# Extended LSTMs for Knowledge Tracing: Peeking Inside the Black Box (Student Abstract)

**Deliang Wang**[1], **Yu Lu**[2], **Gaowei Chen**[1]

[1]Faculty of Education, The University of Hong Kong, Hong Kong, China
[2]Faculty of Education, Beijing Normal University, Beijing, China
wdeliang@connect.hku.hk, luyu@bnu.edu.cn, gwchen@hku.hk

## Abstract

This paper proposes extended Long Short-Term Memory (LSTM) networks for the knowledge tracing task and employs explainable AI methods to address interpretability issues. Specifically, we developed an extended LSTM-based model to automatically diagnose students' knowledge states. We then leveraged three interpreting methods—gradient sensitivity, gradient*input, and Deep SHAP—to explain the model's predictions by computing input contributions. The results demonstrate that the proposed model outperforms DKT, and the three methods effectively explain its predictions. Additionally, we identified three key insights into the model's working mechanisms.

## Introduction

Knowledge tracing (KT) involves leveraging students' sequential exercise-answer records and applying machine learning methods to model students' knowledge states and predict their performance on future exercises. This technique has been integrated into intelligent tutoring systems to facilitate personalized learning. Initially, researchers employed Markov chains and logistic regression to construct KT models (e.g., Bayesian knowledge tracing and performance factor analysis) for assessing students' knowledge levels in various skills. Despite the development of multiple variants aimed at better aligning with students' exercise data, these models still exhibited performance limitations.

Subsequently, deep neural networks, such as recurrent neural networks, attention mechanisms, and graph neural networks, have been introduced to the KT domain and have garnered significant attention from researchers due to their superior ability to capture and represent inherent information. Among these deep learning-based KT models, deep knowledge tracing (DKT) (Piech et al. 2015) is recognized as a pioneering model. DKT utilized the long short-term memory (LSTM) network to encode learners' sequential exercise-answer data, model their knowledge states, and forecast their future performance. In comparison to traditional KT models, DKT demonstrated substantially improved performance and could be employed to infer skill relationships (Piech et al. 2015). However, criticisms regarding DKT's inner LSTM structure, have emerged, including its use of a memory vector to summarize knowledge states across all skills and its inefficiency in model training.

As highlighted by Beck et al. (2024), LSTMs face challenges such as an inability to revise storage decisions due to the sigmoid gate structure, limited storage capabilities because of the scalar memory vector, and a lack of parallelizability due to memory mixing. To tackle these issues, Beck et al. (2024) proposes the utilization of exponential gating and novel memory structures, leading to the extension of the original LSTM into sLSTM, mLSTM, and xLSTM architectures. The xLSTMs have been evaluated in language modeling tasks and have demonstrated comparable performance to state-of-the-art Transformers. However, it remains uncertain whether these extended LSTMs can effectively address the KT task. Furthermore, the inherent black box nature of deep neural networks raises concerns related to interpretability and trust, a challenge that also applies to extended LSTMs for knowledge tracing.

Therefore, this study advocates for the use of extended LSTMs to develop a novel KT model (i.e., sLSTM-KT) for evaluating students' knowledge states. Additionally, the study employs explainable AI techniques to elucidate the model's functioning and provide users with insights into the process of knowledge state diagnosis, aiming to enhance user trust and acceptance in practical applications.

## Method

**Model Structure.** Given a sequence of a student's exercise-answer records, denoted as $\{(e_1, a_1), ..., (e_n, a_n)\}$, where $e_i$ and $a_i$ represent the exercise and the correctness of the answer at time step $i$, sLSTM-KT aims to predict the learner's probability of answering question $e_{n+1}$ correctly. The underlying principle of the sLSTM-KT model is illustrated in the following equations. Different from traditional LSTM employing the sigmoid function, the sLSTM unit leverages an exponential function to regulate the input gate, as depicted in Equation (2). This modification renders the gate more sensitive to new input and better revises stored memory. Given that the exponential function may generate large values causing overflows, the extended LSTM stabilizes the input and forget gates by introducing an additional state $m_t$, as shown in Equations (3), (4), and (5). Furthermore, a normalizer state $n_t$ is incorporated to normalize the exponentiated cell state $C_t$, as illustrated in Equations (8) and (9).

More comprehensive technical details can be seen in Beck et al. (2024). Due to page constraints, we focus on the construction and explanation of the sLSTM for knowledge tracing, rather than providing an exhaustive account of all extended LSTMs (e.g., mLSTM and xLSTM).

$$\mathrm{f_t} = \sigma\left(\mathbf{W}_{\mathrm{fh}}\mathrm{h_{t-1}} + \mathbf{W}_{\mathrm{fx}}\mathbf{x_t} + \mathrm{b_f}\right) \quad (1)$$

$$\mathrm{i_t} = \exp\left(\mathbf{W}_{\mathrm{ih}}\mathrm{h_{t-1}} + \mathbf{W}_{\mathrm{ix}}\mathbf{x_t} + \mathrm{b_i}\right) \quad (2)$$

$$m_t = \max\left(\log\left(\mathrm{f}_t\right) + m_{t-1}, \log\left(\mathrm{i}_t\right)\right) \quad (3)$$

$$\mathrm{i}'_t = \exp\left(\log\left(\mathrm{i}_t\right) - m_t\right) \quad (4)$$

$$\mathrm{f}'_t = \exp\left(\log\left(\mathrm{f}_t\right) + m_{t-1} - m_t\right) \quad (5)$$

$$\widetilde{\mathrm{C_t}} = \tanh\left(\mathbf{W}_{\mathrm{ch}}\mathrm{h_{t-1}} + \mathbf{W}_{\mathrm{cx}}\mathbf{x_t} + \mathrm{b_c}\right) \quad (6)$$

$$\mathrm{C_t} = \mathrm{f'}_t \odot \mathrm{C_{t-1}} + \mathrm{i'}_t \odot \widetilde{\mathrm{C_t}} \quad (7)$$

$$n_t = \mathrm{f}'_t \odot n_{t-1} + \mathrm{i}'_t \quad (8)$$

$$\widetilde{\mathrm{h_t}} = \mathrm{C_t}/n_t \quad (9)$$

$$\mathrm{o_t} = \sigma\left(\mathbf{W}_{\mathrm{oh}}\mathrm{h_{t-1}} + \mathbf{W}_{\mathrm{ox}}\mathbf{x_t} + \mathrm{b_o}\right) \quad (10)$$

$$\mathrm{h_t} = \mathrm{o_t} \odot \widetilde{\mathrm{h_t}} \quad (11)$$

$$\mathbf{y}_t = \sigma\left(\mathbf{W}_{\mathrm{yh}}\mathrm{h_t} + \mathrm{b_y}\right) \quad (12)$$

**Model Training.** We employ the ASSISTment2009 dataset to construct the sLSTM-KT model. The ASSISTment2009 dataset comprises 325,637 exercise records from 4,151 students, spanning 110 skills. In accordance with previous research guidelines, this study conducts a 5-fold cross-validation, implements early-stopping, and configures the optimizer, batch size, and hidden state dimension as Adam, 100, and 200, respectively. DKT is also trained using these metrics as a baseline.

**Model Explanation.** Three generic explainable AI methods — Gradient Sensitivity (GS) (Li et al. 2015), Gradient*Input (GI) (Kindermans et al. 2019), and Deep SHAP (Lu et al. 2024; Wang et al. 2022) — are employed to elucidate the knowledge state diagnosis from sLSTM-KT by computing the contributions of each exercise-answer record. The higher the contribution of an exercise-answer record, the more significant it is to the prediction.

## Preliminary Experiments and Results

Following a 5-fold cross-validation on the ASSISTment2009 dataset, the sLSTM-KT model achieved an Area Under the Curve (AUC) score of 0.7523 and an accuracy of 0.7202. In comparison, the DKT model attained an AUC score of 0.7481 and an accuracy of 0.7192. These preliminary results demonstrate that sLSTM-KT outperforms DKT slightly in both AUC and accuracy metrics, partially suggesting that the extended LSTM architecture is effective for KT tasks.

To explore the working mechanism of sLSTM-KT, we first employ three interpreting methods to compute the contribution of each input exercise-answer pair to the prediction. Specifically, we divide the test data into sequences of length 15, predict learners' performance on the last exercise of each sequence, and calculate the contribution of preceding exercise-answer pairs. To validate whether these contributions reflect the pairs' importance to the prediction, we select correctly-predicted sequences and sequentially delete exercise-answer pairs based on their contribution value, one by one, until eight pairs are removed. We also perform a random deletion for comparison. As shown in Figure 1, compared to random deletion, removal based on contribution value significantly reduces accuracy from 1, validating the effectiveness of the explanations.

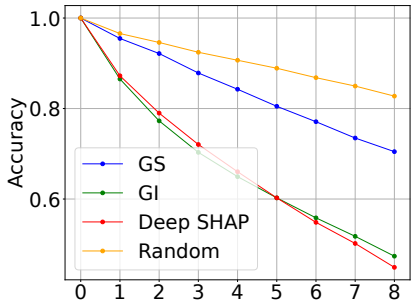


Figure 1: Accuracy change when deleting exercise records from initially correctly predicted sequences.

Using the explanations from Deep SHAP, we examine how sLSTM-KT evaluates learners' knowledge states. Given the input sequence of exercise-answer pairs, we assign each pair a three-tuple value: (correctness, skill, position). Each element in the tuple is binary: correct or wrong, target skill or non-target skill, and close or far. Correct/wrong indicates the answer's accuracy, target/non-target denotes alignment with the skill of the last exercise, and close/far represents whether the pair is among the last 7 pairs or the previous 7. This categorizes exercise pairs into 8 groups. We calculate the mean absolute value for these groups, as shown in Figure 2, yielding the following insights: 1. When skill and position are the same, incorrect pairs likely contribute more than correct ones; 2. When correctness and position are the same, pairs on target skills contribute more than those on non-target skills; 3. When correctness and skill are the same, close pairs contribute more than distant ones.

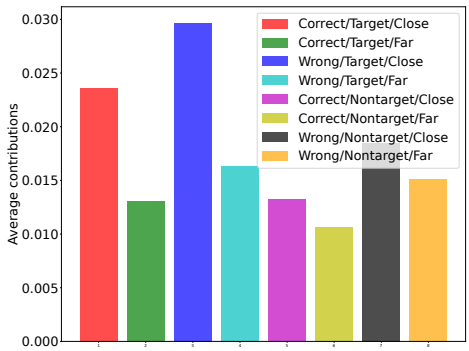


Figure 2: Average contributions among eight groups.

## Acknowledgments

This work was supported by the Natural Science Foundation of China (No. 62277045), and the Hong Kong Research Grants Council, University Grants Committee (Grant No.: 17605221).

## References

Beck, M.; Pöppel, K.; Spanring, M.; Auer, A.; Prudnikova, O.; Kopp, M.; Klambauer, G.; Brandstetter, J.; and Hochreiter, S. 2024. xLSTM: Extended Long Short-Term Memory. *arXiv preprint arXiv:2405.04517*.

Kindermans, P.-J.; Hooker, S.; Adebayo, J.; Alber, M.; Schütt, K. T.; Dähne, S.; Erhan, D.; and Kim, B. 2019. The (un) reliability of saliency methods. *Explainable AI: Interpreting, explaining and visualizing deep learning*, 267–280.

Li, J.; Chen, X.; Hovy, E.; and Jurafsky, D. 2015. Visualizing and understanding neural models in NLP. *arXiv preprint arXiv:1506.01066*.

Lu, Y.; Wang, D.; Chen, P.; and Zhang, Z. 2024. Design and Evaluation of Trustworthy Knowledge Tracing Model for Intelligent Tutoring System. *IEEE Transactions on Learning Technologies*.

Piech, C.; Bassen, J.; Huang, J.; Ganguli, S.; Sahami, M.; Guibas, L.; and Sohl-Dickstein, J. 2015. Deep knowledge tracing. In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1*, 505–513.

Wang, D.; Lu, Y.; Zhang, Z.; and Chen, P. 2022. A generic interpreting method for knowledge tracing models. In *International conference on artificial intelligence in education*, 573–580. Springer.