Front. Digit. Educ., 2025, 2(4): 34
https://doi.org/10.1007/s44366-025-0071-x

RESEARCH ARTICLE

Explainable Few-Shot Knowledge Tracing

Haoxuan Li[a‡], Jifan Yu[b‡], Yuanxin Ouyang[a], Zhuang Liu[c], Wenge Rong[a], Huiqin Liu[b], Juanzi Li[d], Zhang Xiong[a]

[a] National Research Center for Educational Materials, School of Computer Science and Engineering, Beihang University, Beijing 100191, China

[b] School of Education, Tsinghua University, Beijing 100084, China

[c] Smart City College, Beijing Union University, Beijing 100101, China

[d] Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

Abstract Knowledge tracing (KT), aiming at mining students' mastery of knowledge by their exercise records and predicting their performance on future test questions, is a critical task in educational assessment. While researchers achieve tremendous success with the rapid development of deep learning techniques, current KT tasks fall into the cracks from real-world teaching scenarios. Relying on extensive student data heavily and predicting numerical performances solely differ from the settings where teachers assess students' knowledge state from limited practices and provide explanatory feedback. To fill this gap, this study explores a new task formulation, namely, explainable few-shot KT. By leveraging the powerful reasoning and generation abilities of large language models (LLMs), this study then proposes a cognition-guided framework that can track students' knowledge from a few students' records while providing natural language explanations. Experimental results from three widely used datasets show that LLMs can perform comparable or superior to competitive deep KT methods. Finally, this study discusses potential directions and calls for future improvements.

Keywords knowledge tracing, large language models, explainability, educational assessment

1 Introduction

Knowledge tracing (KT) is a well-established problem originated from educational assessment, aiming at modeling students' knowledge mastery dynamically and predicting their future learning performances (Corbett & Anderson, 1994). With the advancement of deep learning, models leveraging recurrent neural networks (RNNs) and attention mechanisms have gradually become mainstream for KT (Choi et al., 2020; Ghosh et al., 2020; Piech et al., 2015a). In recent years, research on KT has shown two notable and promising directions: On the one hand, researchers attempt to incorporate multiple types of side information (e.g., exercise texts (Liu et al., 2019), knowledge concept relationships and mappings (Nakagawa et al., 2019; Pandey & Srivastava, 2020), and students' problem-solving behaviors (Long et al., 2022)) with students' exercise history to more accurately model their knowledge states. On the other hand, students try to reveal the links between latent representations acquired by models and factual data to provide interpretability for KT models (Minn et al., 2022; Tong et al., 2020; Zhu et al., 2023).

Despite the numerous attempts and decent success, the current KT task leaves gaps in reflecting real-world scenarios where teachers evaluate students' knowledge states. First, it relies on extensive student's exercise records to train deep learning KT models to achieve remarkable performance. In contrast, in realistic scenarios, teachers can analyze students' mastery of knowledge from a limited number of practices. Second, unlike teachers can infer students' answers by analyzing and explaining their knowledge mastery, the current task is simplified to only predicting whether a student will answer the test questions correctly, mainly by deep learning sequential predictive models. The "black-box" nature of these models hampers exploring interpretability as they represent student knowledge states as hidden vectors. Apart from

Received February 7, 2025; revised May 15, 2025; accepted June 6, 2025

Yuanxin Ouyang (✉)
E-mail: oyyx@buaa.edu.cn

‡These authors contributed equally to this work and should be considered co-first authors.

the aforementioned gaps, KT models and frameworks encounter difficulties unifying and utilizing the multi-dimensional information collected from learning environments, such as student behaviors, question texts, and knowledge relations. Primarily proposed non-generative sequential models make it challenging for such tasks to come out of numerical prediction and extend to other scenarios, such as open-ended exercising and programming learning.

In recent years, the emergence and widespread utilization of large language models (LLMs) have provided potential solutions to fill the gaps. LLMs' capability to follow complex instructions with only a few examples and provide natural language feedback makes it possible to reform the current KT paradigm. Inspired by the success of LLMs in other fields, this study improves upon the existing KT task formulation and proposes an explainable few-shot KT. As illustrated in Figure 1, compared to traditional KT (TKT), explainable few-shot KT takes a small number of informative student exercise records as input, tracks students' mastery of knowledge, and predicts future performances through reasoning while providing reasonable explanations. Moreover, leveraging LLMs' strong reasoning and generation abilities, the KT task can readily adapt to diverse teaching scenarios with simple adjustments, presenting new opportunities to apply KT models. Three key contributions of this study are as follows: first, analysing the deficiencies of TKT and proposing an explainable few-shot KT task that aligns better with real teaching scenarios; second, introducing a cognition-guided framework that combines LLMs and educational assessment principles to practice explainable few-shot KT; third, adapting three public datasets and conducting experiments with open-source and closed-source LLMs. The results demonstrate that LLMs can perform superior to competitive KT models.

2 Background

2.1 | Knowledge Tracing

2.1.1 Educational Assessment and Bayesian Knowledge Tracing

Educational assessment aims at analyzing students' knowledge states, and an assessment system is generally considered to comprise three main components, including cognition, observation, and interpretation (Pellegrino et al., 2001). Cognition refers to a model of how students represent knowledge. With the introduction of KT concept (Corbett & Anderson, 1994), researchers have estimated students' knowledge states by analyzing their response records. The KT methods consist of two classics, Bayesian KT (BKT) and factor analysis models (Xiong et al., 2024). BKT is a Hidden Markov Model that treats each learner's knowledge state as a binary variable and utilizes Bayesian inference to update the state (Yudelson et al., 2013). In contrast, factor analysis models aim at learning generalized parameters from historical data (Yen & Fitzpatrick, 2006).

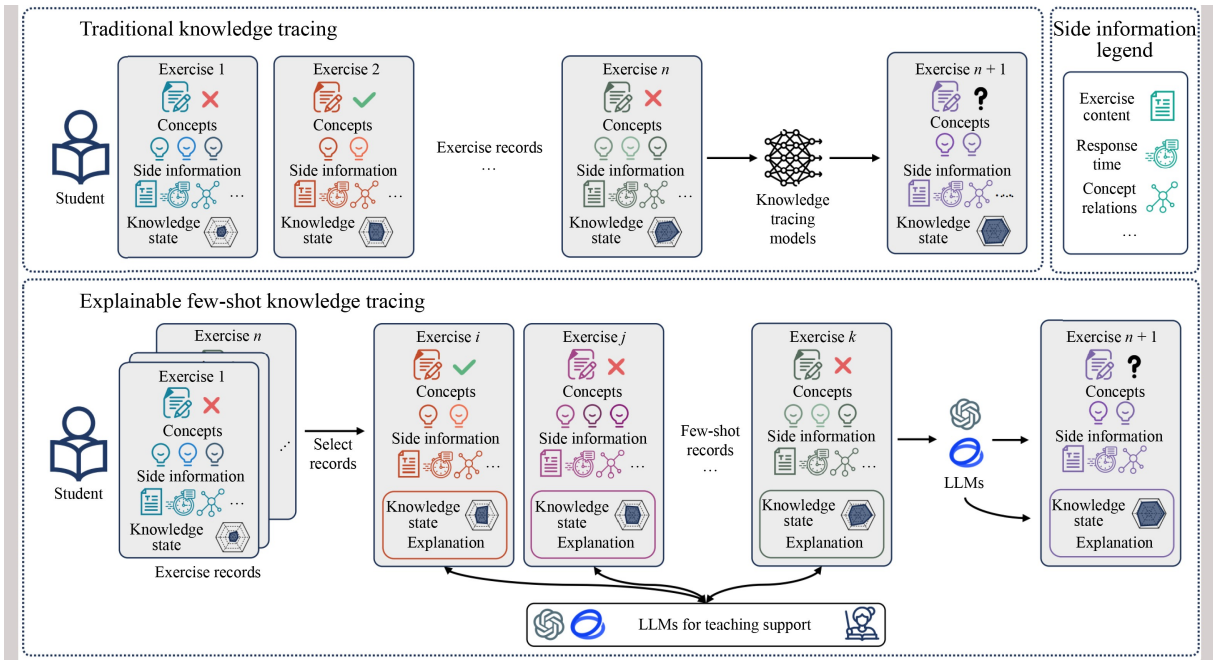

Figure 1 Traditional knowledge tracing and explainable few-shot knowledge tracing. LLMs: large language models.

2.1.2 Deep Knowledge Tracing

Recently, numerous researchers have integrated deep neural networks into KT tasks for their effectiveness and outstanding performance. Piech et al. (2015a) pioneered deep learning for KT using RNNs to process interaction sequences over time and proposed the deep KT task setting. With the success of deep learning techniques (e.g., Word2Vec (Mikolov et al., 2013) and graph neural networks (Kipf & Welling, 2017; Veličković et al., 2017)) in other domains, researchers recognized the potential to leverage these techniques by incorporating auxiliary information of questions (Ghosh et al., 2020; Liu et al., 2019), knowledge concepts (Nakagawa et al., 2019; Tong et al., 2020), and students' learning behaviors (Xu et al., 2023). Moreover, attention-based models were introduced to tackle the computational expense and instability with long sequences of RNNs (Choi et al., 2020; Pandey & Karypis, 2019). While achieving success in performances, the lack of interpretability raised greater attention, as the model should provide transparency and understanding of the reasoning behind learning behaviors over just the outcomes. Models incorporating educational theories, like the Rasch model and the transfer of knowledge, were proposed to enhance interpretability (Bond & Fox, 2013; Ghosh et al., 2020; Tong et al., 2020). Minn et al. (2022) introduced causal relationships within latent features extracted from students' behaviors. Zhu et al. (2023) attempted to introduce causal inference for explanatory KT analysis.

Despite the remarkable success, deep KT tasks remain a few challenges. Most methods demand extensive student exercise logs for model training, aiming at making binary predictions, which differ from the real analyzing scenarios. The "black-box" nature of deep learning models and numerical predictions limits explainability and make it difficult to generalize to other teaching scenarios, such as open-ended KT and programming learning scenarios (Liu et al., 2022; Piech et al., 2012; Piech et al., 2015b).

2.2 | Large Language Models

LLMs typically refer to transformer-based models containing hundreds of billions of parameters with multi-head attention layers stacked in very deep neural networks (Zhao et al., 2023). LLMs can be categorized as open-sourced models, like the LLaMA and GLM series (Du et al., 2021; Touvron et al., 2023), or close-sourced models, like GPT-4. Trained on massive text data, LLMs exhibit solid natural language understanding to follow complex instructions and solve sophisticated problems due to their emergent abilities which are not presenting in small language models but arising in large ones (Wei et al., 2022). Moreover, LLMs can leverage multi-dimensional information for reasoning and generate natural language responses. Currently, researchers have achieved decent success across domains, like weather forecasting (Bi et al., 2022), recommendation (Bao et al., 2023), and medicine (Thirunavukarasu et al., 2023). Advances in LLMs also brought new possibilities for education, where multiple aspects (e.g., teacher assistance, adaptive learning, and learning tools) benefited from the application of LLMs (Li et al., 2024; Wang et al., 2024; Xiong et al., 2024). The accomplishments in other fields indicated that applying LLMs to KT could lead to similar success.

However, of the less exploration is the work of utilizing LLMs to KT. Neshaei et al. (2024) explored extending the sequence modeling capabilities of LLMs to KT. It was found that fine-tuned LLMs outperformed naive baselines and matched BKT, which suggested that further refinements and deeper understanding of their predictive mechanisms could enhance performance. Similarly, Lee et al. (2024) proposed integrating pre-trained language models with KT methodologies, effectively utilizing semantic information within textual data, addressing the cold-start problem and enhancing interpretability significantly. Moreover, Kim et al. (2024) enhanced the precision of KT through the introduction of option-weighting schemes combined with LLMs. Furthermore, Cheng et al. (2024) incorporated LLMs into computerized adaptive testing environments, utilizing the analytical capabilities of LLMs to assess learners' knowledge states and match suitable test items dynamically. Despite the first attempts, it is limited by the original KT task settings, where LLMs cannot handle such extensive student exercise records. It motivates us to explore a KT paradigm for the era of LLMs and to leverage the advantages of LLMs to address the shortcomings of traditional KT settings.

3 Explainable Few-Shot Knowledge Tracing

Deep KT task is formulated as estimating student next state $\hat{S}_{t+1}$ given student records $\mathcal{X}_t$ (Piech et al., 2015a), where t means the time in a temporal context, the set of questions ($\mathcal{Q}$), the set of knowledge concepts ($\mathcal{C}$), and a KT model M_{DP}, denoted as,

$$\mathcal{X}_t = \{x_0, x_1, \ldots, x_t\}, \ \mathcal{F}: \mathcal{Q} \to \mathcal{C}, \tag{1}$$

$$\hat{S}_{t+1} = \arg\max_{\boldsymbol{y}} P(\boldsymbol{y} | M_{\mathrm{DP}}, \mathcal{X}_t, \mathcal{Q}, \mathcal{C}), \tag{2}$$

where $\mathcal{F}$ denotes the mapping relations of exercises and knowledge concepts and $\boldsymbol{y}$ denotes a possible student state, performance label, or action that the model is trying to predict. The explainable few-shot KT is defined by integrating selected student records $\mathcal{X}'_t \in \mathcal{X}_t$,

the given predictive model ($\mathcal{M}$), question set ($\mathcal{Q}'$), and knowledge concept set (C') with extended information to output estimated student states $\hat{S}'_{t+1}$ and generate explanation $\hat{E}$, further formulated as,

$$\hat{S}'_{t+1}, \hat{E} = \arg\max_{\omega,\varphi} P(\omega, \varphi | \mathcal{M}, \mathcal{X}'_t, \mathcal{Q}', C'), \tag{3}$$

where ω denotes the weight of the model that predicts student performance, and φ denotes the weight of the model that explains why students perform as predicted.

3.1 | Overview

To further practice explainable few-shot KT tasks, this study proposes a cognition-guided framework by leveraging LLMs, as depicted in Figure 2, which consists of three indispensable fundamental components originated from assessment systems (Pellegrino et al., 2001): first, observation, denoted as M_O, defining the task scenario and collecting data; second, cognition, represented as M_C, modeling learners' knowledge state and predicting performances; third, interpretation, denoted as M_I, providing explanations of assessing processes. Notably, most existing KT models primarily function as the cognition module. This study unified the cognition and interpretation, enabling them to ingest the diverse data acquired by the observation module and integrating three components to form a cohesive system.

3.2 | Observation

Observation, which refers to the tasks and situations that allows one to observe students' performance (Pellegrino et al., 2001), defines the learning environment within the assessment system, determining factors, such as the type of knowledge acquired by learners and tasks they engage with. It collects multi-dimensional and multi-modal data $\mathcal{X}_{raw}$ from the designated learning environment $\mathcal{E}$, generating the necessary inputs for subsequent assessment processes, and thus comprises two sub-modules, learning data collection M_{dc} and learning data selection M_{ds}.

First, learning data collection determines the data types to be collected based on the task scenario and carries out the collection process. Typically, it involves gathering information like the student's response sequence, including correctness, timestamps, and duration, as well as question-related information, like problem content and knowledge concepts, and forms a structured dataset $\mathcal{X}_c$.

Second, learning data selection curates the processed data $\mathcal{X}_c$ through strategic selection and reorganization, providing the refined inputs $\mathcal{X}_s = \{x_1, x_2, ..., x_s\}$ to the cognition and interpretation modules as required. In the implementation, this study selected several exercise records to generate informative few-shots for LLMs to predict performance. The advantage of LLMs lies in their ability to leverage in-context learning and reasoning, enabling them to extract high-quality insights while seamlessly ingesting diverse inputs. It aligns with real-world instructional scenarios and lays the foundation for fully exploiting the strengths of LLMs for cognition and interpretation.

3.3 | Cognition

Cognition, which is a model of how students represent knowledge and develop competence in the domain (Pellegrino et al., 2001), synthesizes a comprehensive representation of the learners' evolving knowledge state $\hat{K}_s$ and generates predictions $\hat{P}$ from $\mathcal{X}_s$. This module is divided into two sub-modules, knowledge state analysis

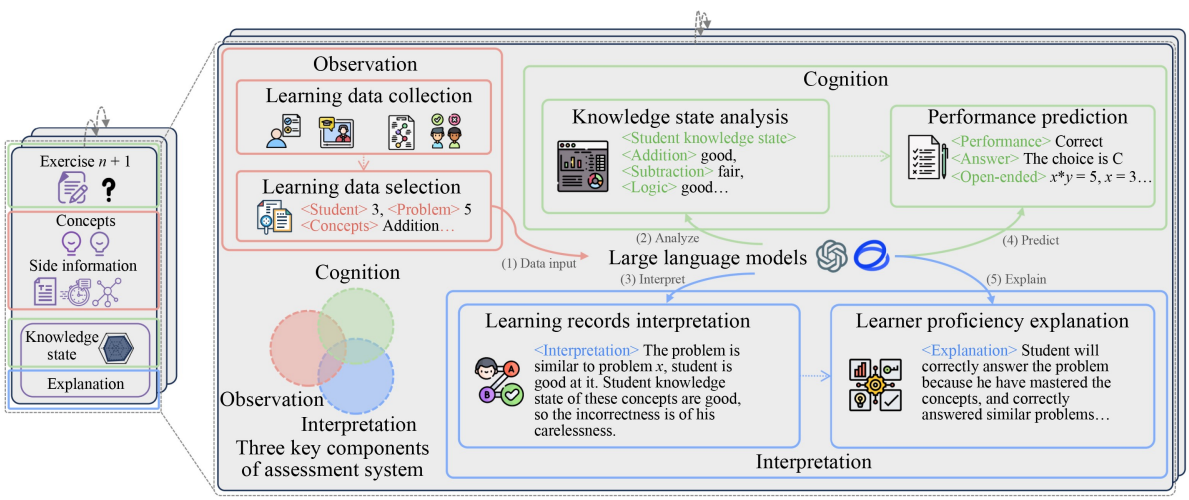

Figure 2 Cognition-guided framework for explainable few-shot knowledge tracing.

M_{C_a} and performance prediction M_{C_p}.

First, knowledge state analysis dynamically analyzes student's mastery of knowledge throughout the practice process by $\mathcal{X}_s$, containing student response records, question information, and behavioral patterns. It generates reliable knowledge state estimates $\hat{K}_s$ as essential references for performance prediction and interpretation, formulated as,

$$\hat{k}_j = M_{C_a}\left(\mathcal{X}_j, \hat{K}_{j-1}, \hat{I}_{j-1}\right), \tag{4}$$

$$M_{C_a}(\cdot) = \arg\max_{\boldsymbol{\omega}} P\left(\boldsymbol{\omega}|\cdot, \boldsymbol{\omega}_{C_a\text{prompts}}\right), \tag{5}$$

where $\hat{k}_j$ is the estimated knowledge state with respect to $\mathcal{X}_j$ at time j and $\hat{K}_{j-1} = \{\hat{k}_1, \hat{k}_2, ..., \hat{k}_{j-1}\}$, $\boldsymbol{\omega}_{C_a\text{prompts}}$ is the prompts designed for knowledge state analysis, and $\hat{I}_{j-1}$ is the set of interpretation of $\hat{k}_1$ to $\hat{k}_{j-1}$. In the implementation, LLMs are asked to generate student mastery of knowledge with ternary value, such as good, fair, or fail, for each concept contained in the exercise the student encounters.

Second, performance prediction forecasts student's performance $\hat{P}$ on predefined environment $\mathcal{E}$ by mining selected data $\mathcal{X}_s$, estimated state $\hat{K}_s$ and interpretation $\hat{I}_s$ from M_I, denoted as,

$$\hat{P} = M_{C_p}\left(\mathcal{X}_s, \hat{K}_s, \hat{I}_s, x_p\right), \tag{6}$$

$$M_{C_p}(\cdot) = \arg\max_{\boldsymbol{\omega}} P\left(\boldsymbol{\omega}|\cdot, \boldsymbol{\omega}_{C_p\text{prompts}}\right), \tag{7}$$

where x_p is the data of exercise to predict and $\boldsymbol{\omega}_{C_p\text{prompts}}$ is the prompts designed for predicting performance. Traditionally, performance is quantified as the probability of correctness or percentage scores. However, by leveraging the generative capabilities of LLMs, this study can extend the prediction to a broader range of learning scenarios less explored, such as open-ended question answering tasks.

3.4 | Interpretation

Interpretation, which is a method for making sense of the data relative to the proposed cognitive model (Pellegrino et al., 2001), leverages $\hat{P}$, $\hat{K}_s$ from the previous modules to generate diagnostic feedback and interpretable analytical insights. These insights facilitate targeted pedagogical interventions to optimize students' educational experience and provide a mechanism to evaluate and justify the validity of the observation module's task design and data selection strategies. The interpretation module comprises two sub-modules: learning records interpretation M_{it} and learner proficiency explanation M_{pe}.

First, learning records interpretation harnesses data $\mathcal{X}_s$ and knowledge estimates $\hat{K}_s$ to furnish natural language explanations $\hat{I}_s$ for learners' historical practice behaviors, formulated as,

$$\hat{i}_j = M_{it}\left(\mathcal{X}_j, \hat{K}_j, \hat{I}_{j-1}\right), \tag{8}$$

$$M_{it}(\cdot) = \arg\max_{\boldsymbol{\omega}} P\left(\boldsymbol{\omega}|\cdot, \boldsymbol{\omega}_{it,\text{prompts}}\right), \tag{9}$$

where $\hat{i}_j$ is the interpretation of student records $\mathcal{X}_j$, $\hat{I}_{j-1} = \{\hat{i}_1, \hat{i}_2, ..., \hat{i}_{j-1}\}$, and $\boldsymbol{\omega}_{it,\text{prompts}}$ is the prompt designed for learning records interpretation. For instance, if a learner demonstrates proficiency in relevant concepts yet provides an incorrect answer, the mistake might be explained by carelessness. Such explanations can refine cognition's analyses and predictions by considering transient factors without unjustly lowering the learner's estimated knowledge state. Conversely, conventional deep learning models often incorrectly penalize occasional errors, producing inaccurate predictions. The flexibility allows for capturing diverse learner behaviors and offering more reliable explanations compared to numerical interpretations from the existing KT models.

Second, learner proficiency explanation integrates $\mathcal{X}_s$ from observation, the estimated knowledge state $\hat{K}_s$, and the explanations $\hat{I}_s$ to provide meaningful insights $\hat{E}$ into the performance predictions of the cognition module. It clarifies the complex interplay among learners' proficiency levels, learning habits, and task performance by situating these predictions within specific instructional scenarios and learner task contexts. The final process can be formulated as,

$$\hat{E} = M_{pe}\left(\mathcal{X}_s, \hat{K}_s, \hat{I}_s, x_p, \hat{P}\right), \tag{10}$$

$$M_{pe}(\cdot) = \arg\max_{\boldsymbol{\omega}} P\left(\boldsymbol{\omega}|\cdot, \boldsymbol{\omega}_{pe,\text{prompts}}\right), \tag{11}$$

where $\hat{E}$ promotes a nuanced explanation of learners' competencies, empowering educators to make timely adjustments to teaching content and cater to individual needs, and $\boldsymbol{\omega}_{pe,\text{prompts}}$ indicates the input prompt of proficiency explanation. For practical prompt design, this study followed the principles proposed by Pellegrino et al. (2001), which posited inferences on student models provided with supportive evidence about assessment tasks. Concurrently, insights from three experienced teaching assistants were collected, thereby aggregating a diverse array of potential evidence sources.

4 Experiments

4.1 | Task Setups

4.1.1 Datasets

This study selected three public datasets, including

FrcSub, MoocRadar (Yu et al., 2023), and XES3G5M (Liu et al., 2023). The detailed statistics of these three datasets are presented in the Electronic Supplementary Material (ESM). The task can integrate multi-dimensional information as input by designing structured textual data and appropriate prompts. Depending on the type of side information incorporated, this study created three different modes, including scant and sparse modes for three datasets, with an additional moderate mode for MoocRadar and XES3G5M, varying degrees of information richness as shown in Figure 3. The scant mode utilizes only the primary student identification document (ID), exercise ID, skill ID, and student interaction records. The sparse mode builds upon scant by incorporating skill representation information. The moderate further includes textual descriptions of exercises over the sparse.

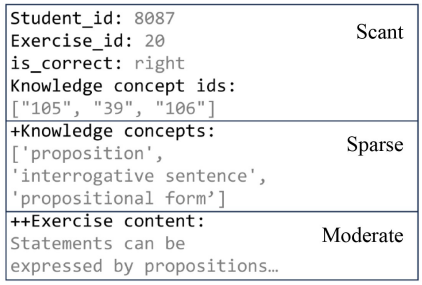

Figure 3 Dataset of different modes.

4.1.2 Metrics

The study collected accuracy, precision, recall, and F1 score as evaluation metrics, as the area under the curve could not be employed since LLMs provided binary predictions. Given that F1 score integrates both precision and recall metrics, using accuracy together with F1 score provides a more comprehensive assessment of model predictive performance. The experimental results of precision and recall metrics are included in the ESM.

4.1.3 Baselines

The study selected six commonly employed and competitive baselines in KT as shown in Table 1: First, deep knowledge tracing (DKT) employs long short-term memory layers to encode the students' knowledge state and predict their performance on exercises (Piech et al., 2015a); second, dynamic key-value memory network designs a static key matrix to capture relations between knowledge components and a dynamic value matrix to track the evolution of students' knowledge (Zhang et al., 2017); third, graph-based knowledge tracing (GKT) leverages graph structure to model interactions between exercises (Nakagawa et al., 2019); fourth, attentive KT utilizes an attention mechanism to characterize temporal distance between questions and student's history interactions (Ghosh et al., 2020); fifth, self-attentive knowledge tracing (SAKT) incorporates a self-attention module to capture latent relations between exercises and students' responses (Pandey & Karypis, 2019); sixth, transformer-based knowledge tracing adopts a transformer architecture to jointly model sequences of exercises and responses (Choi et al., 2020).

Table 1 Comparison of the accuracy and F1 score among baselines in three datasets

Input	Baseline	Dataset					
		FrcSub		MoocRadar		XES3G5M	
		ACC	F1 score	ACC	F1 score	ACC	F1 score
Full-set	DKT	0.7481	0.7514	0.8210	0.8882	0.8355	0.9017
	DKVMN	0.7909	0.8077	0.8147	0.8836	0.8372	*0.9037*
	GKT	0.5480	0.3043	0.7991	0.8772	0.8169	0.8923
	AKT	0.7747	0.7869	0.8194	0.8870	0.8435	**0.9063**
	SAKT	0.7476	0.7389	0.7956	0.8706	0.8298	0.8990
	SAINT	**0.8061**	**0.8201**	*0.8241*	*0.8904*	*0.8399*	0.9044
Few-shot	ChatGLM3-6B	0.6571	0.6496	0.5378	0.6753	0.5434	0.6580
	GLM-4	*0.7939*	*0.7889*	**0.8489**	**0.9052**	**0.8491**	0.8978
	GPT-4	0.7968	0.7471	0.8246	0.9029	0.8176	0.8714

Notes. ACC: accuracy, DKT: deep knowledge tracing, DKVMN: dynamic key-value memory networks, GKT: graph-based knowledge tracing, AKT: attentive knowledge tracing, SAKT: self-attentive knowledge tracing, SAINT: self-attention and intersample attention transformer. The data marked in bold, underlined, and italics mean that the baseline with the best metrics.

4.2 | Overall Performance

This study compared the best performance of ChatGLM3-6B (Du et al., 2021), GLM-4, and GPT-4 across modes of three datasets with considered baselines. Notably, all considered baselines require the full training set to achieve best performances, whereas the proposed datasets in the study only require a few, and such a small amount is far from enough for the baselines. The best three metrics in each column in Table 1 are marked using bold, underlined, and italics. Overall, GLM-4 and GPT-4 achieved superior performance to the baselines on all three datasets. Notably, on the MoocRadar dataset, GLM-4, and GPT-4 outperformed all baselines, showing improvements of 3.01% and 1.66% in accuracy and F1 score, respectively. It demonstrates that leveraging LLMs within explainable few-shot KT can match or surpass conventional deep learning models. In contrast, ChatGLM3-6B did not perform as well as expected,

which could be attributed to the extensive input context. During experiments, this study observed that ChatGLM3-6B frequently struggled to follow instructions, with failure rates of 15.5% on one-hot datasets and 13.7% on sparse datasets, indicating that a fine-tuned smaller model might potentially achieve better performance. GPT-4 and GLM-4 demonstrated no instruction-following failures in these tasks. Specifically, the study presented more comprehensive results and the implementation details to achieve the best performance in the ESM.

4.3 | Analysis

Empirical observations from designed experiments and the potential directions for improving performances when leveraging LLMs for explainable few-shot KT will be discussed.

Exercise text helps a lot; knowledge concepts do a little. This study analyzed the performance of GLM-4 and GPT-4 on different modes in FrcSub and MoocRadar as shown in Figure 4. GLM-4-ACC denotes the accuracy metrics for GLM-4 selecting the first 4 few-shots. It can be observed that the performances substantially improve from sparse to moderate mode in MoocRadar. Integrating only knowledge concepts gains a relatively lower improvement or even a slight decline in FrcSub using GPT-4. It indicates that combining exercise textual information benefits more than knowledge concepts since exercise texts provide more contexts, and concepts provide less than those using IDs. Therefore, a key consideration for boosting performance lies in fully leveraging the existing datasets, formulating them into structured texts, and designing proper prompts that enable LLMs to utilize the additional information effectively.

Increasing the number of few-shots benefits, but too much leads to confusion. This study analyzed GLM-4's performance when using 4, 8, and 16 randomly selected few-shots to explore the impact of different numbers of few-shots on the final results. As shown in Table 2, increasing the number of few-shots leads to performance improvement.

Notably, for the XES3G5M-sparse dataset, the accuracy saw a significant 71.4% improvement from 0.4399 with 4 shots to 0.7542 with 16 shots, and the F1 score achieved an impressive 78.4% enhancement. These results highlight the substantial benefits of utilizing more few-shots, especially for student with long records, as presented in the ESM. However, excessive few-shots would result in an overly long and repeated context, hampering LLMs' capabilities. Even with 4 few-shots, for those that are relatively small, like ChatGLM3-6B in Figure 5, it fails to follow the instructions, and for GLM-4 and GPT-4 in Figure 6, it leads to incorrectly capturing the student behavior information. As a consequence, it may result in generating misguided information. Therefore, developing

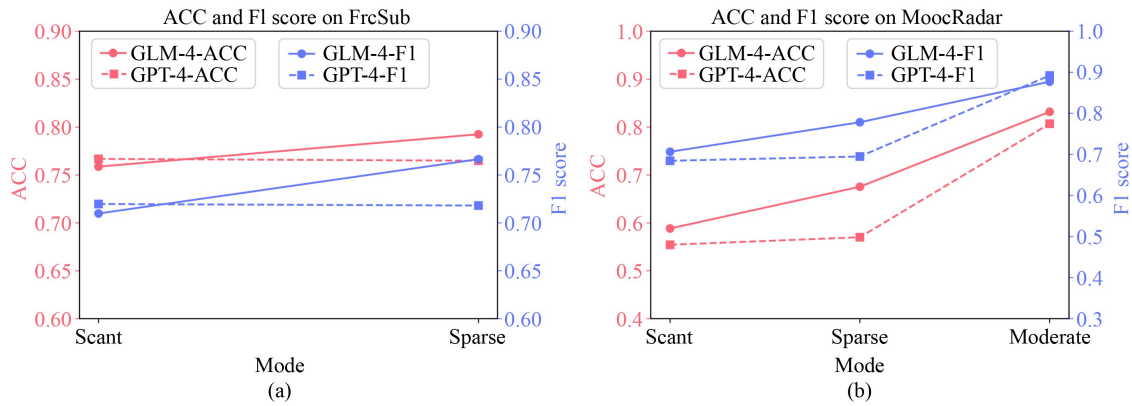

Figure 4 Performance of GLM-4 and GPT-4 on (a) FrcSub and (b) MoocRadar.

Table 2 Performance from different number of few-shots of GLM-4 on three datasets

Number of few-shots of GLM-4	FrcSub-sparse		XES3G5M-sparse		XES3G5M-moderate	
	ACC	F1 score	ACC	F1 score	ACC	F1 score
4	$0.7192^{+0.0}$	$0.7086^{+0.0}$	$0.4399^{+0.0}$	$0.4707^{+0.0}$	$0.6672^{+0.0}$	$0.7592^{+0.0}$
8	$0.7771^{+8.1\%}$	$0.7568^{+6.8\%}$	$0.7057^{+60.4\%}$	$0.7752^{+64.7\%}$	$0.7928^{+18.8\%}$	$0.8623^{+13.6\%}$
16	$0.7939^{+10.4\%}$	$0.7889^{+11.3\%}$	$0.7542^{+71.4\%}$	$0.8395^{+78.4\%}$	$0.8491^{+27.3\%}$	$0.8978^{+18.3\%}$

Note. ACC: accuracy. The subscript data in the line of the number 8 and number 16 of few-shots of GLM-4 refer to the percentage increase relative to the data in the line of the number 4 of few-shots of GLM-4.

ChatGLM3-6B

```
Exercise_id: 6595
Knowledge concepts: ['operation system', 'application software']
Exercise content:
Which of the following cannot be used to describe an operating system …
is_correct: right
Explanation: N/A
Exercise_id: 6597
…
<Exercise to Predict>
Exercise content:
Lecture 20 Deadlocks and Process Communication - 20.3 The Banker's Algorithm can be used to _____ Deadlocks? 'A': …
Knowledge concepts: ["Banker's Algorithm", "Deadlock Avoidance", "Deadlock Detection"]
<Output Predicted is_correct>
exercise_id: 6691, knowledge concepts: ["Banker's Algorithm", "Deadlock Avoidance", "Deadlock Detection"], is_correct: 1
exercise_id: 6692, knowledge concepts: ["Deadlock Detection"], is_correct: 1
Explanation: N/A
```

Figure 5 Case study of ChatGLM3-6B.

GLM-4

```
Exercise_id: 20
…
Exercise_id: 24
Exercise content:
The propositions whose truth values can be determined independently by logic itself are:
…
Knowledge concepts: ['truth value of proposition', 'compound proposition', 'logic']
is_correct: right
Explanation:
1. The new exercise contains the knowledge points: ['proposition', …].
2. The new question is similar to question 2 in the student's record of work, as it also involves the knowledge points 'proposition' and 'interrogative sentence'.
3. Student's Knowledge state:
- 'proposition', good (from previous exercise 2) …
4. The student gets it right, possibly because correct mastery.
Exercise_id: 29
…
<Exercise to Predict>
Exercise content:
Which of the following is not a partition?
'A': …
Knowledge concepts: ["predicate", "basic proposition", "partition"]
<Output Predicted is_correct>
1
Explanation:
1. The new exercise contains the knowledge points: ['predicate', …].
2. Similar to question 24, the student has encountered the knowledge points 'basic proposition' and 'partition' before.
3. Student's Knowledge state:
- 'proposition', good (from previous exercise 24)
- 'interrogative sentence', fair (from previous exercise 24)
…
4. The student gets it right, possibly because correct mastery.
```

GPT-4

```
Exercise_id: 20
…
Exercise_id: 24
Exercise content:
The propositions whose truth values can be determined independently by logic itself are:
…
Knowledge concepts: ['truth value of proposition', 'compound proposition', 'logic']
is_correct: right
Explanation:
1. The new exercise contains 'proposition', ….
2. It's a new question, there is some kind of connection between the previous knowledge points and questions. The knowledge point 'proposition' was also present in the previous question.
3. Student's Knowledge state:
'proposition', fair …
4. The student gets it wrong, possibly due to incorrect mastery. The student may not have fully understood the knowledge points….
Exercise_id: 29
…
<Exercise to Predict>
Exercise content:
Which of the following is not a partition?
'A': …
Knowledge concepts: ["predicate", "basic proposition", "partition"]
<Output Predicted is_correct>
0
Explanation:
1. The new exercise contains 'predicate',….
2. It's a new question, there is some kind of connection between the previous knowledge points and questions. The knowledge point 'basic proposition' was also present in the previous question.
3. Student's Knowledge state:
'proposition', fair, 'interrogative sentence', fair, …
4. The student gets it right, possibly due to correct mastery. The student may have understood the knowledge points involved in the question, leading to the correct answer.
```

Figure 6 Case study of GLM-4 and GPT-4.

effective memory modules enabling LLMs to leverage more few-shots for tracking students' states remains an important direction to explore.

Random few-shots work better in long sequences. This study investigates the impact of different few-shot selection strategies on the final performance. Figure 7 shows the performance of the "first" or "random" selection strategies, using ChatGLM3-6B and GLM-4 on FrcSub-scant and MoocRadar-scant. Generally, the random selection outperforms selecting the first few exercises as few-shots. It is more pronounced in datasets with longer

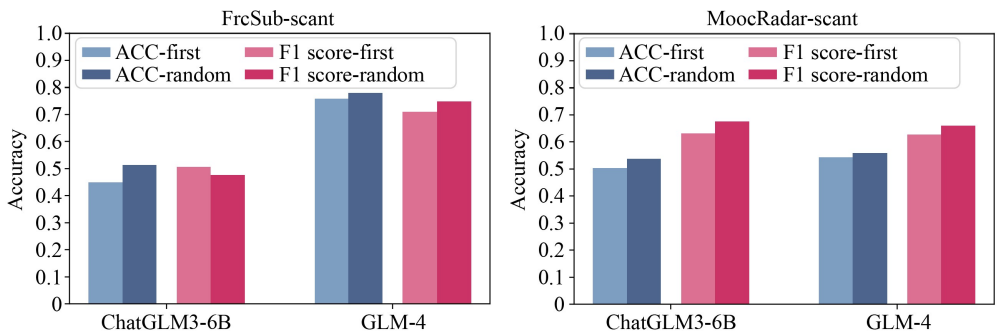

Figure 7 Performance comparison of different few-shots selection strategies. ACC: accuracy.

student interaction records. When student learning histories are extensive, test exercises are more likely to be unrelated to the initial questions, as demonstrated in the ESM. It is worth noting that there remains significant room for improvement in selection strategies. Exploring more optimal selection methods are recommended in this study. For instance, one could select the most recent exercise records, similar exercises to the predicted ones, or utilize retrieval-augmented generation to construct informative few-shots.

4.4 | Evaluation of Explanations

To validate the rationality of the explanation generated by LLMs, this study invited experts with extensive teaching experience to conduct the evaluation. All annotators have at least one semester of teaching assistant experience and have been involved in educational research for at least one year, ensuring that the annotators are familiar with educational assessment. The evaluation of the quality of explanation is divided into objective and subjective metrics. Objectively, researchers asked annotators to directly assess the generated explanations based on five indices, including completeness of the explanation, irrelevant sentences, justifying whether LLMs identify the knowledge concepts, justifying whether LLMs include the new knowledge concept into the student's knowledge state, and correct identification of the student's response. Subjectively, the study provided two sets of explanations to annotators for comparison, including one generated with and one without the guidance of the proposed framework. Annotators are tasked with expressing their opinions on the reasonableness of explanations on two subjective questions. 1,000 explanations generated by GLM-4 from MoocRadar were randomly sampled and more than 800 valid annotations are collected as shown in Table 3 and Figure 8.

Results presented in Table 3 indicate that LLMs exhibit a relatively stable performance, suggesting a strong adherence to the instruction. However, despite the favorable results on objective metrics, Figure 8 shows that nearly half of the annotators perceived the quality of student knowledge state modeling and exercise analysis generated by LLMs to be sub-par, even under guidance. More than 20% of the annotators found the analyses produced by LLMs without guidance to be more reasonable. Moreover, almost 10% of the annotators thought that explanations by two means were not reasonable, indicating that LLMs could hardly handle the tasks of modeling student knowledge and analysis of student exercise behaviors, which required substantial professional experience and reasoning abilities. The analyses they generated should be considered as mere references rather than authoritative

Table 3 Objective metrics results of explanations

Objective metrics	Option 1	Option 2	Option 3
Completeness of the explanation	Yes (**99.4%**)	No (0.6%)	N/A
Irrelevant sentences	Yes (2.4%)	No (**97.6%**)	N/A
Justifying whether LLMs identify the knowledge concepts	Accurate (**90.2%**)	Partially accurate (4.6%)	Not accurate (5.2%)
Justifying whether LLMs include the new knowledge concept into the student's knowledge state	Yes (**87.4%**)	No (12.6%)	N/A
Correct identification of the student's response	Yes (**96.4%**)	No (3.6%)	N/A

Notes. LLMs: large language models. The data marked in bold mean that the baseline with the best performance.

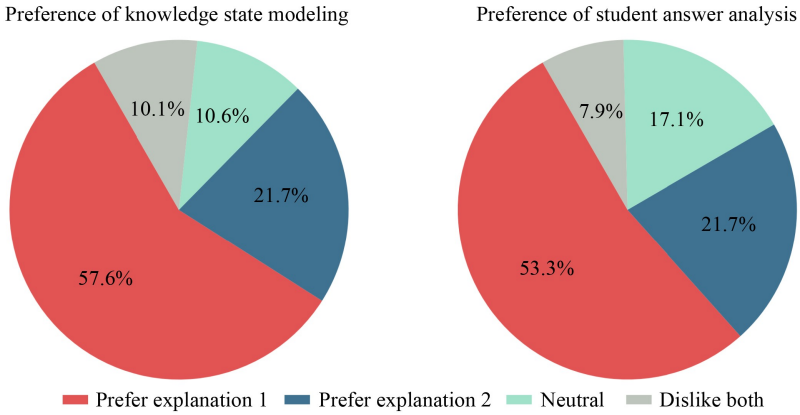

Figure 8 Results of subjective questions of explanations.

judgments. It is worth noting that the Fleiss' Kappa coefficient among annotators approaches zero, a result of the high subjectivity involved in assessing student knowledge states and exercise analyses. Consequently, researchers only present annotators' preferences for the reasonableness of explanations without a ground truth, which can hardly be obtained during the educational practices.

4.5 | Case Study

Examples of all considered LLMs from the MoocRadar-moderate are randomly selected in this study. It involves estimating students' knowledge states in history records, predicting their performances, and providing an explanation. Identifiers are colored to correspond with modules in Figure 6. The content before <Exercise to Predict> contained the previous context, including four few-shots and the LLM's analysis of students' responses based on the selected examples. <Exercise to Predict> contained the information of the test exercise. <Output Predicted is _ correct> represented the LLM's prediction of the students' performance on <Exercise to Predict>, followed by an explanation for the prediction. For detailed prompts and more cases, please refer to the ESM.

4.5.1 ChatGLM3-6B

As shown in Figure 5, knowledge state analysis for each student exercise record is removed to limit the input length for ChatGLM3-6B. However, in many cases, even though it output predictions on all exercises student had encountered, it failed to satisfactorily meet the required output format, which is only 0 or 1 for a single test exercise.

4.5.2 GLM-4 and GPT-4

As illustrated in Figure 6, we highlighted the differences between the outputs of GLM-4 and GPT-4 in grey. We observed that both models are able to follow the instructions and generate formatted explanations in most cases. Differences occur where GPT-4 incorrectly assumed the student had answered Exercise 24 incorrectly and provided an explanation for this assumption. Moreover, when explaining its prediction, GPT-4 failed to recognize the incorrectness of the student's performance and instead offered an explanation suggesting the student had answered correctly. This issue was also present in some cases for GLM-4, possibly due to the models' limited context window to accurately identify such few-shot information and specific information.

5 Conclusions

This study formulated the explainable few-shot KT task to fill the gap between the TKT task and realistic scenarios and proposed a cognition-guided framework to conduct this task. Moreover, this study further demonstrated that LLMs could achieve superior performances to competitive baselines in TKT while providing more natural language explanations under the proposed framework.

This study discussed potential directions for further enhancing LLMs performance on the task, including providing more informative relevant few-shots. Based on this ability, LLMs can understand students' essays and programming codes, even for multi-modal inputs, such as drawings and speeches. By modifying the prompts of modules in the framework and incorporating specific information, it is worthwhile to extend explainable few-shot KT to new tasks in the future, tasks that have been less explored by existing methods, like open-ended question answering and KT programming. Future studies could integrate retrieval-augmented generation to prioritize pedagogically relevant examples and other student experiences in the database to guide LLMs for better understanding of the current situation of students. Moreover, the framework enables expanding existing datasets, providing high-quality analyses and explanations while predicting student performance, which lays the foundation for more explainable KT utilizing LLMs. Researchers can further plan to implement the method on an online learning system. It benefits intelligent tutoring systems and chatbots to understand student knowledge states and respond to student needs to address the reluctance of face-to-face communication.

The requirement for integrating detailed analyses and question-specific results in an elevated level of application program interface consumption despite the reliance on a limited number of shots. Moreover, the efficacy of relatively small models is constrained, compounded by the limitations on text length. It creates a challenge to simultaneously leverage both generated analyses and question-specific information. As educational applications scale to accommodate a larger number of students, reducing costs becomes imperative to support widespread KT. Furthermore, the study has not empirically evaluated the effectiveness of the generated explanations within real-world educational settings. As the proposed framework has not yet been deployed on an actual teaching platform, researchers lack direct analysis and feedback from genuine students' interactions. The absence of practical validation limits understanding of how the explanations

influence student learning outcomes and their overall educational utility. Addressing this gap through real-world implementation and systematic analysis will constitute an important direction for the future research efforts. Additionally, generated explanations and predictions may contain biases or inaccuracies, which could impact teachers' judgments of student states and subsequent decisions. Further efforts are needed to improve the robustness and reliability of the models.

Acknowledgments This work was supported by the National Natural Science Foundation of China (Grant No. 62377002) and the SMP-Zhipu.AI Large Model Cross-Disciplinary Fund (Grant No. 20240211).

Authors Contributions Haoxuan Li was responsible for the writing and revision of this paper, experimental design, and implementation; Jifan Yu was responsible for advising the experimental design and writing the paper; Yuanxin Ouyang was involved in the writing and revision of the paper; Zhuang Liu, Wenge Rong, Huiqin Lin, Juanzi Li, and Zhang Xiong were responsible for advising the writing and providing guidance on the research route of the paper. All authors approved the version to be published and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Conflict of Interest The authors declare that they have no conflict of interest.

Ethics Statements The authors declare that their Institutional Ethics Committee confirmed that no ethical review was required for this study. Written informed consent for participation was not required because all participants' data was anonymized before the statistical analyses were conducted.

Data Availability Statements The authors confirm that all data generated or analyzed during this study are included in this published article.

Electronic Supplementary Material Supplementary material is available in the online version of this article at https://doi.org/10.1007/s44366-025-0071-x and is accessible for authorized users.

References

Bao, K. Q., Zhang, J. Z., Zhang, Y., Wang, W. J., Feng, F. L., & He, X. N. (2023). TALLRec: An effective and efficient tuning framework to align large language model with recommendation. In: *Proceedings of the 17th ACM Conference on Recommender Systems*. Singapore: ACM, 1007–1014.

Bi, K. F., Xie, L. X., Zhang, H. H., Chen, X., Gu, X. T., & Tian, Q. (2022). Pangu-Weather: A 3D high-resolution model for fast and accurate global weather forecast. *arXiv Preprint*, arXiv:2211.02556.

Bond, T. G., & Fox, C. M. (2013). *Applying the Rasch model: Fundamental measurement in the human sciences*. 2nd ed. New York: Psychology Press.

Cheng, C., Zhao, G. H., Huang, Z. Y., Zhuang, Y., Pan, Z. Y., Liu, Q., Li, X., & Chen, E. H. (2024). Towards explainable computerized adaptive testing with large language model. In: *Findings of the Association for Computational Linguistics: EMNLP 2024*. Miami: ACL, 2655–2672.

Choi, Y., Lee, Y., Cho, J., Baek, J., Kim, B., Cha, Y., Shin, D., Bae, C., & Heo, J. (2020). Towards an appropriate query, key, and value computation for knowledge tracing. In: *Proceedings of the 7th ACM Conference on Learning @ Scale*. New York: ACM, 341–344.

Corbett, A. T., & Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, *4*(4), 253–278.

Du, Z. X., Qian, Y. J., Liu, X., Ding, M., Qiu, J. Z., Yang, Z. L., & Tang, J. (2021). GLM: General language model pretraining with autoregressive blank infilling. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. Dublin: ACL, 320–335.

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

Kim, J., Chu, S., Wong, B., & Yi, M. (2024). Beyond right and wrong: Mitigating cold start in knowledge tracing using large language model and option weight. *arXiv Preprint*, arXiv:2410.12872.

Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In: *Proceedings of the 5th International Conference on Learning Representations*. Toulon: ICLR.

Lee, U., Bae, J., Kim, D., Lee, S., Park, J., Ahn, T., Lee, G., Stratton, D., & Kim, H. (2024). Language model can do knowledge tracing: Simple but effective method to integrate language model and knowledge tracing task. *arXiv Preprint*, arXiv:2406.02893.

Li, H., Li, C. L., Xing, W. L., Baral, S., & Heffernan, N. T. (2024). Automated feedback for student math responses based on multi-modality and fine-tuning. In: *Proceedings of the 14th Learning Analytics and Knowledge Conference*. Kyoto: ACM, 763–770.

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

Liu, Z. T., Chen, J. H., Liu, Q. Q., Huang, S. Y., Tang, J. L., & Luo, W. Q. (2022). PYKT: A python library to benchmark deep learning based knowledge tracing models. In: *Proceedings of the 36th International Conference on Neural Information Processing Systems*. New Orleans: ACM, 1347.

Liu, Z. T., Liu, Q. Q., Guo, T., Chen, J. C., Huang, S. Y., Zhao, X. Y., Tang, J. L., Luo, W. Q., & Weng, J. (2023). XES3G5M: A

knowledge tracing benchmark dataset with auxiliary information. In: *Proceedings of the 37th International Conference on Neural Information Processing Systems*. New Orleans: ACM, 1429.

Long, T., Qin, J. R., Shen, J., Zhang, W. N., Xia, W., Tang, R. M., He, X. Q., & Yu, Y. (2022). Improving knowledge tracing with collaborative information. In: *Proceedings of the 15th ACM International Conference on Web Search and Data Mining*. New York: ACM, 599–607.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In: *Proceedings of the 1st International Conference on Learning Representations*. Scottsdale: ICLR.

Minn, S., Vie, J. J., Takeuchi, K., Kashima, H., & Zhu, F. D. (2022). Interpretable knowledge tracing: Simple and efficient student modeling with causal relations. In: *Proceedings of the 36th AAAI Conference on Artificial Intelligence*. AAAI, 12810–12818.

Nakagawa, H., Iwasawa, Y., & Matsuo, Y. (2019). Graph-based knowledge tracing: Modeling student proficiency using graph neural network. In: *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*. Thessaloniki: IEEE, 156–163.

Neshaei, S. P., Davis, R. L., Hazimeh, A., Lazarevski, B., Dillenbourg, P., & Käser, T. (2024). Towards modeling learner performance with large language models. In: *Proceedings of the 17th International Conference on Educational Data Mining*. Atlanta: EDM.

Pandey, S., & Karypis, G. (2019). A self attentive model for knowledge tracing. In: *Proceedings of the 12th International Conference on Educational Data Mining*. Montreal: EDM, 384–389.

Pandey, S., & Srivastava, J. (2020). RKT: Relation-aware self-attention for knowledge tracing. In: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. New York: ACM, 1205–1214.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. Washington: National Academies Press.

Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L., & Sohl-Dickstein, J. (2015a). Deep knowledge tracing. In: *Proceedings of the 29th International Conference on Neural Information Processing Systems*. Montreal: ACM, 505–513.

Piech, C., Huang, J., Nguyen, A., Phulsuksombati, M., Sahami, M., & Guibas, L. (2015b). Learning program embeddings to propagate feedback on student code. In: *Proceedings of the 32nd International Conference on Machine Learning*. Lille: ICML, 1093–1102.

Piech, C., Sahami, M., Koller, D., Cooper, S., & Blikstein, P. (2012). Modeling how students learn to program. In: *Proceedings of the 43rd ACM Technical Symposium on Computer Science Education*. Raleigh: ACM, 153–160.

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. *Nature Medicine*, *29*(8), 1930–1940.

Tong, S. W., Liu, Q., Huang, W., Hunag, Z. Y., Chen, E. H., Liu, C. R., Ma, H. P., & Wang, S. J. (2020). Structure-based knowledge tracing: An influence propagation view. In: *Proceedings of the 2020 IEEE International Conference on Data Mining (ICDM)*. Sorrento: IEEE, 541–550.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., & et al. (2023). LLaMA: Open and efficient foundation language models. *arXiv Preprint*, arXiv:2302.13971.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2017). Graph attention networks. *arXiv Preprint*, arXiv:1710.10903.

Wang, S., Xu, T. L., Li, H., Zhang, C. L., Liang, J., Tang, J. L., Yu, P. S., & Wen, Q. S. (2024). Large language models for education: A survey and outlook. *arXiv Preprint*, arXiv:2403.18105.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., & et al. (2022). Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Xiong, Z., Li, H. X., Liu, Z., Chen, Z. F., Zhou, H., Rong, W. G., & Ouyang, Y. X. (2024). A review of data mining in personalized education: Current trends and future prospects. *Frontiers of Digital Education*, *1*(1), 26–50.

Xu, B. H., Huang, Z. Y., Liu, J. Y., Shen, S. H., Liu, Q., Chen, E. H., Wu, J. Z., & Wang, S. J. (2023). Learning behavior-oriented knowledge tracing. In: *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. Long Beach: ACM, 2789–2800.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In: Brennan, R. L., ed. *Educational measurement*. 4th ed. Westport: Praeger, 111–153.

Yudelson, M. V., Koedinger, K. R., & Gordon, G. J. (2013). Individualized Bayesian knowledge tracing models. In: *Proceedings of the 16th International Conference on Artificial Intelligence in Education*. Memphis: Springer, 171–180.

Yu, J. F., Lu, M. Y., Zhong, Q. Y., Yao, Z. J., Tu, S. Q., Liao, Z. S., Li, X. Y., Li, M. L., Hou, L., Zheng, H. T., & et al. (2023). MoocRadar: A fine-grained and multi-aspect knowledge repository for improving cognitive student modeling in MOOCs. In: *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Taipei: ACM, 2924–2934.

Zhang, J. N., Shi, X. J., King, I., & Yeung, D. Y. (2017). Dynamic key-value memory networks for knowledge tracing. In: *Proceedings of the 26th International Conference on World Wide Web*. Perth: ACM, 765–774.

Zhao, W. X., Zhou, K., Li, J. Y., Tang, T. Y., Wang, X. L., Hou, Y. P., Min, Y. Q., Zhang, B. C., Zhang, J. J., Dong, Z. C., & et al. (2023). A survey of large language models. *arXiv Preprint*, arXiv:2303.18223.

Zhu, J., Ma, X. D., & Huang, C. Q. (2023). Stable knowledge tracing using causal inference. *IEEE Transactions on Learning Technologies*, *17*, 124–134.