RAG-KT: Cross-platform Explainable Knowledge Tracing with Multi-view Fusion Retrieval Generation

Zhiyi Duan[1], Hongyu Yuan[1], Rui Liu[1]*,
[1]Inner Mongolia University, Hohhot, China,
duanzy@imu.edu.cn, 22509014@mail.imu.edu.cn, imucslr@imu.edu.cn

Abstract

Knowledge Tracing (KT) infers a student's knowledge state from past interactions to predict future performance. Conventional Deep Learning (DL)-based KT models are typically tied to platform-specific identifiers and latent representations, making them hard to transfer and interpret. Large Language Model (LLM)-based methods can be either ungrounded under prompting or overly domain-dependent under fine-tuning. In addition, most existing KT methods are developed and evaluated under a same-distribution assumption. In real deployments, educational data often arise from heterogeneous platforms with substantial distribution shift, which often degrades generalization. To this end, we propose **RAG-KT**, a retrieval-augmented paradigm that frames cross-platform KT as reliable context constrained inference with LLMs. It builds a unified multi-source structured context with cross-source alignment via **Question Group** abstractions and retrieves complementary rich and reliable context for each prediction, enabling grounded prediction and interpretable diagnosis. Experiments on three public KT benchmarks demonstrate consistent gains in accuracy and robustness, including strong performance under cross-platform conditions.

1 Introduction

Knowledge Tracing (KT) is a fundamental task in educational data mining that models a student's knowledge state from historical responses to predict future performance (Corbett and Anderson, 1994). As a cornerstone of personalized learning, KT has been extensively studied and has made notable progress (Shen et al., 2024). Despite these advances, a critical gap remains between common KT research settings and real-world deployment.

Conventional Deep Learning (DL)-based KT approaches often rely on closed, pre-defined identifiers, which limits their ability to generalize to unseen questions, new knowledge concepts, or different platforms (Bhattacharjee and Wayllace, 2025). Moreover, these models typically encode knowledge states as latent vectors, which hinders interpretability and constrains their usefulness for pedagogical decision-making (Ding and Larson, 2019; Minn et al., 2022).

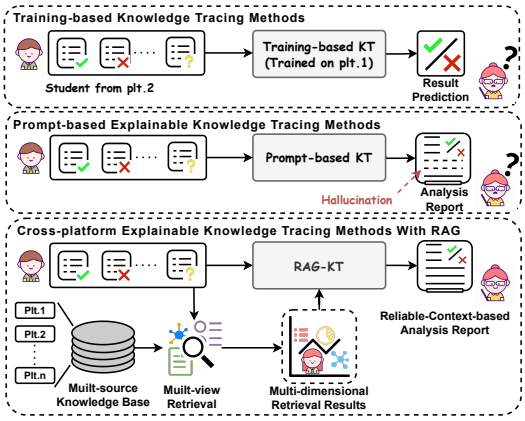

Figure 1: Comparison of KT paradigms. Traditional models struggle with cross-platform shifts, while prompt-based LLMs are prone to hallucinations. RAG-KT overcomes these by leveraging multi-view retrieval to enable reliable, grounded inference.

Recently, Large Language Models (LLMs) have opened new opportunities for improving interpretability and flexibility (Achiam et al., 2023; Bai et al., 2023; Liu et al., 2024; Hao et al., 2025). With their natural language reasoning capability, LLMs can extend KT beyond numeric prediction to interpretable diagnosis by generating human-readable analytical reports (Wang et al., 2024; Li et al., 2025b). However, directly applying LLMs to KT faces a dilemma. Prompt-based methods are flexible, yet they often lack platform-specific learning signals and can therefore hallucinate or produce generic explanations that are not grounded in an individual's learning history (Li et al., 2024;

*Corresponding author

Findings of the Association for Computational Linguistics: ACL 2026, pages 5217–5232

Duan et al., 2025). In contrast, fine-tuning LLMs on interaction data can improve in-domain performance, but it typically strengthens dependence on the source platform's distribution, sacrifices cross-platform generalization, and incurs high retraining costs whenever the data source changes (Li et al., 2025b; Chen et al., 2024; Li et al., 2025a).

Overall, both DL-based KT and current LLM-based solutions are commonly developed and assessed under a same-distribution assumption. This assumption is largely inherited from the fact that most benchmark datasets and training pipelines are collected under single-platform. In practice, however, educational data are generated across many heterogeneous platforms. Even when these platforms cover similar knowledge concepts, their student populations, interaction patterns, and question representations can still differ substantially, resulting in non-trivial distribution shift.

We argue that to handle the complexity of cross-platform data, KT must go beyond the single-distribution assumption. The key challenges lie in two aspects: (i) how to organize cross-platform data and unify heterogeneous information into a single, unified representation, and (ii) how to obtain information-rich and reliable context from it. To this end, we propose RAG-KT, a retrieval-augmented (Lewis et al., 2020; Zhao et al., 2024) knowledge tracing paradigm. As shown in Fig. 1, our proposed framework addresses the interpretability and cross-platform capability challenges by augmenting LLMs with structured retrieval and external knowledge integration.

Specifically, RAG-KT builds a unified heterogeneous graph that integrates interaction signals across platforms. To mitigate identifier inconsistency and enable cross-source alignment, we introduce **Question Group** nodes as an intermediate abstraction layer that aggregates questions with similar instructional attributes. This design allows structured reasoning even when question IDs are not shared across platforms. Building on this representation, we further design a multi-dimensional retrieval mechanism to collect complementary context for the current prediction step. The retrieved context is then organized into structured, context-grounded prompts that constrain the LLM to produce the final prediction and an interpretable analysis report. Experimental results demonstrate that the retrieval-augmented structured context in RAG-KT consistently improves prediction accuracy and robustness. In summary, our main contributions are as follows:

- We propose RAG-KT, the first retrieval-augmented paradigm for applying LLMs to KT, aiming to address the limitations of existing KT methods in interpretability and cross-platform generalizability.
- We construct a multi-source heterogeneous knowledge base and design a multi-view fusion retrieval framework to extract task-relevant knowledge, and generate an explainable analysis report via structured reliable-context-based prompting, thereby enhancing LLM performance on KT.
- We conduct comprehensive experiments on three widely used datasets, and the results show that RAG-KT achieves superior prediction accuracy, produces interpretable analytical reports, and remains robust under cross-platform scenarios.

2 Related Work

2.1 Structured Knowledge Modeling in KT

Early models use binary knowledge states and predefined skill mappings (Corbett and Anderson, 1994). Subsequent neural models encode interactions as embeddings but lack semantic structure and cross-platform generalization (Piech et al., 2015; Zhang et al., 2017). To address this, recent works have introduced knowledge graphs and use GNNs (Scarselli et al., 2008) to capture skill dependencies (Nakagawa et al., 2019; Liu et al., 2020). More recently, methods enhance KT by modeling temporal dynamics or phase-wise learning trajectories over evolving graph structures (Cheng et al., 2024). However, these methods typically assume fixed, well-aligned datasets and do not address the challenges of integrating heterogeneous interactions from multiple platforms. Forcing integration will only affect the quality of relation extraction, which will directly affect prediction accuracy.

2.2 LLM-Enhanced KT Methods

Recent studies have explored LLMs to enhance knowledge tracing, aiming to improve interpretability and address cold-start issues (Li et al., 2025c). Some approaches directly fine-tune LLMs on student interaction data (Li et al., 2025b; Jung et al., 2024; Wang et al., 2025). Meanwhile, others use few-shot or zero-shot prompting strategies (Li et al.,

2024; Kim et al., 2024; Duan et al., 2025). Despite their promise, many existing LLM-based KT methods operate in closed or shallow contexts, lacking structured grounding from educational knowledge. As a result, they may suffer from inconsistency or hallucinations, especially under sparse or out-of-distribution scenarios. In contrast, our RAG-KT framework grounds LLM inference in a structured, heterogeneous knowledge graph and retrieves personalized, multi-dimensional context. This design ensures both interpretability and generalization, especially in low-resource or unseen settings.

3 Methodology

3.1 Overall Framework

The overall framework of RAG-KT is shown in Fig. 2. It begins by constructing a Multi-source Heterogeneous Knowledge Base $\mathcal{B}$ including Multi-source Knowledge Graph $\mathcal{G}$ and Multi-dimensional Interaction Repository $\mathcal{R}$, which unify educational data from diverse platforms and align semantically equivalent entities across datasets. On top of this, a Multi-view Fusion Retrieval mechanism is applied to extract relevant contextual information for given inputs. The retrieved information is then formatted into a Structured Reliable-Context-based Prompt, allowing a frozen LLM to perform zero-shot prediction and generate interpretable reliable-context-based reports. Together, these components enable RAG-KT to achieve robust and explainable KT.

3.2 Multi-source Heterogeneous Knowledge Base Construction

The goal of this stage is to construct a unified and extensible Multi-source Knowledge Graph $\mathcal{G}$ and Multi-dimensional Interaction Repository $\mathcal{R}$ that integrate data from multiple sources, so as to better model the relationships between students, questions, and knowledge concepts from different sources and serve as the foundation for subsequent multi-view retrieval. Unlike existing heterogeneous knowledge graphs for KT, our $\mathcal{G}$ comprises six types of nodes: Knowledge Concepts (K), Question Groups (QG), Questions (Q), Students (S), Ability Levels (A), and Difficulty Levels (D). These nodes are connected via six corresponding types of edges, including $K-K$, $S-A$, $S-QG$, $QG-D$, $QG-K$, and $Q-QG$. Together, these components form a richly structured multi-relational graph.

We first collect multiple publicly available KT datasets, then estimate students' ability levels and question difficulties using the 2-parameter logistic Item Response Theory (IRT-2PL) model:

$$P_{s,q} = \frac{1}{1 + e^{-a_q(\theta_s - b_q)}} \tag{1}$$

where θ_s denotes the ability of student s, b_q is the difficulty of question q, and a_q is the question's discrimination parameter. The resulting θ_s and b_q are normalized and categorized into three levels (e.g., Low, Medium, High) using a standard deviation-based thresholding rule:

$$\text{Level}(x) = \begin{cases} \text{Low}, & x \le \mu - \sigma \\ \text{Medium}, & \mu - \sigma < x < \mu + \sigma \\ \text{High}, & x \ge \mu + \sigma \end{cases} \tag{2}$$

To enhance the structural modeling of domain knowledge, we compile a collection of knowledge concepts and their Predecessor-Successors or Associative relationships from K-12 mathematics textbooks. This forms what we call the Basic KC Graph $\mathcal{G}_{\mathcal{KC}}$, a subgraph composed of K and $K-K$.

Since K representations vary across datasets, we design a KC matching pipeline to align dataset-specific terms with those in the $\mathcal{G}_{\mathcal{KC}}$. For concept pairs with high semantic similarity, direct mapping is performed. For low-similarity cases, we leverage a LLM to review and align ambiguous concepts to the most semantically appropriate node in the $\mathcal{G}_{\mathcal{KC}}$, thereby ensuring consistency and coverage across datasets. Detailed KC matching methods will be provided in the Appendix A.

Since the data for graph construction is multi-source, the resulting questions are diverse, which makes Q alignment necessary. Therefore, we introduce a unified intermediate node type QG, defined by a unique pairing of knowledge concept and difficulty level. Each QG is constructed as:

$$\text{QG}_{i,j} = \{q \mid D_q = D_i, K_q = K_j\} \tag{3}$$

This design allows questions from different datasets with the same semantic intent to be aligned, enabling downstream tasks to generalize across question sources. In parallel with $\mathcal{G}$ construction, we build a structured $\mathcal{R}$ that stores student interaction history in the form of triplets $(S, QG, \text{Correctness})$, which are further enriched by other dimensions such as concept-level and difficulty-level performance. This $\mathcal{R}$ provides the contextual source for later retrieval.

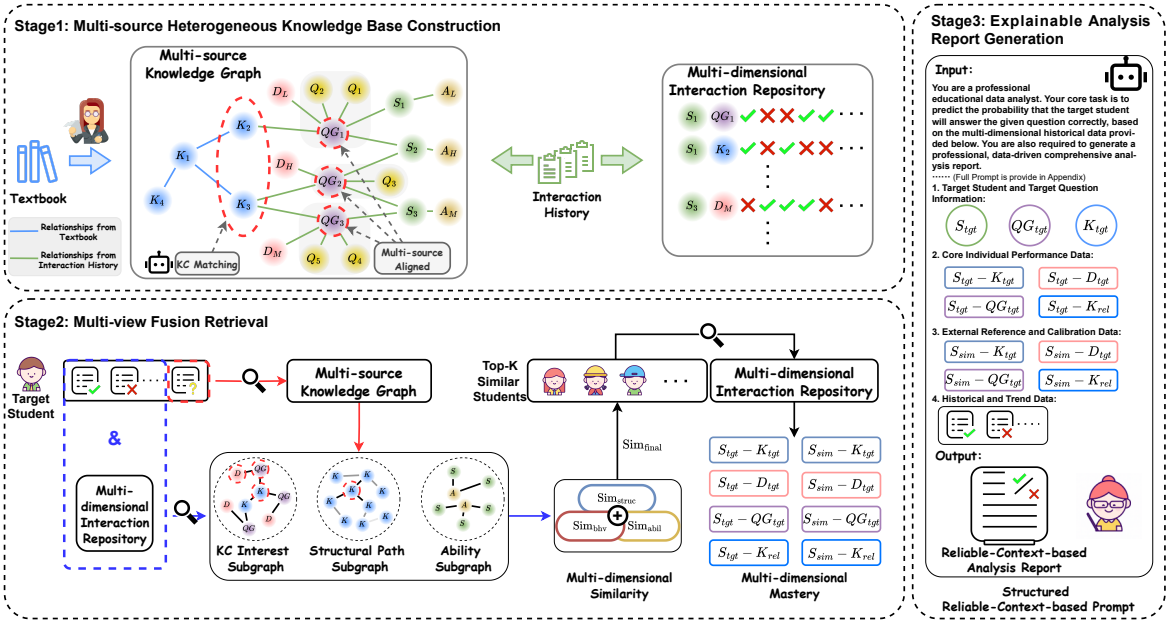

Figure 2: The RAG-KT framework, notation like S_{tgt} ˇ K_{tgt} indicates the target student's performance on the target concept, while S_{sim} ˇ K_{tgt} represents similar students' aggregated performance on that dimension. QG is a collection of semantically equivalent Q from different data sources.

3.3 Multi-view Fusion Retrieval

Retrieving high-quality and task-relevant context is crucial for enabling the LLM to perform accurate zero-shot prediction in KT. Thus, this stage aims to retrieve high-quality, task-relevant contextual information from the $\mathcal{B}$, enabling zero-shot prediction by the LLM. Each prediction target consists of a target question Q_{tgt} and the historical interaction sequence of the target student:

$$S_{tgt} = \{(q_1, r_1), \ldots, (q_n, r_n)\}, r_i \in \{0, 1\} \quad (4)$$

To support retrieval, we first identify the target knowledge concept K_{tgt} associated with Q_{tgt} via semantic matching over $\mathcal{G}$. To support fine-grained reasoning over different facets, we extract three task-specific subgraphs from $\mathcal{G}$, which correspond to different views. They capture relevant information from structural, behavioral, and ability perspectives to provide complementary views that enable more accurate and interpretable retrieval. Each subgraph is paired with a corresponding similarity computation module to identify top-k similar students S_{sim} whose interactions are used to form the final prompt context.

KC Interest Subgraph for Behavior-based Similarity. The KC Interest Subgraph contains all nodes within n-hop neighborhood of K_{tgt}, including related concepts K_{rel}, question group QG_{tgt}, and difficulty D_{tgt} nodes. This subgraph defines the shared semantic scope for comparing student behaviors. We construct a behavior vector $\vec{b}_s$ for each student s over the subgraph nodes. For each node n, the behavior score is defined as:

$$\text{Score}_s^n = [\alpha \cdot \text{Acc}_s^n + (1-\alpha) \cdot \text{DWA}_s^n] \cdot \text{Conf}_s^n \quad (5)$$

where $\alpha \in [0, 1]$ is a configurable weight (default 0.5), Acc_s^n is the historical accuracy, and DWA_s^n is the Dynamic Weighted Accuracy (Fox et al., 2002):

$$\text{DWA}_s^n = \frac{\sum_{i=1}^{N} \beta^{N-i} \cdot r_i}{\sum_{i=1}^{N} \beta^{N-i}}, \quad \beta \in (0, 1) \quad (6)$$

with N attempts on node n and β controlling the recency decay (default 0.8). Conf_s^n is the confidence score, we define the confidence score as the product of sample sufficiency and performance stability, where the former captures the quantity of rich and reliable context and the latter reflects the consistency of behavior. More attempts and more stable recent performance give higher confidence. This design ensures that confidence increases only when a student has answered sufficiently many questions with stable performance. The vector $\vec{b}_s$ is:

$$\vec{b}_s = [\text{Score}_s^{n_1}, \text{Score}_s^{n_2}, \ldots, \text{Score}_s^{n_T}] \quad (7)$$

The behavior similarity is then computed using cosine similarity:

$$\text{Sim}_{\text{bhv}} = 1 - \frac{\vec{b}_{tgt} \cdot \vec{b}_s}{\|\vec{b}_{tgt}\| \cdot \|\vec{b}_s\|} \quad (8)$$

Structural Path Subgraph for Structure-based Similarity. This subgraph includes all K nodes reachable from K_{tgt} via $K - K$ edges in the domain graph $\mathcal{G}_{\mathcal{KC}}$. It captures prerequisite and associative relationships around the target concept. For each student s, we compute a structural path relevance score that reflects how closely the student's learned knowledge concepts are connected to the target concept K_{tgt} within the domain graph $\mathcal{G}_{KC}$. We define the student's structural path score as the average of the inverse path lengths:

$$\text{Score}^s_{\text{struc}} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\text{len}(K_{tgt}, K^s_i)} \tag{9}$$

Where K^s_i denotes the set of knowledge concepts that student s has interacted with, and N is size of the set. $\text{len}(\cdot, \cdot)$ denotes the shortest path length between two nodes in $\mathcal{G}_{KC}$. We normalize the absolute score difference into a similarity score:

$$\text{Sim}_{\text{struc}} = 1 - \frac{|\text{Score}^P_{tgt} - \text{Score}^P_s|}{\text{Score}^P_{max} - \text{Score}^P_{min}} \tag{10}$$

This reflects how well the student's concept path aligns structurally with the target concept.

Ability Subgraph for Ability-based Similarity. The Ability Subgraph includes all S and A nodes and their $S - A$ links from the graph construction phase. Each student's ability θ_s is estimated via IRT-2PL and normalized to $[0, 1]$. We define ability similarity as:

$$\text{Sim}_{\text{abil}} = 1 - |\theta_{tgt} - \theta_s| \tag{11}$$

Fusion and Peer Retrieval. The final similarity score is computed via weighted fusion:

$$\begin{aligned} \text{Sim}_{\text{final}} = \lambda_1 \cdot \text{Sim}_{\text{bhv}} + \lambda_2 \cdot \text{Sim}_{\text{struc}} \\ + \lambda_3 \cdot \text{Sim}_{\text{abil}} \end{aligned} \tag{12}$$

with $\lambda_1 + \lambda_2 + \lambda_3 = 1$ as configurable hyperparameters. We retrieve the top-k most similar students based on $\text{Sim}_{\text{final}}$, and extract their interaction records from $\mathcal{R}$. For each dimension (e.g., K, D, QG), we compute their aggregate performance as:

$$\text{Perf}^d_s = (\text{Acc}^d_s, \text{DWA}^d_s, \text{Attempts}^d_s, \text{Conf}^d_s) \tag{13}$$

The same features are also computed for the target student, forming a structured, multi-perspective context Ctx that is encoded into the LLM prompt for subsequent prediction and explanation.

3.4 Explainable Analysis Report Generation

To address the critical need for interpretable outcomes in educational practice, this stage guides the LLM to transform the retrieved context into a structured, human-readable report through a Structured Reliable-Context-based Prompt. Considering that our method does not involve fine-tuning, prompts thus have a significant impact on the performance of LLMs, making it crucial to design a high-quality prompt. A simple prompt example is shown in the right of Fig. 1, and the complete prompt is provided in the Appendix B.

Input Structure. The LLM prompt consists of four blocks: (1) target student and question metadata (ability θ_{tgt}, concept K_{tgt}, difficulty D_{tgt}, etc.); (2) individual performance metrics across multiple dimensions; (3) peer and similar student aggregates obtained via multi-view fusion retrieval; (4) historical answer trajectory for trend analysis.

Reasoning Framework. The LLM is guided by a reasoning scaffold incorporating: (1) context reliability (e.g., attempt counts); (2) positive and negative attribution factors (e.g., peer success, unstable accuracy); (3) conflict resolution using structural and behavioral consistency; (4) calibration against peer datasets. This promotes transparent and consistent prediction generation.

Output Format. The report includes: (1) predicted probability and qualitative judgment; (2) student ability and knowledge mastery summary; (3) reliable-context-based explanation detailing decision rationale and risk analysis. This enables actionable feedback for personalized instruction.

Given the diversity of actual teaching scenarios, qualitative analysis can better adapt to different educational contexts, capture nuanced learning states that quantitative data alone may miss, and make the feedback more intuitive and actionable for educators and students.

4 Experiment

4.1 Experimental Configuration

We evaluate RAG-KT on three public KT datasets: **ASSIST09** and **ASSIST12**, collected from the ASSISTments platform (Feng et al., 2009), and **DBE-KT22** (Abdelrahman et al., 2022), from an online course at the Australian National University. Additionally, we use the **Eedi** (Wang et al., 2020) dataset from the NeurIPS 2020 Education Challenge to assess performance in a fully cold-start setting, where all students, questions, and platform are unseen.

Table 1: Main results on three benchmark datasets. Bold values denote the best performance; bold and underlined values indicate the second-best. RAG-KT$_{\text{LLM}}$ employs different LLMs as the final prediction model.

Type	Baselines	ASSIST09			ASSIST12			DBE-KT22		
		ACC	AUC	F1	ACC	AUC	F1	ACC	AUC	F1
DL-based Methods	DKT	72.18	81.82	78.12	70.86	69.12	74.98	71.08	72.26	74.52
	DKT+	72.64	82.47	78.65	71.34	70.16	75.42	71.76	73.42	75.06
	AT-DKT	72.62	82.96	79.16	71.58	71.28	75.84	72.32	74.18	75.48
	DKVMN	71.32	81.28	77.92	70.51	68.73	74.76	71.47	72.03	74.65
	Deep-IRT	71.96	80.98	77.84	71.02	69.19	75.03	72.04	72.81	75.09
	AKT	73.76	84.11	80.28	73.09	72.79	77.82	73.61	**75.29**	77.48
	GKT	72.23	81.53	78.23	70.96	69.07	75.11	72.06	73.11	75.03
LLM-based Methods (Prompting)	HISE-KT	79.12	**84.31**	82.03	**73.77**	72.33	79.51	70.82	72.40	79.94
	EFKT	64.85	61.10	73.59	63.03	66.79	66.83	63.50	68.28	66.44
LLM-based Methods (Fine-tuning)	EPLF	70.13	81.27	71.60	70.30	69.64	75.82	74.63	72.72	72.38
	LLM-KT	78.68	83.55	81.03	72.23	72.57	78.73	77.48	75.27	78.25
	2T-KT	74.55	81.32	80.65	72.80	71.60	75.45	75.43	74.42	78.86
	CIKT	75.17	82.27	80.27	72.63	71.89	77.67	76.38	74.73	77.14
Ours	RAG-KT$_{\text{GPT-4o}}$	78.34	82.57	83.20	72.50	72.87	79.37	76.50	74.53	85.74
	RAG-KT$_{\text{Qwen-Plus}}$	**79.20**	83.97	**85.06**	72.60	**73.35**	**80.68**	**78.40**	73.69	**86.60**
	RAG-KT$_{\text{DeepSeek-R1}}$	**80.00**	**85.74**	**85.75**	**74.80**	**73.89**	**82.60**	**78.89**	**76.32**	**86.80**

To ensure leakage-free evaluation, we adopt a student-level disjoint split. After segmenting each student's history into sub-sequences of length 25 and using the last interaction in each sub-sequence as the prediction target, we randomly sample 1,000 test sequences and assign all interactions from the corresponding students to the test set; all remaining students are used for training. Hence, no student appears in both splits. Moreover, all reported results are averaged over five independent random splits/samplings.

We implement RAG-KT using three LLM backbones: GPT-4o, Qwen-Plus, and DeepSeek-R1 to evaluate generalizability across model families. The constructed knowledge graph includes 317 concepts, 593 prerequisite-successor relations, and 932 associative links, 34,171 students, 951 question groups, and over 3.3 million interaction records. For the multi-dimensional retrieval stage, we set the weights of behavior, structure, and ability similarities to $\lambda_1 : \lambda_2 : \lambda_3 = 4 : 3 : 3$ (which is determined through an experiment, result as shown in Table 7), emphasizing behavioral alignment while retaining structural and ability-aware reasoning. The KC Interest Subgraph uses a 2-hop neighborhood, and the maximum path length in the Structural Path Subgraph is capped at 10. Peer retrieval selects the top-2 most similar users per query. All models follow the same preprocessing and evaluation protocols for fair comparison.

4.2 Baselines

To comprehensively evaluate the effectiveness of RAG-KT, we compare our method against a diverse set of baselines, including traditional deep learning-based KT models and recent LLM-enhanced approaches. The baselines are grouped into the following categories:

- **DL-based Methods**: DKT (Piech et al., 2015), DKT+ (Yeung and Yeung, 2018), AT-DKT (Liu et al., 2023), DKVMN (Zhang et al., 2017), Deep-IRT (Yeung, 2019), AKT (Ghosh et al., 2020), GKT (Nakagawa et al., 2019).
- **LLM-based Methods (Prompting)**: HISE-KT (Duan et al., 2025), EFKT (Li et al., 2024), LOKT (Kim et al., 2024).
- **LLM-based Methods (Fine-tuning)**: LLM-KT (Wang et al., 2025), EPLF (Neshaei et al., 2024), CIKT (Li et al., 2025b), CLST (Jung et al., 2024), 2T-KT (Li et al., 2025a).

Specific descriptions of each baseline are provided in the Appendix C.

4.3 Main Results

Tab. 1 presents the main evaluation results across three benchmark datasets. Overall, our proposed RAG-KT framework significantly outperforms all baseline models, including both traditional deep learning approaches and recent LLM-based

Table 2: Ablation results on three benchmark datasets. Each row removes or alters a key component of the full RAG-KT framework to assess its impact.

Methods	ASSIST09			ASSIST12			DBE-KT22		
	ACC	AUC	F1	ACC	AUC	F1	ACC	AUC	F1
Full	**80.00**	**85.74**	**85.75**	**74.80**	**73.89**	**82.60**	**78.89**	**76.32**	**86.80**
w/o Similar Students	77.28	81.07	83.60	69.53	71.47	78.13	77.40	73.98	84.17
w/o Question Groups	76.73	84.25	83.32	70.01	72.33	78.67	77.94	75.50	85.16
w/o KC Interest Subgraph	77.46	**85.49**	**84.23**	70.01	73.06	78.54	77.40	75.23	85.08
w/o Structural Path Subgraph	78.50	84.16	84.89	72.40	71.54	80.67	76.71	**75.98**	85.32
Random Similar Students	**78.54**	83.62	84.03	**73.36**	**73.14**	**81.56**	**78.38**	75.72	**86.43**
w/o Retrieval	68.50	62.37	74.35	64.20	67.41	71.35	64.40	68.58	72.65

KT methods. Specifically, RAG-KT$_{\text{DeepSeek-R1}}$ achieves the highest scores across all three datasets, demonstrating strong robustness and generalization. The second-best results are consistently achieved by RAG-KT$_{\text{Qwen-Plus}}$ in most cases, further validating the stability of our multi-view retrieval-enhanced design across different LLM backbones. Compared to traditional KT methods, which suffer from limited context-awareness and low adaptability in cross-domain settings, our framework provides substantial performance gains. Moreover, in contrast to LLM-based KT methods that rely solely on internal reasoning, our framework benefits from structured external grounding via graph and peer retrieval, yielding not only higher prediction accuracy but also enhanced interpretability. These results affirm the effectiveness of our retrieval-augmented framework in bridging structured modeling and LLM reasoning for KT.

4.4 Ablation Results

The ablation study results, as shown in the Tab. 2, provide clear rich and reliable context of the importance of each key component within the RAG-KT framework:

First, removing the Similar Students module (w/o Similar Students) leads to a substantial drop in performance across all datasets. This highlights the critical role of retrieving ability-aligned and structurally similar students for enhancing prediction. Second, removing Question Group nodes (w/o Question Groups) also results in consistent performance degradation. As Question Groups serve as an abstraction that unifies knowledge concepts with difficulty levels across platforms, their absence weakens the model's generalization and retrieval alignment capabilities. Third, regarding retrieval subgraphs, both the KC Interest Subgraph and Structural Path Subgraph show notable impact

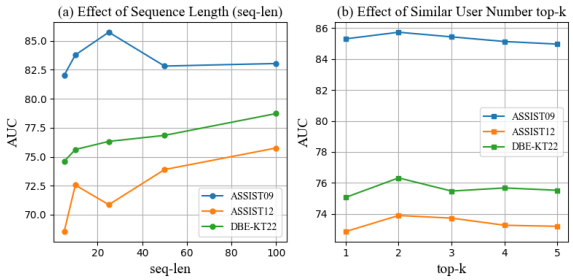

Figure 3: Parameter sensitivity analysis of RAG-KT on three datasets.

when ablated. This demonstrates the value of semantic neighborhood modeling and verifies the effectiveness of structured reasoning based on graph connectivity.

The Random Similar Students setting, which replaces the similarity with random selection, consistently underperforms compared to the full model. This demonstrates the necessity of our designed similarity functions that integrate structural, behavioral, and ability signals. Lastly, removing the entire retrieval module (w/o Retrieval) leads to a sharp drop in performance, achieving results comparable to existing prompting methods. Experimental data confirms that this is the most critical component, fundamentally highlighting the value of extra context in KT.

In summary, all components of RAG-KT contribute positively to performance, with Question Groups and similarity-based retrieval playing especially vital roles in enhancing both prediction accuracy and interpretability.

4.5 Analysis about HyperParams

Fig. 3 presents the AUC performance of the proposed RAG-KT framework under different hyperparameter settings. In subplot (a), trends vary across datasets. ASSIST09 peaks at seq-len = 25,

Table 3: Cold-start Results

Type	Baselines	Eedi		
		ACC	AUC	F1
DL-based Methods	DKT	49.54	50.23	49.61
	DKT+	49.98	50.62	49.84
	DKVMN	51.58	51.20	52.64
	AKT	53.55	49.69	47.68
LLM-based Methods (Cold-start)	LOKT	**65.88**	**65.52**	**57.90**
	CLST	61.45	63.60	42.23
Ours	RAG-KT$_{\text{DeepSeek-R1}}$	**68.00**	**74.40**	**70.43**

whereas ASSIST12 and DBE-KT22 generally improve as the sequence length increases, achieving their highest AUC at seq-len = 100. This indicates that while moderate context suits ASSIST09, the other datasets benefit significantly from longer historical sequences. In subplot (b), we investigate the effect of the number of similar students (top-k) retrieved in the retrieval stage. Overall, performance remains stable when top-k ranges from 2 to 4, with a slight peak at top-k = 2, suggesting that including a small but relevant peer set provides the most beneficial calibration. Too many peers may introduce irrelevant or noisy patterns, slightly degrading performance. These results confirm that our framework is relatively robust to hyperparameter settings within reasonable ranges. Full results is provided in the Appendix D.

4.6 Cold-start Results

As shown in Table 3, our proposed RAG-KT framework consistently outperforms both traditional deep learning models and recent LLM-based methods tailored for cold-start scenarios under the fully cold-start setting, where both students and questions are entirely unseen. In this setting, our constructed knowledge base and heterogeneous graph contain NO information from the Eedi dataset, and the only cross-platform signal comes from retrieval via the Question Group nodes. This demonstrates the robustness and strong generalization ability of our method. By grounding predictions in a unified heterogeneous knowledge graph and leveraging multi-view fusion retrieval, our framework effectively mitigates challenges such as domain shift, data sparsity, and lack of prior interaction history. These issues severely limit the performance of other knowledge graph-based methods in cold-start environments.

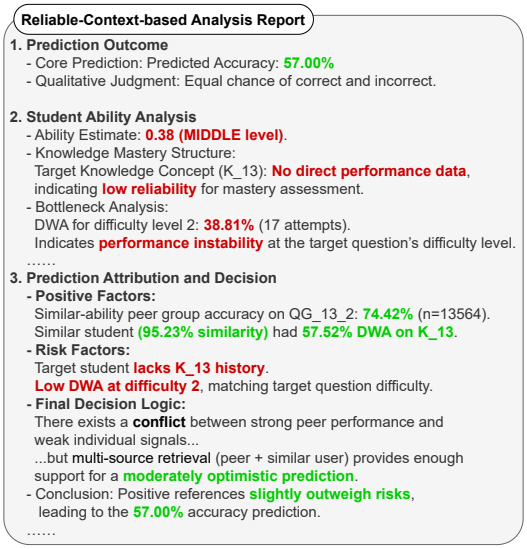

Figure 4: A case for our RAG-KT. The student answers this question correctly. The red font represents positive analysis and the green font represents negative analysis.

Table 4: Scores for report quality evaluation on four dimensions: Explainability, Readability, Educational Usefulness, and Rigorousness

Baselines	Len.	Exp.	Read.	Edu.	Rig.	Total
CIKT	690	4.22	**4.68**	**4.50**	3.83	17.23
EFKT	82	2.70	2.69	2.40	2.26	10.05
HISE-KT	350	**4.72**	4.48	4.35	**4.53**	**18.08**
RAG-KT$_{\text{DeepSeek-R1}}$	297	**4.90**	**4.56**	**4.88**	**4.90**	**19.24**

4.7 Case Study

We present a representative cold-start prediction case in Fig. 4. The model resolved the conflict between weak personal performance and strong external signals through a reasoning process that prioritized consistent multi-source rich context. This highlights the system's context-aware decision mechanism, where structural alignment with peers and similar students enables accurate inference even in sparse data settings. Full example is provided in the Appendix E.

4.8 Analysis about Report Quality

To evaluate the quality of generated diagnostic reports, we adopt a mixed evaluation protocol that combines LLM-based scoring and expert human judgment. The detailed scoring mechanism is given in the Appendix F. For automatic evaluation, we use DeepSeek-R1 (Liu et al., 2024) as an LLM-based evaluator to score generated reports. For human evaluation, we invite three education ex-

perts to conduct an independent evaluation. The raters are graduate students (M.S./Ph.D.) specializing in AI for Education, each with prior research experience. The final human evaluation scores are obtained by averaging across raters. Due to the fact that pedagogical usefulness and report rigorousness are ultimately intended for human educators and learners, we report a final hybrid score that assigns a higher weight to human evaluation. Specifically, we combine the human and LLM-based scores with a 0.7:0.3 weighting scheme.

To ensure a comprehensive comparison, we selected three representative LLM-based baselines: EFKT as the pioneer of LLM-driven explainable KT, CIKT as the representative of fine-tuning KT framework, and HISE-KT as the state-of-the-art method in interpretable KT. As shown in Tab. 4, our proposed RAG-KT framework achieves the highest scores across most dimensions. Although our framework achieves the best overall quality, CIKT slightly outperforms it in readability. The reason for this is likely CIKT's core prompting methodology, which explicitly instructs the LLM to generate prose-style summaries organized by knowledge concepts. This naturally yields longer, more narrative outputs that provide detailed explanations, making them easier to read even if they are less data-dense. However, our framework strikes a better balance between brevity and informativeness across all dimensions.

5 Discussion of Cost and Efficiency

Despite its strong performance, RAG-KT introduces additional inference overhead compared with lightweight conventional KT models, mainly due to its retrieval-augmented pipeline and LLM-based reasoning stage. As shown in Table 5, RAG-KT requires no additional training cost, while achieving the best AUC of 85.74 with a moderate inference latency of approximately 8 seconds per sample. Our profiling analysis shows that most of this time is spent on the final LLM inference stage, which takes around 6 seconds on average, while graph traversal and multi-view similarity computation contribute only a smaller portion of the overall latency. Compared with recent prompt-based explainable KT methods such as HISE-KT, which requires about 19 seconds per sample and achieves an AUC of 84.31, RAG-KT is substantially more efficient while also delivering better predictive performance. Although its latency remains higher than that of some traditional KT models, RAG-KT provides an important practical advantage beyond prediction accuracy: it additionally generates structured, evidence-grounded diagnostic reports, offering interpretability and actionable feedback for real educational use. This makes the framework especially suitable for scenarios such as after-class diagnosis, homework analysis, and teacher-facing learning support, where informative feedback is often more valuable than strict real-time response. In deployment, the inference cost can be further reduced by serving an open-source LLM locally or replacing the current backbone with a smaller model.

Table 5: Cost and Efficiency

Methods	Traing hours (h)	Inference latency (s)	AUC
EFKT	/	5	61.10
HISE-KT	/	19	84.31
CIKT	24	3	82.27
EPLF	4	1	81.27
LLM-KT	10	3	83.55
2T-KT	20	2.5	82.27
RAG-KT	/	8	85.74

6 Conclusion

In this paper, we proposed RAG-KT, a retrieval-augmented knowledge tracing framework that integrated a multi-source heterogeneous knowledge graph with LLMs. To enable cross-platform unification, RAG-KT introduced Question Group as an intermediate abstraction that aligned semantically and pedagogically similar questions across platforms into a shared representation. By leveraging multi-view retrieval based on this unified space, RAG-KT achieved state-of-the-art performance and better interpretability on multiple datasets. Notably, it maintained strong accuracy and interpretability even in fully cold-start scenarios, thereby demonstrating superior cross-platform generalization and real-world deployability.

Limitations

Despite the promising performance and interpretability of RAG-KT, we acknowledge several limitations that require future investigation. Compared to lightweight DL-based models, RAG-KT involves a multi-step process comprising graph retrieval and LLM generation, which incurs higher latency and API costs. In scenarios with extremely sparse data or poorly defined concept taxonomies where effective semantic alignment is impossible,

the retrieval module may fail to extract meaningful context, potentially degrading performance. Future work will explore distilling RAG-KT into smaller, local models to balance efficiency and performance.

Acknowledgments

This research of Zhiyi Duan was funded by the National Natural Science Foundation of China (No. 62567005), and Natural Science Foundation of Inner Mongolia Autonomous Region of China (No. 2025MS06004). This research of Rui Liu was funded by the General Program (No.62476146) of the National Natural Science Foundation of China, the Young Elite Scientists Sponsorship Program by CAST (2024QNRC001), the Outstanding Youth Project of Inner Mongolia Natural Science Foundation (2025JQ011), the Key R&D and Achievement Transformation Program of Inner Mongolia Autonomous Region (2025YFHH0014), the Central Government Fund for Promoting Local Scientific and Technological Development (2025ZY0143).

References

Ghodai Abdelrahman, Sherif Abdelfattah, Qing Wang, and Yu Lin. 2022. Dbe-kt22: A knowledge tracing dataset based on online student evaluation. *arXiv preprint arXiv:2208.12651*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jinze Bai, Shuai Bai, Yunfei Chu, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Indronil Bhattacharjee and Christabel Wayllace. 2025. Cold start problem: An experimental study of knowledge tracing models with new students. *arXiv preprint arXiv:2505.21517*.

Songlin Chen, Weicheng Wang, Xiaoliang Chen, Peng Lu, Zaiyan Yang, and Yajun Du. 2024. Llama-lora neural prompt engineering: A deep tuning framework for automatically generating chinese text logical reasoning thinking chains. *DATA INTELLIGENCE*, 6(2):375–408.

Ke Cheng, Linzhi Peng, Pengyang Wang, Junchen Ye, Leilei Sun, and Bowen Du. 2024. Dygkt: Dynamic graph learning for knowledge tracing. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 409–420.

Albert T Corbett and John R Anderson. 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction*, pages 253–278.

Xinyi Ding and Eric C Larson. 2019. Why deep knowledge tracing has less depth than anticipated. *International Educational Data Mining Society*.

Zhiyi Duan, Zixing Shi, Hongyu Yuan, and Qi Wang. 2025. Hise-kt: Synergizing heterogeneous information networks and llms for explainable knowledge tracing with meta-path optimization. *arXiv preprint arXiv:2511.15191*.

Mingyu Feng, Neil Heffernan, and Kenneth Koedinger. 2009. Addressing the assessment challenge with an online system that tutors as it assesses. *User modeling and user-adapted interaction*, pages 243–266.

Dieter Fox, Wolfram Burgard, and Sebastian Thrun. 2002. The dynamic window approach to collision avoidance. *IEEE robotics & automation magazine*, pages 23–33.

Aritra Ghosh, Neil Heffernan, and Andrew S Lan. 2020. Context-aware attentive knowledge tracing. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2330–2339.

Xiaoshuai Hao, Lei Zhou, and 1 others. 2025. Mimo-embodied: X-embodied foundation model technical report. *arXiv preprint arXiv:2511.16518*.

Heeseok Jung, Jaesang Yoo, Yohaan Yoon, and Yeonju Jang. 2024. Clst: Cold-start mitigation in knowledge tracing by aligning a generative language model as a students' knowledge tracer. *arXiv preprint arXiv:2406.10296*.

JongWoo Kim, SeongYeub Chu, Bryan Wong, and Mun Yi. 2024. Beyond right and wrong: Mitigating cold start in knowledge tracing using large language model and option weight. *arXiv e-prints*, pages arXiv–2410.

Patrick Lewis, Ethan Perez, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Haoxuan Li, Jifan Yu, Yuanxin Ouyang, Zhuang Liu, Wenge Rong, Juanzi Li, and Zhang Xiong. 2024. Explainable few-shot knowledge tracing. *arXiv preprint arXiv:2405.14391*.

Linqing Li, Zhifeng Wang, Joemon M Jose, and Xuri Ge. 2025a. Llm supporting knowledge tracing leveraging global subject and student specific knowledge graphs. *Information Fusion*, page 103577.

Runze Li, Siyu Wu, Jun Wang, and Wei Zhang. 2025b. Cikt: A collaborative and iterative knowledge tracing framework with large language models. *arXiv preprint arXiv:2505.17705*.

Zongxi Li, Zijian Wang, Weiming Wang, Kevin Hung, Haoran Xie, and Fu Lee Wang. 2025c. Retrieval-augmented generation for educational application: A systematic survey. *Computers and Education: Artificial Intelligence*, page 100417.

Aixin Liu, Bei Feng, Bing Xue, Wang, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Yunfei Liu, Yang Yang, Xianyu Chen, Jian Shen, Haifeng Zhang, and Yong Yu. 2020. Improving knowledge tracing via pre-training question embeddings. *arXiv preprint arXiv:2012.05031*.

Zitao Liu, Qiongqiong Liu, Jiahao Chen, Shuyan Huang, Boyu Gao, Weiqi Luo, and Jian Weng. 2023. Enhancing deep knowledge tracing with auxiliary tasks. In *Proceedings of the ACM web conference 2023*, pages 4178–4187.

Sein Minn, Jill-Jênn Vie, Koh Takeuchi, Hisashi Kashima, and Feida Zhu. 2022. Interpretable knowledge tracing: Simple and efficient student modeling with causal relations. In *Proceedings of the AAAI conference on artificial intelligence*, pages 12810–12818.

Hiromi Nakagawa, Yusuke Iwasawa, and Yutaka Matsuo. 2019. Graph-based knowledge tracing: modeling student proficiency using graph neural network. In *IEEE/WIC/aCM international conference on web intelligence*, pages 156–163.

Seyed Parsa Neshaei, Richard Lee Davis, Adam Hazimeh, Bojan Lazarevski, Pierre Dillenbourg, and Tanja Käser. 2024. Towards modeling learner performance with large language models. *arXiv preprint arXiv:2403.14661*.

Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J Guibas, and Jascha Sohl-Dickstein. 2015. Deep knowledge tracing. *Advances in neural information processing systems*, 28.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2008. The graph neural network model. *IEEE transactions on neural networks*, pages 61–80.

Shuanghong Shen, Qi Liu, Zhenya Huang, Yonghe Zheng, Minghao Yin, Minjuan Wang, and Enhong Chen. 2024. A survey of knowledge tracing: Models, variants, and applications. *IEEE Transactions on Learning Technologies*, pages 1858–1879.

Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S Yu, and Qingsong Wen. 2024. Large language models for education: A survey and outlook. *arXiv preprint arXiv:2403.18105*.

Zichao Wang, Angus Lamb, Evgeny Saveliev, and 1 others. 2020. Instructions and guide for diagnostic questions: The neurips 2020 education challenge. *arXiv preprint arXiv:2007.12061*.

Ziwei Wang, Jie Zhou, Qin Chen, Min Zhang, Bo Jiang, Aimin Zhou, Qinchun Bai, and Liang He. 2025. Llm-kt: Aligning large language models with knowledge tracing using a plug-and-play instruction. *arXiv preprint arXiv:2502.02945*.

Chun-Kit Yeung. 2019. Deep-irt: Make deep learning based knowledge tracing explainable using item response theory. *arXiv preprint arXiv:1904.11738*.

Chun Kit Yeung and Dit Yan Yeung. 2018. Addressing two problems in deep knowledge tracing via prediction-consistent regularization. In *Proceedings of the 5th ACM Conference on Learning @ Scale*, pages 5:1–5:10. ACM.

Jiani Zhang, Xingjian Shi, Irwin King, and Dit-Yan Yeung. 2017. Dynamic key-value memory networks for knowledge tracing. In *Proceedings of the 26th international conference on World Wide Web*, pages 765–774.

Suifeng Zhao, Tong Zhou, Zhuoran Jin, Hongbang Yuan, Yubo Chen, Kang Liu, and Sujian Li. 2024. Awecita: Generating answer with appropriate and well-grained citations using llms. *DATA INTELLIGENCE*, 6(4):1134–1157.

A KC Matching Methodology

To address the challenge of varied terminologies for the same KC across different datasets (e.g., Pythagorean Theorem vs. Gougu Theorem), we implemented a pipeline to align all dataset-specific KCs (KC^d) with a canonical set of concepts in our pre-constructed $\mathcal{G}_{\mathcal{KC}}$ (KC^b). To achieve this, we designed and employed a two-stage hybrid matching pipeline.

Embedding-based Automatic Matching: First, we utilize a embedding model to generate vector representations for all concepts in both KC^b and KC^d. We then compute the cosine similarity for each (KC^d_j, KC^b_i) pair. If a pair exhibits the highest similarity score for a given KC^d_j and this score exceeds a threshold of 0.85, we consider it a high-confidence match. This threshold is the best split point obtained by testing on a small validation set.

LLM-based Semantic Alignment: For pairs that fall below the threshold, or for more nuanced cases involving semantic equivalence despite lexical differences, we leverage a Large Language Model (LLM), such as GPT-4, for deep semantic alignment. We engineered a structured prompt that instructs the LLM to determine if two concepts are equivalent based on four strict criteria:

1. Pedagogical Equivalence: Are the concepts pedagogically equivalent or do they have a very high degree of content overlap?

2. Syllabus Coherence: Are they typically classified under the same specific module in an educational syllabus?

3. Core Skill Identity: Do they teach the same fundamental mathematical essence or target the same core skills?

4. Exclusion of Weak Relations: The relationship must be one of equivalence, not merely topical relevance or partial overlap.

We explicitly evaluate alignment accuracy by first asking domain experts to manually align the KCs between the source and target platforms as ground truth. We then run our KC Match procedure and compare its outputs against this expert annotation. The resulting alignment accuracies are as shown in Table 6.

Through this carefully designed prompt, the LLM utilizes its extensive domain knowledge to accurately map the remaining ambiguous concepts.

Table 6: KC Match Accuracy.

Source	Target	Accuracy
ASSIST09	ASSIST12	96.33
ASSIST09	DBE-KT22	92.67
ASSIST12	DBE-KT22	96.50

This two-stage hybrid approach ensures that our KC alignment process is both efficient for straightforward cases and accurate for complex semantic challenges.

B Structured Reliable-Context-based Prompt

The complete Structured Reliable-Context-based Prompt is shown in Fig. 5.

C Baselines

Specific descriptions of each baseline:

- **DL-based Methods**: These methods represent mainstream deep learning models for knowledge tracing:
  - **DKT**: A foundational RNN-based KT model that encodes student response sequences for prediction.
  - **DKT+**: An extension of DKT that incorporates regularization and reconstruction losses to mitigate forgetting.
  - **AT-DKT**: Integrates attention mechanisms to better capture dependencies in student sequences.
  - **DKVMN**: A memory-augmented model that represents student knowledge states using key-value memory networks.
  - **Deep-IRT**: Combines IRT principles with deep learning to improve modeling of student ability and item properties.
  - **AKT**: Employs self-attention mechanisms and concept-aware modeling to capture temporal and contextual dynamics.
  - **GKT**: Incorporates a graph structure to model skill dependencies and student progress.
- **LLM-based Methods (Prompting)**: These methods employ LLMs in zero-shot or few-shot settings without parameter updates:

  - **EFKT**: Tracks knowledge states through few-shot prompting and generates natural language explanations.
  - **LOKT**: Assigns differentiated weights to multiple-choice options to better capture students' knowledge mastery, and serves as a comparison method for the fully cold-start setting in this study.
  - **HISE-KT**: Synergizes heterogeneous information networks with LLMs, employing LLM-powered meta-path optimization and similar student retrieval to achieve accurate zero-shot prediction and reliable-context-based explanations.

- **LLM-based Methods (Fine-tuning)**: These models involve explicit training or fine-tuning of LLMs for KT tasks:
  - **LLM-KT**: Proposes a plug-and-play prompting approach combining behavioral traces and textual context.
  - **EPLF**: Evaluates LLMs' zero-shot and fine-tuning ability to perform KT. We choose its fine-tuning setting for comparing.
  - **CIKT**: Collaboratively fine-tunes two LLMs, a predictor and an analyst, to generate and use interpretable knowledge state descriptions.
  - **CLST**: Reformulates the KT task as a natural language modeling problem using LLMs, and is used in this work to evaluate performance under fully cold-start scenarios.
  - **2T-KT**: Leverages LLMs to simulate a teacher's thinking mode combined with knowledge graphs to address the new knowledge concept prediction problem.

D HyperParams Results

The complete HyperParams results are shown in Table 7.

E Case Study

As shown in Fig. 6 this section details A case study demonstrating the RAG-KT compared to other models (EFKT and CIKT) in a cold-start scenario. The student answers this question correctly.

F Scoring Mechanism

The quality of the generated analysis reports was evaluated based on four dimensions: Explainability, Readability, Educational Usefulness, and Rigorousness. Each dimension was scored on a 1 to 5 scale, with detailed criteria provided in the Table 8.

Structured Reliable-Context-based Prompt

Task Description:
You are a professional educational data analyst. Your core task is to predict the probability that the target student will answer the given question correctly, based on the multi-dimensional historical data provided below. You are also required to generate a professional, data-driven comprehensive analysis report.
Your analysis should go beyond simple numerical matching, delve deeply into the meaning behind the data, evaluate the reliability of the evidence, and identify key factors influencing the prediction result as well as potential risks.

Concept Definitions:
- Ability: An assessment of the student's overall competence, used to identify and validate similar student groups.
- Difficulty: A numerical indicator of how challenging a question or question group is. Higher values indicate greater difficulty, requiring deeper and broader knowledge from the student.
- Discrimination: An index measuring how effectively a question distinguishes between students of varying ability levels. A highly discriminative item results in significantly higher accuracy for high-ability students compared to low-ability ones; otherwise, the item is considered poor at differentiation.
- Knowledge Point: The smallest cognitive unit in the subject matter, serving as the foundation for a student's knowledge structure.
- Question Group: A set of questions organized around a specific knowledge point and difficulty level.
- Accuracy: The number of correct responses divided by the total number of attempts within a specific question group or knowledge point. It reflects historical overall performance without distinguishing between recent and older records.
- Dynamic Weighted Accuracy (DWA): An advanced accuracy metric that assigns greater weight to more recent responses, simulating learning and forgetting curves. Compared to standard accuracy, it more accurately reflects the student's current true level. High DWA indicates good recent performance; low DWA may reflect recent struggles or forgotten knowledge.
- Attempt Count: When the sample size is small (e.g., fewer than 3 attempts), the calculated accuracy and related metrics may be unstable due to randomness and should be interpreted with caution.
- Confidence for DWA: A score indicating the reliability of DWA, derived from a combination of attempt count and the consistency of past performance.
- Attempt History: A time-ordered sequence of responses for a specific question group or knowledge point, visually demonstrating the student's learning trajectory and performance fluctuations.
- Similar User: A group of students whose learning behavior patterns closely match the target student's, based on response history, ability scores, and other features. Their performance provides critical contextual references for the target student's prediction.
- Similarity: A measure of behavioral resemblance between students. The higher the value, the greater the reference value of the peer's data.
- Learning Momentum: A dynamic indicator reflecting the student's recent learning status and improvement trends. It is evaluated primarily in two ways:
    - DWA vs. Accuracy: When DWA is significantly higher than accuracy, it indicates recent performance improvement above long-term average, suggesting positive learning momentum. Conversely, a lower DWA may suggest forgetting or decline.
    - Attempt History: Sequences like [wrong, wrong, correct, correct, correct] clearly show a shift from errors to sustained correct responses, serving as strong signals of positive learning momentum.
- Associated Knowledge Point: Other knowledge points logically related to the target knowledge point, which are further divided into:
    - Predecessor: Foundational knowledge points that must be mastered before learning the target point, forming the logical base for deep understanding.
    - Associative: Knowledge points related in content but without a strict prerequisite relationship; mastering them aids in analogical understanding.

Output Requirements:
The <result> section is the final professional analysis report. Its language must be professional, concise, and data-driven, avoiding colloquial expressions. All analytical judgments must cite key supporting data explicitly. Do not include meaningless parentheses.

Output Format:
<result>
1. Prediction Outcome
- Core Prediction: Clearly state the predicted probability of the student answering the question correctly (e.g., Predicted Accuracy: XX%).
- Qualitative Judgment: Based on the accuracy in the core prediction, select the most appropriate level from the five defined categories. Use the exact descriptions below:
    - >80%: Expected to answer correctly with stability.
    - 60%-79%: High likelihood of answering correctly.
    - 40%-59%: Equal chance of correct and incorrect.
    - 20%-39%: High risk of incorrect response.
    - <20%: Expected to struggle with the question.
2. Student Ability Analysis
- Overall Ability Assessment:
    - Briefly describe the student's ability level (e.g., high, above average) and compare it to peer groups, clearly stating their relative position (e.g., at the average level, slightly above, slightly below).
- Knowledge Mastery Structure:
    - Target Knowledge Point: Clearly assess the student's mastery of the target knowledge point. You **must** state the **reliability** of this assessment (high/low), citing attempt count and confidence score.
    - Instructional Suggestion: If mastery is low or reliability is weak, recommend specific actions (e.g., increase exposure through medium-difficulty items, reinforce with visual examples, remedial instruction).
- Bottleneck Analysis:
    - Use DWA data across difficulty levels to determine whether the student exhibits a disconnect between knowledge and application, and whether this question may trigger that challenge.
    - Instructional Suggestion: If an application bottleneck is detected, suggest targeted training (e.g., scaffolding with graduated difficulty, multi-step problem tasks, integration with real-world contexts).
3. Prediction Attribution and Decision
- Key Positive Factors: List 1–2 core pieces of evidence supporting the student's likelihood of answering the question correctly, with corresponding data.
- Key Risk Factors: List 1–2 core risks that may lead to an incorrect answer, with corresponding data.
- Final Decision Logic: This is the core of the report and must clearly explain how the model balanced positive and risk factors.
    - Conflict Clarification: If a signal conflict exists (e.g., "strong knowledge mastery but low ability bottleneck"), it must be explicitly stated first.
    - Principle Application: Clearly explain which principle underpinned your decision.
    - Conclusion Formation: Based on the above decisions, synthesize and report the final predicted probability.
- Instructional Summary: Briefly summarize the instructional focus based on this prediction (e.g., "Maintain challenge level and monitor transfer to adjacent skills," or "Pause progression; reinforce pre-requisites and mid-tier problems before retry").
</result>
The final professional analysis report must be contained within <result></result> tags.

Retrieval Results:
1. Target Student and Target Question Information:
- Target student's ability estimate:
- Target student's ability level:
- Target question's associated question group:
- Target question's difficulty:
- Target question's discrimination:
- Knowledge point assessed by the target question:
2. Core Individual Performance Data:
- Performance on target knowledge point:
- Performance by difficulty level:
- Performance on target question group:
- Performance on related knowledge points:
3. External Reference and Calibration Data:
- List of similar students:
- Similar students' performance on the target knowledge point:
- Similar students' performance at the target difficulty level:
- Similar students' performance on the target question group:
- Similar students' performance on related knowledge points:
- Performance on the target question group by all students with similar ability to the target student:
4. Historical and Trend Data:
- Target student's answer history:

Figure 5: Structured Reliable-Context-based Prompt.

Interaction History

(K_150, x-y graph reading, D_2, wrong), (K_150, x-y graph reading, D_2, wrong), (K_63, changes in the direction of figures, D_2, correct), (K_166, distance formula, D_2, correct),
(K_150, x-y graph reading, D_2, wrong), (K_166, distance formula, D_2, wrong), (K_150, x-y graph reading, D_2, wrong), (K_166, distance formula, D_2, wrong), (K_150, x-y graph reading, D_2, wrong),
(K_150, x-y graph reading, D_2, wrong), (K_63, changes in the direction of figures, D_2, correct), (K_63, changes in the direction of figures, D_2, correct), (K_166, distance formula, D_2, wrong),
(K_150, x-y graph reading, D_1, wrong), (K_166, distance formula, D_2, correct), (K_150, x-y graph reading, D_3, wrong), (K_150, x-y graph reading, D_3, wrong), (K_150, x-y graph reading, D_1, wrong),
(K_63, changes in the direction of figures, D_2, wrong), (K_150, x-y graph reading, D_1, wrong), (K_166, distance formula, D_2, correct), (K_166, distance formula, D_2, wrong),
(K_150, x-y graph reading, D_3, wrong), (K_150, x-y graph reading, D_3, correct)

Target Question

(K_13, greatest common factor, D_2)

EFKT

1. Knowledge points involved in the new question
- The new exercise contains: K_13 (greatest common factor)

2. Link between the new question and previous work
- It's a new question.
- There is no direct connection between the previous knowledge points (x-y graph reading, changes in direction of figures, distance formula) and this question (greatest common factor). The topic is distinct and has not appeared in the student's history.

3. Student's updated knowledge state
Based on historical performance:
- K_150 (x-y graph reading): bad (1 correct / 12 attempts; consistent errors, last correct was isolated)
- K_63 (changes in direction of figures): fair (3 correct / 4 attempts; mostly successful but one recent error)
- K_166 (distance formula): bad (3 correct / 7 attempts; inconsistent, last attempt wrong)
- K_13 (greatest common factor): bad (no prior attempts; new topic with no evidence of mastery)

4. Analysis of student's performance on the new question
-The student gets it wrong, likely because of incorrect mastery.
- Reasons:
- The student has no recorded history with K_13 (greatest common factor), indicating no prior exposure or practice.
- Given the student's overall struggle with math topics (e.g., K_150 and K_166 at "bad" mastery), it is unlikely they mastered this new concept spontaneously.
- Carelessness is less probable as the error aligns with foundational gaps in related mathematical operations (e.g., arithmetic fundamentals).

Summary:
The student lacks foundational mastery of K_13, leading to an expected error. Targeted practice on greatest common factors is recommended.

CIKT

1. x-y graph reading (K_150)
- Attempts: 13 questions
- Performance:
- D_1 difficulty: 3 attempts, all incorrect. D_2 difficulty: 6 attempts, all incorrect. D_3 difficulty: 4 attempts, 1 correct (most recent), 3 incorrect.
- Analysis:
- The student struggled consistently across all difficulty levels, with 12 consecutive incorrect responses initially.
- The single correct response (on the last attempt, D_3) suggests a potential breakthrough after repeated practice, though mastery remains uncertain.
- Overall, this topic is a significant weakness requiring focused reinforcement.

2. changes in the direction of figures (K_63)
- Attempts: 4 questions
- Performance: D_2 difficulty: 3 correct, 1 incorrect (most recent attempt).
- Analysis:
- Strong initial performance (3 correct) but the latest attempt was incorrect, indicating possible inconsistency or lapse in understanding.
- The student handles D_2 questions well generally but may need practice to solidify reliability.

3. distance formula (K_166)
- Attempts: 7 questions
- Performance: D_2 difficulty: 3 correct, 4 incorrect (including the most recent attempt).
- Analysis:
- Inconsistent results: Correct responses interspersed with errors, showing partial understanding but lack of stability.
- The final attempt was incorrect, suggesting unresolved gaps or application issues under pressure.

Next Question
- Topic: "Greatest common factor" (K_13)
- Difficulty: D_2
- Note: This topic has not appeared in the student's history. Performance here will establish a baseline for this new knowledge area.

Overall Summary
- Strengths: Moderate competence in "changes in the direction of figures" (K_63), though recent errors warrant monitoring.
- Weaknesses:
- Severe difficulties in "x-y graph reading" (K_150) despite extensive practice. The lone late success is promising but inconclusive.
- Inconsistency in "distance formula" (K_166), indicating fragile knowledge application.
- Recommendations:
- Priority intervention for K_150 (x-y graph reading), especially D_1/D_2 fundamentals.
- Reinforce K_63 and K_166 with mixed-difficulty questions to improve reliability.
- Closely monitor performance on the new topic (K_13) to assess baseline understanding.

Conclusion:
The student exhibits partial grasp of K_63 and K_166 but faces critical challenges in K_150. The recent correct response in K_150 (D_3) may signal emerging understanding, yet the error-prone history demands caution. Targeted practice on weaker areas and consistent evaluation of new topics are advised.
Result: Wrong

RAG-KT

Reliable-Context-based Analysis Report:

1. Prediction Outcome
- Core Prediction: Predicted Accuracy: 57.00%
- Qualitative Judgment: Equal chance of correct and incorrect.

2. Student Ability Analysis
- Overall Ability Assessment: The target student's ability estimate is 0.38, at the MIDDLE level, which is consistent with the average level of peer groups with similar ability.
- Knowledge Mastery Structure: Target Knowledge Point (S_13, greatest common factor): There is no direct performance data (Dynamic Weighted Accuracy, Accuracy, Attempt Count, and Confidence are all unavailable), so the assessment of mastery is unreliable (low reliability).
- Bottleneck Analysis: The student's Dynamic Weighted Accuracy for difficulty level 2 is 38.81 (17 attempts, Confidence 49.89); for difficulty level 3, it is 29.08 (4 attempts, Confidence 40.00). This shows a performance decline with increasing difficulty, indicating a knowledge application bottleneck. The target question has a difficulty level of 2, which may trigger this challenge given the student's low DWA at this difficulty.
- Instructional Suggestion: Implement graduated problem-solving scaffolds focusing on multi-step GCF applications and real-world contextualization to bridge knowledge-application gaps.

3. Prediction Attribution and Decision
- Key Positive Factors:
- Students with similar ability to the target student have a 74.42% accuracy on QG_13_2 (sample size 13564), providing strong positive reference due to the large sample.
- A similar student (U_asst12190138, similarity 95.23) has a 57.52% DWA on S_13 (8 attempts, Confidence 77.35), reflecting reliable performance on the target knowledge point.
- Key Risk Factors:
- The target student has no direct performance data on S_13, lacking evidence for core knowledge mastery.
- The student's DWA at difficulty level 2 (target question difficulty) is 38.81 (Confidence 49.89), indicating unstable recent performance at this difficulty.
- Final Decision Logic:
- Conflict Clarification: A conflict exists between the high accuracy of similar ability groups on QG_13_2 (74.42%) and the target student's low DWA at difficulty level 2 (38.81%); meanwhile, the similar student's S_13 performance contrasts with the target student's lack of S_13 data.
- Principle Application: The decision prioritizes comprehensive reference integration, balancing large-sample peer data, similar user performance, and individual difficulty-level performance.
- Conclusion Formation: Positive factors (peer group accuracy, similar user's S_13 DWA) slightly outweigh risk factors (no S_13 data, low individual DWA at difficulty 2), resulting in a 57.00% predicted accuracy.
- Instructional Summary: Proceed with target question but pair with real-time feedback. Prioritize foundational reinforcement of S_13 through tiered difficulty problems before advancing to associative concepts. Monitor application bottlenecks via DWA trends at difficulty 2.

Figure 6: Case Study.

Table 7: Complete HyperParams results.

HyperParams	Value	ASSIST09			ASSIST12			DBE-KT22		
		ACC	AUC	F1	ACC	AUC	F1	ACC	AUC	F1
seq_len	5	76.77	82.03	83.62	69.34	68.54	78.52	76.68	74.59	84.58
	10	78.48	**83.78**	85.15	70.37	72.55	78.32	78.19	75.62	85.81
	25	**80.00**	**85.74**	**85.75**	70.37	70.86	79.28	**78.89**	76.32	**86.80**
	50	78.28	82.82	85.14	**74.80**	**73.89**	**82.6**	**77.98**	**76.84**	85.59
	100	**79.90**	83.04	**86.04**	**75.83**	**75.75**	**82.76**	**78.89**	**78.72**	**86.60**
top_k	1	**79.79**	85.31	85.55	73.79	72.85	**82.24**	77.47	75.06	85.22
	2	**80.00**	**85.74**	**85.75**	**74.80**	**73.89**	**82.60**	**78.89**	**76.32**	**86.80**
	3	79.51	**85.44**	85.29	**74.20**	**73.72**	81.17	**78.51**	75.47	**85.70**
	4	79.10	85.14	85.06	74.16	73.26	80.04	78.00	**75.67**	85.33
	5	78.80	84.97	**85.82**	73.18	73.19	80.88	77.65	75.52	85.25
$\lambda_1 : \lambda_2 : \lambda_3$	1 : 1 : 1	79.21	**85.41**	85.13	73.95	73.52	81.98	78.13	**76.15**	86.24
	2 : 1 : 1	**79.72**	85.30	**85.45**	**74.51**	**73.66**	**82.31**	**78.54**	75.98	**86.59**
	4 : 3 : 3	**80.00**	**85.74**	**85.75**	**74.80**	**73.89**	**82.60**	**78.89**	**76.32**	**86.80**
	3 : 4 : 3	79.58	85.15	85.24	74.33	73.28	82.07	78.21	75.74	86.22
	3 : 3 : 4	79.61	85.23	85.39	74.45	73.41	82.15	78.43	75.81	86.48

Table 8: Scoring Mechanism.

Dimension	Score	Description
Explainability	1	Lacks any reasoning or presents a chaotic causal chain.
	2	Vague causal logic that is difficult to understand.
	3	Partially valid reasoning but lacks overall coherence.
	4	The reasoning chain is largely complete and causal relationships are clear.
	5	Fully reveals the reasoning process, accurately explaining why a prediction was made.
Readability	1	Chaotic, obscure, or incomprehensible language.
	2	Verbose, with a disorganized and messy structure.
	3	Largely clear language, but with logical leaps or excessive jargon.
	4	Clear and well-structured expression with appropriate use of terminology.
	5	Fluent, logically coherent, and concise language that is easy for the reader to understand.
Educational Usefulness	1	Offers no educational value and contains only generic statements.
	2	Identifies issues too vaguely to provide guidance for students or teachers.
	3	Points out specific problems and provides a preliminary analysis.
	4	Clearly pinpoints a student's issues and offers insightful feedback.
	5	Provides highly targeted and actionable suggestions that are significantly helpful for teaching and learning.
Rigorousness	1	Content is subjective, reasoning is arbitrary, and lacks any factual support.
	2	Provides some context, but the argumentation is vague and unconvincing.
	3	Generally context-based, but lacks detail or has a loose logical structure.
	4	Supported by sufficient context, with meticulous logic and clear details.
	5	Features a rigorous reasoning structure where all conclusions are explicitly supported by clear context, with no logical fallacies.