Predictive Student Modelling in an Online Reading Platform

Effat Farhana,[1] **Teomara Rutherford,** [2] **Collin F. Lynch** [3]

[1] Vanderbilt University
[2] University of Delaware
[3] North Carolina State University
effat.farhana@vanderbilt.edu, teomara@udel.edu, cflynch@ncsu.edu

Abstract

Use of technology-enhanced education and online learning systems has become more popular, especially since the onset of the COVID-19 pandemic. These systems capture a rich array of data as students interact with them. Predicting student performance is an essential part of technology-enhanced education systems to enable the generation of hints and provide recommendations to students. Typically, this is done through use of data on student interactions with questions without utilizing important data on the temporal ordering of students' other interaction behavior, (e.g., reading, video watching). In this paper, we hypothesize that to predict students' question performance, it is necessary to (i) consider other learning activities beyond question-answering and (ii) understand how these activities are related to question-solving behavior. We collected middle school physical science students' data within a K12 reading platform, Actively Learn. This platform provides reading-support to students and collects trace data on their use of the system. We propose a transformer-based model to predict students' question scores utilizing question interaction and reading-related behaviors. Our findings show that integrating question attempts and reading-related behaviors results in better predictive power compared to using only question attempt features. The interpretable visualization of transformer's attention can be helpful for teachers to make tailored interventions in students' learning.

Introduction

Predicting students' performance on question-answering tasks is a fundamental challenge in education. Accurate predictions may be used to drive tailored instruction, automated feedback, and adaptive testing among other things. Prior research on modeling question-answering has primarily focused using students' prior question attempts and skill-specific models such as Learning Factors Analysis (LFA) (Chi et al. 2011), knowledge tracing (KT) (Corbett and Anderson 1994), Q-matrices (Barnes 2005), and Item Response Theory (IRT) (Hambleton, Swaminathan, and Rogers 1991). These approaches rely on mappings from questions to skills and a trace of the students past question interactions to assess student abilities and estimate the question difficulties.

Although these approaches have been successful, they only utilize a small subset of the available information. Modern learning environments including intelligent tutoring systems (ITS) record a rich array of student-system interaction data including reading, video viewing, and website visits, as well as more complex self-regulation and problem-solving actions such as information lookup and goal-setting (Zimmerman 2000). Recent work has shown that these data can be utilized to support effective student modeling. Zhang et al. (Zhang et al. 2017b) for example, showed that incorporating additional features in a deep knowledge tracing (DKT) model, such as time of first attempt, number of attempts, and whether the first action was a hint or question attempt improved performance. Similarly, Mongkhonvanit and colleagues tracked students' video interactions along with question solving behaviors to predict future question performance in a massive open online course (MOOCs) (Mongkhonvanit, Kanopka, and Lang 2019). They did not, however, consider how these interactions were related to the students' question attempts. Although all these approaches use additional features concatenated to the question features, none of these studies have considered each action as a temporal event. In one recent study, Choi et al. emphasized the above-mentioned lack of studies combining both question solving and interaction behavior. To support incorporation of fine-grained details of learning, the authors released large-scale student-interaction data, Ednet (Choi et al. 2020) containing students' question attempts and other interactions, such as video watching, choosing a response, or reading a passage as temporal events.

Our study stems from the same motivation as that of Choi et al. (Choi et al. 2020). We emphasize that learning is a dynamic process and performance on a question not only depends on previous question attempts but also other learning activities. Consider Figure 1 in this regard. Previous KT approaches takes account previous question attempts, Q_1 to Q_4 to predict a student's question performance Q_5 (**Case 1**). However, a student may perform other learning activities within the system, such as video watching, reading, and highlighting. **Case 2** illustrates just such a scenario where a student performed one highlighting and one annotation prior to attempting Q_5. We hypothesize that to fully understand students' performance on a question, we need to consider (i) that their question-solving and other interaction behaviors may influence learning and (ii) how these behaviors are related to the question attempt. In this study, we fo-

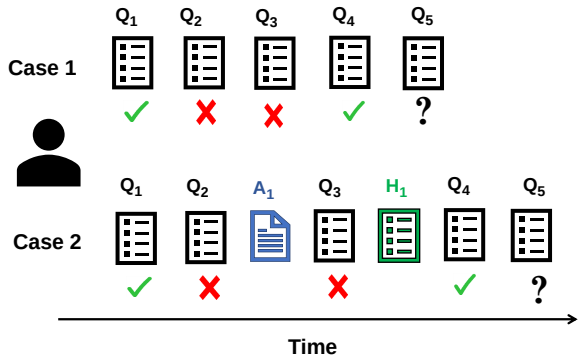

Figure 1: Motivating Example. Q: Question Attempt, A: Annotation, H: Highlighting. Previous KT approaches consider **Case 1** scenarios. Proposed approach can handle both **Case 1** and **Case 2** scenarios.

cus on predicting students' performance by taking account of the two criteria above. Our analysis is focused on Actively Learn (AL), a popular online reading platform used in US Schools. We investigated how students' question-solving behaviors and interactive reading-associated features contribute to their learning. We identified three reading-related features within the AL system that are likely to relate to question-solving, namely annotating (Makany, Kemp, and Dror 2009), highlighting (Winne et al. 2017), and vocabulary lookup (Biemiller and Slonim 2001).

Reading to learn is a complex cognitive process. In order to understand a written text, a learner first needs to decode the text into letters, words, and sentences (Perfetti 1985; Shankweiler et al. 1999). Then, the reader must comprehend the meaning of these units and connect the meaning with their prior knowledge (Kintsch and Walter Kintsch 1998; Kintsch 1988; Magliano et al. 2005). Reading to learn is an integral skill in domain-specific learning, such as science (Norris and Phillips 2003). Students in the United States may not be adequately prepared to engage in this skill. According to the bi-yearly National Assessment of Educational Progress (NAEP) of 2019, only 37% 8th-graders were proficient in academic reading. Moreover, this percentage was below that of 2017 [1]. This may be due to the fact that students do not get enough opportunities to learn and practice reading at school (Gomez, Herman, and Gomez 2007).

Digital reading platforms can help to to address this problem. These platforms allow students to engage with reading materials while teachers monitor their activities and provide adaptive support. Our research was conducted in the context of one such platform, AL. This paper makes the following contributions:

- Contrary to previous approaches that rely only on question-solving actions, we consider temporal ordering of question-solving behavior along with other student interaction behaviors.
- Our features are theoretically grounded in the learning sciences literature. Our findings may be helpful for the creation of reading interventions.
- Our transformer-based model provides explanatory ability by attention–thus it can be helpful to teachers and students to understand model's prediction.

[1] https://nces.ed.gov/nationsreportcard/reading/

Related Work

Knowledge Tracing. (KT) In order to model students' understanding of a domain we must track or trace their knowledge. Existing KT models fall into two broad categories: cognitive diagnostic and deep learning-based models. The first category centers around the Bayesian knowledge tracing (BKT) (Corbett and Anderson 1994). It uses students' historical exercise records to estimate the probability that a student has mastered a specific skill. The Rasch model evaluates students' latent ability level needed to obtain a 50% chance of accurately answering a question. The descendent of Rasch models are learning factor analysis (LFA) (Cen, Koedinger, and Junker 2006) and performance factor analysis (PFA) (Pavlik, Cen, and Koedinger 2009). All these methods use student mastery or proficiency metrics to predict students' performance.

Piech et al. proposed deep knowledge tracing (DKT) using a recurrent neural network (Piech et al. 2015). The DKT model achieved 25% gain in area under curve (AUC) on two real world datasets compared to statistical prediction models. Since then, researchers proposed variations of DKT methods. One limitation to the original DKT was it did not consider skills associated with each question. Zhang and colleagues (Zhang et al. 2017a) proposed a model utilizing Dynamic Key-Value Memory Networks (DKVMN) to capture skill-set associated with questions. The original BKT and DKT does not capture students' forgetting behavior. Forgetting curve theory states that the students' memory decay exponentially with time (Ebbinghaus 2013). Recent models, such as DKT-Forget, integrates forgetting behavior considering elapsed time between current event and previous event, and elapsed time between events with similar skill tagging (Nagatani et al. 2019). Shin et al. (Shin et al. 2021), incorporated two types of temporal features: total amount of time spent on a question and the time gap between consecutive question attempts. Wang et al. (Wang et al. 2021) studied temporal cross-skill effect in KT methods. They argued that the mastery of a skill is not only dependent on similar types of skills but also is influenced by other skills.

Pandey et al. proposed an *attention-based knowledge tracing* model, SAKT (Pandey and Karypis 2019). In attention-based models, a positional embedding is used to track sequential behavior instead of a recurrent neural network. An enhancement of the SAKT model is RKT (Pandey and Srivastava 2020)–which takes into account time decaying forgetting behavior and question relationships in prediction. Ghosh et al. (Ghosh, Heffernan, and Lan 2020) proposed a time-decaying monotonic attention, AKT, to capture the importance of previous question-solving behavior. Ghosh et al. argued that when students face a question, past question-solving behavior from unrelated concepts or those are not recent temporally may not be relevant.

The AL system does not incorporate skill-tagging or a student's total time spent on a question-solving. Thus, we integrate the time-decaying feature in our attention-based model, similar to the AKT and RKT.

Reading Behaviors. Reading-to-learn is a self-regulated learning (SRL) activity (Michalsky 2013). SRL refers to planning, monitoring, and controlling activities during learning (Zimmerman 1989, 2000). We identified three reading-related SRL activities within the AL platform considering previous literature: annotating (Makany, Kemp, and Dror 2009), highlighting(Winne et al. 2017), and vocabulary lookups (Biemiller and Slonim 2001).

Contextual Embeddings. Representing text is a fundamental research problem in natural language processing. These methods include unsupervised approaches, such as Word2Vec (Mikolov et al. 2013) and Glove (Pennington, Socher, and Manning 2014). Supervised methods, such as Elmo (Peters et al. 2018), generate word embedding taking the sentence context. Recently, transformer-based methods, such as bidirectional encoder representations from transformers (BERT) (Devlin et al. 2019) and its variants, such as SentenceBERT (Reimers et al. 2019) have shown promising results in contextual word embedding to represent sentences. In our preliminary analysis, we found Universal Sentence Encoder (USE) (Cer et al. 2018) was better performing than SentenceBERT. Thus, we applied USE (Cer et al. 2018) to encode textual representations in our study.

Question Relationship Modelling. Su et al. (Su et al. 2018) proposed an exercise-enhanced deep knowledge tracing, EERNN. According to the authors, exercise texts may semantically represent underlying knowledge concepts. An extended version of the study was done by the same authors proposing EKT (Liu et al. 2019). As we do not have questions-to-skillset mapping, our approach is similar to the RKT and EERNN methods–which utilize cosine similarities between current and previous exercises to discover exercise relationships. In addition, we also incorporate two other types of relationships: (i) cosine similarities between the current exercise and previous SRL action texts and (ii) cosine relationship between the current question's response text and previous SRL action texts.

Problem Formulation

The AL system catalogs K-12 curriculum-integrated reading articles. Teachers can create assignments using provided articles or by using their own reading materials. AL assignments contain text-embedded questions that can be multiple choice and short answer questions. Questions are graded on a scale of zero to four. The system provides reading support for students, allowing actions such as highlighting, annotating, and vocabulary lookup.

We aggregated students' actions into a unified transaction log. Table 1 shows a hypothetical example of our log. Columns 'AID' and 'Time' refer to unique assignment ID and sorted timestamps of students' activities within the system, respectively. In the above example, S1 highlighted once (at T1), took a note (at T2), looked up vocabulary once (at T4), and attempted three questions (at T3, T5, and T6, respectively). The 'Action Text' column contains the textual description of corresponding actions. For example, the action text at timestamp T1 and T5 contain textual descriptions of S1's highlighted text and attempted question, respectively. We formulate our task as follows:

AID	Time	Action	Action ID	Action Text
A1	T1	Highlight	h1	Highlighted Text
A1	T2	Annotation	a1	Note Text
A1	T3	**Question**	q1	Ques.Text
A1	T4	Vocab. Lookup	v1	Vocab text
A1	T5	**Question**	q2	Ques. Text
A1	T6	**Question**	q3	Ques. Text

Table 1: A Hypothetical Student, S1's, Action Log in the AL

Definition 1 ***(PROBLEM DEFINITION).*** *Given the log of students' actions and textual contents of each action, our goal is to predict students' question performance at time t_T considering students' previous actions within an assignment with timestamps $t_1, t_2, t_3, ..., t_{T-1}$.*

Scope. In past studies with DKT and its variations, question-answering sequences were longer than ours (e.g., 65.9 questions per student (Su et al. 2018)). In contrast, the scope of our study is within a reading-comprehension assignment–which has shorter sequence lengths. Detailed statistics of our dataset are in Table 2.

Modelling

Figure 2 presents a high-level overview of our model.

Input Embedding

We encode actions to a fixed dimensional input interaction vector, **x**. The vector **x** comprises the following parts:

- *Text Embedding.* We use the USE to represent each action text into a *d = 512* dimensional vector, **E** .

$$\mathbf{E} = USE(ActionText) \in \mathbb{R}^d \tag{1}$$

- *Action Type Embedding.* We use an embedding size *d = 512* to encode four different action types including question attempts and three types of SRL behaviors.
- *Score Encoding.* We extend score of each question to a vector size *d*. Otherwise, if the action is an SRL, we create a vector *d* with dummy input -1 as score.
- *Response Encoding.* If the action is a question attempt, we compute cosine similarity value between each question and corresponding student's response text. We extend the value to a vector size *d*. Otherwise, if the action is an SRL, we create a vector *d* with input 0.

We concatenate these four feature vectors to form an input interaction vector for each timestamp i, $\mathbf{x_i} \in \mathbb{R}^{4d}$.

In our model, we use the interaction vector as key and values when computing attention. Our query vector is the text embedding of actions, $\mathbf{E} \in \mathbb{R}^d$. To match the interaction vector's dimension to the query vector, we apply a feed forward network (FFN) with a ReLU activation to convert the $\mathbb{R}^{4d}$ dimensional vector into a $\mathbb{R}^d$ dimensional one.

$$\hat{\mathbf{x}}_\mathbf{i} = FFN(\mathbf{x_i}) \in \mathbb{R}^d \tag{2}$$

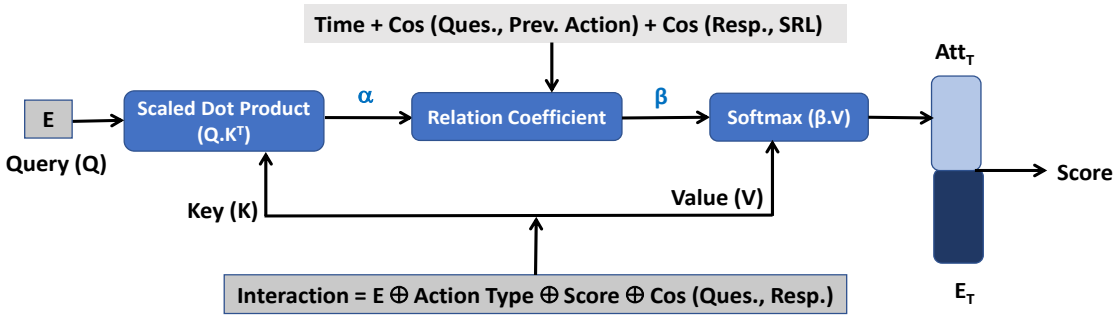

Figure 2: Proposed Model: High-level Overview of Methodology

A student's interaction sequence is represented as $\mathbf{X} = [\hat{\mathbf{x_1}}, \hat{\mathbf{x_2}}, \hat{\mathbf{x_3}}, ..., \hat{\mathbf{x}}_{\mathbf{T-1}}]$.

Note that we do not use any positional embedding (Vaswani et al. 2017) as our time-decaying relation (next section) incorporates inter-event information. Our approach is similar to (Ghosh, Heffernan, and Lan 2020; Gu 2021).

Forgetting, Question and Response Relation

We integrate the elapsed time, question relation, and response relation with the model's attention (next subsection).

- *Time Relation.* This component takes into account students' forgetting behavior. We took the exponential of the elapsed time Δ_i between the T_{th} action timestamp, t_T and i_{th} previous action's timestamp, t_i. The elapsed time sequence of a student is

$$\mathbf{R_{Time}} = [exp(\Delta_1), exp(\Delta_2), ..., exp(\Delta_{T-1})] \quad (3)$$

- *Question Relation.* This component computes the cosine similarities between the T_{th} question embedding, $\mathbf{E_T}$ and a previous action (question attempt or an SRL) text embedding at t_i, $\mathbf{E_i}$. If the T_{th} action is an SRL, the value is 0. The question relation sequence of a student is

$$\mathbf{R_{Ques}} = [g(E_T, E_1), g(E_T, E_2), ..., g(E_T, E_{T-1})] \quad (4)$$

where $g(\mathbf{E_T}, \mathbf{E_i}) = cos(\mathbf{E_T}, \mathbf{E_i})$, if the T_{th} action is a question attempt or otherwise 0.

- *Question Response Relation.* This component is only applicable for the SRL dataset (Table 2). It computes cosine similarities between question response text embedding at time T, $\mathbf{RespE_T}$ and a previous SRL action embedding at t_i, $\mathbf{E_i}$. The question response relation sequence of a student is

$$\mathbf{R_{Resp}} = [f_1, f_2, f_3, ..., f_{T-1}] \quad (5)$$

where

$$f_i = \begin{cases} cos(\mathbf{RespE_T}, \mathbf{E_i}) & \text{action } t_i \text{ is an SRL \& } t_T \text{ action is a ques. attempt} \\ 0 & \text{otherwise} \end{cases}$$

We combine equations 3, 4, and 5 and take the softmax.

$$\mathbf{R} = \text{Softmax}(\mathbf{R_{Time}} + \mathbf{R_{Ques}} + \mathbf{R_{Resp}}) \quad (6)$$

Attention

First, we describe the basic scaled-dot product attention (Vaswani et al. 2017) and then our integration of relation to the attention. Let, $\mathbf{W^Q}$, $\mathbf{W^K}$, and $\mathbf{W^V}$ denote projection matrices for query, key, and value space, respectively with dimension $\mathbb{R}^{d*d}$. Let, $\mathbf{q_i}$ be the query vector of a student's question attempted at time i. The attention, $\mathbf{Att_i}$, is

$$\alpha_{i,j} = \text{Softmax}(\frac{\mathbf{q_i W^Q}.(\mathbf{k_j W^K})^\mathbf{T}}{d}), \text{all} j < i$$
$$\mathbf{Att_i} = \sum_{j<i} \alpha_{i,j}.\mathbf{v_j W^V} \quad (7)$$

where $\mathbf{k_j}$ and $\mathbf{v_j}$ denote respectively key and value vectors for previous actions, $j < i$. In transformer-based KT models, keys and values are past events, so $j < i$ (Pandey and Karypis 2019). The $\mathbf{Att_i}$ value denotes the relevance of each past interaction with the question attempted at time i.

In our model, we combine $\alpha_{i,j}$ with the relation coefficient, $\mathbf{R}$, from Equation 6 as follows:

$$\beta_{i,j} = \lambda \alpha_{i,j} + (1 - \lambda)\mathbf{R_j} \quad (8)$$

where R_j is the j_{th} coefficient of $\mathbf{R}$ and λ is a trainable parameter in our model.

And finally, we compute the attention $\mathbf{Att_i}$ by multiplying $\beta_{i,j}$ with our value (interaction) vectors

$$\mathbf{Att_i} = \sum_{j<i} \beta_{i,j}.\hat{\mathbf{x_j}} \mathbf{W^V} \quad (9)$$

Multihead Attention. We compute multihead attention from different semantic subspaces (Vaswani et al. 2017). The process computes attention value on each head as described above. The final output is a linear transformation of concatenated attention values on each head.

Prediction

The predication layer concatenates the attention output, $\mathbf{Att_T}$ and T_{th} question embedding, $\mathbf{E_T}$. Then we pass this concatenated input through a fully connected layer to generate prediction. In this study, we formulated the prediction task as both a binary classification and a regression problem, similar to the work of Su et al. (Su et al. 2018). For the classification task, the output passes through a Sigmoid function

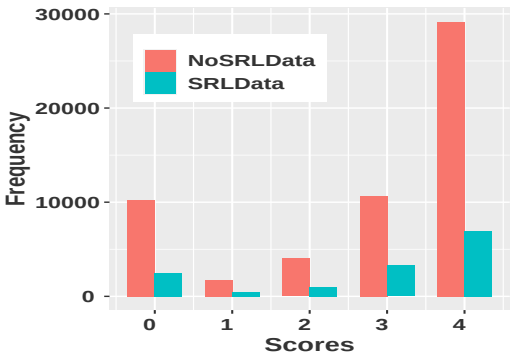

Figure 3: Histogram of Scores

to generate the probability that a student answered a question correctly. If a student's actual score at time t is r_t and predicted score is $\hat{r_t}$, then the loss for a specific student is $\mathcal{L} = -\sum_{t=1}^{T} (r_t \log \hat{r_t}$) + (1-r_t)log(1-$\hat{r_t}$). For the regression task, we minimize the mean squared error (MSE) loss between $\hat{r_t}$ and r_t.

Experiments

Dataset Preparation

We conducted our study with anonymized middle school physical science data collected from AL in 2018. We produced three different AL datasets:

- D1: Students only performed question attempts in an assignment but no SRL behaviors were captured. All entries in this dataset are **Case 1** in Figure 1.
- D2: Students performed question attempts and one or more SRL behaviors in an assignment. All entries in this dataset are **Case 2** in Figure 1.
- D2.1: This dataset is a subset of D2. We only consider question attempts from a student's action sequence ignoring SRL actions. For example, if we remove SRL actions from **Case 2** in Figure 1, the resulting action sequence will be only question-solving attempts.

Table 2 shows our dataset statistics. Number of interactions includes question attempts and SRL. We evaluate all models on D1 and D2.1 and only our proposed model on D2 dataset. We discard student sequences with less than two entries. Observing dataset statistics in Table 2, we selected previous action sequence lengths 8, 10, and 15 for D1, D2.1, and D2 datasets to predict a question's score at time T.

Baselines and Metrics

As our dataset does not contain questions to skill mapping, we selected DKT models which do not require skill information. We compared our model with two LSTM-based models, the DKT (Piech et al. 2015), EERNN (Su et al. 2018), and one attention-based model, RKT (Pandey and Srivastava 2020). For classification formulation, given the skewed nature of our dataset in Figure 3, we formulated it as a binary classification problem. We considered score = 4 as one and any other score as zeros. We use AUC and accuracy as evaluation metrics. For the regression task, we report MSE and mean average error (MAE) as evaluation metrics.

	Datasets		
	No SRL (D1)	SRL (D2)	Removed SRL (D2.1)
No. assign.	754	425	378
No. ques.	1,934	1,292	1,260
No. students	8,060	1,796	1,680
No. interactions	55,185	20,831	14,043
Seq. Mean (SD)	5.05 (3.18)	9 (4.9)	6.4 (3.6)
Seq. Median	5	9	6

Table 2: Dataset Statistics

Implementation

We applied the standard 5-fold cross-validation (CV) at the student level to evaluate all models. We used 20% of the training data as validation. For fair comparison among models, the dataset split remained the same across all models.

We used the USE (Cer et al. 2018) for textual encoding. USE can take phrases, sentences, and short paragraphs as inputs and encodes inputs into a fixed-length vector of 512. For all models, we used a batch size of 200, embedding size of 512, and 300 epochs. We applied early stopping if the validation AUC did not increase (or MAE did not decrease for regression) over five epochs. We also reimplemented DKT and EERNN and used the author's provided source code for RKT (Pandey and Srivastava 2020). For our proposed model and RKT, we used learning rate = $1e^{-3}$ and number of attention heads = 4. For DKT and EERNN, we tuned hyperparameters on the validation set for each fold: dropout = $\{0.20, 0.33, 0.66\}$ and learning rate = $\{1e^{-4}, 5e^{-4}, 5e^{-3}\}$.

For our regression task, we adapted all baseline models, as authors used those with datasets with binary values. In the EERNN and RKT, authors first extended the binary score to the equal length of the question embedding before appending it to question embedding. Similarly, we extended a question's score $[0-4]$ to a vector of length 512 and appended it with the embedding. DKT has an input vector length of $5Q$, where Q is the number of questions in the dataset and five possible scores per question.

Results

Question Score Prediction. Table 3 and 4 present classification and regression results, respectively. Our model has the highest classification accuracy and AUC value on D1 and D2.1 datasets. For the regression task, our model has the lowest MSE for D1 and D2.1 datasets. Considering the MAE value, the RKT has the lowest value on the D1 dataset. The vanilla DKT model performed poorly for both tasks, except for D2.1 dataset classification. As we do not have skill labeling, we used question identifiers to model DKT. DKT models with a large number of questions and a few observations per question decrease the model's performance, as noted by Sonker et al. (Sonkar et al. 2020).

Dataset	Method	Accuracy	AUC
D1	DKT	0.54	0.55
	EERNN	0.73	0.78
	RKT	0.71	0.73
	Proposed	**0.77**	**0.84**
D2	Proposed	0.78	0.85
D2.1	DKT	0.62	0.65
	EERNN	0.75	0.75
	RKT	0.65	0.64
	Proposed	**0.77**	**0.83**

Table 3: Classification: Mean of 5-fold CV on test dataset.

Dataset	Method	MSE	MAE
D1	DKT	9.92	2.55
	EERNN	2.13	1.16
	RKT	2.35	**1.02**
	Proposed	**2.11**	1.04
D2	Proposed	1.75	0.98
D2.1	DKT	9.92	2.78
	EERNN	1.98	1.09
	RKT	2.22	1.03
	Proposed	**1.95**	**1.01**

Table 4: Regression: Mean of 5-fold CV on test dataset.

Our model's performance on D2 and D2.1 datasets show that combining SRL features and question attempts result in better predictive power in both predictive tasks compared to only including question attempts.

Ablation Studies. We performed a series of ablation studies to understand the contribution of each component in prediction. We present our classification and regression ablation results in Tables 5 and 6, respectively.

Ques. Relation and Resp. SRL Relation cover cases where we removed Equations 4 and 5 respectively from Equation 6. Similar to Ghosh et al. (Ghosh, Heffernan, and Lan 2020) and Gu (Gu 2021), we also conducted an ablation study to compare the sinusoidal positional encoding and proposed time-decaying feature. $Time_{sin}$ refers to removing Equation 3 and integrating the fixed sinosoidal positional embedding (Vaswani et al. 2017) to the interaction vector, $\mathbf{\hat{x}_i}$.

Tables 5 and 6 show that removing question and response SRL relation components decreased the model's performance. For both classification and regression tasks, removing the response SRL relation leads to poorer performance when compared to the question relation. However, removing the time-decaying feature and incorporating the sinusoidal encoding improves the model's performance. One explanation could be that the attention model was developed for natural language processing tasks incorporating the word embedding and positional embedding in an additive manner (Vaswani et al. 2017). Thus, adding positional encoding to the USE embedding led to better performance. Secondly, we used an additive term to combine the time-decaying feature in attention in Equation 8. Ghosh et al. stated that using a multiplicative time-decaying feature to their attention model led to better performance compared to additive one

Dataset	Method	Accuracy	AUC
D1	Ques. Relation	0.73	0.78
	Time$_{sin}$	0.78	0.86
	Full Model (Table 3)	0.77	0.84
D2.1	Ques. Relation	0.75	0.74
	Time$_{sin}$	0.78	0.83
	Full Model (Table 3)	0.77	0.83
D2	Ques. Relation	0.78	0.79
	Resp. SRL Relation	0.74	0.77
	Time$_{sin}$	0.78	0.86
	Full Model (Table 3)	0.78	0.85

Table 5: Classification Ablation.

Dataset	Method	MSE	MAE
D1	Ques. Relation	2.42	1.31
	Time$_{sin}$	1.79	0.92
	Full Model (Table 4)	2.11	1.04
D2.1	Ques. Relation	2.24	1.08
	Time$_{sin}$	1.80	0.94
	Full Model (Table 4)	1.95	1.01
D2	Ques. Relation	1.97	1.06
	Resp. SRL Relation	2.19	1.16
	Time$_{sin}$	1.94	1.06
	Full Model (Table 4)	1.75	0.98

Table 6: Regression Ablation.

(see Section 3.2 (Ghosh, Heffernan, and Lan 2020)). We opted for the additive operation to avoid any additional computational cost and integrate students' forgetting behavior. Although the sinusoidal encoding results in better performance, it captures sequential event positions and does not capture the inter-event duration.

Attention Visualization. For the visualization purpose, we present an example of attention weight averaged on four heads in Figure 4. The student covered eight actions. Table 7 presents the timestamp, action type, action text, and score received on the question. We exclude T3-T5 entries for better illustration. We describe the model's classification score prediction on the question attempted at T8.

First, Figure 4 shows the model puts more weight on recent actions T5- T7 compared to distant ones T1-T4. This results from the model's time-decaying feature. Second, we observe from Table 7 the student scored zeroes on questions attempted at T6 and T7. The question submitted at T8 is a resubmission of T6. Thus, the model captured previous scores on questions to predict the question score at T8 = 0. Third, we observe that the student has used the term "atoms" in response to T8–which they looked up as a vocabulary word ("atoms") at T1. Equation 5 captures the relationship between a question's response text and any previous SRL. Although the student used the vocabulary lookup in their response, Figure 4 shows that our model has put less weight on T1 action compared to T5-T7. We conclude this is a result of the time-decaying feature.

Overall, our model can provide information about (i) how past question attempts impact the current question's score

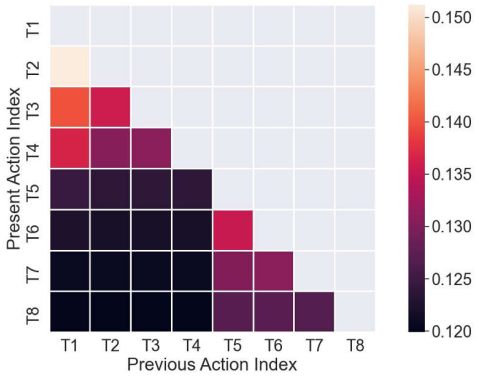

Figure 4: Attention Weight Visualization on Test Fold (Classification). Weight is Averaged on Four Heads.

Time (Type)	Action Text (Score)
T1 (V)	atoms (N/A)
T2 (V)	cesium (N/A)
T3-T5	...
T6 (Q)	Why is a telephone considered an example of matter? (0)
T7 (Q)	What is the difference between mass and volume? (0)
T8 (Q)	Why is a telephone considered an example of matter? (*True Score = 0*, *Predicted Score= 0*) Resp.: "because it has weight ,it takes up space and is made of **attoms** and molecules"

Table 7: Student Action Sequence of Figure 4. V = Vocabulary Lookup, Q = Question Attempt.

and (ii) how a student's SRL behaviors before attempting a question are related to the (a) question text and (b) submitted answer text.

Discussion and Implications

We outline possible implications below.

For Researchers. Although researchers have proposed dozens of variations on the basic DKT model (see Section Related Work), these enhancements only consider the students' past problem-solving behavior to predict future question performance. We address this gap by utilizing past question-answering and reading-related SRL features in our predictive model. Our experiments on D2 and D2.1 datasets show that combining temporal ordering of SRL features along with question attempts leads to better predictive performance than considering only question attempts.

Teaching and Learning. In a teaching context, a model's interpretation is important for intervention. By providing an estimate of student performance that is tied to specific study habits and SRL features we can: (i) better identify what contributed to students' performance and (ii) what study habits students can engage in to support better results. As our model captures temporal data, a next step can be designing a recommendation system for students' personalized education.

Reading Intervention. Effective reading interventions are a central goal for science education in the United States, as seen by the emphasis within the new science education framework proposed by National Research Council *"...Reading, interpreting, and producing text are fundamental practices of science in particular, and they constitute at least half of engineers' and scientists' total working time."* (NRC 2012). Prior researchers have identified that reading in science is different from reading in other domains, because (i) science reading involves diagrams, charts, and multimodal sources (Yore 2012) and (ii) academic science text is difficult to comprehend (Buehl 2017) due to high lexical density (i.e., the ratio of content words to overall words). Our study explores how science-related reading behaviors and question-answering can be modeled–an important step toward analyzing this area of education research.

Conclusions

In this study, we investigated middle school students' performance on question attempts and reading-related interactions within an online reading platform, Actively Learn. We combine theory from learning science to identify reading-related student interaction—SRL behaviors (Zimmerman 1989), such as annotating, highlighting, and vocabulary lookup. Our transformer-based predictive model integrates the temporal ordering of question-solving and students' interactions —an area that is underexplored. Our findings show that by combining temporal ordering of question attempts with other interaction behaviors we can yield better predictive power when compared to only taking question-attempt actions. The attention visualization provides explanations by displaying weights of different actions during prediction. We believe findings of our study will be helpful for students, teachers, and researchers in AI and education.

Acknowledgements

This research is partially supported by NSF #1821475 "Concert: Coordinating Educational Interactions for Student Engagement" Collin F. Lynch, Tiffany Barnes, and Sarah Heckman (Co-PIs). We would like to thank Actively Learn for providing study data.

References

Barnes, T. 2005. The Q-matrix method: Mining student response data for knowledge. In *American Association for Artificial Intelligence 2005 Educational Data Mining Workshop*, 1–8. Pittsburgh, PA: AAAI Press.

Biemiller, A.; and Slonim, N. 2001. Estimating root word vocabulary growth in normative and advantaged populations: Evidence for a common sequence of vocabulary acquisition. *Journal of educational psychology*, 93(3): 498.

Buehl, D. 2017. *Developing readers in the academic disciplines*. Stenhouse Publishers.

Cen, H.; Koedinger, K.; and Junker, B. 2006. Learning Factors Analysis – A General Method for Cognitive Model Evaluation and Improvement. In Ikeda, M.; Ashley, K. D.;

and Chan, T.-W., eds., *Intelligent Tutoring Systems*, 164–175. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-540-35160-3.

Cer, D.; Yang, Y.; Kong, S.-y.; Hua, N.; Limtiaco, N.; John, R. S.; Constant, N.; Guajardo-Cespedes, M.; Yuan, S.; Tar, C.; et al. 2018. Universal Sentence Encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 169–174.

Chi, M.; Koedinger, K. R.; Gordon, G. J.; Jordan, P. W.; and VanLehn, K. 2011. Instructional Factors Analysis: A Cognitive Model For Multiple Instructional Interventions. In Pechenizkiy, M.; Calders, T.; Conati, C.; Ventura, S.; Romero, C.; and Stamper, J. C., eds., *Proceedings of the 4th International Conference on Educational Data Mining, Eindhoven, The Netherlands, July 6-8, 2011*, 61–70. www.educationaldatamining.org. ISBN 978-90-386-2537-9.

Choi, Y.; Lee, Y.; Shin, D.; Cho, J.; Park, S.; Lee, S.; Baek, J.; Bae, C.; Kim, B.; and Heo, J. 2020. Ednet: A large-scale hierarchical dataset in education. In *International Conference on Artificial Intelligence in Education*, 69–73. Springer.

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

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Ebbinghaus, H. 2013. Memory: A contribution to experimental psychology. *Annals of neurosciences*, 20(4): 155.

Ghosh, A.; Heffernan, N.; and Lan, A. S. 2020. Context-aware attentive knowledge tracing. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2330–2339.

Gomez, L.; Herman, P.; and Gomez, K. 2007. Integrating text in content-area classes: Better supports for teachers and students. *Voices in Urban Education*, 14: 22–29.

Gu, Y. 2021. Attentive Neural Point Processes for Event Forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 7592–7600.

Hambleton, R. K.; Swaminathan, H.; and Rogers, H. J. 1991. *Fundamentals of item response theory*, volume 2. Sage.

Kintsch, W. 1988. The use of knowledge in discourse processing: A construction-incrementation model'. *Psychological Review*, (85): 363–394.

Kintsch, W.; and Walter Kintsch, C. 1998. *Comprehension: A paradigm for cognition*. Cambridge university press.

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

Magliano, J. P.; Todaro, S.; Millis, K.; Wiemer-Hastings, K.; Kim, H. J.; and McNamara, D. S. 2005. Changes in reading strategies as a function of reading training: A comparison of live and computerized training. *Journal of Educational Computing Research*, 32(2): 185–208.

Makany, T.; Kemp, J.; and Dror, I. E. 2009. Optimising the use of note-taking as an external cognitive aid for increasing learning. *British Journal of Educational Technology*, 40(4): 619–635.

Michalsky, T. 2013. Integrating skills and wills instruction in self-regulated science text reading for secondary students. *International Journal of science education*, 35(11): 1846–1873.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; and Dean, J. 2013. Distributed Representations of Words and Phrases and Their Compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, 3111–3119. Red Hook, NY, USA: Curran Associates Inc.

Mongkhonvanit, K.; Kanopka, K.; and Lang, D. 2019. Deep knowledge tracing and engagement with moocs. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, 340–342.

Nagatani, K.; Zhang, Q.; Sato, M.; Chen, Y.-Y.; Chen, F.; and Ohkuma, T. 2019. Augmenting knowledge tracing by considering forgetting behavior. In *The world wide web conference*, 3101–3107.

Norris, S. P.; and Phillips, L. M. 2003. How literacy in its fundamental sense is central to scientific literacy. *Science education*, 87(2): 224–240.

NRC. 2012. *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. National Academies Press.

Pandey, S.; and Karypis, G. 2019. A Self-Attentive Model for Knowledge Tracing. *International Educational Data Mining Society*.

Pandey, S.; and Srivastava, J. 2020. RKT: Relation-Aware Self-Attention for Knowledge Tracing. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 1205–1214.

Pavlik, P. I.; Cen, H.; and Koedinger, K. R. 2009. Performance Factors Analysis –A New Alternative to Knowledge Tracing. 531–538. NLD: IOS Press. ISBN 9781607500285.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Perfetti, C. A. 1985. *Reading ability.* oxford university Press.

Peters, M.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2227–2237.

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.

Reimers, N.; Gurevych, I.; Reimers, N.; Gurevych, I.; Thakur, N.; Reimers, N.; Daxenberger, J.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Shankweiler, D.; Lundquist, E.; Katz, L.; Stuebing, K. K.; Fletcher, J. M.; Brady, S.; Fowler, A.; Dreyer, L. G.; Marchione, K. E.; Shaywitz, S. E.; et al. 1999. Comprehension and decoding: Patterns of association in children with reading difficulties. *Scientific studies of reading*, 3(1): 69–94.

Shin, D.; Shim, Y.; Yu, H.; Lee, S.; Kim, B.; and Choi, Y. 2021. *SAINT+: Integrating Temporal Features for EdNet Correctness Prediction*, 490–496. New York, NY, USA: Association for Computing Machinery. ISBN 9781450389358.

Sonkar, S.; Waters, A. E.; Lan, A. S.; Grimaldi, P. J.; and Baraniuk, R. G. 2020. qDKT: Question-Centric Deep Knowledge Tracing. In *Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020)*, 677–681.

Su, Y.; Liu, Q.; Liu, Q.; Huang, Z.; Yin, Y.; Chen, E.; Ding, C.; Wei, S.; and Hu, G. 2018. Exercise-enhanced sequential modeling for student performance prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2435–2443.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Wang, C.; Ma, W.; Zhang, M.; Lv, C.; Wan, F.; Lin, H.; Tang, T.; Liu, Y.; and Ma, S. 2021. Temporal Cross-Effects in Knowledge Tracing. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, WSDM '21, 517–525. New York, NY, USA: Association for Computing Machinery. ISBN 9781450382977.

Winne, P. H.; Nesbit, J. C.; Ram, I.; Marzouk, Z.; Vytasek, J.; Samadi, D.; and Stewart, J. 2017. Tracing Metacognition by Highlighting and Tagging to Predict Recall and Transfer. *AERA Online Paper Repository*.

Yore, L. D. 2012. Science literacy for all: More than a slogan, logo, or rally flag! In *Issues and challenges in science education research*, 5–23. Springer.

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

Zhang, L.; Xiong, X.; Zhao, S.; Botelho, A.; and Heffernan, N. T. 2017b. Incorporating rich features into deep knowledge tracing. In *Proceedings of the fourth (2017) ACM conference on learning@ scale*, 169–172.

Zimmerman, B. J. 1989. Models of self-regulated learning and academic achievement. In *Self-regulated learning and academic achievement*, 1–25. Springer.

Zimmerman, B. J. 2000. Attaining self-regulation: A social cognitive perspective. In *Handbook of self-regulation*, 13–39. Elsevier.