Evaluating Deep Sequential Knowledge Tracing Models for Predicting Student Performance

Joel Mandlazi
School of Computer Science
and Applied Mathematics
University of the Witwatersrand
Johannesburg, South Africa
thabo.mandlazi1@students.wits.ac.za

Ashwini Jadhav
Science Teaching and Learning Unit
Faculty of Science
University of the Witwatersrand
Johannesburg, South Africa
Ashwini.Jadhav@wits.ac.za

Ritesh Ajoodha
School of Computer Science
and Applied Mathematics
University of the Witwatersrand
Johannesburg, South Africa
Ritesh.Ajoodha@wits.ac.za

***Abstract*—One of the key priorities in all instructional environments is to ensure that students recognise their learning mechanisms and pathways. Knowledge Tracing (KT), the task of modelling student knowledge from their learning history, is an important problem in the field of Artificial Intelligence in Education (AIEd) and has numerous applications in the development of interactive and adaptive learning technologies. KT can be utilised to understand each student's distinct learning style, particular needs, and ability levels.**

We trained and evaluated the performance of Knowledge Tracing models on the ASSISTments dataset and EdNet-KT1 dataset. This study revealed that deep learning models for knowledge tracing (Deep Knowledge Tracing (DKT), Dynamic Key-Value Memory Network (DKVMN), and Attentive Knowledge Tracing (AKT)) outperform the Markov process model (Bayesian Knowledge Tracing). We also observed that AKT and DKT go hand in hand with predicting whether or not the following question will be answered correctly or incorrectly by the student.

***Index Terms*—Knowledge Tracing, Artificial Intelligence in Education**

I. Introduction

High dropout rates have challenged educational institutions to improve their teaching techniques with the aim of keeping students motivated. The Department of Higher Education and Training (DHET) acknowledges that dropout rates are high and that throughput rates are low [1]. One of the causes of these high dropout rate is students' poor performance (mainly caused by students not understanding certain concepts prior to exam/test) and a weak academic background.

Despite the fact that South African universities provide free tutoring and counselling to students [1], it should be noted that there is a misalignment between the university's teaching styles and the students' learning styles, which results in poor student success. This may be a significant factor in the low throughput rate. The misalignment between university teaching styles and students' learning styles could be remedied using Knowledge Tracing (KT), the task of modelling student knowledge over time based on their experience in educational applications in the past [2], [3]. KT seeks to identify a student's learning state based on prior experience and then include appropriate hints and a personalised series of practice questions based on individual strengths and weaknesses [4].

In this age of big data, we all leave individual information footprints, resulting in an abundance of data [5]. With the development of Data Science and the the availability of Intelligent Tutoring Systems (ITS), data-driven models that aim to understand the dynamic nature of student behaviour through data interactions have become popular. KT models can be used to understand each student's unique learning behaviour, individual needs, and skill-levels.

Since the early development of KT methods prior to 2010, KT has been regarded as important in the field of Artificial Intelligence in Education (AIEd). A variety of KT methods have been developed including Bayesian Knowledge Tracing (BKT), Deep Knowledge Tracing (DKT), Dynamic Key-Value Memory Networks (DKVMN), and Attentive Knowledge Tracing (AKT) amongst others ([1]–[4]). The dispute over which strategies are most effective has emerged and remains largely unresolved due to the lack of a public, large-scale benchmark dataset that reflects a wide range of student behaviours and can fully exploit the potential of existing cutting-edge data-driven models.

The comparison of KT models is now possible thanks to the introduction of the EdNet dataset [6], a large-scale hierarchical dataset of various student behaviours collected from a multi-platform self-study solution combined with an AI teaching system. The goal of this work is to train and evaluate the performance of KT models (described in III-B) in order to determine the dependencies and relationships between solved questions in a dataset and use this information to estimate the likelihood that a student will properly answer an exam problem that they have not yet seen.

II. Literature Review

Researchers went on to resolve the raging argument over whether KT models are better at forecasting the probability of a student properly answering the next question. Table I summarises the relevant literature aiming at predicting student performance. Bayesian Knowledge Tracing (BKT) has been widely applied in educational researches and various ITS. According to [3][4] and [11], BKT assumes that student knowledge is represented as binary variable, given a skill, either the student masters the skill or does not. In BKT observations are also binary, either the student answers the

TABLE I: Review of key papers on Knowledge Tracing

	Datasets							Mod-els			
Author(s)	Synthetic	ASSISTments	Static	Simulated	KDD	Khan Maths	Other	BKT	DKT	DKVMN	SAKT
Piech, Spencer, Huang, *et al.* [3]		✓		✓		✓		✓	✓		
Zhang, Shi, King, *et al.* [4]	✓	✓	✓					✓	✓	✓	
Pandey and Karypis [7]	✓	✓	✓						✓	✓	✓
Khajah, Lindsey, and Mozer [8]	✓	✓	✓				✓	✓	✓		
Minn, Yu, Desmarais, *et al.* [9]		✓					✓	✓	✓		
Xiong, Zhao, Van Inwegen, *et al.* [10]	✓				✓			✓	✓		

exercise correctly or incorrectly.

Lee and Brunskill [12] examined the possibility of individualising all four BKT parameters. The student-specific parameters were fitted differently than in [13]. They just fit per-student BKT parameters for each student, rather than fitting per-skill and per-student BKT parameters that would later be combined. Furthermore, the goodness of fit by Lee and Brunskill [12] individualised models was not discussed in their paper. Their focus was to investigate if the individualised model would arrange less or more practice opportunities than the standard BKT skill-specific model when used in an ITS.

The original BKT research [13] suggested that operationalising the discussed individualised BKT model could be difficult. Lee and Brunskill [12] reached the reasonable conclusion that employing customised model parameters could save time for stronger students while allocating more time to difficult pupils. Their research, however, was founded on the notion that customised BKT models predict student data better, which was not tested.

Yudelson, Koedinger, and Gordon [11], implemented individualised BKT models capable of enhancing and predicting student progress on ITS. In contrast to the usual BKT [13], their approach did not necessitate changing the underlying HMM. Their intriguing discovery was that adding student-specific probability of learning ($p - learn$) benefits model accuracy more than adding student-specific probability of initial-mastery ($p - init$)..

Xiong, Zhao, Van Inwegen, *et al.* [10] compared two well studied KT models (Performance Factors Analysis (PFA) Model and BKT) with the emerging DKT on 5 different datasets. Which demonstrated that DKT does not perform overwhelmingly well on ASSISTment datasets, even when well prepared. However, the overall performance of DKT is certainly better than PFA and BKT. Furthermore, Piech, Spencer, Huang, *et al.* [3], also compared the state-of-the-art BKT to DKT using 3 different datasets and came to the conclusion that indeed DKT does outperform BKT. It should be noted that Piech, Spencer, Huang, *et al.* [3] did not compare DKT to PFA although they used the ASSISTments dataset.

Similar to previous work AKT uses raw embedding of raw questions. Ghosh, Heffernan, and Lan [2], compared AKT with several baseline KT models, including BTK + [11], DKT [3], DKT+(which is an improved DKT with regularisation on prediction consistency [14]), DKVMN [4] and SAKT [7], and found that AKT perform better in most datasets (ASSISTMents2009, ASSISTMents2015 and ASSISTments2017). However, when it comes to the and Static2011 dataset DKT+ marginally outperform AKT, this could be as a result that the Statics20211 is a small datasets.

Contrary to Zhang, Shi, King, *et al.* [4], Ghosh, Heffernan, and Lan [2] found that DKT outperforms DKVMN and that AKT out performs its variant SAKT. This discovery implies that attention mechanisms are more adaptive than recurrent neural networks, making them better suited to capture the rich information seen in large-scale real-world learner response datasets.

Pandey and Karypis [7] evaluated state-of-the-art deep knowledge tracing models against SAKT on different datasets (Sythetic-5, ASSISTMents2009, ASSISTMents2015, ASSISTChallenge and Static2011). In their findings, SAKT outperformed all the KT models in all the datasets except on the ASSISTChallenge dataset, DKT performed at par with SAKT. This could be due to the fact that ASSITChallenge dataset is the most dense dataset of all real-world datasets.

III. Methodology

To evaluate which KT model(s) best predict the probability of a student answering the next question correctly. We implement BKT, DKT, DKVMN and AKT which are discussed in III-B. The KT models are implement using the benchmark ASSISTments dataset and the recently released EdNet-KT1 dataset, discussed in III-A.

A. Data

The datasets are made up entirely of binary problems (where the label 1 means correct and 0 means incorrect). The knowledge component is the most essential aspect of the model among all the variables available in the datasets. When we train, we use it to find the relationship and dependencies between the problems. We processed all the datasets by removing all the missing values in our attributes. Furthermore, we only used three parameters to train our KT models: the user identity (user_id), the Knowledge component ID (skill_id), and the binary variable (answer_id) that shows whether the problem was answered properly or wrongly.

a) ASSISTments: dataset collected during the 2009–2010 school year. This dataset comes from skill builder (mastery learning) problem sets, in which a learner is regarded to have mastered a skill if they achieve a particular condition (usually three right answers in a row), and no further questions are given after that. Due to memory constraints, we chose 200 000 binary problems at random, which are classed as the knowledge state; an overview of the datasets we employed is presented in table II.

TABLE II: Overview of ASSISTments Datasets

		TOTAL
ASSISTments	Answered Problems	200000
	Students	4000
	Exercise Tags	69

b) EdNet-KT1: dataset, has been collected since April 18, 2017 using this question-response style. The fact that the questions arrive in bundles is one of EdNet's most notable features. EdNet is composed of a total of 131,441,538 interactions collected from 784,309 students of Santa since 2017 [6]. However, for the sake of his study, we used 200 000, which was chosen at random due to memory constraints. The dataset is broken down in the table III.

TABLE III: Overview of EdNet Datasets

		TOTAL
EdNet-KT1	Answered Problems	200000
	Students	4000
	Exercise Tags	88

B. Models

The KT models described below were trained in batches of 50 on Noam Optimizer using 10 epochs, with the goal of determining which KT models best predict whether or not the student will answer the following question correctly, given the previous engagement with exercises. The following KT models are described below: Hidden Markov BKT, Memory Augmented Networks DKVMN, LSTM-RNN DKT and Self-Attentive AKT.

a) Bayesian Knowledge Tracing: It is a popular approach to model student knowledge as a latent variable. The latent variable is modified based on how well observed student opportunities to apply the skill are correct. This modelling technique is known as a Hidden Markov Model (HMM) a special case of BKT. HMM is used to update the probabilities as a learner answers tags of exercises correctly or incorrectly. BKT model are created on the assumption that once a skill is mastered, it is never forgotten [3].

BKT uses four types of model parameters. $p(L_0)(also\ known\ as\ p-init)$, initial probability that the students knows the skill a prior. $p(T)(also\ known\ as\ p-transit)$, probability that a student's knowledge of a talent will progress from not understanding the skill to understanding the skill. $p(S)(also\ known\ as\ p-slip)$, probability that a student may make a mistake when using a mastered skill.. $p(G)(also\ known\ as\ p-guess)$, probability that a students applies a skill not known correctly.

Equations that follows are used to update the student knowledge skills, given that parameters are set of all skills:

$$p(L_1)_u^k = p(L_0)^k\ , \tag{1}$$

$$p(L_{t+1}|observation = 1)_u^k = \frac{}{p(L_t)_u^k\ (1-p(s)^k)p(L_t)_u^k\ (1-p(s)^k)+(1-p(L_t)_u^k)\ p(G)^k}\ , \tag{2}$$

$$p(L_{t+1}|observation = 0)_u^k = \frac{}{p(L_t)_u^k\ (1-p(s)^k)p(L_t)_u^k\ (1-p(s)^k)+(1-p(L_t)_u^k)\ p(G)^k}\ , \tag{3}$$

$$p(L_{t+1})_u^k = p(L_{t+1}|observation)_u^k + (1 - p(L_{t+1}|observation)k_u\ p(T)^k)\ , \tag{4}$$

$$p(C_{t+1})_u^k = p(L_t)_u^k\ (1 - p(S)k_u + (1 - p(L_t)k_u\ p(G)^k)\ , \tag{5}$$

Equation 1, is the p-init parameter for that skill is used to set the initial probability of $student_u$ mastering $skill_k$. The conditional probability is calculated using equation 2 or equation 3, depending on whether the $student_u$ applied $skill_k$ correctly or incorrectly. Equation 4, is a conditional probability used to update the probability of students' mastery level given a skill. Equation 5, is used to calculate the probability that $student_u$ will apply $skill_k$ correctly given a new exercise.

b) Deep Knowledge Tracing: (DKT) is the recent adoption of recurrent neural nets (RNNs) in the field of AIEd. DKT achieved a drastic improvement over the state-of-the-art BKT [13] and the results of it have demonstrated to be able to discover the latent structure in skill concept and can be used for curriculum optimisation [3].

When it comes to neural networks, the term 'deep' usually refers to the use of different processing layers; in DKT, the term 'deep' refers to the network's recurrent structure and the 'depth' of information over time. Using large vectors of artificial neurons, this neural net family reflects latent knowledge state as well as its temporal dynamics, and it allows the latent variable to be expressed.

The well-known issue of vanishing and ballooning gradients plagues traditional RNNs. The usual activation functions and cumulative back-propagation error signals either diminish rapidly or grow out of bounds while developing a deep neural net. Specifically, they either decay or develop exponentially ('vanish' or 'explode').

The long short-term memory (LSTM) model [15] is proposed to tackle the problem of vanishing gradients and achieves exceptional results on a range of previously unlearnable tasks. LSTM is a recurrent neural network variant that includes LSTM units in addition to standard RNN units. To determine when and which old information to forget and which recent information is important to remember, LSTM units use two distinct gates: $forget$ and $input$ gates.

c) Dynamic Key-Value Memory Networks: (DKVMN) takes advantage of the relationship between concepts as well as the ability to trace each concept state. The DKVMN model associates each exercise with the underlying concepts and keeps track of each concept's state. At each timestamp, the attempted exercise's knowledge of the associated concept states is updated [4].

DKVMN stores idea representations in a static matrix called key, while student knowledge of each concept is saved and updated in a dynamic matrix called value. Because learning is not a static process, the network with two static memory matrices is insufficient for KT.

M^t stands for memory, and it is a $N \times d$ matrix, where N represents the number of memory sites and d represents the embedding size. The input is xt at each timestamp t. The write weight w_t^w and read weight w_t^r are calculated using the embedding vector x_t. The model assumes that when a student responds the same way to an exercise that has been saved in memory, the model is correct, and x_t is written to the previously used memory locations, and when a new exercise arrives or the student responds differently, x_t is written to the least recently used memory locations.

When a student attempts a DKVMN exercise, the weighted sum of all memory slots in the value matrix is calculated by taking the softmax activation of the inner product between x_t and each key slot $M^k(i)$, and the student mastery over related concepts is retrieved as a weighted sum of all memory slots in the value matrix:

$$r_t = \sum_{i=1}^{N} w_t M_t^v(i) \ , \tag{6}$$

$$w(i) = Softmax(x_t^T M^k(i)) \ , \tag{7}$$

The calculated read content r_t is treated as a summary of the students' mastery level of an exercise. Finally to predict the student performance:

$$f_t = Tanh(W_1^T [r_t, x_t] + b_1), \ p_t = Sigmoid(W^T 2 f_t + b_1) \ , \tag{8}$$

p_t is the probability that exercise e_t is answered correctly.

The value matrix is updated by the model according to the correctness of the student's answer, then computes an erase vector e_t and an add vector a_t as:

$$e_t = Sigmoid(E^T + b_e), \ a_t = Tanh(D^T v_t + b_a) \ , \tag{9}$$

where the transformation matrices $E, D \in \mathbb{R}^{d_v \times d_v}$.

The memory vectors from the value of the component $M_t^v(i)$ from the previous timestamp are modified as follows:

$$M_t^v(i) = M_t^{\bar{v}}(i) + w_t(i) a_t \ , \tag{10}$$

$$M_t^{\bar{v}}(i) = M_t^v(i)[1 + w_t(i) e_t] \ , \tag{11}$$

d) Attentive Knowledge Tracing: (AKT), is a versatile attention-based neural network model that blends a number of unique, interpretable model components inspired by cognitive and psychometric models with an attention-based neural network model [2]. AKT is a variant of Self-Attention based Knowledge Tracing (SAKT), which models a student's interaction history (without the use of RNNs) and predicts their performance on the next exercise by taking into account relevant exercises from their previous interactions [7].

The AKT technique consists of two self-attentive encoders, one for queries and one for knowledge acquisition, a single attention-based knowledge retriever, and a feed-forward response prediction model. AKT learns context-aware representations of the questions and responses using the two self-attentive encoders. The $firstencoder$ is the question encoder, which generates changed, contextualized representations of each question based on the sequence of questions the learner has previously practiced on. Similarly, the second encoder is known as the $knowledgeencoder$ because it generates changed, contextualized representations of the knowledge learned by the learner while answering prior questions.

C. Evaluation

When predicting whether or not a student will answer the next question correctly, the Area Under Curve (AUC) and Accuracy (ACC) metrics were used to evaluate the KT models (described in section III-B). This was due to the fact that predicting student performance was regarded as a binary classification problem. The evaluation metrics used in this paper were also used in [3], [7].

IV. Results and Discussion

Table IV, showing the overview of the datasets (the ASSISTments with 200 000 answered problems, 69 exercise tags and 4000 students and EdNet-KT1 with 200 000 answered problems, 69 exercise tags and 4000 students), were used to implement and evaluate performance of KT models. Each dataset was divided into 60% training, 20% validation and 20% testing. Figure 1, shows validation AUC in each epoch of training. Figure 1a, shows validation AUC when training using ASSISTments dataset with an average validation loss 0.47 for all models and figure1b shows validation when training using EdNet-KT1 dataset with an average validation loss 0.46 for deep learning models for KT models. The overall comparison of KT Models when predicting whether or not the student will answer the next question correctly is shown in table V. Figure 2, shows that a Markov process method BKT performed worst against the deep learning KT models in both dataset. DKVMN shows a noticeable improved performance against BKT in both datasets, however, it does not perform better than DKT and AKT. We also notice the difference in performance between DKT and AKT, as DKT outperforms AKT on the ASSISTment dataset, however the converse happens on the EdNet-KT1 dataset. The toe-to-toe between DKT and AKT could be due to the fact that AKT models student history without the use of RNNs.

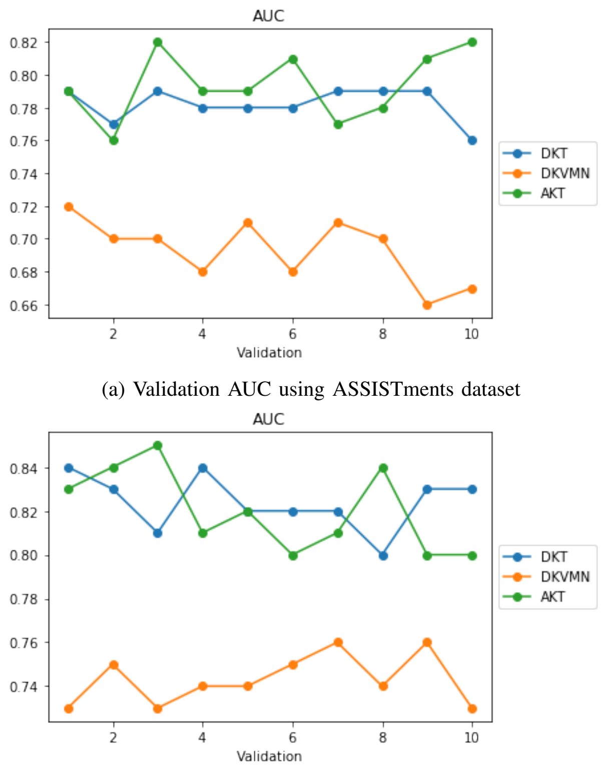
TABLE IV: Overview of Datasets

Dataset	Overview		
	Students	Exercise Tags	Answers
ASSISTments	4000	69	200K
EdNet-KT1	4000	88	200K

TABLE V: Overall Performance of models

Dataset	AUC		
	DKT	DKVMN	AKT
ASSISTments	0.74	0.68	0.72
EdNet-KT1	0.77	0.71	0.79

According to the findings, attention-based neural networks beat alternative sequence encoder techniques such as Markov,



(a) Validation AUC using ASSISTments dataset

(b) Validation AUC using EdNet-KT1 dataset

Fig. 1: Validation AUC

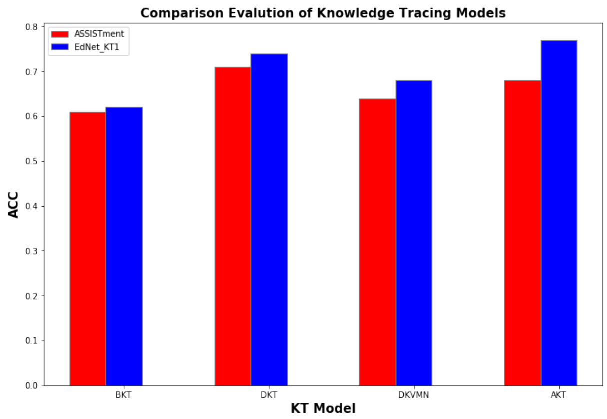


Fig. 2: Performance comparison using Accuracy of Knowledge Tracing Models.

RNN, LSTM, and Memory Augmented Networks when given enough data. Incorporating contextual data, such as the relationship between exercises and subject knowledge, as well as student forget behaviour, helps to improve performance even after the enormous dataset is available.

V. Conclusion

In this research, we looked into the performance of various Knowledge Tracing models. Using the ASSISTment and EdNet-KT1 datasets, we discovered that deep learning models for Knowledge Tracing, such as DKT, DKVMN, and AKT, outperform Markov based models like BKT. On the EdNet-K1 dataset, it was also discovered that AKT outperforms DKT. On the ASSISTment dataset, however, the opposite transpired. As a result, deep learning models for Knowledge Tracing could be used to help educational institutions provide individualised study materials to help students absorb knowledge concepts more effectively. The findings of this study may be studied further given enough processing power and memory.

References

[1] Department of Higher Education and Training, *2000 to 2016 First Time Entering Undergraduate Cohort Studies For Public Higher Education Institutions*, https://www.dhet.gov.za/HEMIS/2000 TO 2016 FIRST TIME ENTERING UNDERGRADUATE COHORT STUDIES FOR PUBLIC HEIs.pdf.

[2] A. Ghosh, N. Heffernan, and A. S. Lan, *Context-aware attentive knowledge tracing*, 2020. arXiv: 2007.12324 [cs.LG].

[3] C. Piech, J. Spencer, J. Huang, S. Ganguli, M. Sahami, L. Guibas, and J. Sohl-Dickstein, *Deep knowledge tracing*, 2015. arXiv: 1506.05908 [cs.AI].

[4] J. Zhang, X. Shi, I. King, and D.-Y. Yeung, "Dynamic key-value memory networks for knowledge tracing," in *Proceedings of the 26th International Conference on World Wide Web*, International World Wide Web Conferences Steering Committee, 2017, 765–774. DOI: 10.1145/3038912.3052580.

[5] F. Pedro, M. Subosa, A. Rivas, and P. Valverde, "Artificial intelligence in education: Challenges and opportunities for sustainable development," 2019.

[6] Y. Choi, Y. Lee, D. Shin, J. Cho, S. Park, S. Lee, J. Baek, C. Bae, B. Kim, and J. Heo, *Ednet: A large-scale hierarchical dataset in education*, 2020. arXiv: 1912.03072 [cs.CY].

[7] S. Pandey and G. Karypis, *A self-attentive model for knowledge tracing*, 2019. arXiv: 1907.06837 [cs.LG].

[8] M. Khajah, R. V. Lindsey, and M. C. Mozer, "How deep is knowledge tracing?" *CoRR*, vol. abs/1604.02416, 2016. arXiv: 1604.02416. [Online]. Available: http://arxiv.org/abs/1604.02416.

[9] S. Minn, Y. Yu, M. C. Desmarais, F. Zhu, and J.-J. Vie, "Deep knowledge tracing and dynamic student classification for knowledge tracing," in *2018 IEEE International conference on data mining (ICDM)*, IEEE, 2018, pp. 1182–1187.

[10] X. Xiong, S. Zhao, E. G. Van Inwegen, and J. E. Beck, "Going deeper with deep knowledge tracing.," *International Educational Data Mining Society*, 2016.

[11] M. V. Yudelson, K. R. Koedinger, and G. J. Gordon, "Individualized bayesian knowledge tracing models," in *Artificial Intelligence in Education*, Springer Berlin Heidelberg, 2013, pp. 171–180, ISBN: 978-3-642-39112-5.

[12] J. I. Lee and E. Brunskill, "The impact on individualizing student models on necessary practice opportunities.," *International Educational Data Mining Society*, 2012.

[13] A. T. Corbett and J. R. Anderson, "Knowledge tracing: Modeling the acquisition of procedural knowledge," *User modeling and user-adapted interaction*, vol. 4, no. 4, pp. 253–278, 1994.

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

[15] J. Schmidhuber and S. Hochreiter, "Long short-term memory," *Neural Comput*, vol. 9, no. 8, pp. 1735–1780, 1997.