Enhancing Length Generalization for Attention Based Knowledge Tracing Models with Linear Biases

Xueyi Li[1,5], **Youheng Bai**[1], **Teng Guo**[1], **Zitao Liu**[1*], **Yaying Huang**[1], **Xiangyu Zhao**[2], **Feng Xia**[3], **Weiqi Luo**[1] and **Jian Weng**[4]
[1]Guangdong Institute of Smart Education, Jinan University
[2]School of Data Science, City University of Hong Kong
[3]School of Computing Technologies, RMIT University
[4]College of Information Science and Technology, Jinan University
[5]School of Intelligent Systems Science and Engineering, Jinan University
lixueyi@stu2021.jnu.edu.cn, baiyouheng27@outlook.com, {tengguo, liuzitao, huangyaying, lwq}@jnu.edu.cn, xianzhao@cityu.edu.hk, f.xia@ieee.org, cryptjweng@gmail.com

Abstract

Knowledge tracing (KT) is the task of predicting students' future performance based on their historical learning interaction data. With the rapid advancement of attention mechanisms, many attention based KT models are developed. However, existing attention based KT models exhibit performance drops as the number of student interactions increases beyond the number of interactions on which the KT models are trained. We refer to this as *the length generalization of KT model*. In this paper, we propose **stableKT** to enhance length generalization that is able to learn from short sequences and maintain high prediction performance when generalizing on long sequences. Furthermore, we design a multi-head aggregation module to capture the complex relationships between questions and the corresponding knowledge components (KCs) by combining dot-product attention and hyperbolic attention. Experimental results on three public educational datasets show that our model exhibits robust capability of length generalization and outperforms all baseline models in terms of AUC. To encourage reproducible research, we make our data and code publicly available at https://pykt.org.

1 Introduction

Knowledge tracing (KT) is a sequential prediction task that utilizes the historical learning interaction data of students to predict their responses to future questions. This is achieved by modeling students' mastery of knowledge, i.e., knowledge states, as they interact with learning platforms such as massive open online courses (MOOCs) and intelligent tutoring systems. Solving the KT task can empower teachers to better guide students who need further attention or recommend personalized learning materials. This is crucial for the development of next-generation intelligent and personalized education.

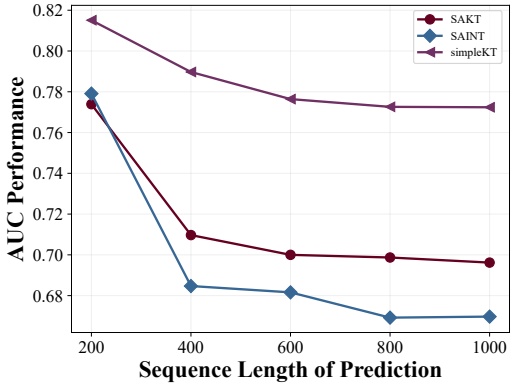

Figure 1: The AUC performance of KT models with different sequence lengths.

Recently, with the rapid development of attention mechanisms [Vaswani *et al.*, 2017], many attention based KT models are developed such as SAKT [Pandey and Karypis, 2019], SAINT [Choi *et al.*, 2020] and simpleKT [Liu *et al.*, 2023b]. To build an effective attention based KT model, a major design decision is the length of student interaction sequences at training stage, denoted L herein, which has been equivalent to the length of student interaction sequences at prediction stage. When computing the attention scores, incorporating more student interactions (achieved by a larger L) in the context window improves estimations of student knowledge mastery at prediction stage. However, longer interaction sequences are more expensive to train on [Press *et al.*, 2022]. We refer to this as *the length generalization of KT model*, which denotes a KT model's capability to continue performing well as the number of student interactions increases beyond the number of interactions on which the KT model is trained.

While in some cases the attention based KT models trained on short sequences can be directly applied on longer interac-

*Corresponding author.

tion sequences at prediction stage, the performance degrades as the length of sequences increases (Shown in Figure 1). We train attention based KT models, such as SAKT [Pandey and Karypis, 2019], SAINT [Choi *et al.*, 2020] and simpleKT [Liu *et al.*, 2023b], on student interaction sequences with a fixed length of 200 and evaluate on sequences of varying lengths, i.e., 200, 400, 600, 800 and 1000. In Figure 1, these attention based KT models display lower prediction performance on longer sequences compared to shorter ones, which poses a significant challenge when generalizing the well-trained KT models to students who have long historical interaction sequences.

Furthermore, in real-world educational scenarios, there are intricate relationships between questions and their associated knowledge components (KCs)[1]. Effectively capturing these relationships may significantly boost the performance of KT models [Cui *et al.*, 2023]. However, most existing attention based KT models, such as SAKT [Pandey and Karypis, 2019], SAINT [Choi *et al.*, 2020] and simpleKT [Liu *et al.*, 2023b], simply rely on the standard dot-product attention function, which computes the similarity between two student interactions by taking their inner product. Such inner product fails to model the complex and hidden structural properties of questions and their associated KCs.

Therefore, in this paper, we present KT solutions that are able to help models learn from short sequences and generalize well to longer sequences at prediction stage, and at the same time effectively capture complex relationships between questions and their associated KCs when learned from a collection of real-world student interaction data. Our work focuses on the refinements of a popular attention based KT baseline model, i.e., simpleKT [Liu *et al.*, 2023b].

Briefly, the simpleKT model explicitly captures question-specific variations of the individual differences among questions covering the same set of KCs and uses the standard dot-product attention function to extract the time-aware information embedded in the student learning interactions. However, when learning a standard simpleKT model from real-world educational datasets characterized by variations in interaction sequence lengths, several crucial questions arise: (1) Since the performance of attention based KT models notably drops on longer sequences, *how can we maintain consistent and stable prediction performance across student interaction sequences of varying lengths?* (2) Due to the fact that the questions and their associated KCs may contain intricate relationships, *how can we effectively capture such complex and structural relationships?*

In this work we address the above issues by introducing a novel KT model, i.e., **stableKT** that

- is able to learn from short sequences and maintain stable and consistent performance when generalizing on long sequences by biasing query-key attention scores with penalties that are proportional to query-key distances.
- captures complex relationships between questions and their associated KCs by computing their similarities using the depth of their lowest common ancestor in a hierarchy.
- supports accurate estimations of student knowledge state and response predictions.

Our stableKT model builds upon the standard simpleKT model and enhances its length generalization with linear biases applied to attention scores. It utilizes multi-head aggregation module to capture individual differences and complex hierarchical relationships. To ensure fair comparisons with recently developed deep learning based KT (DLKT) models, we choose to follow a publicly available standardized KT task evaluation protocol [Liu *et al.*, 2022]. We conduct comprehensive and rigorous experiments on three public datasets, and the results show that our stableKT model is able to greatly enhance the length generalization and improve the prediction performance in terms of AUC.

2 Background and Related Works

2.1 SimpleKT

The simpleKT is a popular and widely used KT baseline model that captures question-specific variations of the individual differences among questions covering the same set of KCs [Liu *et al.*, 2023b]. It utilizes the standard dot-product attention function to simplify the sophisticated student knowledge state estimation. The definitions of the simpleKT model are as follows:

$$
\begin{aligned}
(\boldsymbol{x}_t, \boldsymbol{y}_t) &= \text{InteractionEncoder}(\boldsymbol{q}_t, \boldsymbol{c}_t, \boldsymbol{r}_t) \\
\mathbf{Q} = \boldsymbol{x}_{t+1}; \mathbf{K} &= \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_t\}; \mathbf{V} = \{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_t\} \\
\boldsymbol{h}_{t+1} &= \text{SelfAttention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) \\
\hat{\boldsymbol{r}}_{t+1} &= \text{PredictionLayer}(\boldsymbol{h}_{t+1}, \boldsymbol{x}_{t+1})
\end{aligned}
$$

where $\boldsymbol{q}_t$, $\boldsymbol{c}_t$, $\boldsymbol{r}_t$ represent question, KC and response respectively at the (t)-th time step. The InteractionEncoder is an encoder that characterizes the latent factor of question difficulty. $\hat{\boldsymbol{r}}_{t+1}$ and $\boldsymbol{h}_{t+1}$ denote prediction result and the extracted knowledge state respectively at the $(t+1)$-th time step. The PredictionLayer is a two-layer fully connected network.

2.2 Related Works

Attention Based KT Models

Attention based KT models utilize attention mechanisms to capture relationships among student interactions. SAKT is the first research work that adopted a self-attention network to predict students' future performance [Pandey and Karypis, 2019]. Since then, many KT models use attention based network to capture the potential relationships between student interactions. Ghosh et al. proposed a monotonic attention mechanism, building upon the dot-product attention function, which computes attention weights with exponential time-related decay [Ghosh *et al.*, 2020]. Huang et al. incorporated a k-selection module designed to choose relevant historical interactions based on the highest dot-product attention scores [Huang *et al.*, 2023]. Im et al. represented forgetting behaviors as linear biases in their approach [Im *et al.*, 2023]. Yin et al. designed a temporal and cumulative attention to diagnose students' knowledge proficiency from each question mastery state [Yin *et al.*, 2023].

[1] A knowledge component (KC) is a generalization of everyday terms like concept, principle, fact, or skill.

Length Generalization
The capability of length generalization allows a KT model to continue performing well as the length of student interaction sequences increases beyond the length of interaction sequences on which the KT model is trained. Position embedding plays an important role in length generalization [Press *et al.*, 2022; Chi *et al.*, 2023; Qin *et al.*, 2024]. Sinusoidal position embedding employs sinusoidal functions with either non-learned or learnable parameters to generate position embedding and combines the position embedding with input embeddings [Vaswani *et al.*, 2017; Jacob *et al.*, 2019]. Different from Sinusoidal position embedding, Rotary position embedding computes embedding by sinusoidal functions with queries and keys instead of input embedding [Su *et al.*, 2021]. T5 position embedding provides position embedding by adding a learned, shared bias to each query-key score before the softmax operation of attention [Raffel *et al.*, 2020].

However, the KT models utilizing the above position embeddings both experience performance drops at prediction stage when applied to longer sequences. Inspired by [Press *et al.*, 2022], we enhance the length generalization of our stableKT model by biasing query-key attention scores with penalties that are proportional to query-key distance. Additionally, different from existing attention based KT models, our stableKT model designs a multi-head aggregation module that combines dot-product attention and hyperbolic attention to capture a more complex relationship between questions and their associated KCs.

3 The StableKT Framework

3.1 Problem Definition

Our objective is to develop a KT model $\boldsymbol{\mathcal{M}}$ that is able to learn from short student interaction sequences and maintain high prediction performance when applied on longer sequences. We refer to this as *the length generalization of KT model*, which is defined as follows:

Definition 1 (Length Generalization of KT Model)**.** *Given a student interaction dataset $\boldsymbol{\mathcal{D}}$, a KT model $\boldsymbol{\mathcal{M}}$, if for any l_p that $l_p > l_t$, there is,*

$$\frac{|auc_p(\boldsymbol{\mathcal{M}}, \boldsymbol{\mathcal{D}}) - auc_t(\boldsymbol{\mathcal{M}}, \boldsymbol{\mathcal{D}})|}{auc_t(\boldsymbol{\mathcal{M}}, \boldsymbol{\mathcal{D}})} < \epsilon$$

then KT model $\boldsymbol{\mathcal{M}}$ is considered to have the capability of the length generalization, where auc_p and auc_t denote the AUC scores on sequences with length l_p and l_t on prediction and training sequences respectively and ϵ is a small positive constant.

3.2 The Framework Overview

In this section, we present the framework overview of our stableKT model (Shown in Figure 2) that consists of five components: (1) interaction encoding module that explicitly uses a scalar to characterize the latent factor of question difficulty (See Section 3.3); (2) hyperbolic attention module that captures complex relationships between questions and their associated KCs (See Section 3.4); (3) length generalization module that enhances KT model prediction performance on longer sequences (See Section 3.5); (4) multi-head aggregation module that utilizes both the hierarchy-aware similarity score and the standard dot-product attention score (See Section 3.6); and (5) prediction module that uses a two-layer fully connected network to make prediction (See Section 3.7).

3.3 Interaction Encoding Module

Due to the fact that there are various difficulty levels between questions covering the same set of KCs, it is crucial to effectively represent student interactions. Similar to simpleKT [Liu *et al.*, 2023b], we encode the interactions as follows:

$$\boldsymbol{z}_{\boldsymbol{c}_t} = \mathbf{W}_c \cdot \boldsymbol{e}_{\boldsymbol{c}_t}; \quad \boldsymbol{a}_{\boldsymbol{r}_t} = \mathbf{W}_r \cdot \boldsymbol{e}_{\boldsymbol{r}_t}$$
$$\boldsymbol{x}_t = \boldsymbol{z}_{\boldsymbol{c}_t} \oplus \boldsymbol{f}_{\boldsymbol{q}_t} \odot \boldsymbol{v}_{\boldsymbol{c}_t}; \quad \boldsymbol{y}_t = \boldsymbol{z}_{\boldsymbol{c}_t} \oplus \boldsymbol{a}_{\boldsymbol{r}_t}$$

where $\boldsymbol{z}_{\boldsymbol{c}_t}$ and $\boldsymbol{a}_{\boldsymbol{r}_t}$ denote the latent representations of KC $\boldsymbol{c}_t$ and student response $\boldsymbol{r}_t$ on question $\boldsymbol{q}_t$. $\boldsymbol{e}_{\boldsymbol{c}_t}$ and $\boldsymbol{e}_{\boldsymbol{r}_t}$ represent the original s-dimensional and 2-dimensional one-hot vectors of the corresponding KC and response respectively. $\mathbf{W}_c \in \mathbb{R}^{d \times s}$ and $\mathbf{W}_r \in \mathbb{R}^{d \times 2}$ are learnable linear transformation operations. $\boldsymbol{x}_t$ denotes the augmented embedding of KC $\boldsymbol{c}_t$ with difficulty vector $\boldsymbol{f}_{\boldsymbol{q}_t}$. $\boldsymbol{v}_{\boldsymbol{c}_t}$ represents the question-centric variation of $\boldsymbol{q}_t$ covering this KC $\boldsymbol{c}_t$. $\boldsymbol{q}_t$ denotes the question answered at the (t)-th timestamp. $\boldsymbol{y}_t$ represents the embedding of interaction with response $\boldsymbol{r}_t$. $\odot$ and $\oplus$ represent the element-wise product and addition operators respectively.

3.4 Hyperbolic Attention Module

Existing attention based KT models mostly employ the dot-product attention function to capture relationships between questions and their associated KCs in Euclidean space. The attention scores $\mathbf{S}_{\text{dot}}$ of dot-product attention are calculated by taking the inner product of query $\mathbf{Q}$ and key $\mathbf{K}$ as:

$$\mathbf{S}_{\text{dot}} = \frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}$$

where $\sqrt{d}$ is a scaling factor and the $\mathbf{K}^T$ represents the transpose of $\mathbf{K}$.

However, in real-world educational datasets, there exists a tree-like hierarchical relationship between questions and their associated KCs, which is a challenge to capture using dot-product attention in Euclidean space. Following the previous work of [Yu *et al.*, 2023; Tseng *et al.*, 2023], we measure a hierarchy-aware similarity score of query $\mathbf{Q}$ and key $\mathbf{K}$ in hyperbolic space that is well-suited for embedding tree-like structures. Intuitively, the radius of a hyperbolic ball is directly proportional to its volume, which grows exponentially. Similarly, the number of leaves grows exponentially with respect to depth in a tree.

Hyperbolic Mapping
Since the hyperbolic space cannot be isometrically embedded into Euclidean space, similar to [Gulcehre *et al.*, 2019; Tseng *et al.*, 2023], in this work we use the Poincaré half-space model to represent the hyperbolic space by a subset of Euclidean space. In the hyperbolic space represented by

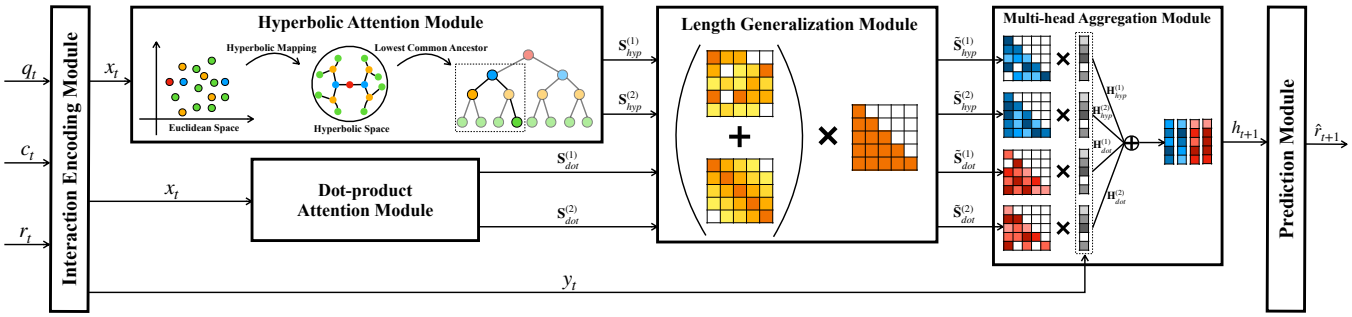

Figure 2: The overview of the proposed stableKT framework.

Poincaré half-space model, ideal points, points at infinity and geodesics[2] all have particularly nice Euclidean forms and details can be found at [Gulcehre *et al.*, 2019; Yu *et al.*, 2023; Tseng *et al.*, 2023].

Here, we discuss mappings of student interactions from Euclidean space to the hyperbolic space. Specifically, let $\mathbf{X}_d$ be the last dimension of $\mathbf{X}$ and $\mathbf{X}_{[:-1]}$ be the first $d-1$ dimensions of $\mathbf{X}$. Similar to [Tseng *et al.*, 2023], we transform the query $\mathbf{Q}$ and key $\mathbf{K}$ matrices in the standard dot-product attention via the penumbral mapping function defined as follows:

$$\Psi(\mathbf{X})_{[:-1]} = \mathbf{X}_{[:-1]} \frac{\alpha}{1+\exp\left(-\mathbf{X}_d\right)}$$
$$\Psi(\mathbf{X})_d = \frac{\alpha}{1+\exp\left(-\mathbf{X}_d\right)}$$

where α is the mapping coefficient. Therefore, we obtain the mapped query matrix $\hat{\mathbf{Q}}$ and key matrix $\hat{\mathbf{K}}$ in the hyperbolic space, i.e.,

$$\hat{\mathbf{Q}} = [\Psi(\mathbf{Q})_{[:-1]}; \Psi(\mathbf{Q})_d]; \quad \hat{\mathbf{K}} = [\Psi(\mathbf{K})_{[:-1]}; \Psi(\mathbf{K})_d]$$

Similarity in Hyperbolic Space

In the hyperbolic space, the similarity between two data points reflects their hierarchical relationships. More specifically, we associate two points by the depth of their lowest common ancestor (LCA) in the cone partial ordering in hyperbolic cones, which is analogous to finding their LCA in a latent tree and captures how divergent two points are. We compute the attention scores in the hyperbolic space following the penumbral attention definition defined in [Tseng *et al.*, 2023]. In particular, if Equation (1) holds true, as follows:

$$\left(\left\|\Psi(\mathbf{Q})_{[:-1]} - \Psi(\mathbf{K})_{[:-1]}\right\| - \sqrt{\alpha^2 - \Psi(\mathbf{Q})_d^2}\right)^2 + \Psi(\mathbf{K})_d^2 < \alpha^2 \quad (1)$$

the hierarchy-aware similarity score $\mathbf{S}_{\text{hyp}}$ in hyperbolic space is computed by:

$$u = \frac{\sqrt{\alpha^2 - \Psi(\mathbf{Q})_d^2} + \sqrt{\alpha^2 - \Psi(\mathbf{K})_d^2} - \left\|\Psi(\mathbf{Q})_{[:-1]} - \Psi(\mathbf{K})_{[:-1]}\right\|}{2}$$
$$\mathbf{S}_{\text{hyp}} = \exp(-\gamma \max\left(\Psi(\mathbf{Q})_d, \Psi(\mathbf{K})_d, \sqrt{\alpha^2 - u^2}\right))$$

[2]The shortest path between two points.

where γ is a scaling coefficient and $\|\cdot\|$ denotes the L2-norm distance. If Equation (1) does not hold true, the hierarchy-aware similarity score $\mathbf{S}_{\text{hyp}}$ is computed by:

$$v = \frac{\left\|\Psi(\mathbf{Q})_{[:-1]} - \Psi(\mathbf{K})_{[:-1]}\right\|^2 + \Psi(\mathbf{Q})_d^2 - \Psi(\mathbf{K})_d^2}{2\left\|\Psi(\mathbf{Q})_{[:-1]} - \Psi(\mathbf{K})_{[:-1]}\right\|}$$
$$\mathbf{S}_{\text{hyp}} = \exp(-\gamma\sqrt{v^2 + \Psi(\mathbf{K})_d^2})$$

3.5 Length Generalization Module

To help stableKT continue performing well as the number of student interactions increases, inspired by [Press *et al.*, 2022], we design a length generalization module that penalizes query-key attention scores with linear biases. Specifically, the linear biases implicitly contain relative position information of student interaction sequences by leveraging query-key distance in a computationally friendly way, which does not include learnable parameters and can avoid attention based KT model overfitting the position embeddings during training. We formulate the length generalization module $\boldsymbol{g}$ as follows:

$$\boldsymbol{g}(\mathbf{S}, \mathbf{B}, \mathbf{C}) = \text{Softmax}\{(\mathbf{S} \oplus \mathbf{B}) \odot \mathbf{C}\} \quad (2)$$

where the matrices $\mathbf{S}$, $\mathbf{B}$ and $\mathbf{C}$ denote attention scores (from either hyperbolic attention module or dot-product attention module), linear biases and causal mask respectively. The Softmax$\{\cdot\}$ represents softmax function.

The matrix $\mathbf{B}$ in Equation (2) penalizes attention scores with linear biases. Each element of matrix $\mathbf{B}$ is computed by:

$$b_{mn} = -|m-n| \cdot 2^{-8\frac{i}{H}}$$

where b_{mn} denotes the element at the (m)-th row and (n)-th column of matrix $\mathbf{B}$. The $2^{-8\frac{i}{H}}$ denotes a coefficient that adjusts the attention scores for the (i)-th attention head out of H attention heads.

The matrix $\mathbf{C}$ in Equation (2) ensures that attention based KT models cannot peek the future interaction sequences of students. Each element of matrix $\mathbf{C}$ is computed by:

$$c_{mn} = \begin{cases} 1, & m \geq n \\ 0, & \text{otherwise} \end{cases}$$

where c_{mn} denotes the element at the (m)-th row and (n)-th column of $\mathbf{C}$.

3.6 Multi-head Aggregation Module

To effectively capture the intricate relationships between questions and their KCs, inspired by [Pan *et al.*, 2022; Wang *et al.*, 2023; Li *et al.*, 2023], we design a multi-head aggregation module by concatenating dot-product attention head $\mathbf{H}_{\text{dot}}$ and hyperbolic attention head $\mathbf{H}_{\text{hyp}}$:

$$\mathbf{H}_{\text{dot}}^{(i)} = \boldsymbol{g}(\mathbf{S}_{\text{dot}}, \mathbf{B}, \mathbf{C}) \cdot \mathbf{V}$$
$$\mathbf{H}_{\text{hyp}}^{(j)} = \boldsymbol{g}(\mathbf{S}_{\text{hyp}}, \mathbf{B}, \mathbf{C}) \cdot \mathbf{V}$$
$$\boldsymbol{h}_{t+1} = \text{Concat}(\{\mathbf{H}_{\text{dot}}^{(i)}\}, \{\mathbf{H}_{\text{hyp}}^{(j)}\}) \cdot \mathbf{W}_h$$

where $\mathbf{H}_{\text{dot}}^{(i)}$ and $\mathbf{H}_{\text{hyp}}^{(j)}$ denote the (i)-th dot-product attention head and the (j)-th hyperbolic attention head respectively. $\mathbf{V}$ represents the value of attention. $\mathbf{W}_h \in \mathbb{R}^{d \times s}$ is a learnable linear transformation operation.

3.7 Prediction Module

We use a two-layer fully connected network to make prediction and optimize the prediction function by minimizing the binary cross-entropy loss between the ground-truth response $\boldsymbol{r}_{t+1}$ and the prediction probability $\hat{\boldsymbol{r}}_{\boldsymbol{t+1}}$ as follows:

$$\hat{\boldsymbol{r}}_{t+1} = \sigma(\phi(\mathbf{W}_2 \cdot \phi(\mathbf{W}_1 \cdot [\boldsymbol{h}_{t+1}; \boldsymbol{x}_{t+1}] + \boldsymbol{b}_1) + \boldsymbol{b}_2))$$
$$\boldsymbol{\mathcal{L}} = -\sum_t (\boldsymbol{r}_{t+1} \cdot \log \hat{\boldsymbol{r}}_{t+1} + (1 - \boldsymbol{r}_{t+1}) \cdot \log(1 - \hat{\boldsymbol{r}}_{t+1}))$$

where σ, ϕ denote Sigmoid and ReLU function. $\boldsymbol{b}_1$, $\boldsymbol{b}_2$, $\mathbf{W}_1$, $\mathbf{W}_2$ are trainable parameters.

4 Experiments

We present the details of our experiment settings and the corresponding results in this section. We conduct comprehensive analyses and investigations to illustrate the effectiveness of our stableKT model.

4.1 Datasets

We select three public real-world educational datasets to evaluate the effectiveness of our model.

- Algebra 2005-2006 (AL2005)[3]: The AL2005 dataset stems from KDD Cup 2010 EDM Challenge which includes 13-14 year-old students' interactions with Algebra questions. It has detailed step-level student responses to the mathematical problem. In our experiments, we use the concatenation of the problem name and step name as a unique question.
- Bridge to Algebra 2006-2007 (BD2006)[3]: The BD2006 dataset, similar to AL2005 dataset, consists of mathematical problems from logs of students' interactions with intelligent tutoring systems. The unique question construction of BD2006 dataset is similar to AL2005 dataset.
- NeurIPS2020 Education Challenge (NIPS34)[4]: The NIPS34 dataset is provided by NeurIPS 2020 Education Challenge. We use the dataset of Task 3 & Task 4 to evaluate our models. It contains students' answers to mathematics questions from Eedi which millions of students interact with daily around the globe.

To ensure reproducibility in our experiments, we rigorously follow the data pre-processing steps suggested in [Liu *et al.*, 2022]. We filter out student sequences that are shorter than 3 interactions. Data statistics are summarized in Table 1.

Dataset	# of interactions	# of students	# questions	# of KCs
AL2005	607,021	574	173,113	112
BD2006	1,817,458	1,145	129,263	493
NIPS34	1,382,678	4,918	948	57

Table 1: Data statistics of three widely used datasets.

4.2 Baselines

We compare our stableKT model with the following state-of-the-art DLKT models to evaluate the effectiveness of our approach:

- DKT [Piech *et al.*, 2015] uses a LSTM layer to encode the students' knowledge state for predicting their response performances.
- DKVMN [Zhang *et al.*, 2017] exploits two memory networks to extract the relationships between different KCs and students' knowledge states.
- GKT [Nakagawa *et al.*, 2019] casts the knowledge structure as a graph and reformulates the KT task as a time series node-level classification problem via a graph neural network.
- SAKT [Pandey and Karypis, 2019] leverages a self-attention mechanism to capture the relationships between question and KCs. It employs question embeddings as queries and utilizes interaction embeddings as both keys and values.
- SAINT [Choi *et al.*, 2020] employs a Transformer-based encoder-decoder architecture to handle students' question and response sequences.
- AKT [Ghosh *et al.*, 2020] introduces a monotonic attention to enhance self-attention by considering the students' forgetting behaviors.
- ATKT [Guo *et al.*, 2021] exploits adversarial perturbations to the interaction embeddings to enhance robustness of the model.
- LPKT [Shen *et al.*, 2021] uses a learning gate to distinguish students' absorptive capacity of knowledge and forgetting gate to model the decline of students' knowledge over time.
- simpleKT [Liu *et al.*, 2023b] uses dot-product attention to extract the time-aware information embedded in student learning interactions.

[3]https://pslcdatashop.web.cmu.edu/KDDCup

[4]https://eedi.com/projects/neurips-education-challenge

- DKT-AT [Liu *et al.*, 2023a] performs two auxiliary learning tasks, including question tagging prediction and individualized prior knowledge prediction task, to enhance the predictive capability of DKT.
- sparseKT [Huang *et al.*, 2023] incorporates a k-selection module to select relevant historical interactions with the highest attention scores to improve the robustness of attention based KT models.
- FoLiBiKT [Im *et al.*, 2023] is an attention based KT model that represents forgetting behaviors as linear biases decoupled from the question correlation.
- DTransformer [Yin *et al.*, 2023] exploits a two-level framework to explicitly diagnose learner's knowledge states and increase stability of knowledge state diagnosis by contrastive learning.

4.3 Experimental Setting

To evaluate the length generalization of KT models, all models are trained on student interaction sequences with the fixed length of 200 and evaluated on sequences with the length of 200, 400, 600, 800 and 1000, respectively. We perform standard 5-fold cross-validation for every combination of models and datasets. We choose to use early stopping when the performance is not improved after 10 epochs. For each hyperparameter combination, we use the Adam optimizer to train the models up to 200 epochs. We adopt the Bayesian search method to find the best hyperparameters for each fold. The embedding dimension, the hidden state dimension, the two dimension of the prediction layers are both set to [64, 256]. The learning rate, dropout rate and random seed are set to [1e-3, 1e-4, 1e-5], [0.05, 0.1, 0.3, 0.5] and [42, 3407] respectively. The scaling coefficient γ and mapping coefficient α are both set to [0.1, 0.3, 0.5, 0.7, 0.9, 1.0, 2.0, 5.0]. Similar to existing works [Liu *et al.*, 2022; Piech *et al.*, 2015; Ghosh *et al.*, 2020], we use the AUC to evaluate the KT prediction performance.

Model	Length of Interaction Sequences				
	200	400	600	800	1000
DKT	0.8149±0.0011	0.8150±0.0011	0.8150±0.0011	0.8149±0.0011	0.8149±0.0011
DKVMN	0.8054±0.0011	0.8039±0.0014	0.8030±0.0016	0.8025±0.0017	0.8023±0.0018
GKT	0.8110±0.0009	0.8111±0.0009	0.8111±0.0009	0.8111±0.0009	0.8111±0.0009
SAKT	0.7899±0.0036	0.6743±0.0023	0.6691±0.0030	0.6677±0.0024	0.6666±0.0018
SAINT	0.7715±0.0018	0.6691±0.0110	0.6589±0.0021	0.6539±0.0017	0.6551±0.0016
AKT	0.8306±0.0019	0.8277±0.0030	0.8258±0.0038	0.8241±0.0045	0.8227±0.0051
ATKT	0.7995±0.0023	0.7816±0.0025	0.7641±0.0039	0.7523±0.0047	0.7446±0.0050
LPKT	0.8268±0.0004	0.8216±0.0019	0.8107±0.0104	0.7990±0.0181	0.7891±0.0197
simpleKT	0.8210±0.0014	0.7808±0.0078	0.7763±0.0055	0.7535±0.0263	0.7655±0.0169
DKT-AT	0.8246±0.0019	0.8238±0.0019	0.8235±0.0019	0.8233±0.0020	0.8233±0.0020
sparseKT	0.8080±0.0030	0.7628±0.0091	0.7557±0.0082	0.7546±0.0118	0.7573±0.0090
FoLiBiKT	0.8310±0.0010	0.8288±0.0007	0.8272±0.0014	0.8256±0.0017	0.8242±0.0020
DTransformer	0.8188±0.0025	0.8156±0.0025	0.8137±0.0028	0.8123±0.0030	0.8112±0.0033
stableKT (Ours)	**0.8351±0.0008**	**0.8349±0.0008**	**0.8348±0.0009**	**0.8348±0.0009**	**0.8348±0.0009**

(a) Performance comparisons in terms of AUC on AL2005 dataset.

Model	Length of Interaction Sequences				
	200	400	600	800	1000
DKT	0.8015±0.0008	0.8015±0.0008	0.8015±0.0008	0.8015±0.0008	0.8015±0.0008
DKVMN	0.7983±0.0009	0.7956±0.0009	0.7936±0.0010	0.7925±0.0012	0.7919±0.0014
GKT	0.8046±0.0008	0.8047±0.0009	0.8047±0.0009	0.8047±0.0010	0.8047±0.0010
SAKT	0.7739±0.0015	0.7097±0.0056	0.7000±0.0042	0.6987±0.0035	0.6962±0.0044
SAINT	0.7791±0.0018	0.6847±0.0035	0.6816±0.0027	0.6692±0.0037	0.6697±0.0024
AKT	0.8208±0.0007	0.8187±0.0008	0.8168±0.0010	0.8155±0.0012	0.8144±0.0014
ATKT	0.7889±0.0008	0.7641±0.0028	0.7370±0.0041	0.7142±0.0042	0.6963±0.0040
LPKT	0.8056±0.0008	0.8014±0.0021	0.7965±0.0029	0.7939±0.0031	0.7923±0.0031
simpleKT	0.8151±0.0006	0.7897±0.0046	0.7764±0.0124	0.7726±0.0090	0.7724±0.0088
DKT-AT	0.8104±0.0009	0.8098±0.0008	0.8095±0.0007	0.8092±0.0006	0.8089±0.0006
sparseKT	0.8087±0.0079	0.7518±0.0080	0.7452±0.0090	0.7277±0.0150	0.7408±0.0082
FoLiBiKT	0.8199±0.0008	0.8171±0.0007	0.8145±0.0011	0.8125±0.0016	0.8110±0.0020
DTransformer	0.8093±0.0009	0.8052±0.0020	0.8023±0.0029	0.8002±0.0035	0.7985±0.0039
stableKT (Ours)	**0.8252±0.0003**	**0.8250±0.0003**	**0.8249±0.0003**	**0.8248±0.0003**	**0.8248±0.0003**

(b) Performance comparisons in terms of AUC on BD2006 dataset.

Model	Length of Interaction Sequences				
	200	400	600	800	1000
DKT	0.7689±0.0002	0.7689±0.0002	0.7689±0.0002	0.7689±0.0002	0.7689±0.0002
DKVMN	0.7673±0.0004	0.7673±0.0004	0.7673±0.0004	0.7672±0.0004	0.7672±0.0004
GKT	0.7689±0.0024	0.7689±0.0025	0.7689±0.0025	0.7689±0.0025	0.7689±0.0025
SAKT	0.7525±0.0009	0.7331±0.0013	0.7329±0.0011	0.7330±0.0011	0.7330±0.0011
SAINT	0.7895±0.0009	0.7708±0.0009	0.7703±0.0012	0.7700±0.0012	0.7700±0.0012
AKT	0.8033±0.0003	0.8030±0.0004	0.8028±0.0004	0.8028±0.0004	0.8028±0.0004
ATKT	0.7665±0.0001	0.7630±0.0005	0.7620±0.0006	0.7619±0.0006	0.7619±0.0006
LPKT	0.8004±0.0003	0.7997±0.0005	0.7993±0.0006	0.7992±0.0007	0.7992±0.0006
simpleKT	0.8035±0.0000	0.7952±0.0017	0.7961±0.0012	0.7960±0.0012	0.7960±0.0012
DKT-AT	0.7816±0.0002	0.7815±0.0002	0.7815±0.0002	0.7815±0.0002	0.7815±0.0002
sparseKT	0.8034±0.0013	0.7918±0.0021	0.7881±0.0026	0.7851±0.0029	0.7826±0.0029
FoLiBiKT	0.8032±0.0002	0.8029±0.0003	0.8028±0.0003	0.8028±0.0003	0.8028±0.0003
DTransformer	0.7994±0.0003	0.7988±0.0003	0.7985±0.0003	0.7985±0.0003	0.7985±0.0003
stableKT (Ours)	**0.8059±0.0004**	**0.8060±0.0004**	**0.8060±0.0004**	**0.8060±0.0004**	**0.8060±0.0004**

(c) Performance comparisons in terms of AUC on NIPS34 dataset.

Table 2: Performance comparisons in terms of AUC. The best AUCs are in **bold** and the second-best AUCs are underlined.

4.4 Results

Overall Performance

We report the average AUC and the standard deviations across 5 folds. Table 2 shows the overall performance. From Table 2, we have the following observations: (1) Our stableKT model maintains stable and consistent prediction performance across different length of student interaction sequences in all three datasets. In comparison to sequences of length 200, the performance of most baseline models degrades notably as the length of sequences increases on three datasets. For example, simpleKT model exhibits significant drop in performance at sequences of length 1000 compared to sequences of length 200, reaching up to 5.55% on AL2005 dataset. This indicates that our stableKT model has the capability to generalize the well-trained KT model to students who have long historical interaction sequences. (2) At the same length of student sequences, our stableKT model outperforms all the baseline models in all three datasets and improves the AUC of the original simpleKT model by up to 6.93% on AL2005 dataset, 5.24% on BD2005 dataset and 1.00% on NIPS34 dataset at sequences of length 1000. This indicates that our model can better capture complex relationships between questions and their associated KCs by using a multi-head aggregation module. (3) The DKT and GKT do not exhibit notable performance drops among AL2005, BD2006 and NIPS34 datasets, but they achieve far less AUC than our model. Since these models do not rely on attention mechanisms, they typically do not face length generalization challenges. However, compared to our stableKT model, these models struggle to effectively model the knowledge states of students.

Impact on Different Position Embeddings

As mentioned in Section 2, exploiting complex position embeddings, such as Sinusoidal, T5 and Rotary, usually hurts attention based KT models' capability of length generalization. We conduct experiments on our stableKT model with Sinusoidal, T5, Rotary and our length generalization module respectively. Figure 3 shows the results of our stableKT model with different position embeddings. From Figure 3, we have the following observations: (1) At different lengths of stu-

dent interaction sequences, our length generalization module maintains stable and consistent high prediction performance while other position embeddings exhibit notable drops in performance. This indicates that our length generalization module can effectively enhance the length generalization of KT models than other position embeddings. (2) At the same length of student interaction sequences, using different position embeddings also affects the prediction performance of KT models, and our length generalization module can effectively improve the prediction performance.

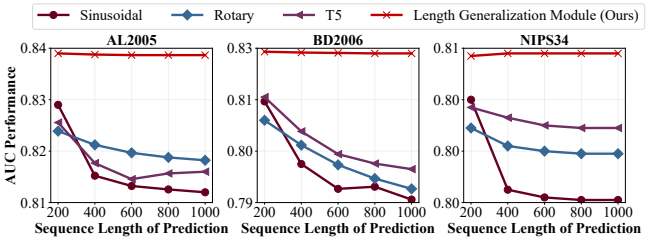

Figure 3: Performance analysis with different position embeddings.

Impact on Different Attention Mechanisms

We visualize the impact on dot-product attention and hyperbolic attention, as shown in Figure 4 and Figure 5. From that, we have the following observations: (1) Dot-product attention and hyperbolic attention, in Figure 4, focus on different information of questions. For example, when predicting the 19-th question, dot-product attention emphasizes information from questions 16, 17 and 18 while hyperbolic attention focuses on information from questions 0, 2, 3, 4, 5, 8, 13 and 18. This indicates that our multi-head aggregation module is able to utilize both dot-product attention and hyperbolic attention to capture more information of questions. (2) To further explore the characteristics of questions focused by hyperbolic attention, we visualize the questions that hyperbolic attention focuses on when predicting the 19-th question. As shown in Figure 5, the questions focused by hyperbolic attention, such as 0, 2, 3, 4, 5, 8, 13 and 18, exhibit a hierarchical relationship in terms of KCs, indicating that hyperbolic attention is able to capture hierarchical relationships of student interaction sequences. The Qs represents questions index and KCs denotes the questions' associated KCs index. The Maths[3] represents that the index 3 of KCs is Math, similarly for others.

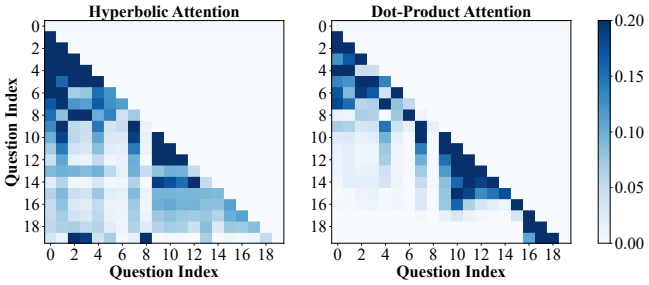

Figure 4: Visualization of both hyperbolic attention and dot-product attention. The Question index represents questions answered by a specific student.

Qs	KCs	Hierarchical Relationships of KCs
0	[3], [32], [211], [42]	
2	[3], [32], [144], [204]	
3	[3], [32], [144], [204]	
4	[3], [32], [211], [42]	
5	[3], [32], [144], [204]	
8	[3], [32], [219], [37]	
13	[3], [32], [144], [204]	
18	[3], [32], [219], [37]	
19	[3], [32], [144], [204]	

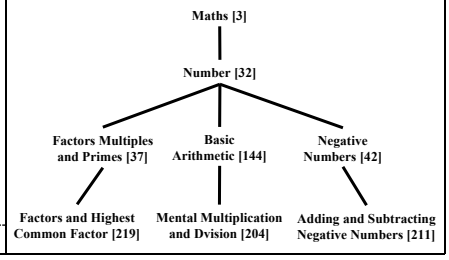

Figure 5: Hierarchical relationships captured by hyperbolic attention.

Ablation Study

We systematically examine the effect of two key components in our stableKT model by constructing four model variants in Table 3. The MA represents the multi-head aggregation module. The LG denotes the length generalization module. The w/o means excluding such module from stableKT model. Please note that the stableKT w/o MA & LG is equivalent to the vanilla simpleKT model. From Table 3, we have the following observations: (1) Compared to other variants, stableKT model maintains stable and consistent high prediction performance in all cases. This empirically verifies the importance of both length generalization module and multi-head aggregation module. (2) When comparing stableKT w/o LG to stableKT w/o MA & LG, we can observe that if the model lacks the length generalization module, its performance notably degrades as the length of sequences increases. Furthermore, the stableKT w/o MA & LG experiences more performance drops compared to stableKT w/o LG. This indicates that the multi-head aggregation module also has a positive impact on length generalization.

Models	Length of Interaction Sequences				
	200	400	600	800	1000
stableKT	**0.8351±0.0008**	**0.8349±0.0008**	**0.8348±0.0009**	**0.8348±0.0009**	**0.8348±0.0009**
stableKT w/o MA	0.8317±0.0010	0.8314±0.0011	0.8312±0.0011	0.8312±0.0011	0.8312±0.0011
stableKT w/o LG	0.8236±0.0036	0.7933±0.0099	0.7962±0.0087	0.7736±0.0266	0.7834±0.0140
stableKT w/o MA & LG	0.8210±0.0014	0.7808±0.0078	0.7763±0.0055	0.7535±0.0263	0.7655±0.0169

Table 3: Component analysis of stableKT model.

5 Conclusion

In this paper, we propose stableKT model to enhance length generalization for standard attention based KT model. Specifically, our stableKT model is able to learn from short sequences and maintain stable performance when generalizing on long sequences. Furthermore, we design a multi-head aggregation module to effectively capture individual differences and complex hierarchical relationships between questions and their associated KCs. Experimental results on three real-world educational datasets demonstrate that our stableKT model has the capability of length generalization and outperforms a wide range of state-of-the-art DLKT models in terms of AUC.

Acknowledgements

This work was supported in part by National Key R&D Program of China, under Grant No. 2022YFC3303600 and in part by Key Laboratory of Smart Education of Guangdong Higher Education Institutes, Jinan University (2022LSYS003).

References

[Chi *et al.*, 2023] Ta-Chung Chi, Ting-Han Fan, Alexander Rudnicky, and Peter Ramadge. Dissecting transformer length extrapolation via the lens of receptive field analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, Toronto, Canada, July 2023.

[Choi *et al.*, 2020] Youngduck Choi, Youngnam Lee, Junghyun Cho, Jineon Baek, Byungsoo Kim, Yeongmin Cha, Dongmin Shin, Chan Bae, and Jaewe Heo. Towards an appropriate query, key, and value computation for knowledge tracing. In *Proceedings of the 7th ACM Conference on Learning At Scale*, Virtual Conference, August 2020.

[Cui *et al.*, 2023] Jiajun Cui, Zeyuan Chen, Aimin Zhou, Jianyong Wang, and Wei Zhang. Fine-grained interaction modeling with multi-relational transformer for knowledge tracing. *ACM Transactions on Information Systems*, 41(4):1–26, January 2023.

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

[Gulcehre *et al.*, 2019] Caglar Gulcehre, Misha Denil, Mateusz Malinowski, Ali Razavi, Razvan Pascanu, Karl Moritz Hermann, Peter Battaglia, Victor Bapst, David Raposo, Adam Santoro, et al. Hyperbolic attention networks. In *Proceedings of the 7th International Conference on Learning Representations*, New Orleans, LA, USA, April 2019.

[Guo *et al.*, 2021] Xiaopeng Guo, Zhijie Huang, Jie Gao, Mingyu Shang, Maojing Shu, and Jun Sun. Enhancing knowledge tracing via adversarial training. In *Proceedings of the 29th ACM International Conference on Multimedia*, Virtual Conference, October 2021.

[Huang *et al.*, 2023] Shuyan Huang, Zitao Liu, Xiangyu Zhao, Weiqi Luo, and Jian Weng. 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*, Taipei, Taiwan, China, July 2023.

[Im *et al.*, 2023] Yoonjin Im, Eunseong Choi, Heejin Kook, and Jongwuk Lee. Forgetting-aware linear bias for attentive knowledge tracing. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, Birmingham, UK, October 2023.

[Jacob *et al.*, 2019] Devlin Jacob, Chang Ming-Wei, Lee Kenton, and Toutanova Kristina. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, MN, USA, June 2019.

[Li *et al.*, 2023] Guanxin Li, Jingang Shi, Yuan Zong, Fei Wang, Tian Wang, and Yihong Gong. Learning attention from attention: efficient self-refinement transformer for face super-resolution. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence*, Macao, SAR, China, August 2023.

[Liu *et al.*, 2022] Zitao Liu, Qiongqiong Liu, Jiahao Chen, Shuyan Huang, Jiliang Tang, and Weiqi Luo. pykt: A python library to benchmark deep learning based knowledge tracing models. In *Proceedings of 36th Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, New Orleans, LA, USA, November 2022.

[Liu *et al.*, 2023a] Zitao Liu, Qiongqiong Liu, Jiahao Chen, Shuyan Huang, Boyu Gao, Weiqi Luo, and Jian Weng. Enhancing deep knowledge tracing with auxiliary tasks. In *Proceedings of the ACM Web Conference 2023*, Austin, TX, USA, April 2023.

[Liu *et al.*, 2023b] Zitao Liu, Qiongqiong Liu, Jiahao Chen, Shuyan Huang, and Weiqi Luo. simpleKT: A simple but tough-to-beat baseline for knowledge tracing. In *Proceedings of the 11th International Conference on Learning Representations*, Kigali, Rwanda, May 2023.

[Nakagawa *et al.*, 2019] Hiromi Nakagawa, Yusuke Iwasawa, and Yutaka Matsuo. Graph-based knowledge tracing: Modeling student proficiency using graph neural network. In *Proceedings of the 2019 IEEE/WIC/ACM International Conference on Web Intelligence*, Thessaloniki, Greece, October 2019.

[Pan *et al.*, 2022] Zizheng Pan, Jianfei Cai, and Bohan Zhuang. Fast vision transformers with hilo attention. In *Proceedings of the 36th Conference on Neural Information Processing Systems*, New Orleans, LA, USA, November 2022.

[Pandey and Karypis, 2019] Shalini Pandey and George Karypis. A self-attentive model for knowledge tracing. In *Proceedings of the 12th International Conference on Educational Data Mining*, Montréal, Canada, July 2019.

[Piech *et al.*, 2015] Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J Guibas, and Jascha Sohl-Dickstein. Deep knowledge tracing. In *Proceedings of the 28th Conference on Neural Information Processing Systems*, Montreal, Quebec, Canada, December 2015.

[Press *et al.*, 2022] Ofir Press, Noah Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *Proceedings of the 10th International Conference on Learning Representations*, Virtual Conference, April 2022.

[Qin *et al.*, 2024] Zhen Qin, Yiran Zhong, and Hui Deng. Exploring transformer extrapolation. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, Vancouver, Canada, February 2024.

[Raffel *et al.*, 2020] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(140):1–67, February 2020.

[Shen *et al.*, 2021] Shuanghong Shen, Qi Liu, Enhong Chen, Zhenya Huang, Wei Huang, Yu Yin, Yu Su, and Shijin Wang. Learning process-consistent knowledge tracing. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, Virtual Conference, August 2021.

[Su *et al.*, 2021] Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021.

[Tseng *et al.*, 2023] Albert Tseng, Tao Yu, Toni J.B. Liu, and Christopher De Sa. Coneheads: Hierarchy aware attention. In *Proceedings of the 37th Conference on Neural Information Processing Systems*, New Orleans, LA, USA, December 2023.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st Conference on Neural Information Processing Systems*, Long Beach, CA, USA, December 2017.

[Wang *et al.*, 2023] Zhiwei Wang, Junlin Xian, Kangyi Liu, Xin Li, Qiang Li, and Xin Yang. Dual-view correlation hybrid attention network for robust holistic mammogram classification. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence*, Macao, SAR, China, August 2023.

[Yin *et al.*, 2023] Yu Yin, Le Dai, Zhenya Huang, Shuanghong Shen, Fei Wang, Qi Liu, Enhong Chen, and Xin Li. Tracing knowledge instead of patterns: Stable knowledge tracing with diagnostic transformer. In *Proceedings of the ACM Web Conference 2023*, Austin, TX, USA, April 2023.

[Yu *et al.*, 2023] Tao Yu, Toni JB Liu, Albert Tseng, and Christopher De Sa. Shadow cones: Unveiling partial orders in hyperbolic space. *arXiv preprint arXiv:2305.15215*, 2023.

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