205–1214.

[17] Ben Peters, Vlad Niculae, and André FT Martins. 2019. Sparse Sequence-to-Sequence Models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 1504–1519.

[18] Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J Guibas, and Jascha Sohl-Dickstein. 2015. Deep knowledge tracing. *Advances in Neural Information Processing Systems* 28 (2015).

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* 30 (2017).

[20] Chenyang Wang, Weizhi Ma, Min Zhang, Chuancheng Lv, Fengyuan Wan, Huijie Lin, Taoran Tang, Yiqun Liu, and Shaoping Ma. 2021. Temporal cross-effects in knowledge tracing. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 517–525.

[21] Zichao Wang, Angus Lamb, Evgeny Saveliev, Pashmina Cameron, Yordan Zaykov, José Miguel Hernández-Lobato, Richard E Turner, Richard G Baraniuk, Craig Barton, Simon Peyton Jones, et al. 2020. Instructions and Guide for Diagnostic Questions: The NeurIPS 2020 Education Challenge. *ArXiv preprint* abs/2007.12061 (2020). https://arxiv.org/abs/2007.12061

[22] Chun-Kit Yeung and Dit-Yan Yeung. 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.

[23] Jiani Zhang, Xingjian Shi, Irwin King, and Dit-Yan Yeung. 2017. Dynamic key-value memory networks for knowledge tracing. In *Proceedings of the 26th International Conference on World Wide Web*. 765–774.

[24] Guangxiang Zhao, Junyang Lin, Zhiyuan Zhang, Xuancheng Ren, Qi Su, and Xu Sun. 2019. Explicit sparse transformer: Concentrated attention through explicit selection. *arXiv preprint arXiv:1912.11637* (2019).