研究与开发

SATKT：基于Transformer的稀疏注意力知识追踪模型

诸葛斌[1]，肖梦凡[1]，汪盈[1]，董黎刚[1]，洪金珠[2]，蒋献[1]
（1. 浙江工商大学信息与电子工程学院（萨塞克斯人工智能学院），浙江 杭州 310018；
2. 浙江工商大学统计与数据科学学院，浙江 杭州 310018）

摘　要：随着在线教育的迅速发展，准确建模学习者的知识状态并为其提供个性化支持已成为实现高效教学的关键。知识追踪旨在基于学生的历史答题数据，动态建模其潜在知识状态。近年来，尽管Transformer模型以其卓越的序列数据处理能力被引入知识追踪领域，但其在处理长序列时仍面临计算复杂度高、预测准确率有限等挑战。对此，提出一种融合稀疏注意力机制的Transformer知识追踪模型——SATKT（sparse attention Transformer knowledge tracing）模型。该模型通过问题知识嵌入模块，从题目和知识点两个层面全面刻画学生的知识状态演化过程；引入稀疏注意力机制，有选择性地关注关键交互信息；结合混合损失函数，实现多目标联合优化，进一步提升模型的收敛稳定性与预测精度。在4个知识追踪数据集上的实验结果表明，SATKT在AUC指标上相较现有主流模型平均提升约1.33%，在ACC与RMSE指标上也表现更优，展现出更高的预测准确率与更强的泛化能力。本研究为智慧教育场景下的个性化学习分析提供了一种有效且可扩展的新思路。

关键词：知识追踪；Transformer；智慧教育；稀疏注意力机制
中图分类号：TP301
文献标志码：A
doi: 10.11959/j.issn.1000−0801.DXKX250469

SATKT: a sparse attention knowledge tracing model based on Transformer

Zhuge Bin[1], Xiao Mengfan[1], Wang Ying[1], Dong Ligang[1], Hong Jinzhu[2], Jiang Xian[1]
1. School of Information and Electronic Engineering (Sussex Artificial Intelligence Institute), Zhejiang Gongshang University, Hangzhou 310018, China
2. School of Statistics and Data Science, Zhejiang Gongshang University, Hangzhou 310018, China

收稿日期：2025−07−22；**修回日期：**2026−01−08
通信作者：洪金珠，hongjz@zjgsu.edu.cn
基金项目：浙江省领雁项目（No.2025C02024）；浙江省教育厅一般项目（No.Y202558980）；桐乡通用人工智能研究院项目（No.TAGI2-A-2024-0005）；浙江省高等教育“十四五”第二批研究生教学改革项目（No.JGCG2024225）；2024年度浙江工商大学高等教育研究课题（No.Xgy2419）；浙江省大学生科技创新活动计划（新苗人才计划）项目(No.2025R408B075)
Foundation Items: Lingyan Project of Zhejiang Province (No. 2025C02024), General Research Project of Zhejiang Provincial Department of Education (No. Y202558980), Tongxiang General Artificial Intelligence Research Institute (No. TAGI2-A-2024-0005), Zhejiang Province Higher Education “14th Five-Year Plan” Second Batch of Postgraduate Teaching Reform Projects(No. JGCG2024225), 2024 Zhejiang Gongshang University Higher Education Research Project (No.Xgy2419), Zhejiang University Students’ Scientific and Technological Innovation Project (No. 2025R408B075)

Abstract: With the rapid development of online education, how to accurately model learners' knowledge status and provide personalized support has become a key part of realizing efficient teaching systems. Knowledge tracing aims to model students' potential knowledge status by using their historical answer data. In recent years, the Transformer model has been introduced into the field of knowledge tracing for its excellent sequence data processing capabilities, but its computational complexity is high when processing long sequences, and there is still room for further improvement in prediction accuracy. To address this, a Transformer-based knowledge tracing model integrating sparse attention mechanism, named SATKT (sparse attention Transformer knowledge tracing), was proposed. In this model, a question-knowledge embedding module was employed to comprehensively capture the evolution of students' knowledge states from both the question level and the knowledge concept level. A sparse attention mechanism was introduced to selectively focus on key interaction information. Furthermore, a hybrid loss function was incorporated to achieve multi-objective joint optimization, thereby further improving the convergence stability and prediction accuracy of the model. Experimental results on four knowledge tracing datasets show that SATKT achieves an average improvement of approximately 1.33% in AUC compared with existing mainstream models, and also performes better in terms of ACC and RMSE, demonstrating higher prediction accuracy and stronger generalization ability. This study provides an effective and scalable new approach for personalized learning analysis in smart education scenarios.

Key words: knowledge tracking, Transformer, smart education, sparse attention mechanism

0　引言

在智慧教育加速发展的背景下，信息技术正不断重塑传统教育模式，深刻推动教育范式的变革。尽管智慧教育展现出显著的变革潜力，但在实际落地过程中仍面临诸多挑战。知识追踪（knowledge tracing，KT）作为智慧教育系统中的关键组成部分，已广泛应用于学生建模、个性化推荐与智能反馈等核心任务中[1]。其基本目标是分析学生的历史学习行为，动态建模其潜在的知识状态，从而预测学生未来的作答表现。具体而言，给定一个学生的学习交互序列$X_0, X_1, \cdots, X_T$，学生的每次交互可以表示为(q_t, r_t)，其中，q_t表示学习者在时间步t回答的试题，$r_t \in \{0, 1\}$表示作答结果的正误。知识追踪示意图如图1所示，学习者依次作答五道题（q_1, q_2, q_3, q_4, q_5），其中前四道题的作答结果(r_1, r_2, r_3, r_4)已知，第五道题的作答结果q_5为预测目标。知识追踪模型利用学习者的历史作答记录，推断其当前对相关知识点的掌握情况，并预测其在q_5上的作答结果。通过知识追踪，教育系统能够动态评估学生的知识状态，从而提供个性化的学习建议与教学干预，以提升学习效果。随着在线教育平台的广泛应用与学习数据的激增，知识追踪模型不仅需要具备更高的预测精度，还应具备良好的可解释性、稳定性和可扩展性，才能在真实教育场景中发挥智能决策作用[2]。

图1　知识追踪示意图

早期的知识追踪方法，如贝叶斯知识追踪和深度知识追踪，在建模学生的知识演化过程中发挥了重要作用。然而，这些方法在捕捉复杂知识结构、建模长期依赖关系等方面存在明显不足。为克服上述局限，近年来Transformer架构被引入KT任务中。凭借其强大的并行计算能力和出色的全局依赖关系建模表现，Transformer在智慧教育任务中已取得显著进展[3]。具体而言，Transformer通过自注意力机制允许模型在序列的任意两点间直接建立关联，从而能够在全局视角优化

学习者知识状态的预测。然而，全注意力机制的主要缺点在于，它需要对序列中所有元素进行复杂计算，尤其在处理长序列时效率低下，并可能导致信息过载。

针对上述挑战，本文提出了一种新型的知识追踪模型——基于Transformer的稀疏注意力知识追踪（sparse attention Transformer knowledge tracing，SATKT）模型。该模型通过引入问题知识嵌入、稀疏注意力机制和混合损失函数算法，不仅在处理大规模学习者互动数据时更为高效，而且在预测学习者知识状态方面更为准确。

本文的主要创新点如下。

（1）问题知识嵌入：结合问题得分嵌入与知识参数嵌入，有效整合学习者的作答行为与问题知识点特征，从而有效捕捉学习者在问题和知识层面的动态状态变化。

（2）稀疏注意力机制：在Transformer的注意力机制中融入稀疏机制，通过设定的阈值选择最重要的注意力权重，实现高效稳定的知识状态估计和跟踪。

（3）融合损失函数：提出一种融合加权二元交叉熵损失（binary cross-entropy loss，BCE Loss）、知识一致性损失和正则化损失的新型损失函数，以优化模型训练效果并增强对序列数据的学习能力。

（4）广泛的实证评估：在4个开源知识追踪数据集上对本文所提模型进行了评估，并与其他主流模型进行对比，结果显示本文所提模型在多项评价指标上均有所提升。

1 相关工作

知识追踪作为个性化教育的关键任务之一，在教育心理学和教育数据挖掘领域得到了广泛关注[4]。知识追踪通过分析学生的历史学习互动数据，刻画学生潜在知识状态的动态变化过程，并在此基础上预测其未来的学习表现[5-6]。

1.1 传统知识追踪模型

早期研究以贝叶斯知识追踪（Bayesian knowledge tracing，BKT）[7]和项目反应理论（item response theory，IRT）模型[8]为代表。BKT将学习状态抽象为是否掌握的隐变量，而IRT通过项目特征曲线建立学生能力与题目属性之间的函数关系。然而，传统模型在跨题差异、长时依赖和个体化刻画方面存在限制，难以充分反映实际学习过程的复杂性。

1.2 基于深度学习的知识追踪模型

随着深度学习的发展，Piech等[9]提出的深度知识追踪（DKT）模型首次利用循环神经网络对学生的历史交互序列进行建模，显著提高了知识掌握状态预测的准确率。此后，大量研究在DKT的基础上引入记忆机制、注意力机制和时间建模等结构，以增强模型的表征能力和个性化程度[10]。Zhang等[11]提出的DKVMN（dynamic key-value memory network）模型利用动态键值来存储和更新学生的知识掌握状态；Jin等[12]提出的SKVMN（sequential key-value memory network）模型在DKVMN的基础上引入序列处理机制，能够捕捉学生在学习过程中的序列信息。QIKT[13]模型则以问题中心的认知表示，结合交互编码器、知识获取模块和问题解决模块，综合建模学生的知识状态和学习过程。DKT+forget[14]模型在DKT的基础上引入遗忘概念，通过建模学生个体的遗忘曲线来模拟知识随时间的衰减；KVFKT[15]模型将基于注意力机制的知识概念嵌入与基于时间间隔和遗忘曲线的遗忘量化模块相结合，增强了模型的可解释性和准确性。STHKT[16]模型采用三元异构图和霍克斯过程（Hawkes process）来捕捉概念、问题和学生之间的时空依赖关系，从而实现更精准的知识追踪。尽管这些模型在不同层面提升了知识追踪的性能，但仍存在一些不足。例如，基于循环神经网络的模型在建模长序列依赖时易出现梯度消失或信息遗失问题，限制

了模型对长时学习行为的捕捉能力。

1.3　基于 Transformer 的知识追踪模型

Transformer 是一种基于自注意力机制的序列建模架构，由 Vaswani 等人于 2017 年提出[17]。与传统的循环神经网络相比，Transformer 完全抛弃了循环结构，转而借助多头注意力机制，在全局范围内捕捉序列中任意位置之间的依赖关系，从而具备更强的并行计算能力与长距离依赖建模性能。

受 Transformer 架构的启发，注意力机制被引入知识追踪领域，使模型能够更好地捕捉学生历史交互中的关键信息。Pandey 等[18]提出的 SAKT 模型利用自注意力机制建模学生历史交互中的长期依赖关系，提升了未来答题表现的预测性能。Choi 等[19]提出的分离式自注意力神经知识追踪模型——SAINT，将题目与学生响应分别作为输入进行建模，从而提升了模型性能。Shin 等[20]进一步提出了 SAINT+模型，通过引入题目难度、学生答题时长等辅助特征，丰富了学生–题目交互的表示。Yin 等[21]提出的 DTransformer 模型结合对比学习算法，提升了知识状态诊断的稳定性，降低了对特定学习模式的敏感性。Zhan 等[22]提出的 Bi-CL4KT 模型基于双向 Transformer 和对比学习机制，借助 Cloze 任务和数据增强提升知识追踪的表征能力和预测精度。尽管基于 Transformer 的知识追踪模型在捕捉长时依赖与并行计算方面表现优异，但其全注意力机制在长序列场景中仍存在计算开销大、注意力权重分散等局限，不利于实现对学习行为的高效与可解释建模。针对上述不足，本文提出一种基于 Transformer 的稀疏注意力知识追踪模型（SATKT），以实现更高效、精准的知识状态建模。

2　SATKT 模型

本文提出的SATKT模型主要由以下4个组件——输入模块、问题知识嵌入模块、Transformer 层、输出模块构成，其架构如图 2 所示。输入模块接收原始学习行为序列，包括题目编号、作答得分及知识点标签等多模态数据，并对这些信息进行标准化处理，以转化为模型可解析的格式。问题知识嵌入模块采用双通道嵌入策略，将题目编号与得分序列转换为高维表示；同时引入知识参数嵌入，以增强对知识层次特征的建模能力。Transformer 层基于多头注意力机制，分别对问题层和知识层的序列特征进行建模。其中，知识层引入稀疏注意力机制，仅保留高权重的注意力连接，从而显著降低计算复杂度，同时确保对关键依赖关系的有效捕捉。输出模块将问题层和知识

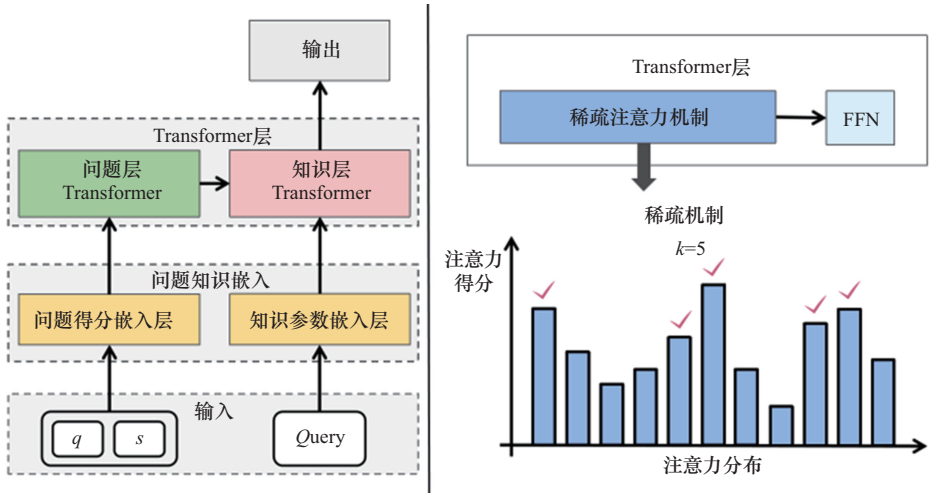


图 2　SATKT 模型架构

层的特征进行融合，并通过多层前馈网络对融合特征进行非线性变换，最终经由Sigmoid函数生成学生对下一题目的作答正确概率预测。模型通过优化由加权二元交叉熵损失、知识一致性损失和正则化损失组成的混合损失函数，实现了对学生动态知识状态的高效、鲁棒建模。

2.1 问题知识嵌入

问题知识嵌入模块充分考虑了教育场景中题目特征与知识状态的异构性，其核心任务是采用可学习的嵌入矩阵，将题目编号映射至连续向量空间，并利用线性变换将作答编码为与题目嵌入维度匹配的修正向量，进而将学习者的历史作答行为与知识点特征转换为统一的嵌入表示，为后续的序列建模和知识状态诊断提供基础。该模块由两部分构成：问题与得分嵌入层（question & score embedding layer）和知识参数嵌入层（knowledge parameter embedding layer）。二者协同作用，能够动态捕捉学习者在不同题目和知识点上的掌握情况，进而支持个性化的学习分析。

问题与得分嵌入层的作用是将问题序列与学习者的回答得分转换为高维的嵌入表示。问题嵌入的表达式如式（1）所示。

$$q_{\text{emb}} = \text{Embedding}(q) \tag{1}$$

其中，q表示问题编号，q_{emb}是问题映射到高维空间的嵌入表示。

结合回答得分，问题与得分嵌入层能够进一步提升对学生知识点掌握情况的建模精度。问题与得分嵌入的表达式如式（2）所示。

$$s_{\text{emb}} = \text{Embedding}(s) + q_{\text{emb}} \tag{2}$$

其中，s表示问题编号，s_{emb}是结合问题与得分的嵌入表示。

知识参数嵌入层通过初始化一组固定的知识点向量，将知识点嵌入模型，并结合问题嵌入表示诊断学生对各个知识点的掌握程度。该层的关键任务是通过捕捉学生在不同知识点上的掌握情况，帮助模型更准确地进行个性化学习分析。知识点嵌入层的输入是一个知识点的向量表示，而该向量被作为查询输入到知识级注意力计算中。知识点参数和知识层的掌握表达式如式（3）所示。

$$Z_{\text{knowledge}} = \text{Attention}\left(K_{\text{params}}, q_{\text{emb}}, q_{\text{emb}}\right) \tag{3}$$

其中，$Z_{\text{knowledge}}$表示学习者对知识点的掌握程度，K_{params}表示知识点的参数向量。

通过融合问题嵌入与知识参数向量，模型能够实现对学习者知识点掌握情况的全面建模。

问题与得分嵌入的表示s_{emb}和学习者对知识点的掌握程度表示$Z_{\text{knowledge}}$共同作为后续模块的输入，用于生成针对学习者下一答题行为的预测结果。该问题与知识嵌入模块借助上述两个部分的紧密结合，全面捕捉学习者对问题和知识点的掌握情况。

2.2 稀疏注意力机制

注意力机制是SATKT模型中的核心组成部分，能够有效捕捉序列数据中的依赖关系。模型中的多头自注意力机制通过计算查询（Query）、键（Key）和值（Value）之间的相似性来获得注意力得分，从而对输入序列进行加权求和，以捕捉不同位置之间的相关性。

多头注意力机制并行计算多个注意力头的结果并进行拼接，使模型在每个时间步能够计算其与序列中所有其他时间步的注意力权重，从而实现跨步信息聚合，更好地捕捉序列中不同位置间的依赖关系。

为进一步提升长序列建模的效率，SATKT在多头注意力机制中引入了稀疏注意力机制。该机制通过稀疏化处理，仅保留每个时间步得分最高的部分注意力连接，从而显著减少计算量。具体而言，稀疏注意力机制会筛选注意力得分，只保留与当前时间步最相关的前k个交互，而不是与所有时间步进行计算。

在模型计算复杂度方面，标准的全注意力机

制要求每个时间步与所有其他时间步进行交互。这意味着，对于序列长度为n，嵌入维度为d的输入，其计算注意力分数的计算复杂度为$O(n^2)$，注意力输出与键和值的矩阵乘法复杂度是$O(n^2\times d)$。在引入稀疏注意力后，模型只选择每个时间步相关的k个重要交互，而不是所有的n个交互。因此，稀疏注意力的计算复杂度从$O(n^2)$降至$O(k\times n)$，整体复杂度变为$O(k\times n\times d)$，显著提升了长序列场景下的计算效率。

其注意力得分计算表达式如式（4）所示。

$$S(\boldsymbol{Q},\boldsymbol{K})=\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}} \tag{4}$$

其中，$\boldsymbol{Q}$、$\boldsymbol{K}$分别表示查询（Query）、键（Key）矩阵，由输入特征经过线性映射得到。$\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}$表示查询矩阵与键矩阵转置后的点积，用于计算不同位置之间的相关性，T表示矩阵转置。d_k表示键向量（Key）的维度。

为筛选注意力得分并仅保留最显著的部分，采用Top$k$稀疏策略。具体做法是先对注意力得分进行排序，选择每个查询位置对应的前k个最高得分，其计算过程如式（5）所示。

$$S^{(k)}=\mathrm{Top}k(\boldsymbol{S},k) \tag{5}$$

其中，$\boldsymbol{S}$表示所有注意力得分部分的候选集合，Top$k(\cdot)$为选取前k个最大值的运算，k为预设的超参数。

在Top$k$函数中，通过设置阈值将非前k部分的得分设为负无穷大，从而在后续计算中屏蔽这些不重要位置的影响，如式（6）所示。

$$\tilde{S}_{i,j}=\begin{cases}S_{i,j}, & S_{i,j}\in\boldsymbol{S}_{i,j}^{(k)}\\ -\infty, & \text{其他}\end{cases} \tag{6}$$

其中，$S_{i,j}$表示第i个查询对第j个键的注意力得分，$\boldsymbol{S}_{i,j}^{(k)}$为对应查询的前$k$个最大得分组成的集合。

接着，对稀疏化后的得分矩阵$\boldsymbol{S}_{i,j}^{(k)}$进行softmax归一化，再与值矩阵$\boldsymbol{V}$相乘，得到最终的注意力输出，如式（7）所示。

$$O=\mathrm{softmax}(\boldsymbol{S}_{i,j}^{(k)})\cdot\boldsymbol{V} \tag{7}$$

其中，$\boldsymbol{V}$表示对应的值向量矩阵。

这种稀疏机制能够选择性地聚焦于重要部分，在提升模型预测精度的同时，显著增强了长序列处理能力，使其更适应大规模数据集下的高效建模需求。

2.3　混合损失函数

为应对教育数据中普遍存在的类别不平衡问题，并提高模型的泛化能力，本文提出了一种混合损失函数联合优化框架。该损失函数结合了加权二元交叉熵损失和知识一致性损失。加权二元交叉熵损失主要用于处理类别不平衡问题，而知识一致性损失则有助于模型在多样化的学习场景中保持知识表示的稳定性，从而提高模型的泛化性能。

（1）加权二元交叉熵损失

针对类别分布不均的情况，加权二元交叉熵损失通过给每个样本分配不同的权重来减少模型对多数类别的偏倚。其表达式如式（8）所示。

$$L_{\mathrm{BCE}}=-\frac{1}{N}\sum_{i=1}^{N}w_i\left[y_i\log\left(\hat{y}_i\right)+\left(1-y_i\right)\log\left(1-\hat{y}_i\right)\right] \tag{8}$$

其中，w_i为第i个样本的权重，y_i和$\hat{y}_i$分别为真实标签和预测值。

通过调节权重，加权二元交叉熵损失能够有效平衡不同类别样本在训练中的贡献，避免模型偏向多数类别，进而提高模型在不平衡数据集上的表现。

在具体实现中，正样本（$y_i=1$）的权重设置为1，负样本权重设置为一个可调的参数w_-。通过增加w_-，模型在训练过程中会更加关注负样本，从而缓解类别不平衡对模型训练效果的影响。本研究在训练过程中对w_-进行了网格搜索（候选范围为[1.0, 1.5]），以确定权重。当$w_-=1.2$时，模型在训练稳定性和测试性能之间取得了较好的平衡。因此，在后续实验中固定采用该值。

（2）知识一致性损失

知识一致性损失旨在约束模型学习到的知识表示在不同的时间步之间保持一致性。就学生作答场景数据而言，学生在某一时间点对某个知识点的掌握状态应与后续时间点的掌握状态一致。为度量知识表示之间的一致性，知识一致性损失通过计算特征向量之间的余弦相似性来评估不同时间步的知识状态一致性：

$$L_{\text{consistency}}=\frac{\lambda}{B\times T}\sum_{i=1}^{B\times T}\left(1-\cos\left(z_i,z_j\right)\right)^2 \quad (9)$$

其中，B和T分别表示批量大小和序列长度，z_i和z_j分别表示不同时间步的特征向量，λ为一致性损失的权重系数。

知识一致性损失通过鼓励知识表示一致性，能够有效防止过拟合，提升模型泛化能力。

模型的总损失函数由加权二元交叉熵损失、知识一致性损失和正则化损失三部分组成：

$$L_{\text{total}}=L_{\text{BCE}}+L_{\text{consistency}}+\lambda_{\text{reg}}L_{\text{reg}} \quad (10)$$

其中，L_{reg}为正则化损失，λ_{reg}为正则化损失得的权重。

正则化损失有助于防止模型过拟合，而加权二元交叉熵损失与知识一致性损失则分别用于提高模型的分类准确性和增强模型的知识表示一致性。

通过上述混合损失函数的设计，SATKT模型能够有效地处理教育数据中类别不平衡与表示一致性问题，提升模型在不同任务和数据集上的预测表现，并增强模型的泛化能力。

3 实验设计与结果分析

3.1 数据集

本研究在4个真实的在线教育数据集进行仿真实验，包括3个知识追踪领域的经典公开数据集及本研究团队自主研发的豆豆云助教小程序中的“思想道德修养与法律基础”课程数据集。数据集信息见表1，包括学生数、题目数、交互记录数。

（1）assist09：源于2009—2010年的Assistment在线辅导系统，包含4 217名学习者的活动记录、26 688道数学题目和123个知识点，总计346 860条交互记录。平均每位学习者的交互序列长度为82.25。

表1 数据集信息

数据集名称	学生数/名	题目数/个	交互记录数/条
assist09	4 217	26 688	346 860
algebra05	574	1 084	809 694
statics	331	1 223	142 124
doudouyun	955	1 118	345 517

（2）algebra05：来自KDD Cup 2010教育数据挖掘挑战赛，涵盖2005—2006年间中学生在代数课程上的答题情况。该数据集共包括574名学习者、1 084个题目、138个知识点，总共809 694条交互记录，平均每名学习者的交互序列长度达1 410.62。

（3）statics：来自卡耐基梅隆大学2011年秋季学期的工程力学课程，记录了学生在该课程中的答题数据。它包含331名学习者、1 223个题目、154个知识点，共142 124条交互记录，平均每名学习者的交互序列长度约为429.38。

（4）doudouyun：采集自本研究团队自主研发的智慧在线教育系统豆豆云助教小程序的做题记录，数据库中有两千多万条历史学习数据。本文采用大学本科阶段“思想道德修养与法律基础”课程的大学生答题数据，包含955名学生者、345 517条交互记录，平均每名学习者的交互序列长度约为361.8。

为深入理解模型在答题行为中处理正负样本的能力，对划分后的训练集与测试集分别进行了统计，计算了各个数据集中答题结果为“正确”（标签为1，即正样本）与“错误”（标签为0，即负样本）的比例。各数据集正样本比例见表2，为模型在处理数据不平衡问题提供了直观依据。

表2　各数据集正样本比例

数据集	训练集正样本比例	测试集正样本比例
assist09	61.21%	64.06%
algebra05	79.08%	84.29%
statics	70.20%	70.51%
doudouyun	49.86%	49.75%

3.2　实验细节

3.2.1　实验环境及参数设置

实验环境操作系统为Ubuntu 22.04.4 LTS，编程语言为python3.8.20，模型实现基于PyTorch1.9.0和CUDA 10.2。SATKT模型在4个数据集下的实验参数设置见表3，其中包含了嵌入维度、训练轮次、学习率等在内的主要参数的取值。为防止模型在训练过程中发生过拟合，本实验采用早停机制，其耐心值（early_stop）设为10，即当验证集性能在连续10个训练周期中未显著提升时，提前终止训练。此举有助于节省计算资源并增强模型泛化能力。

为进一步提高训练效率，针对交互序列长度不一致的问题，本文引入了基于固定窗口长度的序列切分机制。通过设置参数seq_len，将原始不等长的交互序列划分为若干长度固定的子序列，使模型在训练过程中可以以固定序列长度进行批量处理，从而避免序列长度差异导致的显存占用不均或模型学习效率降低。

表3　实验参数

参数名称	参数含义	参数值
batch size	每次训练迭代中使用的样本数量	16
d_model	模型中嵌入向量的维度	128
dropout	在训练过程中应用的dropout概率	0.3
learning_rate	学习率	0.001
n_epochs	训练过程中完整通过训练集的次数	100
n_heads	多头注意力机制中的注意力头数量	8
n_layers	模型中的Transformer层数	3
early_stop	早停机制的耐心值	10

3.2.2　评估指标

为评估SATKT模型的有效性，实验采用曲线下面积（area under the curve，AUC）、准确率（accuracy，ACC）和均方根误差（root mean squared error，RMSE）3个指标进行性能评估。

（1）AUC

AUC作为二分类模型性能评估的核心指标，通过计算受试者工作特征曲线（ROC曲线）与横轴围成的面积，综合评价模型在不同分类阈值下对正负样本的判别能力。ROC曲线以真阳性率（true positive rate，TPR）为纵轴、假阳性率（false positive rate，FPR）为横轴，反映分类器在灵敏度与特异性间的权衡关系。AUC的取值范围严格限定于[0.5,1]区间，其中AUC=0.5表示模型没有分类能力，与随机猜测相当；AUC=1则表示模型有完美的分类能力。该指标的优势在于其不受分类阈值影响，能够消除决策边界选择带来的偏差，从而提供对模型全局判别效力的稳健评估，尤其在类别分布不平衡的场景中展现出显著的应用价值。

（2）ACC

准确率定义为模型正确预测样本数与总体样本量的比值，其数学表达如式（11）所示。

$$\mathrm{ACC}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \tag{11}$$

其中，TP为真正例（true positives），TN为真反例（true negatives），FP为假正例（false positives），FN为假反例（false negatives）。ACC的取值在0到1之间，值越大表示模型预测的准确性越高。然而，该指标在类别分布严重偏斜的数据集中可能产生误导性结论，如在负样本占主导的数据集中，模型可能仅预测负样本即可获得较高的ACC值，而无法有效识别正样本。因此，需要结合AUC等多指标构建多维评估体系。

（3）RMSE

RMSE作为回归任务的核心评估指标，通过

计算预测值与真实值间误差的均方根值，量化模型对连续变量的预测偏差。RMSE的值越小，表示模型的预测误差越小，预测性能越好。RMSE对预测误差较大的样本更加敏感，其更适用于对预测精度要求较高的场景。其表达式如式（12）所示。

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2} \tag{12}$$

其中，y_i为真实值，$\hat{y}_i$为预测值，n为样本数量。

通过综合使用以上评估指标，可以从分类判别能力、预测准确性及误差可控制等多个维度评估本文提出模型在不同方面的性能表现，从而验证模型的有效性和可靠性。

3.3 结果分析

3.3.1 对比实验

为验证SATKT模型在知识追踪任务中的性能，本节将其与5个基线模型进行全面比较，以全面评估SATKT在不同任务指标上的表现，并考察其在4类数据集上的适应能力。

实验选取了知识追踪领域的5个代表性模型作为基线模型：DKT模型利用循环神经网络建模学生作答序列中的时间依赖性，捕捉知识状态的动态变化；DKVMN模型引入键值对存储机制，通过动态记忆更新学生对知识点的掌握状态，有效处理非线性知识变化；AKT模型采用注意力机制，对学生作答行为中的重要部分进行加权，以解决长序列中的记忆遗忘问题；SAKT模型则利用Transformer架构的自注意力机制，实现全局序列建模，克服了传统循环神经网络在处理长期依赖时的限制，提升了建模效率和准确性；DTransformer是一个基于Transformer的知识追踪模型，通过时间与累积注意力机制估计题目掌握情况，进而诊断知识点熟练度，并结合对比学习增强知识状态的稳定性与一致性。

SATKT与5个基线模型（DKT、DKVMN、AKT、SAKT、DTransformer）在4个数据集上的平均AUC值见表4。从表4可以看出，SATKT在各数据集上均取得了最优的AUC表现：在assist09数据集上，相比次优模型提升了约2.28%；在algebra05数据集上，相比次优模型提升了0.93%；在statics数据集上，相比次优模型提升了1.40%；在doudouyun数据集上，相比次优模型提升了0.70%。这些结果表明，SATKT在不同规模和特性的教育数据集上均展现出稳健的判别能力与较强的适应性。在assist09数据集上，SATKT的提升幅度最大，取得了最优的AUC表现。这可能是因为该数据集中学生交互序列较短（平均约82条），模型在建模过程中更依赖于对不同学生之间知识差异的捕捉。SATKT基于问题知识嵌入构建细粒度的题目与知识点表示，并借助稀疏注意力机制过滤冗余交互，突出关键性信息，从而有效提升了区分学生答对与答错的能力。综合来看，SATKT在4个数据集上的AUC均优于现有主流模型，既能在短序列数据中展现出强判别力，也能在长序列、复杂知识结构和真实噪声环境下保持稳健表现，体现出良好的泛化能力与应用潜力。

SATKT与基线模型在ACC指标上的表现如图3所示。从整体来看，SATKT在多数数据集上取得了最优或接近最优的结果，体现了其较高的预测精度与稳定性。在assist09数据集上，SATKT

表4 SATKT与5个基线模型在4个数据集上的平均AUC值

数据集	DKT	DKVMN	AKT	SAKT	DTransformer	SATKT
assist09	0.790 8	0.789 1	0.778 3	0.793 9	0.801 4	0.819 7
algebra05	0.755 8	0.785 7	0.775 0	0.743 3	0.782 1	0.793 0
statics	0.818 2	0.826 5	0.793 8	0.818 6	0.827 3	0.838 9
doudouyun	0.694 1	0.698 4	0.641 6	0.740 8	0.741 3	0.746 5

的ACC明显高于所有基线模型，表明其在中等规模、短序列数据条件下能够更准确地刻画学生的知识状态。在algebra05数据集上，各模型的ACC表现非常接近（AUC值为0.82～0.85），但SATKT的结果略低于SAKT。这可能与该数据集正样本比例偏高（超过80%）有关，部分模型因倾向于预测多数类而在ACC指标上获得微弱的表面优势。然而，SATKT在该数据集的AUC指标上表现最优，说明其在类别分布不平衡的情况下仍然能够有效地区分“答对”和“答错”样本；同时其RMSE也保持较低水平，表明SATKT并未一味偏向多数类，而是借助复合损失函数缓解了不平衡带来的过拟合问题。在statics数据集上，SATKT与SAKT和DTransformer的ACC表现相当，并显著优于DKT和AKT。这表明在知识结构复杂、强调高阶认知的数据中，基于Transformer的全局建模能力和注意力机制的优势得到充分体现。在doudouyun数据集上，SATKT的ACC明显优于DKT、DKVMN和AKT，同时也优于SAKT和DTransformer，表明在真实教育场景、数据含有噪声和异质性的条件下，SATKT能够更好地过滤无效信息，从而保持较高的预测准确性。

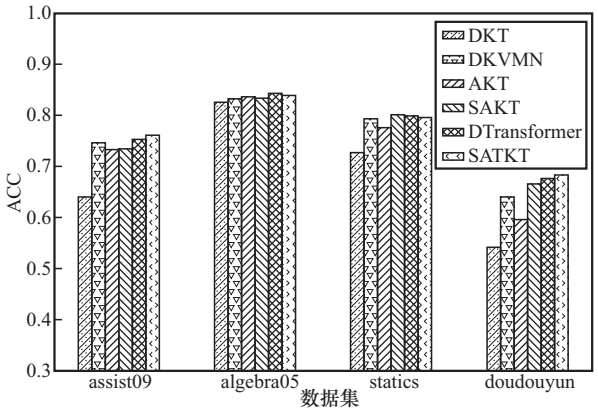


图3　SATKT与基线模型在ACC指标上的表现

SATKT与基线模型在RMSE指标上的表现如图4所示。整体来看，SATKT在所有数据集上均取得了最低或接近最低的RMSE值，表明其预测结果更接近真实值，误差更小，具有更高的稳定性与可靠性。

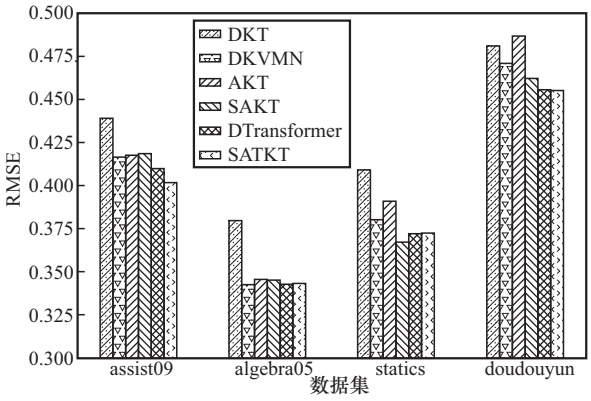


图4　SATKT与基线模型在RMSE指标上的表现

3.3.2　消融实验

为深入验证模型中各模块的有效性，本研究在4个数据集上对SATKT模型及其多个变体进行消融实验。所有实验均保持一致的超参数设置，以准确评估问题知识嵌入、稀疏注意力机制和复合损失函数对模型预测准确性和鲁棒性的贡献。实验设计如下。

（1）SATKT-N：作为基础模型，仅采用标准Transformer结构与BCE Loss，不包含任何优化模块，用于验证模型在无附加改进时的基准性能。

（2）SATKT-QK：该变体移除模型中的稀疏注意力机制和复合损失函数，仅保留问题知识嵌入模块，并基于标准Transformer结构和BCE Loss进行训练，以分析问题知识嵌入对模型性能的独立贡献。

（3）SATKT-SP：该变体保留稀疏注意力机制，但移除了知识嵌入模块和复合损失函数，仅使用标准的BCE Loss进行训练，以评估稀疏注意力机制对模型性能的独立贡献。

（4）SATKT-CL：该变体在SAKT模型中融入复合损失函数（包含知识一致性损失和正则化损失），以分析复合损失函数在模型性能提升中的效果。

消融实验AUC结果见表5。表5表明，与基准模型SATKT-N相比，3种单模块增强变体均带来不同程度的性能提升，说明问题知识嵌入、稀疏注意力机制和复合损失函数对模型性能均有独立贡献。完整的SATKT模型在assist09、algebra05、doudouyun 3个数据集上均取得最佳表现，验证了多模块协同设计的有效性。在statics数据集上，SATKT-CL的表现略优于完整模型。这可能与statics数据集题目数量多、知识结构复杂、需要更细致的特征保留有关。复合损失函数通过强化知识一致性与信息保持能力，在该场景下能够更好地保持知识一致性与细节信息，从而表现优于依赖稀疏注意力机制的完整模型。

表5　消融实验AUC结果

数据集	SATKT-N	SATKT-QK	SATKT-SP	SATKT-CL	SATKT
assist09	0.807 2	0.810 7	0.815 5	0.813 6	0.819 7
algebra05	0.779 1	0.784 7	0.790 4	0.792 3	0.793 0
statics	0.828 6	0.836 3	0.838 0	0.839 2	0.838 9
doudouyun	0.730 1	0.739 2	0.743 0	0.745 3	0.746 5

3.3.3　扩展实验

为深入探究稀疏注意力机制模块中引入的参数k对模型效果的影响，本文以algebra05数据集为代表，在训练过程中设置k依次取3到10的连续整数值，系统研究该参数对模型性能的影响。k对SATKT模型AUC效果影响如图5所示。

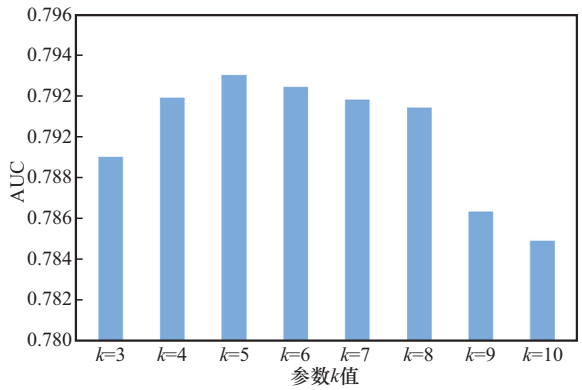


图5　参数k对SATKT模型AUC效果影响

从图5可以看出，当k值从3增至5，模型性能逐步提升，AUC从0.789 0上升至峰值0.793 5；此后继续增大k值，AUC逐渐下降，至k=10时，AUC降至0.785 0左右。该结果表明，较小的k虽有较低的上下文覆盖率，但能有效规避冗余信息干扰，增强局部关注的有效性；而过大的k值则可能引入噪声，导致注意力权重分散，从而降低模型判别能力。对algebra05数据集而言，k的值选取为5时效果最佳，此时稀疏注意力机制保留的关键邻近交互可实现高质量稀疏建模，有效平衡注意力聚焦与信息捕捉能力。

为进一步分析模型内部注意力机制的行为模式，本文对4个注意力头的注意权重进行了可视化分析，algebra05数据集某一批次样本中4个注意力头的注意力热图如图6所示，其横纵坐标分别对应序列中的题目索引，颜色深浅表示注意强度。由图6可以看出，所有热力图均呈现出明显的主对角线强化特征，说明模型主要关注当前题目及其邻近时间步，符合学生答题行为中短期依赖性与知识连贯性的认知规律。同时，部分注意力头中也出现远离对角线的高激活区域，反映出模型在某些情况下能捕捉非连续题目之间的长期依赖，这可能与知识点跳跃、概念迁移等学习行为相关。这表明模型通过多头注意力机制，不仅能够捕捉局部的细粒度时序变化，也能建模跨步长依赖关系，从而提升知识状态建模的完整性与灵活性。

综上所述，扩展实验表明，在引入稀疏注意力机制后，通过合理的参数配置，SATKT模型能够有效提升对学习序列中关键知识状态的关注能力，同时显著降低不必要的计算开销。

4　结束语

针对知识追踪领域在长序列数据中存在的预测准确度不足的问题，本研究提出了基于稀疏注意力机制的Transformer知识追踪模型SATKT。

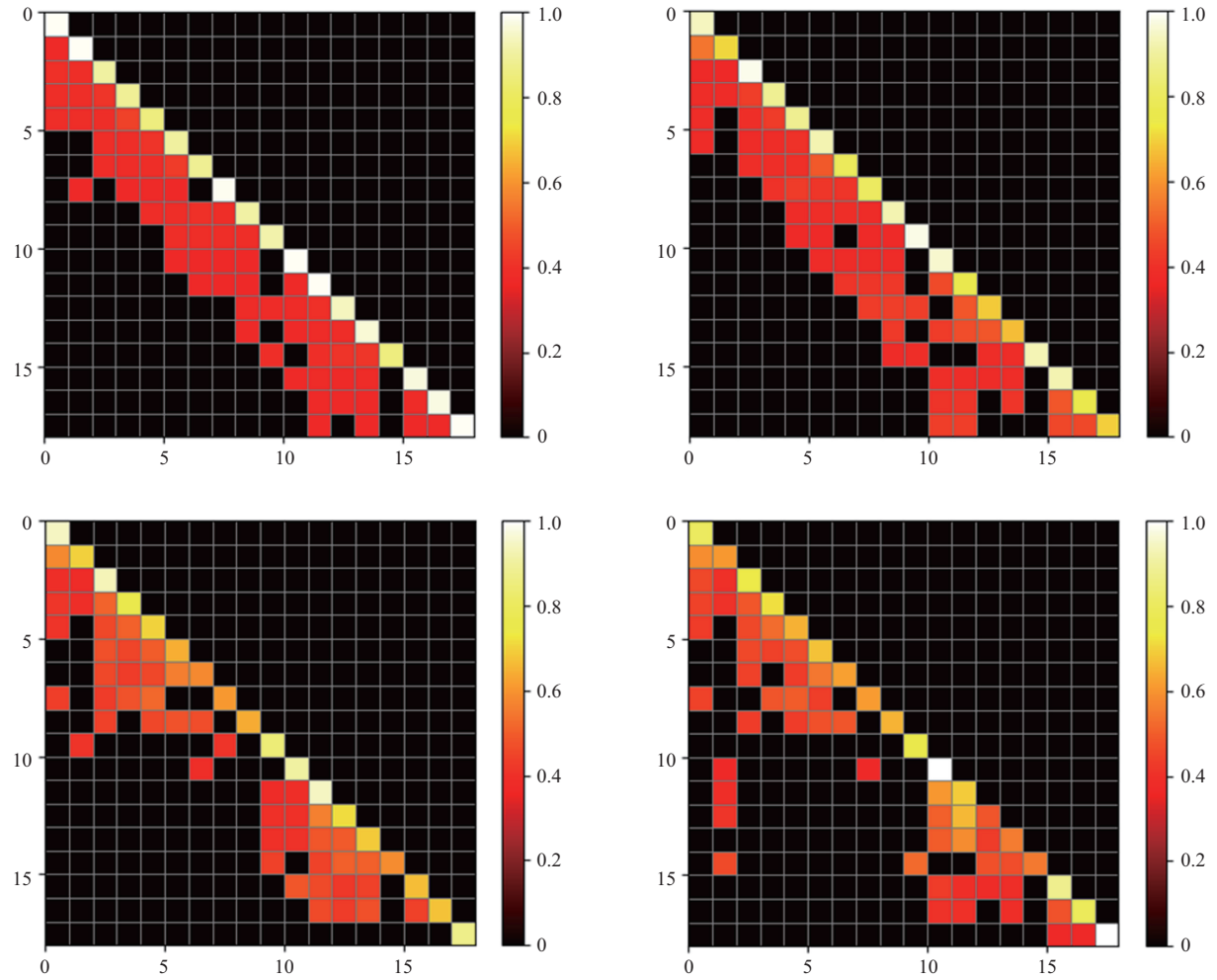


图6　序列中4个注意力头的注意力热图

该模型在自注意力机制的基础上，引入问题知识嵌入模块，以捕捉题目与知识点之间的层次化关系；设计了稀疏注意力机制，有选择性地保留最相关历史交互，减少冗余信息干扰；通过构建混合损失函数平衡模型收敛速度与预测稳定性，从而显著提升了模型的预测性能，也增强了模型的可扩展性和鲁棒性。实验评估结果表明，相较于现有主流模型，SATKT在各项指标上均取得了显著提升，充分验证了其在知识追踪领域的潜力与优越性，展现出在教育数据挖掘与智能教育系统中的广阔应用前景。

尽管如此，本研究仍然存在可进一步改进和拓展的空间。例如，在不同数据集或学习场景下，最优k值可能存在差异，未来可探索动态或自适应的稀疏机制，以降低对人工参数调整的依赖，提升模型在不同环境下的自适应性。

参考文献：

[1] Pirolli P, Kairam S. A knowledge-tracing model of learning from a social tagging system[J]. User Modeling and User-Adapted Interaction, 2013, 23(2): 139-168.

[2] Shen S H, Liu Q, Huang Z Y, et al. A survey of knowledge tracing: models, variants, and applications[J]. IEEE Transactions on Learning Technologies, 2024, 17: 1858-1879.

[3] Park S, Lee D, Park H. Enhancing knowledge tracing with concept map and response disentanglement[J]. Knowledge-Based Systems, 2024, 302: 112346.

[4] 刘淇, 陈恩红, 朱天宇, 等. 面向在线智慧学习的教育数据挖掘技术研究[J]. 模式识别与人工智能, 2018, 31(1): 77-90.

Liu Q, Chen E H, Zhu T Y, et al. Research on educational data mining for online intelligent learning[J]. Pattern Recognition and Artificial Intelligence, 2018, 31(1): 77-90.

[5] Ghosh A, Heffernan N, Lan A S. Context-aware attentive knowledge tracing[C]//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM Press, 2020: 2330-2339.

[6] Yeung C K, Yeung D Y. Addressing two problems in deep knowledge tracing via prediction-consistent regularization[C]// Proceedings of the Fifth Annual ACM Conference on Learning at Scale. New York: ACM Press, 2018: 1-10.

[7] Corbett A T, Anderson J R. Knowledge tracing: Modeling the acquisition of procedural knowledge[J]. User Modelling and User-Adapted Interaction, 1995, 4(4): 253-278.

[8] Embretson S E, Reise S P. Item response theory[M]. Hove: Psychology Press, 2000

[9] Piech C, Bassen J, Huang J, et al. Deep knowledge tracing[C]// Proceedings of the 29th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2015: 505-513.

[10] Abdelrahman G, Wang Q, Nunes B. Knowledge tracing: a survey[J]. ACM Computing Surveys, 2023, 55(11): 1-37.

[11] Zhang J N, Shi X J, King I, et al. Dynamic key-value memory networks for knowledge tracing[C]//Proceedings of the 26th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2017: 765-774.

[12] Abdelrahman G, Wang Q. Knowledge tracing with sequential key-value memory networks[C]//Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2019: 175-184.

[13] Chen J H, Liu Z T, Huang S Y, et al. Improving interpretability of deep sequential knowledge tracing models with question-centric cognitive representations[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2023, 37(12): 14196-14204.

[14] Nagatani K, Zhang Q, Sato M, et al. Augmenting knowledge tracing by considering forgetting behavior[C]//Proceedings of the World Wide Web Conference. New York: ACM Press, 2019: 3101-3107.

[15] Guan Q, Duan X, Bian K, et al. KVFKT: a new horizon in knowledge tracing with attention-based embedding and forgetting curve integration[C]//Proceedings of the 31st International Conference on Computational Linguistics. Abu Dhabi：Association for Computational Linguistics, 2025: 4399-4409.

[16] Li S T, Shen S H, Su Y, et al. STHKT: spatiotemporal knowledge tracing with topological hawkes process[J]. Expert Systems with Applications, 2025, 259: 125248.

[17] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017). New York: ACM Press, 2017: 6000-6010.

[18] Pandey S, Karypis G. A self-attentive model for knowledge tracing[C]//Proceedings of the 12th International Conference on Educational Data Mining (EDM 2019). Montreal, Canada: International Educational Data Mining Society, 2019: 384-389.

[19] Choi Y, Lee Y, Cho J, et al. Towards an appropriate query, key, and value computation for knowledge tracing[C]//Proceedings of the Seventh ACM Conference on Learning @ Scale. New York: ACM Press, 2020: 341-344.

[20] Shin D, Shim Y, Yu H, et al. SAINT+: integrating temporal features for EdNet correctness prediction[C]//Proceedings of the LAK21: 11th International Learning Analytics and Knowledge Conference. New York: ACM Press, 2021: 490-496.

[21] Yin Y, Dai L, Huang Z Y, et al. Tracing knowledge instead of patterns: stable knowledge tracing with diagnostic transformer[C]// Proceedings of the ACM Web Conference 2023. New York: ACM Press, 2023: 855-864.

[22] Zhan H J, Kim J J, Liu G M. Contrastive learning with bidirectional transformers for knowledge tracing[C]//Proceedings of the ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2024: 5040-5044.

[作者简介]

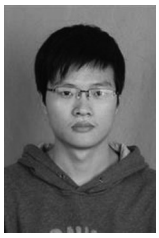
诸葛斌（1976–），男，博士，浙江工商大学信息与电子工程学院教授，主要研究方向为网络和通信技术、互联网技术和网络安全。

肖梦凡（2001–），女，浙江工商大学信息与电子工程学院硕士生，主要研究方向为智慧教育和个性化推荐。

汪盈（2000– ），女，浙江工商大学信息与电子工程学院硕士生，主要研究方向为智慧教育和个性化推荐。

洪金珠（1978– ），女，浙江工商大学统计与数据科学学院讲师，主要研究方向为在线教育、大数据。

董黎刚（1972– ），男，博士，浙江工商大学信息与电子工程学院教授，主要研究方向为智能网络、在线教育。



蒋献（1988– ），男，浙江工商大学信息与电子工程学院讲师、实验员，主要研究方向为在线教育。