HISE-KT: Synergizing Heterogeneous Information Networks and LLMs for Explainable Knowledge Tracing with Meta-Path Optimization

Zhiyi Duan[1], Zixing Shi[1], Hongyu Yuan[1], Qi Wang[2*]

[1]Department of Computer Science, Inner Mongolia University, Hohhot, Inner Mongolia
[2]School of Artificial Intelligence, Jilin University, Changchun, Jilin
duanzy@imu.edu.cn, zixingshi@mail.imu.edu.cn, yuanhongyu_1997@163.com, qiwang@jlu.edu.cn

Abstract

Knowledge Tracing (KT) aims to mine students' evolving knowledge states and predict their future question-answering performance. Existing methods based on heterogeneous information networks (HINs) are prone to introducing noises due to manual or random selection of meta-paths and lack necessary quality assessment of meta-path instances. Conversely, recent large language models (LLMs)-based methods ignore the rich information across students, and both paradigms struggle to deliver consistently accurate and evidence-based explanations. To address these issues, we propose an innovative framework, *HIN-LLM Synergistic Enhanced Knowledge Tracing (HISE-KT)*, which seamlessly integrates HINs with LLMs. HISE-KT first builds a multi-relationship HIN containing diverse node types to capture the structural relations through multiple meta-paths. The LLM is then employed to intelligently score and filter meta-path instances and retain high-quality paths, pioneering automated meta-path quality assessment. Inspired by educational psychology principles, a similar student retrieval mechanism based on meta-paths is designed to provide a more valuable context for prediction. Finally, HISE-KT uses a structured prompt to integrate the target student's history with the retrieved similar trajectories, enabling the LLM to generate not only accurate predictions but also evidence-backed, explainable analysis reports. Experiments on four public datasets show that HISE-KT outperforms existing KT baselines in both prediction performance and interpretability.

Introduction

Knowledge Tracing (KT), as a core task of educational data mining and intelligent education systems, aims to dynamically model the knowledge state of students based on their historical interaction sequences and to predict their performance on future learning content (Corbett and Anderson 1994). Accurate KT models are critical for personalized learning-path recommendation, adaptive learning-resource allocation, and precise instructional interventions, serving as key technological support for improving educational quality and efficiency (Shen et al. 2024; Duan et al. 2024).

Recently, researchers have attempted to leverage heterogeneous information networks (HINs) to enhance both the performance and interpretability of KT models (Shi et al. 2018; Xu et al. 2023). A HIN can naturally integrate various types and levels of entities in educational scenarios (e.g., students, questions, and knowledge concepts) and their rich interaction relationships (e.g., answering, inclusion, and assessment), providing a structured representation framework for modeling complex learning processes (Sun et al. 2024). By defining meta-paths to characterize specific semantic relationship patterns among entities, HINs can effectively capture the implicit and complex associations in learning processes (Sun et al. 2011). However, existing HIN-based KT methods face significant challenges. First, meta-path selection often relies on expert knowledge or random walks, inevitably producing redundant or low-information path instances. These instances increase computational overhead and introduce noise, ultimately degrading model performance (Meng et al. 2015; Liu et al. 2023b). Second, model interpretability is typically limited to the network-structure level (e.g., path weights), making it difficult to provide high-level explanations that are semantically clear and aligned with human cognition. These issues limit the credibility and applicability of such models in practical educational decision-making.

Meanwhile, large language models (LLMs) have demonstrated remarkable performance across various tasks due to their powerful capabilities in semantic understanding, contextual reasoning, and natural language generation (Zhao et al. 2023; Achiam et al. 2023; Zhou et al. 2024; Guo et al. 2025). Preliminary studies applying LLMs to KT have shown encouraging progress. LLMs possess a natural advantage in generating explanations in natural language, which may enhance the interpretability of KT models. However, existing LLM-based KT methods primarily rely on instruction fine-tuning or prompt engineering over answering sequences (Li et al. 2025a; Jung et al. 2024), failing to structurally model the complex higher-order interactions among students, questions, and knowledge concepts (e.g., indirect associations formed via mediating nodes such as students' abilities and question difficulties). This limitation hampers the capture of deeper learning patterns, constrains performance gains, and induces behavioral explanation hallucinations due to evidence scarcity.

To address these limitations, we propose an innovative *HIN-LLM Synergistic Enhanced Knowledge Tracing (HISE-*

*Corresponding author.

KT) framework, which integrates the structured-relationship strengths of HINs with the semantic understanding, reasoning, and natural-language generation capabilities of LLMs to improve both performance and interpretability. Specifically, to represent the complex interaction relationships in the learning process, we first construct a multi-relationship HIN containing multiple node types and multi-dimensional semantic relationships. Then, an LLM-based scoring method is proposed to perform multi-dimensional semantic evaluation of the generated meta-path instances, mitigating the randomness and redundancy of traditional meta-path selection. The Top-K meta-paths with the highest discriminative power and information content are retained to construct an initial candidate student retrieval set. Relative Evaluation Theory argues that assessing a student against similar peers filters out common noise and highlights genuine ability differences (Gibbons and Murphy 1990). Social Comparison Theory further posits that individuals gauge their own competence by comparing themselves to those with comparable traits, yielding the most meaningful benchmarks (Suls and Wheeler 2012). Guided by these, a rule-based collaborative filtering that considers path matching, knowledge state similarity, and historical performance trends is proposed to select the Top-S most relevant similar students. The historical trajectories of these students supply valuable, comparable context for the target learner. Building on this context, a structured prompt is designed to enable zero-shot KT prediction and generate an explainable analysis report, which clarifies the prediction basis, highlights potential learning difficulties, and offers actionable improvement suggestions, thereby enhancing decision transparency. **The major contributions are summarized as follows**:

- We propose HISE-KT, the first KT framework that synergizes structured relational modeling with HINs and generative semantic intelligence with LLMs, establishing a new paradigm for KT.
- We establish an LLM-powered meta-path optimization mechanism, where semantic scoring and selective filtering automatically eliminate noisy paths while preserving meaningful interactions, overcoming redundancy and inefficiency issues in HIN-based KT.
- We develop a rule-based collaborative filtering mechanism inspired by Relative Evaluation and Social Comparison Theory, providing contextual anchors that mitigate LLM hallucinations and enable traceable explanations.
- Experiments on multiple publicly available educational datasets verify that HISE-KT outperforms state-of-the-art baselines in both predictive performance and explanation quality.

Related Work

HIN-Based KT Methods

Recent studies have tried to use HINs to model the relationships in the learning process. STHKT (Li et al. 2025b) constructs heterogeneous graphs and combines topological Hawkes processes with graph convolutional networks to fuse spatiotemporal information. MGEKT (Qiu and Wang 2024) uses meta-paths to capture higher-order semantics in a heterogeneous graph, while integrating a gated attention mechanism to model both student interactions and long-term dependencies. SimQE (Sun et al. 2024) captures the similarity of questions through biased random walks of meta-paths on a weighted HIN to enhance question representation. However, these methods model entities in KT relatively simplistically and do not evaluate the quality of meta-path instances, resulting in redundant samples that hinder the modeling of truly discriminative high-order relationships.

LLM-Based KT Methods

LLM is introduced into KT to improve interpretability due to its excellent semantic understanding and generation capabilities. Some works employ instruction fine-tuning of LLMs to directly learn from student interaction data (e.g., CLST (Jung et al. 2024), LLM-KT (Wang et al. 2025), CIKT (Li et al. 2025a)). LLM-KT performs instruction fine-tuning on the LLMs via Low-Rank Adaptation, encapsulating student response history and question information into structured prompts while integrating sequential behavior embeddings with textual context to enhance knowledge tracing performance (Wang et al. 2025). Meanwhile, other works achieve KT by constructing structured prompts (e.g., EFKT (Li et al. 2024), LOKT (Kim et al. 2024)). EFKT utilizes few-shot prompting to leverage the reasoning and generation capabilities of LLMs for completing KT with limited practice records and generating interpretable natural language prediction reports (Li et al. 2024). However, these methods rely solely on a single student's historical interaction data for KT prediction, without considering the cross-student interaction information, which may lead to hallucinations and reduce the reliability of predictions and explanations.

Methodology

In this section, we introduce HISE-KT, the overall framework is shown in Fig. 1. The framework comprises four sequential modules: (1) MRHIN Construction; (2) LLM-Powered Meta-Path Optimization; (3) Rule-Based Collaborative Filtering; (4) LLM Prediction.

Multi-Relationship Heterogeneous Information Network Construction Module

This module aims to construct a Multi-Relationship Heterogeneous Information Network (MRHIN) to better model the structured relationships between students, questions, and knowledge concepts. Existing HINs often only consider the shallow relationships between students, questions, and knowledge concepts, which ignore student attributes and question attributes. Therefore, we construct MRHIN, which contains five types of nodes: student (U), question (Q), knowledge concept (K), student ability (A), and question difficulty (D). There are four types of edges: question-student ($Q-U$), question-knowledge concept ($Q-K$), question-question difficulty ($Q-D$), and student-student ability ($U-A$). These edges are used to model the deep structured relationships between nodes, forming a multi-relationship heterogeneous information network.

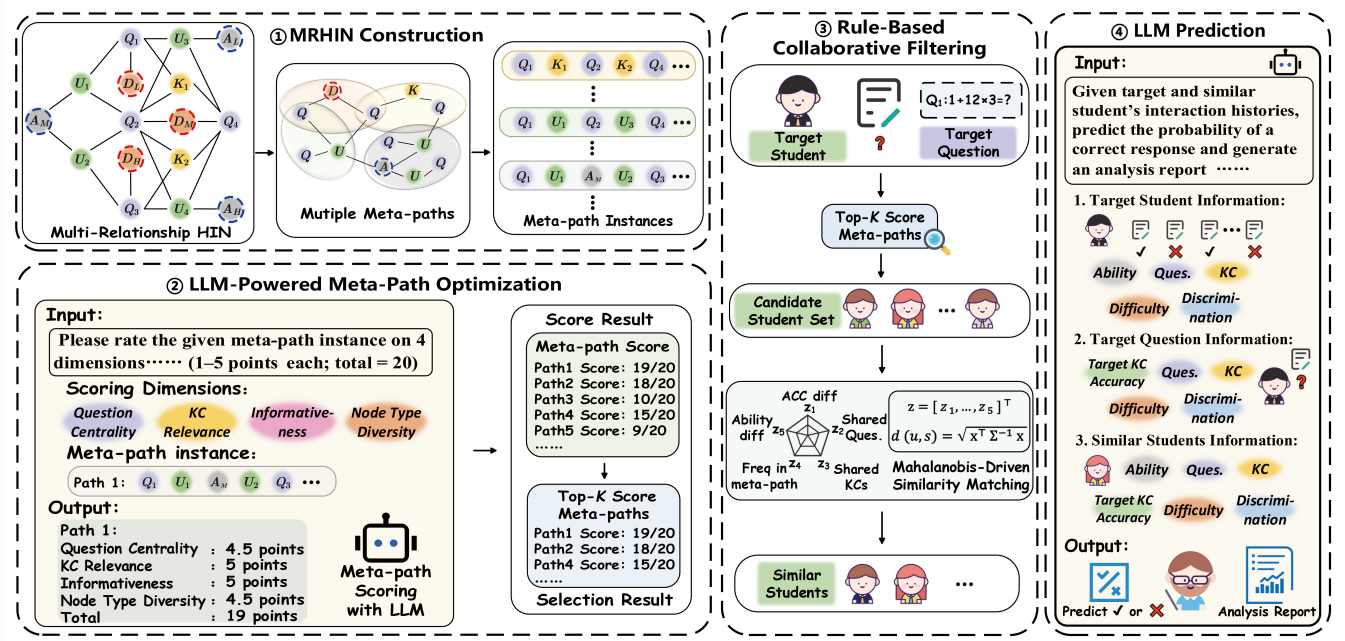


Figure 1: The framework of HISE-KT: (1) Build a multi-relationship HIN, then define and instantiate multiple meta-paths; (2) Employ the LLM to score and select high-quality meta-path instances; (3) Retrieve the target student's similar peers through high-quality meta-path; (4) Input cross-student information into LLM for KT prediction and analysis reports generation.

External Node Construction. This section employs the two-parameter logistic Item Response Theory (IRT-2PL) model (Birnbaum 1968) to estimate each student's latent ability and each question's difficulty. Since raw ability and difficulty parameter values are continuous, we discretize them into three levels (e.g., Low, Medium, High) based on their mean and standard deviation. These levels are then used to construct the external nodes A and D. The detailed IRT-2PL and level division formulas are provided in Appendix A.

Meta-Path Construction. In order to capture the latent high-order relationships among five node types (U, Q, K, A, and D) in MRHIN, we design 14 meta-paths to capture the relationships between these nodes. Since different meta-paths represent different semantics, these meta-paths are divided into two categories: *basic meta-paths* and *composite meta-paths*. *Basic meta-paths* are defined as those that reflect a single semantic and cannot be decomposed into a second independent semantic. We define four *basic meta-paths*, which capture the most direct relationships between nodes:

- *Q-U-Q*: two questions are answered by the same student, capturing the relationship between Q and U.
- *Q-K-Q*: two questions examine the same knowledge concept, capturing the relationship between Q and K.
- *Q-D-Q*: two questions belong to the same question difficulty, capturing the relationship between Q and D.
- *Q-U-A-U-Q*: two questions are answered by students with same ability, capturing the relationship among Q, U, and A.

Composite meta-paths are composed of basic paths. In order to capture the complex cross-semantic relationships between multiple nodes, we define ten meta-paths. Finally, for each question node, N path instances are sampled for every meta-path via an equal-probability random walk on MRHIN. The specific definition of the ten *composite meta-paths* is provided in Appendix B.

LLM-Powered Meta-path Optimization Module

Random walk sampling generates meta-path instances that cover multi-dimensional semantics for each target question but also introduces redundant and noisy links. We employ an LLM to score each meta-path instance across four dimensions and select the Top-K highest scoring paths for each meta-path template. These selected paths better capture the structural relationships in the MRHIN.

Define the instantiated path as $P = (v_0, \ldots, v_L)$, where $v_0 = q_0$ denotes the target question node, $Q(P)$ represents the set of question nodes in the path, k^* is the target knowledge concept, and $\mathrm{dist}(x, y)$ measures the shortest distance between two nodes x and y in the MRHIN. Uppercase letters indicate node types (U, Q, K, A, D), while lowercase letters denote concrete nodes (e.g., q_0). In every formula, "$\propto$" indicates a positive correlation between the score and the meta-path instance quality. The formulaic expressions of the four dimensions are shown below, and the complete scoring prompt is provided in Appendix C.

Question Centrality. When the number of random walk steps increases, subsequent nodes in the path tend to deviate from the target question, resulting in a lack of direct relevance at the question level. We use *Question Centrality* dimension C_q to measure whether the path consistently

maintains a star-like closed structure centered around q_0. The closer and more frequently question nodes revisit q_0, the higher the question centrality, we have:

$$C_q(P) \propto \left(1 - \frac{1}{|Q(P)|} \sum_{q \in Q(P)} \frac{\mathrm{dist}(q_0, q)}{L}\right) \quad (1)$$

where $\mathrm{dist}(q_0, q)$ is the distance between q_0 and q in MRHIN, L is the length of the meta-path instance.

Knowledge Concept Relevance. High-quality paths should closely revolve around the target knowledge concept k^* to highlight semantic relevance at the knowledge level. If the path frequently turns to irrelevant knowledge concepts, it provides limited or even detrimental predictive performance. The *Knowledge Concept Relevance* dimension R_{KC} measures the proportion of questions related to k^* in the path; a higher proportion indicates more focused semantic coherence, we have:

$$R_{\mathrm{KC}}(P) \propto \frac{\left|\{\, q \in Q(P) \mid k^* \in \mathrm{KC}(q)\}\right|}{|Q(P)|} \quad (2)$$

where $\mathrm{KC}(q)$ is the set of knowledge concepts contained in question q.

Informativeness. During random walks, paths may cycle among a few nodes, creating redundant links and diluting effective signals. The *Informativeness* dimension I_{info} encourages paths to continuously introduce new nodes (U, Q, K) to obtain more independent evidence, we have:

$$I_{\text{info}}(P) \propto \frac{\mathrm{distinct}_{\neg q_0, k^*}(P)}{\left|\{\, v \in P \mid \text{type}(v) \in \{U, Q, K\}\}\right|} \quad (3)$$

where $\mathrm{distinct}_{\neg q_0, k^*}(P)$ denotes the unique occurrences of three node types (U, Q, K), while ignoring repeated visits to the initial question node q_0 and target knowledge concept node k^*. A higher proportion of distinct nodes yields a higher informativeness score.

Node Type Diversity. The two external nodes, A and D, can reveal higher-order interactions between student proficiency and question intensity. A complete and balanced level distribution provides rich information for capturing structured relationships in the MRHIN. The *Node Type Diversity* dimension D_{type} encourages paths to simultaneously cover various levels of A and D while avoiding extreme imbalances. The more comprehensive the levels and the more balanced their distribution, the higher the node type diversity score, we have:

$$D_{\text{type}}(P) \propto -\frac{1}{\log|\mathcal{L}|} \sum_{t \in \mathcal{L}} p_t \log p_t \quad (4)$$

where $\mathcal{L} = \{A_{\text{Low}}, A_{\text{Medium}}, A_{\text{High}}, D_{\text{Low}}, D_{\text{Medium}}, D_{\text{High}}\}$, p_t denotes the relative frequency of nodes belonging to Level t.

LLM generates scores for each path of the four dimensions, with a maximum score of 5 points per dimension and a total score of 20 points:

$$S(P) = C_q + R_{\mathrm{KC}} + I_{\text{info}} + D_{\text{type}} \quad (5)$$

where S denotes the total score of a meta-path instance.

Rule-Based Collaborative Filtering Module

The high-quality meta-paths selected by LLM can effectively capture the structural relationships between the five types of nodes (U, Q, K, A, D) in MRHIN. Inspired by Relative Evaluation (Gibbons and Murphy 1990) and Social Comparison Theory (Suls and Wheeler 2012) in educational psychology, we recognize that the performance and ability evaluation of individual learners is often more meaningful and valuable when compared with their similar peers.

Building on these paths, we propose a rule-based collaborative filtering mechanism to retrieve students with similar potential knowledge states to the target student, thereby capturing cross-student interaction information. This module can be divided into three stages: Candidate Student Set Construction, Cross-Student Similarity Encoding, and Mahalanobis-Driven Similarity Matching.

Candidate Student Set Construction. Based on the target student's interaction history, the target question to be tested can be obtained. According to all the high-quality meta-path instances retained under the target question, the candidate student set $\mathcal{C}$ is constructed after removing the duplicate student nodes that are appearing in these paths. $\mathcal{C}$ incorporates student attributes and relevant historical interaction information. Due to the constraints of the meta-path quality assessment, the students in $\mathcal{C}$ have potential connections with the target question.

Cross-Student Similarity Encoding. To quantify the difference between the target student u and a candidate student s, we construct a five-dimensional feature vector as:

$$\mathbf{z}_{u,s} = [\, z_1, z_2, z_3, z_4, z_5 \,]^\top \quad (6)$$

where

$$z_1 = |\theta_u - \theta_s| \quad (7)$$

here, z_1 encodes the difference in student ability θ between u and s. The closer the abilities, the more similar the two students are.

$$z_2 = \frac{c}{|K|} \sum_{k \in K} \left|\mathrm{acc}_u(k) - \mathrm{acc}_s(k)\right| \quad (8)$$

here, c is a scaling constant that controls the decay rate. z_2 encodes the difference in accuracy between u and s on shared knowledge concepts K. A smaller accuracy difference implies greater similarity.

$$z_3 = (1 + N_{\mathrm{Q}})^{-c} \quad (9)$$

here, z_3 encodes the count of shared questions N_{Q} between u and s. The more questions they both answered, the more similar the two students are.

$$z_4 = (1 + N_{\mathrm{K}})^{-c} \quad (10)$$

here, z_4 encodes the count of shared knowledge concepts N_{K} between u and s. A larger overlap indicates greater similarity.

$$z_5 = (1 + f)^{-c} \quad (11)$$

here, z_5 encodes the association strength between the candidate student and the target question, based on their meta-path co-occurrence frequency f. A higher frequency implies a stronger association.

Mahalanobis-Driven Similarity Matching. Considering the differences in scales and correlations among the five-dimensional similarity features, we measure the similarity between u and s using Mahalanobis distance (De Maesschalck, Jouan-Rimbaud, and Massart 2000). This metric can automatically normalize feature dimensions and incorporate covariance information. The similarity distance calculation can be defined as:

$$d(u,s) = \sqrt{\left(\mathbf{z}_{u,s} - \boldsymbol{\mu}\right)^\top \boldsymbol{\Sigma}^{-1} \left(\mathbf{z}_{u,s} - \boldsymbol{\mu}\right)} \tag{12}$$

where $\boldsymbol{\mu}$ is the mean vector of the feature vectors sampled from a large number of student pairs in the training set, and $\boldsymbol{\Sigma}$ is the corresponding covariance matrix, regularised with Schmidt shrinkage to avoid singularity. Finally, we sort all candidate students in ascending order of $d(u,s)$ and select the Top-S students with the smallest distances as the most similar students to the target student.

LLM Prediction Module

Considering that traditional knowledge tracing models can usually only output numerical indicators and lack high-level semantic explanations of prediction results. In order to address the critical needs of both prediction accuracy and interpretability in educational scenarios, this stage integrates cross-student interaction information into the prompt and employs the LLM to perform KT prediction and interpretable analysis report generation. Considering that our method employs a zero-shot prediction approach without fine-tuning the LLM, the quality of the prompt template is particularly critical to the prediction. A simple prompt example is shown in Fig. 1, with the complete version provided in Appendix C.

Structured Prompt Construction. The input information consists of three blocks: (1) *target student information*: including student ability, historical interaction sequence, and question attributes (knowledge concept, difficulty, and discrimination parameters obtained from the IRT-2PL model) within the interactions; (2) *target question information*: including question attributes and the student's historical accuracy on the target KC; (3) *similar student information*: including student ability, historical interaction sequences on the target KC and corresponding accuracy.

Prediction and Explanation Generation. Once the structured prompt is constructed, we invoke the LLM to perform two tasks in a single pass: (1) *answer result prediction*: including the binary result of the answer $(correct/wrong)$ and the confidence probability; (2) *natural language analysis report*: a three-sentence analysis and explanation report interpreting the prediction process.

Experiment

In this section, we conduct extensive experiments to illustrate the effectiveness of our proposed method.

Experimental Settings

Datasets. We evaluate the performance of HISE-KT on four benchmark datasets: **Assistment09** (Feng, Heffernan, and Koedinger 2009), **Slepemapy** (Papoušek, Pelánek, and Stanislav 2016), **Statics2011** (Koedinger et al. 2010) and **Frcsub** (Wu et al. 2015). The statistics of the datasets are in Tab. 1. More details are provided in Appendix D.

Baselines. To evaluate the performance of our proposed method, we compare it with three categories of baselines: Deep Learning (DL)-based, HIN-based, and LLM-based methods. The DL-based methods include DKT (Piech et al. 2015), DKT+ (Yeung and Yeung 2018), DKT-Forget (Nagatani et al. 2019), AT-DKT (Liu et al. 2023a), DKVMN (Zhang et al. 2017), Deep-IRT (Yeung 2019), AKT (Ghosh, Heffernan, and Lan 2020), GKT (Nakagawa, Iwasawa, and Matsuo 2019), SAKT (Pandey and Karypis 2019), and CoKT (Long et al. 2022); the HIN-based methods include PEBG+DKT (Liu et al. 2020), TCL4KT (Sun et al. 2023), SimQE (Sun et al. 2024), and STHKT (Li et al. 2025b); and the LLM-based methods include EPFL (Neshaei et al. 2024) and EFKT (Li et al. 2024).

Implementation Details. We divide the dataset into training, validation, and test sets in a ratio of 8:1:1. The LLM used for meta-path scoring is Qwen-Plus (Bai et al. 2023) and for prediction are Qwen-Plus and DeepSeek-V3 (Liu et al. 2024) without additional fine-tuning. For meta-path instantiation, the number of path instances N sampled under each meta-path for each question is 100, and the path length is 20. The scaling constant c in cross-student similarity encoding is set to 2. Parameter selection for Top-K paths and Top-S similar students is detailed in the parameter sensitivity study.

Main Results

Tab. 2 shows the AUC performance of all compared models (The complete AUC and ACC results are provided in Appendix F). We find that HISE-KT outperforms all three categories of baseline models, indicating the effectiveness of our proposed HISE-KT method. Compared with DL-based methods, our method improves ACC by 0.74%-9.68% and AUC by 2.44%-15.85%, showing that it achieves the highest performance without sequence modeling or fine-tuning. Compared with HIN-based methods, our method improves ACC by 1.93%-7.56% and AUC by 2.95%-11.75%, demonstrating that it selects high-quality meta-paths and extracts more informative structured relationships on MRHIN. Compared with LLM-based methods, our method improves ACC by 1.28%-24.31% and AUC by 7.24%-29.06%, confirming that integrating cross-student information with LLM reason-

Datasets	Students	Questions	KCs	Interactions
Assistment09	3,013	9,795	107	297,575
Slepemapy	5,000	2,727	1,390	615,042
Statics2011	331	633	97	111,468
Frcsub	536	20	8	98,624

Table 1: Statistics of the four processed datasets.

Models	AUC			
	Assistment09	**Slepemapy**	**Statics2011**	**Frcsub**
DKT	0.7452	0.7386	0.7927	0.8695
DKT+	0.7551	0.7426	0.7995	0.8736
DKT-Forget	0.7485	0.7448	0.7852	0.8645
AT-DKT	0.7596	0.7551	0.8024	0.8831
DKVMN	0.7442	0.7726	0.7774	0.9023
Deep-IRT	0.7425	0.7584	0.7841	0.8462
AKT	0.7788	0.7952	0.8242	0.8954
GKT	0.7459	0.7010	0.7888	0.8387
SAKT	0.7133	0.7633	0.7776	0.8469
CoKT	0.8211	0.8164	0.8270	0.9238
PEBG+DKT	0.8183	0.8213	0.8179	0.9187
TCL4KT	0.7918	0.8018	0.8357	0.9079
SimQE	0.8027	0.8269	0.8321	0.9163
STHKT	0.8056	0.8574	0.8295	0.9135
EPFL	0.5827	0.6088	0.7189	0.7446
EFKT	0.7945	0.6843	0.7671	0.8758
Ours				
HISE-KT$_{DS}$	0.8381	0.9383	0.8692	0.9322
HISE-KT$_{QP}$	**0.8703**	**0.9749**	**0.8888**	**0.9482**

Table 2: **AUC** performance on four datasets. All results are the mean of five runs. Best in **bold**, second-best underlined. **HISE-KT$_{DS}$** denotes our method using DeepSeek-V3; **HISE-KT$_{QP}$** denotes our method using Qwen-Plus.

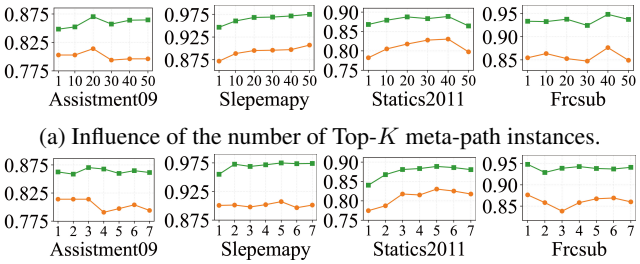
Methods	AUC			
	Assistment09	**Slepemapy**	**Statics2011**	**Frcsub**
Full	0.8703	0.9749	0.8888	0.9482
w/o MSR	0.8593	0.9709	0.8805	0.9399
w/o MSL	0.8473	0.9662	0.8516	0.9321
w/o SimU	0.7734	0.6911	0.8024	0.9177
w/o RSimU	0.7754	0.6937	0.7774	0.8692
w/o IRT	0.8492	0.9668	0.8447	0.9248

Table 3: Ablation results in terms of **AUC** on four datasets.

ing not only yields optimal performance but also enhances interpretability. These results confirm that the synergy between the structured modeling of HIN and the reasoning capabilities of LLM enables accurate and interpretable KT.

Ablation Study

Tab. 3 shows the ablation study results in terms of AUC (The complete AUC and ACC results are provided in Appendix F). To further investigate the contribution of different components in HISE-KT, we design five variants: (1) **w/o MSR** indicates removing the meta-path scorer and using Random-K meta-paths that have not been filtered by the meta-path scorer to replace the Top-K meta-paths; (2) **w/o MSL** indicates using the Lowest-K scoring meta-paths to replace the Top-K scoring meta-paths; (3) **w/o SimU** indicates removing information about similar students during LLM prediction; (4) **w/o RSimU** indicates students randomly selected from the candidate student set are used to replace similar



(a) Influence of the number of Top-K meta-path instances.

(b) Influence of the number of Top-S similar students.

Figure 2: Parameter-sensitivity analysis. Green(upper line) is AUC; orange(bottom line) is ACC.

students obtained through similarity retrieval; (5) **w/o IRT** indicates removing IRT-2PL parameters (e.g., ability, difficulty, discrimination) during KT prediction.

From Tab. 3, we can conclude that removing the meta-path scorer leads to a decline in performance, indicating that high-quality meta-paths are crucial for capturing structured relationships in MRHIN. Moreover, the performance using the Lowest-K scoring paths is inferior to that using Random-K paths, further confirming the importance of the meta-path scorer. When similar student information is not incorporated, the performance drops significantly, indicating that the cross-student interaction information is crucial for KT prediction. Additionally, replacing the similar students retrieved by the rule-based collaborative filtering module with randomly selected ones further degrades performance, suggesting that irrelevant students introduce noise and thereby damage performance. When the IRT information is removed, the LLM cannot assess whether the difficulty and discrimination of the target question match the target student's ability based on such parameter information, consequently affecting the prediction performance. In summary, all components of HISE-KT contribute positively to performance, with the meta-path scorer and similar students information being particularly crucial for effective cross-student interaction modeling.

Parameter Sensitivity

Fig. 2 shows the impact of the number of Top-K meta-path instances and the number of Top-S similar students on prediction performance. As can be seen from Fig. 2a, the prediction performance is poor when the Top-K parameter is selected to be small, indicating that fewer meta-paths cannot mine the rich information between questions and students. When the value of Top-K increases, the prediction performance gradually improves; however, when Top-K becomes too large, the model performance tends to show diminishing returns or slight performance degradation, indicating that relatively low-quality meta-path instances introduce noise that degrades prediction performance.

In Fig. 2b, for the Top-S parameter, different datasets have different optimal numbers of similar students, typically between 3 and 5, whereas the *Frcsub* dataset achieves its best results with Top-$S = 1$. Because the questions in this dataset often involve multiple knowledge concepts, intro-

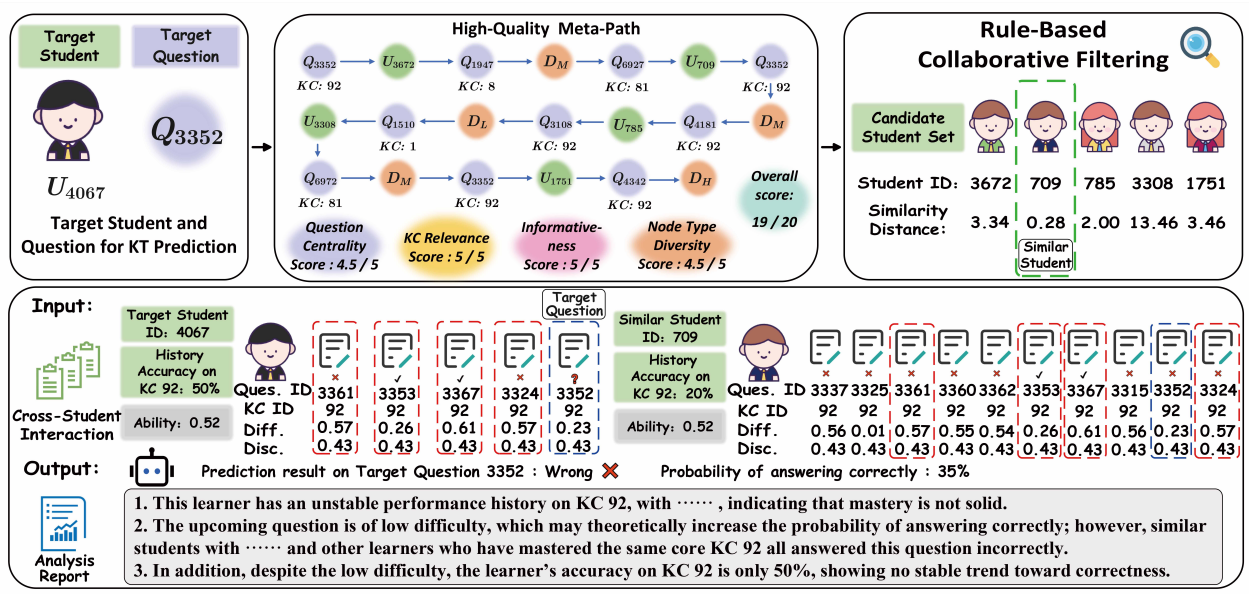


Figure 3: Case study of a target student U_{4067} on question Q_{3352}.

Dimensions	Analysis
Question Centrality	The target question node Q_{3352} appears three times and all other question nodes are located close to it. This star-shaped structure shows that the path remains focused on Q_{3352}, yielding a high score.
KC Relevance	Four of the seven questions directly examine KC_{92} while only a few involve other concepts. KC_{92} appears repeatedly, indicating strong conceptual coherence and yielding a high score.
Informativeness	There are no repeated student, non-target question, or knowledge concept nodes. Each step contributes new information without redundancy, yielding a high score.
Node Type Diversity	The path spans three difficulty levels (D_{low}, D_{medium}, D_{high}) and interleaves them with different question and student nodes, resulting in rich node types and a high diversity score.

Table 4: Analysis of high-quality meta-path instance in the case study.

ducing too many similar students shows inconsistent mastery of multiple knowledge concepts, which tends to interfere with the LLM's judgment. These results demonstrate the relative robustness of HISE-KT to hyperparameter configurations within reasonable ranges.

Case Study

We select a high-quality meta-path instance to demonstrate the entire prediction process. Fig. 3 shows the process, and Tab. 4 lists the quality analysis results of the meta-path instance. First, given the target student U_{4067} and the target question Q_{3352}, we extract the high-quality meta-path instance corresponding to the question. Then, based on this path, we retrieve the most similar student U_{709}, and input the interaction history between U_{4067} and U_{709} into the LLM to complete the prediction of U_{4067}'s performance and generate a three-sentence explainable analysis report. It is noteworthy that U_{4067}'s accuracy on KC_{92} is 50% with significant performance fluctuations. It is difficult to make a reliable judgment solely based on U_{4067}'s own data. Meanwhile, U_{709}'s interaction history completely overlaps with U_{4067}, and the answer result of Q_{3352} is wrong. Based on the cross-student information, the LLM predicts that U_{4067} also answers incorrectly.

Conclusion

In this paper, we propose an innovative framework that synergistically integrates heterogeneous information networks with large language models called HISE-KT to achieve accurate and explainable knowledge tracing. Key innovations include deploying LLMs for automated meta-path quality assessment to eliminate noise, designing an educational psychology-inspired mechanism to retrieve similar student trajectories for enriched context, and leveraging structured prompts to fuse historical data with relational evidence. This integration enables simultaneous prediction and evidence-based explanation generation. Extensive experiments across four public benchmarks demonstrate HISE-KT's superior accuracy and unprecedented interpretability over state-of-the-art baselines. Our research establishes a new paradigm for developing transparent, adaptive educational AI through principled fusion of structural and semantic modeling.

Acknowledgments

This work was funded by the National Natural Science Foundation of China (Nos. 62567005 and 62206107), and Natural Science Foundation of Inner Mongolia Autonomous Region of China (No. 2025MS06004).

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Birnbaum, A. 1968. Some latent trait models and their use in inferring an examinee's ability. *Statistical theories of mental test scores*.

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

De Maesschalck, R.; Jouan-Rimbaud, D.; and Massart, D. L. 2000. The mahalanobis distance. *Chemometrics and intelligent laboratory systems*, 50(1): 1–18.

Duan, Z.; Dong, X.; Gu, H.; Wu, X.; Li, Z.; and Zhou, D. 2024. Towards more accurate and interpretable model: Fusing multiple knowledge relations into deep knowledge tracing. *Expert Systems with Applications*, 243: 122573.

Feng, M.; Heffernan, N.; and Koedinger, K. 2009. Addressing the assessment challenge with an online system that tutors as it assesses. *User modeling and user-adapted interaction*, 19(3): 243–266.

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

Gibbons, R.; and Murphy, K. J. 1990. Relative performance evaluation for chief executive officers. *ILR Review*, 43(3): 30–S.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Jung, H.; Yoo, J.; Yoon, Y.; and Jang, Y. 2024. CLST: Cold-Start Mitigation in Knowledge Tracing by Aligning a Generative Language Model as a Students' Knowledge Tracer. *arXiv preprint arXiv:2406.10296*.

Kim, J.; Chu, S.; Wong, B.; and Yi, M. 2024. Beyond Right and Wrong: Mitigating Cold Start in Knowledge Tracing Using Large Language Model and Option Weight. *arXiv e-prints*, arXiv–2410.

Koedinger, K. R.; Baker, R. S.; Cunningham, K.; Skogsholm, A.; Leber, B.; and Stamper, J. 2010. A data repository for the EDM community: The PSLC DataShop. *Handbook of educational data mining*, 43: 43–56.

Li, H.; Yu, J.; Ouyang, Y.; Liu, Z.; Rong, W.; Li, J.; and Xiong, Z. 2024. Explainable few-shot knowledge tracing. *arXiv preprint arXiv:2405.14391*.

Li, R.; Wu, S.; Wang, J.; and Zhang, W. 2025a. CIKT: A Collaborative and Iterative Knowledge Tracing Framework with Large Language Models. *arXiv preprint arXiv:2505.17705*.

Li, S.; Shen, S.; Su, Y.; Sun, X.; Lu, J.; Mo, Q.; Wu, Z.; and Liu, Q. 2025b. Sthkt: Spatiotemporal knowledge tracing with topological Hawkes process. *Expert Systems with Applications*, 259: 125248.

Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Liu, Y.; Yang, Y.; Chen, X.; Shen, J.; Zhang, H.; and Yu, Y. 2020. Improving knowledge tracing via pre-training question embeddings. *arXiv preprint arXiv:2012.05031*.

Liu, Z.; Liu, Q.; Chen, J.; Huang, S.; Gao, B.; Luo, W.; and Weng, J. 2023a. Enhancing deep knowledge tracing with auxiliary tasks. In *Proceedings of the ACM web conference 2023*, 4178–4187.

Liu, Z.; Zhang, S.; Zhang, J.; Jiang, M.; and Liu, Y. 2023b. HeteEdgeWalk: a heterogeneous edge memory random walk for heterogeneous information network embedding. *Entropy*, 25(7): 998.

Long, T.; Qin, J.; Shen, J.; Zhang, W.; Xia, W.; Tang, R.; He, X.; and Yu, Y. 2022. Improving knowledge tracing with collaborative information. In *Proceedings of the fifteenth ACM international conference on web search and data mining*, 599–607.

Meng, C.; Cheng, R.; Maniu, S.; Senellart, P.; and Zhang, W. 2015. Discovering meta-paths in large heterogeneous information networks. In *Proceedings of the 24th international conference on world wide web*, 754–764.

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

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

Neshaei, S. P.; Davis, R. L.; Hazimeh, A.; Lazarevski, B.; Dillenbourg, P.; and Käser, T. 2024. Towards modeling learner performance with large language models. *arXiv preprint arXiv:2403.14661*.

Pandey, S.; and Karypis, G. 2019. A self-attentive model for knowledge tracing. *arXiv preprint arXiv:1907.06837*.

Papoušek, J.; Pelánek, R.; and Stanislav, V. 2016. Adaptive geography practice data set. *Journal of Learning Analytics*, 3(2): 317–321.

Piech, C.; Bassen, J.; Huang, J.; Ganguli, S.; Sahami, M.; Guibas, L. J.; and Sohl-Dickstein, J. 2015. Deep knowledge tracing. *Advances in neural information processing systems*, 28.

Qiu, L.; and Wang, L. 2024. Knowledge tracing through enhanced questions and directed learning interaction based on multigraph embeddings in intelligent tutoring systems. *IEEE Transactions on Education*.

Shen, S.; Liu, Q.; Huang, Z.; Zheng, Y.; Yin, M.; Wang, M.; and Chen, E. 2024. A survey of knowledge tracing: Models, variants, and applications. *IEEE Transactions on Learning Technologies*, 17: 1858–1879.

Shi, C.; Hu, B.; Zhao, W. X.; and Yu, P. S. 2018. Heterogeneous information network embedding for recommendation. *IEEE transactions on knowledge and data engineering*, 31(2): 357–370.

Suls, J.; and Wheeler, L. 2012. Social comparison theory. *Handbook of theories of social psychology*, 1: 460–482.

Sun, J.; Du, S.; Liu, Z.; Yu, F.; Liu, S.; and Shen, X. 2023. Weighted heterogeneous graph-based three-view contrastive learning for knowledge tracing in personalized e-learning systems. *IEEE Transactions on Consumer Electronics*, 70(1): 2838–2847.

Sun, J.; Du, S.; Zhou, J.; Yuan, X.; Shen, X.; and Liang, R. 2024. Question Embedding on Weighted Heterogeneous Information Network for Knowledge Tracing. *ACM Transactions on Knowledge Discovery from Data*, 19(1): 1–28.

Sun, Y.; Han, J.; Yan, X.; Yu, P. S.; and Wu, T. 2011. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. *Proceedings of the VLDB Endowment*, 4(11): 992–1003.

Wang, Z.; Zhou, J.; Chen, Q.; Zhang, M.; Jiang, B.; Zhou, A.; Bai, Q.; and He, L. 2025. LLM-KT: Aligning Large Language Models with Knowledge Tracing using a Plug-and-Play Instruction. *arXiv preprint arXiv:2502.02945*.

Wu, R.-z.; Liu, Q.; Liu, Y.; Chen, E.; Su, Y.; Chen, Z.; and Hu, G. 2015. Cognitive Modelling for Predicting Examinee Performance. In *IJCAI*, 1017–1024.

Xu, J.; Huang, X.; Xiao, T.; and Lv, P. 2023. Improving knowledge tracing via a heterogeneous information network enhanced by student interactions. *Expert Systems with Applications*, 232: 120853.

Yeung, C.-K. 2019. Deep-IRT: Make deep learning based knowledge tracing explainable using item response theory. *arXiv preprint arXiv:1904.11738*.

Yeung, C.-K.; and Yeung, D.-Y. 2018. Addressing two problems in deep knowledge tracing via prediction-consistent regularization. In *Proceedings of the fifth annual ACM conference on learning at scale*, 1–10.

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

Zhao, W. X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2).

Zhou, P.; Pujara, J.; Ren, X.; Chen, X.; Cheng, H.-T.; Le, Q. V.; Chi, E.; Zhou, D.; Mishra, S.; and Zheng, H. S. 2024. Self-discover: Large language models self-compose reasoning structures. *Advances in Neural Information Processing Systems*, 37: 126032–126058.