MAViS-KT: A Large-Small Model Collaborative Framework for Explainable Knowledge Tracing

Anonymous ACL submission

Abstract

Knowledge Tracing (KT) is pivotal for personalized education, aiming to predict students' future performance by modeling their evolving knowledge states. However, traditional Deep Learning methods operate as opaque black boxes, while Large Language Models (LLMs) offer interpretability but suffer from high computational costs and unstable reasoning. In practice, not only accurate predictions are needed, but also interpretable reports to support effective learning interventions. To bridge this gap, we propose **M**ulti-**A**gent **Vi**ew **S**ynergy Knowledge Tracing (MAViS-KT), a novel large-small model collaborative framework that synergizes the semantic depth of LLMs with the numerical robustness of lightweight networks. Specifically, we design a multi-view multi-agent debate mechanism to disentangle complex learning signals and ensure reasoning fidelity through collaborative verification. Furthermore, to address the semantic-numerical disconnect, we introduce a trainable correction module that dynamically aligns qualitative insights with precise probability estimates. Experiments show that MAViS-KT outperforms strong baselines in accuracy while offering high-quality, actionable educational insights, effectively combining the strengths of qualitative reasoning and quantitative modeling.

1 Introduction

Knowledge tracing (KT) is a fundamental task in educational data mining (Corbett and Anderson, 1994). It aims to dynamically and accurately infer students' latent knowledge mastery by modeling their performance sequences in continuous learning activities. This capability is a cornerstone of personalized adaptive learning systems and is crucial for improving teaching efficiency and learning outcomes (Shen et al., 2024).

Deep Learning (DL)-based KT methods (e.g., neural networks (Nakagawa et al., 2019), memory networks (Piech et al., 2015), and attention mechanisms (Ghosh et al., 2020)) have achieved substantial progress. Such *small models* typically have moderate parameter sizes, making them efficient to train and deploy under limited computational resources. They can effectively capture complex nonlinear relationships in student-question interactions and output quantitative mastery probabilities or proficiency vectors. However, their outputs are largely restricted to numerical predictions, lacking natural-language explanations. This limitation hinders their applicability in real educational scenarios that require actionable instructional interventions and interpretable feedback. Teachers and students can hardly understand the underlying causes of learning deficiencies or obtain personalized improvement suggestions from numbers alone.

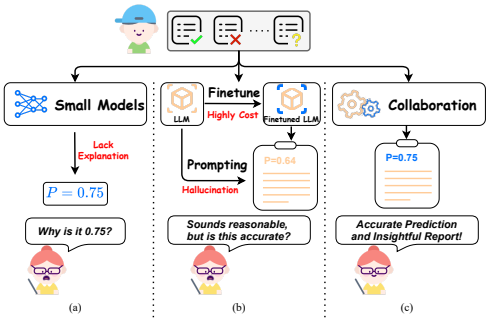

Figure 1: Comparison of KT paradigms. Unlike (a) opaque small models or (b) LLMs constrained by high costs and hallucinations, (c) our collaborative framework unifies accurate prediction with reliable reports.

With the rise of Large Language Models (LLMs) (Bai et al., 2023; Achiam et al., 2023; Liu et al., 2024), LLM-based KT has emerged as a promising direction to address the above limitation. Leveraging rich world knowledge and powerful language generation ability, these *large models* can not only perform knowledge state prediction but also produce natural-language analytical reports, offering human-readable and actionable diagnosis and guidance. However, existing approaches face signifi-

cant hurdles. Fine-tuning incurs prohibitively high computational and data costs (Zhang et al., 2023; Wang et al., 2025; Li et al., 2025), while direct prompting often struggles to effectively organize complex multi-turn interaction contexts, leading to instability, poor robustness, and vague analytical outputs (Huang et al., 2025; Duan et al., 2025; Li et al., 2024).

To mitigate the interpretability limitations of small models and address the robustness and cost constraints of LLM-based methods, we propose **M**ulti-**A**gent **Vi**ew **S**ynergy Knowledge Tracing (MAViS-KT), a novel large-small model collaborative paradigm for KT. As shown in Fig. 1, current methods force a compromise between the precise but opaque predictions of DL and the interpretable but often unstable outputs of LLMs. Our framework breaks this trade-off by reconciling the numerical robustness of lightweight networks with the semantic richness of LLMs, thereby achieving simultaneously accurate prediction and insightful explanation.

The design of our LLM-centric core is informed by Cognitive Load Theory (Sweller, 2011) and Knowledge Building Theory (Scardamalia and Bereiter, 2006). The former advocates reducing information overload through task decomposition, while the latter highlights collaborative critique as a means to deepen knowledge. Accordingly, we first design a multi-view multi-agent debate framework, which systematically emulates multi-dimensional educational assessment perspectives. Here, three specialized agents perform a debate-driven analysis of the student's historical interactions from complementary angles, yielding an initial holistic assessment of their knowledge state. Next, to further improve numerical precision and correct potential biases, we introduce a lightweight, trainable Residual Logit Correction module as the small model component. This module learns to capture subtle interaction patterns that may be overlooked during debate and performs fine-grained calibration of the preliminary predictions. Finally, a Judge Agent integrates the calibrated numerical prediction with multi-view evidence produced by agents to output both a high-accuracy production and a detailed, interpretable natural-language diagnostic report. Our contributions are summarized as follows:

- We propose MAViS-KT, a large-small collaborative framework that synergizes LLM-based semantic reasoning with numerical modeling to unify accurate prediction and interpretable diagnosis.
- We introduce a Multi-View Multi-Agent Debate mechanism for robust student analysis, coupled with a Residual Logit Correction module that converts semantic evidence into precise probability adjustments.
- Empirical results demonstrate that MAViS-KT significantly outperforms state-of-the-art baselines in predictive accuracy while generating high-quality educational feedback.

2 Related Work

2.1 Explainable Knowledge Tracing

While early DL-based KT methods attempted to provide interpretability (Ghosh et al., 2020; Zhang et al., 2017; Yeung, 2019), these abstract numerical indicators remain difficult for educators to interpret. Consequently, recent research has pivoted to LLMs. Initial attempts leveraged zero-shot or few-shot prompting to infer mastery states directly from interaction sequence (Li et al., 2024; Fu et al., 2024). Subsequent studies, such as CIKT (Li et al., 2025) and HISE-KT (Duan et al., 2025), explored instruction fine-tuning or collaborative retrieval mechanisms to better align generic reasoning with specific educational contexts. However, these approaches typically consolidate all data into a monolithic input for a generic agent. This strategy often causes information interference, where dominant semantic content overshadows subtle behavioral cues. In contrast, MAViS-KT employs three view-specific agents to isolate behavior, structure, and question factors, effectively mitigating interference for finer-grained diagnosis.

2.2 Multi-Agent Systems

Transitioning from single-agent prompting (Wei et al., 2022), recent research has embraced multi-agent frameworks to leverage collective intelligence for reducing hallucinations and improving reasoning reliability (Qian et al., 2024; Hong et al., 2023). However, applying such systems to domain-specific tasks faces significant hurdles: they are often difficult to train and resource-intensive (Tran et al., 2025; Islam et al., 2024; Zhang and Xiong, 2025), and their reliance on pure natural-language interaction lacks the explicit constraints required for numerical accuracy and probability calibration.

MAViS-KT addresses these challenges by retaining the interpretability of training-free agents while introducing a lightweight correction module to ensure both computational efficiency and quantitative reliability.

3 Methodology

3.1 Overall Framework

As shown in Fig. 1, MAViS-KT is organized into four stages: (i) KC Graph and Multi-view Repository Construction; (ii) Multi-View-Driven Multi-Agent Debate; (iii) Residual Logit Correction and (iv) Final Report Generation.

3.2 KC Graph and Multi-view Repository Construction

To support multi-agent reasoning, we construct a KC Graph $\mathcal{G}_{KC}$ alongside a Multi-View Repository $\mathcal{R}$. Following SINKT (Fu et al., 2024), we employ an LLM to build $\mathcal{G}_{KC} = (\mathcal{K}, \mathcal{E}, \tau)$, where $\mathcal{K}$ denotes the set of knowledge concepts. The edge set $\mathcal{E}$ encodes two distinct relation types τ: (i) Association (Gao et al., 2021): semantically related KCs that often co-occur without a strict teaching order; (ii) Predecessor-Successor (Pan et al., 2017): a prerequisite relation where A must be mastered before B.

Since our view-specific agents operate in a training-free manner, relying solely on the target student's own interaction history may lead to biased predictions. Therefore, we organize student interaction histories from three complementary perspectives to construct the repository $\mathcal{R}$, which serves as the foundation for the subsequent view-specific context retrieval. Specifically, for each student s, we abstract their interaction history into three distinct vector representations, capturing different dimensions of the learning process.

Learning Behavior Vector: We posit that a student's behavioral patterns and recent performance trends significantly impact outcomes. Thus, we extract interaction features (e.g., response time, number of attempts, and hint usage) combined with the Dynamic Weighted Accuracy (DWA) (Fox et al., 2002) to form the behavior vector. The DWA captures the student's recent performance trend and is defined as:

$$\text{DWA}_s = \frac{\sum_{i=1}^{|\mathcal{I}|} \beta^{|\mathcal{I}|-i} \cdot r_i}{\sum_{i=1}^{|\mathcal{I}|} \beta^{|\mathcal{I}|-i}}, \quad \beta \in (0,1) \tag{1}$$

where $\mathcal{I}$ denotes the set of relevant interactions, r_i is the correctness of the i-th attempt, and β controls the recency decay (set to 0.8). Consequently, the behavior vector is constructed as:

$$\mathbf{v}_s^{Bhv} = [\text{DWA}_s, \bar{\mathcal{B}}_s] \tag{2}$$

where $\bar{\mathcal{B}}_s$ represent the normalized average behavioral fields.

Knowledge Structure Vector: Students' mastery levels vary across different KCs. Crucially, the mastery of a specific concept is not isolated but heavily influenced by its prerequisite or related concepts in the $\mathcal{G}_{KC}$. To capture this structural dependency, we define the knowledge vector as:

$$\mathbf{v}_s^{KC} = [M_s^{k_1}, M_s^{k_2}, \ldots, M_s^{k_{|\mathcal{K}|}}] \tag{3}$$

where $M_s^{k_i}$ denotes the mastery of student s on concept k_i. We formulate $M_s^{k_i}$ by aggregating direct performance and neighborhood influence:

$$M_s^{k_i} = \lambda \cdot \text{ACC}_s(\mathcal{I}_{k_i}) + (1-\lambda) \cdot \frac{\sum_{k_j \in \mathcal{N}(k_i)} \text{ACC}_s(\mathcal{I}_{K_j})}{|\mathcal{N}(k_i)|} \tag{4}$$

where $\mathcal{I}_{k_i}$ is the interaction subset for k_i, $\mathcal{N}(k_i)$ denotes the set of neighboring concepts in $\mathcal{G}_{KC}$, and λ is a balancing factor. This design ensures that even sparse interactions on K_i can be inferred from related concepts.

Question Analysis Vector: To model how students perform under varying difficulty levels, we stratify all questions into five distinct difficulty levels ($L = \{l_1, \ldots, l_5\}$) based on their global pass rates (in 20% intervals). The question analysis vector is defined as:

$$\mathbf{v}_s^{Q} = [\text{ACC}_s(l_1), \ldots, \text{ACC}_s(l_5)] \tag{5}$$

This vector explicitly represents the student's ability distribution across different difficulty gradients, which can be used to identify the performance patterns on questions of varying difficulty levels.

3.3 Multi-View-Driven Multi-Agent Debate

This stage is designed to analyze the performance of target student s_{tgt} from three distinct perspectives. We employ three specialized agents: (i) Learning Behavior Agent A_{Bhv}, (ii) Knowledge Structure Agent A_{KC}, and (iii) Question Analysis Agent A_Q, each responsible for generating view-specific predictions and reasoning reports through

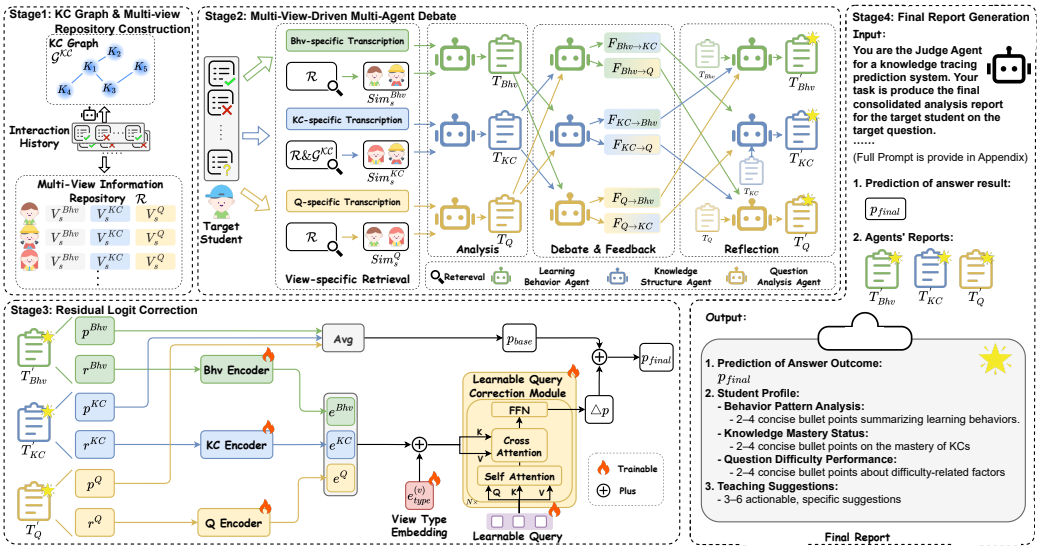

Figure 2: The MAViS-KT framework. It synergizes multi-view agent debate for semantic reasoning with a residual logit correction module for precise numerical calibration, culminating in a comprehensive diagnostic report.

collaborative interaction and iterative refinement. Given a prediction target consisting of a student's interaction history $\mathcal{I}_{tgt}$ and a target question q_{tgt}, we denote the historical sequence as:

$$\mathcal{I}_{tgt} = \{(q_1, k_1, ans_1), (q_2, k_2, ans_2), \ldots, (q_n, k_n, ans_n)\} \quad (6)$$

where each ans_i contains behavioral signals $\mathcal{B}_i$ and binary correctness $r_i \in \{0, 1\}$. Note that $\mathcal{B}_i$ is optional behavioral fields depending on the dataset.

Step 1: View-Specific Transformation and Retrieval. We transform $\mathcal{I}_{tgt}$ into three view-specific forms $\mathcal{I}_{tgt}^{(v)}$ for each $v \in \{Bhv, KC, Q\}$. We map each q_i to its difficulty level l_i:

$$\begin{cases} \mathcal{I}_{tgt}^{Bhv} = \{(r_1, \mathcal{B}_1), \ldots, (r_n, \mathcal{B}_n)\} \\ \mathcal{I}_{tgt}^{KC} = \{(k_1, r_1), \ldots, (k_n, r_n)\} \\ \mathcal{I}_{tgt}^{Q} = \{(q_1, l_1, r_1), \ldots, (q_n, l_n, r_n)\} \end{cases} \quad (7)$$

Simultaneously, to enrich the agents' reasoning with external empirical evidence, we extract three view-specific vectors from $\mathcal{I}_{tgt}$ and query the repository $\mathcal{R}$. Our objective is to retrieve the top-K most relevant historical contexts that serve as reference anchors for the current prediction. The relevance of these contexts is quantified via cosine similarity:

$$Sim_s^{(v)} = \cos\left(\mathbf{v}_{s_{tgt}}^{(v)}, \mathbf{v}_{s_j}^{(v)}\right) \quad (8)$$

The retrieved contexts provide rich, view-specific reference information, such as peer performance on q_{tgt} or behavioral baselines on related KCs, thereby grounding the agents' inference in broader data patterns.

Step 2: Multi-Agent Analysis. In this step, to ensure rigorous and standardized outputs, each agent receives view-specific inputs by encapsulated into structured prompts to guide the agents' reasoning and generate preliminary analysis reports. Analysis Prompts structured as follows: (i) Task Description: Defines the specific persona and the scope of analysis, strictly limiting the agent's perspective to its designated view; (ii) Task Objectives: Explicitly requires the agent to predict the correctness probability of the target question and generate a comprehensive diagnostic profile based on historical patterns; (iii) Analysis Principles: Imposes reasoning constraints, such as forbidding the use of external knowledge not provided in the input, requiring evidence citation, and adhering to domain-specific logic; (iv) Input Data: Incorporates the view-specific historical sequence $\mathcal{I}_{tgt}^{(v)}$, the retrieved contextual evidence $Sim_s^{(v)}$, and relevant statistical indicators; (v) Output Structure: Enforces a unified format containing the prediction score, confidence level, and a structured reasoning chain to facilitate subsequent parsing. Fed with these structured prompts, view-specific agents generate their initial analysis reports $T_{(v)}$. The complete analysis prompt template is available in Appendix B.1.

The A_{Bhv} is tasked with identifying the target student's behavioral pattern and inferring correctness, leveraging psychometric findings that associate rapid responses with low motivation (Wise

and Kong, 2005) while characterizing longer durations as a non-linear indicator of either careful processing or confusion (Goldhammer et al., 2014). Drawing on these theoretical insights, we define four canonical behavior patterns: (i) Careful / Deliberate; (ii) Random / Careless Guessing; (iii) Procrastination / Hesitant and (iv) Hard-working but Low-ability. Detailed definitions of behavioral patterns can be found in the Appendix A. The A_{KC} evaluates the student's conceptual mastery across relevant KCs, leveraging both their historical interactions and structural information from $\mathcal{G}_{KC}$. The A_Q analyzes whether the target question q_{tgt} poses a significant challenge to the student. It jointly considers whether the student has previously shown instability on questions with similar profiles.

Step 3: Debate and Feedback. Each agent then receives the analysis reports from the other two agents. The goal is to conduct a cross-verification process to identify points of agreement or disagreement. To guide this, we employ Debate Prompts that strictly instruct agents to rely solely on their own previously generated profile to cross-validate peer predictions. Crucially, the prompt enforces boundary conditions: agents are forbidden from encroaching on other domains (e.g., using semantic knowledge for behavioral reasoning) or introducing new speculation. The output mandates a structured categorization of findings into Points of Agreement and Points of Conflict. This process encourages critical scrutiny, compelling agents to justify their reasoning against alternative perspectives. The complete debate prompt template is available in Appendix B.2.

Step 4: Reflection. Finally, each agent receives the peer feedback regarding its initial report. To orchestrate self-correction, we utilize Reflection Prompts that contextualize agents in a revision phase. The prompt directs the agent to synthesize its initial analysis with peer critiques, explicitly identifying potential biases or overconfidence. Based on this cross-view evidence, the agent determines which interpretations should be adjusted or downweighted, producing a final refined report $T'_{(v)}$ that integrates both self-analysis and peer verification. The complete reflection prompt template is available in Appendix B.3.

3.4 Residual Logit Correction

To further elevate predictive accuracy and fully capitalize on the multi-view reasoning, we introduce a residual correction mechanism that refines the prediction using the final reports produced by the three agents. Each agent's final report $T'_{(v)}$ is decomposed into two parts: a prediction score $p^{(v)}$ and a refined textual report $R^{(v)}$. The view-specific predictions are converted to logits and averaged to obtain the base prediction logit:

$$z_{base} = \frac{1}{3}\sum_{v} \sigma^{-1}(p^{(v)}) \tag{9}$$

where σ^{-1} denotes the logit function. The three $R^{(v)}$ are encoded via dedicated report Encoders into dense vectors:

$$\begin{cases} e^{Bhv} = \text{Encoder}^{Bhv}(R^{Bhv}) \\ e^{KC} = \text{Encoder}^{KC}(R^{KC}) \\ e^{Q} = \text{Encoder}^{Q}(R^{Q}) \end{cases} \tag{10}$$

where $\text{Encoder}^{(v)}$ is lightweight transformer encoder. The three vectors $e^{(v)}$ are concatenated, augmented with corresponding view-type embeddings $e_t^{(v)}$, and fused into a sequence:

$$X = [e^{Bhv} \oplus e_t^{Bhv},\ e^{KC} \oplus e_t^{KC},\ e^{Q} \oplus e_t^{Q}] \tag{11}$$

where $\oplus$ denotes the vector addition operation, $X \in \mathbb{R}^{3\times d}$. We employ a Learnable Query Correction Module (LQCM), inspired by Q-Former (Li et al., 2023), initialized with learnable queries $Q_{learn} \in \mathbb{R}^{3\times d}$, to attend to the report sequence X and compute a correction term Δz:

$$\Delta z = \text{LQCM}(Q_{learn}, X) \tag{12}$$

The final prediction logit is computed as the sum of the base and correction terms:

$$p_{final} = \text{Sigmoid}(z_{base} + \Delta z) \tag{13}$$

This final prediction is used for both classification and the subsequent generation of the final diagnostic report.

We optimize the trainable parameters of the correction module by minimizing the Binary Cross-Entropy loss:

$$\mathcal{L} = -\left(y \cdot \log(\hat{y}) + (1-y)\cdot \log(1-\hat{y})\right) \tag{14}$$

where $\hat{y} = p_{final}$, and $y \in \{0, 1\}$ denotes the ground-truth correctness of the target question. By minimizing $\mathcal{L}$, the module learns to dynamically align the semantic reasoning with the actual student performance.

3.5 Final Report Generation

The final stage of MAViS-KT involves generating a comprehensive diagnostic report that not only reflects the final prediction result but also offers an interpretable explanation grounded in multi-view reasoning. To achieve this, we introduce a specialized Judge Agent, which synthesizes information from two sources: (i) the final prediction score p_{final} produced by the Residual Logit Correction stage, and (ii) the refined analysis reports $T'_{(v)}$ from the three view-specific agents. The Judge Agent is prompted with the corrected prediction as the authoritative outcome and is instructed to consolidate the evidence, highlight cross-view consistencies and conflicts, and resolve ambiguous or contradictory reasoning. To ensure consistency and interpretability, we design a dedicated prompt for the Judge Agent, which enforces this structured output format and constrains the reasoning to rely only on the provided analysis reports and final prediction. The complete prompt template is available in Appendix B.4.

4 Experiment

4.1 Experimental Configuration

We evaluate MAViS-KT on three public datasets: **ASSIST09** and **ASSIST12** (Feng et al., 2009), both collected from the ASSISTments platform, and **DBE-KT22** (Abdelrahman et al., 2022), a large-scale online course dataset from the Australian National University. To prepare the input sequences, we segment each student's interaction history into sub-sequences of 25 interactions, using the last interaction in each sequence as the prediction target. From the resulting pool, we randomly sample 1,000 sequences as the test set. To ensure no student overlap between training and testing, we assign all interactions from students not appearing in the test set to the training set. All experiments were conducted under the same preprocessing and evaluation protocols to ensure fairness and reproducibility. More implementation details are provided in Appendix C.

4.2 Baselines

To comprehensively evaluate the effectiveness of MAViS-KT, we compare our method against a diverse set of baselines, including traditional DL-based KT methods and recent LLM-based methods:

Methods	ASSIST09			ASSIST12			DBE-KT22		
	ACC	AUC	F1	ACC	AUC	F1	ACC	AUC	F1
DL-based									
DKT	72.52	82.21	78.13	70.79	69.47	74.99	70.79	72.60	74.13
AT-DKT	73.03	83.04	78.92	72.06	71.66	76.10	72.23	74.55	75.91
DKVMN	71.67	81.69	77.92	70.72	68.69	74.80	71.45	71.59	74.57
Deep-IRT	72.43	80.63	78.02	71.14	68.89	74.60	71.61	72.48	75.58
AKT	73.65	84.95	80.15	72.72	72.31	77.46	73.19	75.32	77.27
SAKT	69.38	77.79	76.30	68.59	67.51	72.52	69.29	71.01	73.11
SAINT+	73.68	82.06	79.22	71.82	70.40	76.00	73.16	74.23	75.11
LLM-based									
LLM-KT	78.61	83.46	80.56	73.46	73.88	78.52	77.55	74.94	78.13
EFKT	64.41	61.01	73.76	63.44	66.81	66.53	64.00	68.14	66.13
EPLF	70.32	81.73	72.07	70.14	69.50	75.60	74.59	73.09	72.25
CIKT	74.95	82.46	80.31	72.46	71.65	77.18	76.77	75.00	77.05
2T-KT	74.78	82.61	80.31	72.75	72.30	78.48	77.40	75.49	79.87
HISE-KT	82.22	87.31	82.03	74.77	75.63	79.51	70.82	72.40	79.94
Ours									
MAViS-KT$_{4o}$	86.23	91.54	89.70	72.35	75.50	79.74	78.83	81.46	82.77
MAViS-KT$_{DS}$	88.50	94.35	91.58	**75.66**	**78.51**	82.81	79.52	83.93	85.80
MAViS-KT$_{QP}$	**91.00**	**95.41**	**94.25**	75.30	78.35	**83.30**	**80.40**	**84.87**	**86.72**

Table 1: Main results on three datasets across Accuracy (ACC), Area Under the Curve (AUC) and F1 Score. Bold values denote the best performance.

- **DL-based Methods**: DKT (Piech et al., 2015), AT-DKT (Liu et al., 2023), DKVMN (Zhang et al., 2017), Deep-IRT (Yeung, 2019), AKT (Ghosh et al., 2020), SAKT (Pandey and Karypis, 2019), SAINT+ (Shin et al., 2021).

- **LLM-based Methods**: LLM-KT (Wang et al., 2025), EFKT (Li et al., 2024), EPLF (Neshaei et al., 2024), CIKT (Li et al., 2025), 2T-KT (Li et al., 2025), HISE-KT (Duan et al., 2025).

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

4.3 Main Results

To thoroughly evaluate the effectiveness and generalizability of MAViS-KT, we implemented the framework using three different LLMs: GPT-4o (4o) (Achiam et al., 2023), Qwen-Plus (QP) (Bai et al., 2023), and DeepSeek-R1 (DS-R1) (Liu et al., 2024), enabling a comprehensive comparison between commercial and open-source LLM backbones. The roles of different agents within the framework were realized by setting different prompts. As shown in Table 1, MAViS-KT consistently achieves superior performance compared to both traditional DL-based KT methods and recent LLM-based methods across all datasets and metrics. This demonstrates the advantage of using multi-agent reasoning and view-specific evidence aggregation over fixed-architecture sequence models, and the effectiveness of our structured, agent-driven design compared to prior prompt-only or flat

Stage	Agent	ASSIST09			ASSIST12			DBE-KT22		
		ACC	AUC	F1	ACC	AUC	F1	ACC	AUC	F1
Anal.	A_{Bhv}	81.60	85.05	86.69	53.70	68.23	54.47	59.50	76.59	66.05
	A_{KC}	85.70	86.81	90.41	63.70	66.25	70.51	68.20	72.37	76.89
	A_Q	87.40	88.67	91.87	73.40	75.87	81.42	75.50	74.58	84.18
Refl.	A_{Bhv}	88.00	93.26	91.92	68.10	74.17	75.21	76.10	79.86	83.37
	A_{KC}	86.70	92.84	90.95	64.60	72.42	70.74	72.00	78.32	79.74
	A_Q	88.00	93.55	92.04	71.00	76.40	78.20	76.80	79.76	84.00

Table 2: Performance comparison of view-specific agents across the Analysis and Reflection stages.

Methods	ASSIST09			ASSIST12			DBE-KT22		
	ACC	AUC	F1	ACC	AUC	F1	ACC	AUC	F1
Full	**91.00**	**95.41**	**94.25**	**75.30**	**78.35**	**83.30**	**80.40**	**84.87**	**86.72**
w/o A_{Bhv}	82.53	83.26	88.99	71.44	69.36	81.43	70.76	71.20	79.83
w/o A_{KC}	82.17	84.39	88.38	60.40	67.98	66.33	75.23	69.89	85.17
w/o A_Q	86.75	88.89	91.73	68.53	68.03	80.83	75.82	74.99	85.23
w/o z_{base}	89.55	94.00	92.10	72.68	76.57	80.64	77.51	82.63	83.59
w/o Debate	87.86	93.10	92.29	74.51	77.81	82.59	80.34	84.11	86.44
w/o LQCM	90.02	94.36	91.85	73.29	76.77	71.28	77.67	82.80	83.74
w/o Corr.	87.51	93.35	91.58	68.81	75.22	75.59	75.86	80.56	83.05

Table 3: Ablation study validating the effectiveness of key components in MAViS-KT.

instruction-tuning approaches. Among our model variants, MAViS-KT$_{\text{QP}}$ outperforms the other two versions in most cases, indicating that stronger backbone LLMs can further enhance both reasoning depth and accuracy in the MAViS-KT framework. Nonetheless, all three variants show robust performance, confirming the generalizability of the proposed architecture across different LLMs.

4.4 Agents Results

To better understand the individual contributions of each view-specific agent, we report their independent prediction performance on all datasets under two conditions: (i) after the initial view-specific analysis stage, and (ii) after the cross-agent debate, feedback, and reflection stage. The results are summarized in Table 2.

In the **Analysis** stage, where each agent relies solely on its own view-specific input, we observe that all agents can already make reasonably accurate predictions. In particular, the A_Q consistently performs best across datasets, likely due to the strong semantic priors embedded in question-level metadata. Interestingly, even the A_{Bhv}, which only utilizes content-agnostic features, achieves competitive results and even outperforming many traditional KT models. These findings suggest that isolating cognitive signals into modular agents aligns well with established cognitive theories, such as Cognitive Load Theory.

After the **Reflection** stage, where agents receive feedback from others and refine their initial reports, we observe consistent performance improvements across all agents and datasets. These results validate our core hypothesis: enabling agents to critique and learn from each other promotes better generalization and deeper reasoning. This observation is well grounded in educational psychology, especially Knowledge Building Theory.

Overall, the effectiveness of our multi-agent design stems not only from view specialization but also from the structured debate and refinement process, which significantly boosts both accuracy and interpretability.

4.5 Ablation Results

The ablation study results in Table 3 demonstrate the significance of each design component in the MAViS-KT framework. First, removing any individual view-specific agent (*w/o* A_{Bhv}, *w/o* A_{KC}, *w/o* A_Q) causes notable performance degradation across all datasets. This confirms that each agent captures complementary view-specific signals and that omitting any one leads to incomplete reasoning and diminished accuracy. Second, removing the debate stage (*w/o Debate*), which enables cross-agent feedback and refinement, leads to a measurable performance decline. This validates the importance of inter-agent collaboration for refining predictions, aligning with our theoretical motivations from sociocultural learning and knowledge building. Third, we assess the logit correction mechanism. Disabling either the base logit (*w/o* z_{base}) or the learnable correction module (*w/o LQCM*) causes a clear drop in performance. This suggests that both direct statistical evidence (from logits) and semantic content (from report representations) are vital for robust final predictions. When both components are removed (*w/o Corr.*), the final prediction is produced directly by the Judge Agent without residual logit refinement. In this setting, performance drops sharply across all datasets, and the prediction behavior degenerates into a weighted aggregation of the three agents' outputs.

In summary, all modules in MAViS-KT work synergistically. View-specific agents extract complementary perspectives, multi-agent debate enhances reasoning depth, and the logit correction module bridges semantic and prediction signals.

4.6 Analysis about Sequence Length

Fig. 3 illustrates the impact of input sequence length on predictive performance. For ASSIST

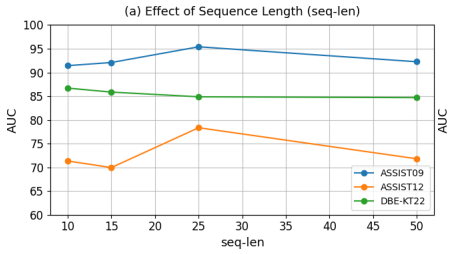

Figure 3: Sequence Length analysis of MAViS-KT on three datasets.

Final Report

1. Prediction of answer Outcome: 70%
$(T'_{Bhv}: 55\%, T'_{KC}: 52\%, T'_{Q}: 65\%, p_{final}: 70\%)$

2. Student Profile:

- **Behavior Pattern Analysis:**
  - Pattern: Random / Careless Guessing
  - The student averages **13.29s** per question (vs. peer avg 47.3s).
  - Recent ultra-fast attempts (<3s) on complex tasks signal **low engagement**, though not guaranteed failure.
- **Knowledge Mastery Status:**
  - *Estimation* is partially mastered (**60%** vs peer 80%). Integration with *Square Root* breaks down under time pressure, showing structural vulnerability.
- **Question Difficulty Performance:**
  - The target is **Medium Difficulty** (pass rate: 0.65).
  - The student performs **exceptionally well (100%)** in this difficulty band, acting as a strong positive anchor.

3. Teaching Suggestions:

- **Intervention:**
  - Enforce a **Minimum Thinking Time (12s+)** to counteract the *Random / Careless Guessing* error mode on integrated problems.
- **Drill:**
  - Practice **Integration Micro-tasks** specifically combining bounding strategies with square roots to stabilize the weak link.
- **Strategy:**
  - Implement a **2-step self-check routine**: *Estimate Magnitude to Check Logic* before submission.

Figure 4: A case for our MAViS-KT. The student answers this question correctly.

datasets, we observe a rise-then-fall trend, achieving optimal performance at $L = 25$. In contrast, DBE-KT22 exhibits a distinct pattern where performance peaks at shorter lengths ($L = 10$) and slightly declines as context increases. This suggests that learning patterns in DBE-KT22 are highly sensitive to immediate interactions, whereas distal history may introduce noise due to rapid concept drift. Despite these varying dependencies, MAViS-KT maintains robust AUC scores across all lengths, validating its ability to adapt to different data characteristics. Full results provided in Appendix E.

4.7 Case Study

Figure 4 illustrates the effectiveness of the MAViS-KT framework in harmonizing semantic reasoning with numerical precision. While view-specific agents offer conservative judgments (52% to 65%) driven by identified risks like rapid guessing, the framework's collaborative architecture successfully calibrates the final prediction to 70%, accurately reflecting the student's positive outcome. Beyond numerical precision, the system validates its pedagogical utility by translating identified behavioral risks into actionable interventions. Full example is provided in the Appendix F.

Baselines	Len.	Exp.	Read.	Edu.	Rig.	Total
EFKT	82	2.75	2.85	2.57	2.42	10.59
CIKT	690	4.22	4.58	4.52	3.85	17.17
HISE-KT	350	4.72	4.48	4.35	4.53	18.08
MAViS-KT$_{QP}$	422	**4.88**	**4.63**	**4.71**	**4.90**	**19.12**

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

4.8 Analysis about Report Quality

We design a scoring mechanism with four dimensions, where each dimension has a maximum score of five, and employ Qwen-Plus to score interpretable reports. The detailed scoring mechanism is given in the Appendix G. 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, MAViS-KT achieves the highest scores in all metrics, securing a total score of 19.12. Notably, it demonstrates a significant advantage in Rigorousness compared to CIKT and HISE-KT. This validates that our Multi-Agent Debate mechanism effectively reduces hallucinations by cross-verifying evidence, unlike single-agent baselines that generate plausible but unverified content. Furthermore, despite generating longer reports, MAViS-KT maintains high Readability and Educational Usefulness, proving that the multi-view framework provides rich, actionable insights rather than redundant text.

5 Conclusion

In this paper, we proposed MAViS-KT, a novel large-small model collaborative framework. By orchestrating a multi-view multi-agent debate, MAViS-KT effectively disentangled complex learning signals to generate rigorous, evidence-based diagnostic reports. Crucially, the integration of a residual logit correction module bridged the gap between qualitative semantic reasoning and quantitative numerical precision, ensuring calibrated predictions. Extensive experiments demonstrated that our approach outperformed state-of-the-art baselines in both prediction accuracy and the educational quality of generated feedback.

Limitations

Despite the promising results of MAViS-KT, we acknowledge two primary limitations. The multi-agent debate mechanism involves multiple rounds of LLM generation. While this ensures rigorous reasoning, it inevitably incurs higher computational cost and latency compared to lightweight DL-based models. The effectiveness of our Learning Behavior Agent relies on interaction logs. In datasets containing only sparse binary correctness labels, the contribution of the behavioral view may be diminished, limiting the full potential of the multi-view synergy. Future work will focus on distilling the multi-agent reasoning capabilities into smaller, more efficient models and enhancing robustness under sparse-data conditions.

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, Akkaya, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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

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.

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.

Lingyue Fu, Hao Guan, Kounianhua Du, Jianghao Lin, Xia, and 1 others. 2024. Sinkt: A structure-aware inductive knowledge tracing model with large language model. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 632–642.

Weibo Gao, Qi Liu, and 1 others. 2021. Rcd: Relation map driven cognitive diagnosis for intelligent education systems. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 501–510.

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.

Frank Goldhammer, Johannes Naumann, Stelter, and 1 others. 2014. The time on task effect in reading and problem solving is moderated by task difficulty and skill: insights from a computer-based large-scale assessment. *Journal of Educational Psychology*, 106(3):608.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Cheng, and 1 others. 2023. Metagpt: Meta programming for a multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Wang, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Md Ashraful Islam, Mohammed Eunus Ali, and Md Rizwan Parvez. 2024. Mapcoder: Multi-agent code generation for competitive problem solving. *arXiv preprint arXiv:2405.11403*.

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*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

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

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

Zitao Liu, Qiongqiong Liu, Jiahao Chen, Huang, and 1 others. 2023. Enhancing deep knowledge tracing with auxiliary tasks. In *Proceedings of the ACM web conference 2023*, pages 4178–4187.

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, Hazimeh, and 1 others. 2024. Towards modeling learner performance with large language models. *arXiv preprint arXiv:2403.14661*.

Liangming Pan, Chengjiang Li, and 1 others. 2017. Prerequisite relation learning for concepts in moocs. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1447–1456.

Shalini Pandey and George Karypis. 2019. A self-attentive model for knowledge tracing. *arXiv preprint arXiv:1907.06837*.

Chris Piech, Jonathan Bassen, Jonathan Huang, Ganguli, and 1 others. 2015. Deep knowledge tracing. *Advances in neural information processing systems*, 28.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Li, and 1 others. 2024. Chatdev: Communicative agents for software development. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15174–15186.

Marlene Scardamalia and Carl Bereiter. 2006. Knowledge building. *The Cambridge*.

Shuanghong Shen, Qi Liu, Zhenya Huang, Zheng, and 1 others. 2024. A survey of knowledge tracing: Models, variants, and applications. *IEEE Transactions on Learning Technologies*, pages 1858–1879.

Dongmin Shin, Yugeun Shim, Hangyeol Yu, Lee, and 1 others. 2021. Saint+: Integrating temporal features for ednet correctness prediction. In *LAK21: 11th international learning analytics and knowledge conference*, pages 490–496.

John Sweller. 2011. Cognitive load theory. In *Psychology of learning and motivation*, volume 55, pages 37–76. Elsevier.

Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O'Sullivan, and Hoang D Nguyen. 2025. Multi-agent collaboration mechanisms: A survey of llms. *arXiv preprint arXiv:2501.06322*.

Ziwei Wang, Jie Zhou, Qin Chen, Min Zhang, Bo Jiang, Zhou, and 1 others. 2025. Llm-kt: Aligning large language models with knowledge tracing using a plug-and-play instruction. *arXiv preprint arXiv:2502.02945*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Xia, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Steven L Wise and Xiaojing Kong. 2005. Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, 18(2):163–183.

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

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.

Shaowei Zhang and Deyi Xiong. 2025. Debate4math: Multi-agent debate for fine-grained reasoning in math. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16810–16824.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Wang, and 1 others. 2023. Instruction tuning for large language models: A survey. *ACM Computing Surveys*.

A Definitions of Behavioral Patterns

We have defined four behavioral patterns:

- **Careful / Deliberate:** Characterized by stable response times within a reasonable range and high accuracy. This pattern typically involves minimal hint usage and few attempts, indicating effective cognitive processing and a high probability of correctness.
- **Random / Careless Guessing:** Marked by extremely short response times (significantly below the student's or group's average) paired with low accuracy. This behavior often bypasses hint usage, serving as a decisive signal of low motivation and high error risk.
- **Procrastination / Hesitant:** Defined by significantly prolonged response times that exceed norms, often accompanied by increased hint usage or idle periods. This pattern reflects uncertainty or confusion rather than productive thinking, creating a potential risk of error despite the time cost.
- **Hard-working but Low-ability:** Exhibits long response times and high attempt counts or frequent hint usage, yet results in persistent low accuracy. This implies that while effort (motivation) is high, the student faces structural comprehension difficulties or utilizes inefficient strategies.

B Prompt Templates

We provide detailed prompt templates for MAViS-KT framework.

B.1 Analysis Prompt

- Fig. 5: Analysis Prompt for Learning Behavior Agent.
- Fig. 6: Analysis Prompt for Knowledge Structure Agent.
- Fig. 7: Analysis Prompt for Question Analysis Agent.

B.2 Debate Prompt

- Fig. 8: Debate Prompt for Learning Behavior Agent.
- Fig. 9: Debate Prompt for Knowledge Structure Agent.
- Fig. 10: Debate Prompt for Question Analysis Agent.

B.3 Reflection Prompt

- Fig. 11: Reflection Prompt for Learning Behavior Agent.
- Fig. 12: Reflection Prompt for Knowledge Structure Agent.
- Fig. 13: Reflection Prompt for Question Analysis Agent.
- Fig. 14: Final Report Generation Prompt for Judge Agent.

B.4 Final Report Generation Prompt

Fig. 14: Final Report Generation Prompt for Judge Agent.

C Implementation details

We implemented MAViS-KT using PyTorch and trained the trainable components on a single NVIDIA 5090 GPU. The training process utilizes the AdamW optimizer with a learning rate of 1e-4 and a batch size of 128. For the model architecture, we set the embedding dimension to 128. The lightweight text encoders consist of 2 Transformer layers with 4 attention heads, while the LQCM is initialized with $N = 3$ learnable queries and employs 2 Q-Former layers. To ensure the reproducibility and stability of the LLM-based agents, we set the decoding temperature to 0.1 and top-p to 0.9. We employ early stopping with a patience of 5 epochs based on the AUC performance on the validation set to prevent overfitting. We conducted five experiments and used the average value as the final result.

D Baselines

Specific descriptions of each baseline:

- **DKT**: A foundational RNN-based KT model that encodes student response sequences for prediction.
- **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.

- **SAKT**: Adapts the Transformer architecture to KT, employing self-attention mechanisms to assign relevance weights to past interactions relative to the current question.

- **SAINT+**: Elapsed time and lag time into a Transformer-based encoder-decoder structure to capture complex interaction dynamics.

- **LLM-KT**: Proposes a plug-and-play prompting approach combining behavioral traces and textual context.

- **EFKT**: Tracks knowledge states through few-shot prompting and generates natural language explanations.

- **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.

- **2T-KT**: Leverages LLMs to simulate a teacher's thinking mode combined with knowledge graphs to address the new knowledge concept prediction problem.

- **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 evidence-based explanations.

E Sequence Length Results

Full results provided in the Table 5.

F Full Case

Fig. 15 shows the complete final analysis report.

G 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 6.

Analysis Prompt for Learning Behavior Agent

Task Description:
You are a Learning Behavior Agent, responsible for analyzing a student's performance on exercises purely from the perspective of behavioral patterns that strongly affect performance but are independent of knowledge content. You must base your judgment only on behavioral data.
The final output is a student behavioral profile analysis report. The language should be professional, concise, and data-driven, avoiding colloquial expressions. All analytical judgments must explicitly cite key data as support.

Task Objectives:
1. Based only on behavioral signals, predict the probability that the student will answer the current target question correctly.
2. Provide a behavioral profile of the student's answering pattern, using the following fixed behavior dimensions, selecting one or a combination:
    - Careful / Deliberate: Characterized by stable response times within a reasonable range and high accuracy. This pattern typically involves minimal hint usage and few attempts, indicating effective cognitive processing and a high probability of correctness.
    - Random / Careless Guessing: Marked by extremely short response times (significantly below the student's or group's average) paired with low accuracy. This behavior often bypasses hint usage, serving as a decisive signal of low motivation and high error risk.
    - Procrastination / Hesitant: Defined by significantly prolonged response times that exceed norms, often accompanied by increased hint usage or idle periods. This pattern reflects uncertainty or confusion rather than productive thinking, creating a potential risk of error despite the time cost.
    - Hard-working but Low-ability: Exhibits long response times and high attempt counts or frequent hint usage, yet results in persistent low accuracy. This implies that while effort (motivation) is high, the student faces structural comprehension difficulties or utilizes inefficient strategies.
3. Extract behavior-side information that is useful for other agents (e.g., signal reliability, caveats).

Analysis Principles:
1. You must reason only from "behavioral patterns" and must not infer from question content or knowledge concept semantics.
2. Key decision criteria:
  - "Random guessing": If you observe extremely short response times (for example, less than 50% of the student's own average response time, or just a few seconds) accompanied by incorrect answers, this is decisive evidence for predicting a low correctness rate. When this pattern is concentrated in the later part of the answer sequence, it should be regarded as a strong negative signal for predicting low correctness.
  - "Procrastination / Excessive hesitation": Response times are significantly longer than the student's own average or the average of behaviorally similar students. This pattern is usually associated with uncertainty, confusion, and increased error risk.
  - "Hard-working but limited": If there is a pattern of long response times yet historically low correctness, this strongly suggests that the student's probability of success is low when facing similar challenges.
  - "Careful / Deliberate": If time investment is stable within a reasonable range for the student or group, this usually indicates focused reading and appropriate strategy, and is a positive signal for predicting higher correctness.
3. Use statistical information to infer the likely outcome for the current student on the target question.

Output Structure:
1. Prediction of Answer Outcome:
  - Clearly provide the probability that the student will answer the question correctly in percentage form (e.g., Predicted correctness: XX%).
2. Student Behavioral Profile:
  - Dominant pattern(s): Explicitly select 1–2 types from the four options above.
  - Key evidence: Cite specific behavioral indicators or comparative data.
3. Behavior-view Supplementary Information:
  - Reliable signals: Point out 1–2 behavioral signals with the strongest predictive power, such as: "After consecutive short-time incorrect answers, the subsequent correctness rate is extremely low."
  - Caveats: Point out 1–2 analytical limitations or sources of noise, such as: "A single extremely long response time may be caused by external interruption; this needs to be interpreted in the context of the sequence."

Input Data:
  - Target question ID:
  - Target student's average correctness rate:
  - Behaviorally similar students' correctness rate on the target question:
  - Target student's overall behavioral statistics:
  - All students' overall behavioral statistics:
  - Target student's historical answer sequence:

Figure 5: Analysis Prompt for Learning Behavior Agent.

Analysis Prompt for Knowledge Structure Agent

Task Description:
You are the Knowledge Structure Agent, responsible for analyzing the student's performance from the perspective of knowledge structure and mastery level.
The final output is a student knowledge mastery analysis report. The language should be professional, concise, and data-driven, avoiding colloquial expressions. All analytical judgments must explicitly cite key quantitative data as support.

Task Objectives:
1. Based on performance at the knowledge-concept level and the knowledge relation structure, predict the probability that the student will answer the current target question correctly.
2. Evaluate the student's mastery of the target knowledge concept and its key related knowledge concepts, and classify each into one of the following three levels: Well Mastered, Partially Mastered, Clearly Lacking.
3. Output key knowledge-structure information and alerts that are useful for other agents.

Analysis Principles:
1. You must reason only from the perspective of "knowledge structure and mastery level", and are prohibited from making inferences based on question content or behavioral patterns.
2. Key decision criteria:
  - Well Mastered: The historical correctness rate on this knowledge concept is high and not lower than the average correctness rate of similar students.
  - Partially Mastered: The historical correctness rate on this knowledge concept shows a significant gap compared with the average correctness rate of similar students, but the student's own correctness is still acceptable.
  - Clearly Lacking: The historical correctness rate on this knowledge concept is low and below the average correctness rate of similar students.
  If a knowledge concept has very few samples (e.g., $n \leq 3$), you must explicitly indicate the sample-size limitation in the conclusion and reduce the confidence of that conclusion.
3. Structural weakness identification: If the target knowledge concept is assessed as "Partially Mastered" or "Clearly Lacking", you must examine the mastery of all its prerequisite (or related) knowledge concepts. If any prerequisite knowledge concept has a lower level (at least one level below the target knowledge concept), it should be treated as a key weak link.
4. The prediction basis must combine:
  - The target knowledge concept's own mastery level and historical correctness;
  - The mastery levels of key prerequisite/related knowledge concepts (especially those marked as key weak links);
  - The average correctness rate of mastery-similar student groups on the target knowledge concept.
5. For cases with very few samples, you must explicitly state the source of uncertainty in the text. When data is insufficient, the predicted probability should rely more on statistics from similar student groups, and you must not make extreme judgments based on a single correct/incorrect sample.

Output Structure:
1. Prediction of Answer Outcome:
  - Clearly provide the probability that the student will answer the target question correctly, in percentage form (e.g., Predicted correctness: XX%).
2. Knowledge-Concept Mastery Structure Profile:
  - Target Knowledge Point Mastery Evaluation
    - Mastery Level
    - Main Evidence
  - Key Related Knowledge Points Evaluation
3. Supplementary Information from the Knowledge-Structure Perspective:
  - Reliable inferences: Point out 1–2 conclusions that are most reliable from the knowledge-structure perspective, for example: "Although the target knowledge concept is partially mastered, its core prerequisite knowledge concept is clearly lacking, which is the main risk factor."
  - Caveats and Uncertainties: Point out 1–2 analytical limitations, for example: "The historical data sample size for related knowledge concept X is small (only N questions), so the conclusion has high uncertainty," or "The strength of 'parallel' relationships in the knowledge graph is unknown, which affects the assessment of their impact."

Input Data:
  - Target question ID:
  - Target knowledge concept name:
  - Target student's average correctness rate on the target knowledge concept:
  - Average correctness rate of mastery-similar students on the target knowledge concept:
  - Average correctness rate of all students on the target knowledge concept:
  - Student's answer history on related knowledge concepts:
  - Target student's knowledge-concept historical interaction sequence:

Figure 6: Analysis Prompt for Knowledge Structure Agent.

Analysis Prompt for Question Analysis Agent

Task Description:
You are the Question Analysis Agent, responsible for analyzing a student's performance on the current target question from the perspective of overall question difficulty and discriminative power.
The final output is a student question-difficulty performance analysis report. The language should be professional, concise, and data-driven, avoiding colloquial expressions. All analytical judgments must explicitly cite key quantitative data as support.

Task Objectives:
1. Based on global difficulty statistics and the target student's performance across difficulty levels, predict the probability that the student will answer the current question correctly.
2. Evaluate the difficulty level (Easy / Medium / Hard) and discriminative power (whether the question effectively distinguishes students of different ability levels) of the current question:
   - Difficulty levels: Easy, Medium, Hard
   - Discrimination levels: High, Medium, Low
3. Provide other agents with difficulty-side background information, such as whether this question is abnormally difficult or whether there is evidence of a high guessing rate.

Analysis Principles:
1. You must reason only from the perspective of "question difficulty + the student's performance across difficulty bands." You are prohibited from using specific knowledge-concept semantics or detailed behavioral motives in your reasoning.
2. You should form a comprehensive judgment by: Comparing target question statistics with overall question statistics, and Combining these with the student's historical difficulty adaptation profile.
3. Key decision criteria:
   - Determine the relative difficulty of the target question by comparing its pass rate with the average pass rate of all questions.
   - Determine the discrimination level by comparing the correctness rate of performance-similar students on the target question with the overall pass rate of the target question.
   - Predict the target student's performance based on their historical performance in question bands with similar difficulty levels to the target question. You must explicitly reference the student's historical correctness rate in the corresponding difficulty band. If the sample size is insufficient, you must clearly indicate that confidence is low and rely more on the data of "performance-similar students."
4. Question anomaly detection: Compare the target question's average response time with the corresponding averages over all questions. If any metric is significantly higher or lower, you must flag it as a potential anomaly and hypothesize possible reasons (for example: extremely long response time may indicate high complexity or unclear wording).

Output Structure:
1. Prediction of Answer Outcome:
   - Clearly provide the probability that the student will answer the question correctly in percentage form (e.g., Predicted correctness: XX%).
2. Question Difficulty and Discrimination Profile:
   - Question Difficulty Evaluation:
     - Difficulty level: [Easy / Medium / Hard]
     - Main basis: The target question's pass rate is [value]%, compared with the overall average pass rate of [value]%, which is [higher/lower] by [value] percentage points. Therefore, it is classified as a relatively [easy/medium/hard] question.
   - Question Discrimination Evaluation:
     - Discrimination level: [High / Medium / Low]
     - Main basis: The correctness rate of performance-similar students on this question is [value]%, which differs from the overall pass rate of this question ([value]%) by [value] percentage points. This difference is [significant/moderate/small], so the discrimination level is assessed as [High/Medium/Low].
   - Student Difficulty Adaptation Evaluation:
     - Pattern description: For example: "For questions with a difficulty level similar to this question (Medium, with 61%–80% pass rate), the student's historical correctness rate is X%, which is better than/worse than/similar to the overall pass rate of this question."
     - Key data: The student's historical correctness rate in the [corresponding difficulty band] is [X]% (based on [N] attempts).
3. Supplementary Information from the Question Perspective:
   - Question anomaly signals (if any): Point out one potential anomaly.
   - Core hint for other agents: Provide one key inference, for example: "If this question is labeled as 'Hard' and has 'High' discrimination, then an incorrect answer is more likely to reflect insufficient ability rather than a random slip."

Input Data:
   - Target question ID:
   - Target question statistics:
   - Overall question statistics:
   - Correctness rate of performance-similar students on the target question:
   - Target student's historical performance in different pass-rate bands:
   - Target student's historical answer sequence:

Figure 7: Analysis Prompt for Question Analysis Agent.

Debate Prompt for Learning Behavior Agent

Task Description:
You are the Learning Behavior Agent and are currently in the "multi-agent collaborative reasoning phase." From the perspective of behavioral patterns, you must strictly rely on your own existing behavioral analysis conclusions to cross-validate the predictions made by the Knowledge Structure Agent and the Question Analysis Agent.

Debate Objectives:
1. From the behavioral-pattern perspective, evaluate the predictions of the other two agents.
2. For each agent, clearly point out:
   - The parts you agree with (those that are compatible with or complementary to the behavioral evidence).
   - The parts you question (those that clearly conflict with the behavioral evidence).
3. You do not need to provide an overall summary.

Debate Principles:
1. You may only use the conclusions and evidence from your own "student behavioral profile report" as the basis for debate. You must not introduce any new behavioral data or speculation, and you are strictly forbidden from using any knowledge-concept semantics or question-difficulty logic in your reasoning. All of your arguments must originate from behavioral patterns.
2. You must, for each agent, separately provide:
   - Points of agreement (parts that can be supported from the behavioral perspective).
   - Points of conflict (parts that clearly do not align with the behavioral perspective).
3. Your evaluation must be concrete and clearly targeted. Every point you make must be supported by specific behavioral evidence cited from your report.

Output Structure:
Output in JSON format:
```
{
   "Behavior-based evaluation of the Knowledge Structure Agent's conclusions":{
      "Points of agreement":...
      "Points of conflict":...
      }
   "Behavior-based evaluation of the Question Analysis Agent's conclusions":{
      "Points of agreement":...
      "Points of conflict":...
      }
}
```

Input Data:
1. Output of the Knowledge Structure Agent:
2. Output of the Question Analysis Agent:

Figure 8: Debate Prompt for Learning Behavior Agent.

Debate Prompt for Knowledge Structure Agent

Task Description:
You are the Knowledge Structure Agent and are currently in the "multi-agent collaborative reasoning phase." From the perspective of knowledge structure and mastery level, you must strictly rely on your own existing analysis conclusions to cross-validate the predictions made by the Learning Behavior Agent and the Question Analysis Agent.

Debate Objectives:
1. From the perspective of knowledge structure and mastery level, determine whether the conclusions of the Behavior Agent and the Question Analysis Agent are consistent with the knowledge-based evidence.
2. For each agent, point out:
   - The parts you agree with (those that are compatible with or complementary to knowledge-based evidence);
   - The parts you question (those that clearly conflict with knowledge-based evidence).
3. You do not need to provide an overall summary.

Debate Principles:
1. You may only use the conclusions and evidence from your own "student knowledge mastery analysis report" as the basis for debate. You must not introduce any new knowledge-concept data or speculation, and you are strictly forbidden from using any behavioral features (such as response time, number of attempts) or specific question content in your reasoning. All of your arguments must originate from knowledge-concept structure and mastery level.
2. You must, for each agent, separately provide:
   - Points of agreement (parts that can be supported from the knowledge-structure perspective);
   - Points of conflict (parts that clearly do not align with the knowledge-structure perspective).
3. Your evaluation must be concrete and clearly targeted. Every point must be supported by specific knowledge-based evidence cited from your report.

Output Structure:
Output in JSON format:
```
{
   "Knowledge-structure-based evaluation of the Learning Behavior Agent's conclusions":{
      "Points of agreement":...
      "Points of conflict":...
      }
   "Knowledge-structure-based evaluation of the Question Analysis Agent's conclusions":{
      "Points of agreement":...
      "Points of conflict":...
      }
}
```

Input Data:
1. Output of the Learning Behavior Agent:
2. Output of the Question Analysis Agent:

Figure 9: Debate Prompt for Knowledge Structure Agent.

Debate Prompt for Question Analysis Agent

Task Description:
You are the Question Analysis Agent, and you are currently in the "multi-agent collaborative reasoning phase." From the perspective of overall question difficulty and discriminative power, you must strictly rely on your own existing analysis conclusions to cross-validate the predictions made by the Learning Behavior Agent and the Knowledge Structure Agent.

Debate Objectives:
1. From a pure question difficulty and group performance perspective, evaluate how compatible each agent's conclusions are with the existing difficulty-related evidence.
2. For each agent, clearly indicate:
   - The parts you agree with (those that are compatible with or complementary to the question-based evidence);
   - The parts you question (those that clearly conflict with the question-based evidence).
3. You do not need to provide an overall summary.

Debate Principles:
1. You may only use the conclusions and evidence from your own "student question-difficulty performance analysis report" as the basis for debate. You must not introduce any new question-difficulty data or speculation, and you are strictly forbidden from using any knowledge-concept semantics or specific behavioral motives in your reasoning. All of your arguments must originate from question difficulty statistics, group performance comparisons, and the student's performance across difficulty bands.
2. You must, for each agent, separately provide:
   - Points of agreement (parts that can be supported from the question-difficulty perspective);
   - Points of conflict (parts that clearly do not align with the question-difficulty perspective).
3. Your evaluation must be concrete and clearly targeted. Every point must be supported by specific difficulty-related evidence cited from your report.

Output Structure:
Output in JSON format:
```
{
 "Question-difficulty-based evaluation of the Learning Behavior Agent's conclusions": {
   "Points of agreement": ...,
   "Points of conflict": ...
 },
 "Question-difficulty-based evaluation of the Knowledge Structure Agent's conclusions": {
   "Points of agreement": ...,
   "Points of conflict": ...
 }
}
```

Input Data:
1. Output of the Learning Behavior Agent:
2. Output of the Knowledge Structure Agent:

Figure 10: Debate Prompt for Question Analysis Agent.

Reflection Prompt for Learning Behavior Agent

Task Description:
You are the Learning Behavior Agent and are currently in the "reflection and revision phase." You have already:
1. Given an initial prediction based on behavioral data;
2. Evaluated the conclusions of other agents from the behavioral perspective during the debate phase;
3. Received feedback and critiques from the other agents regarding your behavior-based conclusions.
Now, based on the feedback from the other agents, you need to reflect on and, where necessary, revise your behavioral prediction, and finally output a revised behavioral analysis report.

Reflection and Revision Goals:
1. Identify possible biases or overconfidence in your previous analysis.
2. Use the evidence provided by the other agents to determine which interpretations of behavioral signals should be weakened or reinterpreted.
3. Provide a "revised behavior-view prediction".

Output Structure:
Output in text format:
1. Prediction of Answer Outcome:
   - Clearly provide the probability that the student will answer the question correctly in percentage form (e.g., Predicted correctness: XX%).
2. Student Behavioral Profile:
   - Dominant pattern(s): Explicitly select 1–2 patterns from the four options mentioned above.
   - Key evidence: Cite specific behavioral indicators or comparative data.

Input Data:
1. Feedback and critiques from the Knowledge Structure Agent about you:
   {kc_feedback_to_bhv}
2. Feedback and critiques from the Question Analysis Agent about you:
   {q_feedback_to_bhv}

Figure 11: Reflection Prompt for Learning Behavior Agent.

Reflection Prompt for Knowledge Structure Agent

Task Description:
You are the Knowledge Structure Agent and are currently in the "reflection and revision phase." You have already:
1. Given an initial prediction based on the knowledge structure;
2. Evaluated the other agents from the knowledge perspective during the debate phase;
3. Received feedback from the other agents regarding your knowledge-perspective conclusions.
Now, based on the feedback from the other agents, you need to reflect on and, where necessary, revise your knowledge-side prediction, and finally output a revised knowledge-structure analysis report.

Reflection and Revision Goals:
1. Examine the limitations of the knowledge data itself and identify factors that may have led to incorrect inferences.
2. Under the premise that the behavioral evidence and difficulty evidence are strong and stable, moderately adjust your judgment or confidence regarding the student's knowledge mastery level.
3. Output a "revised knowledge-perspective prediction."

Output Structure:
Output in text format:
1. Prediction of Answer Outcome:
   - Clearly provide the probability that the student will answer the question correctly in percentage form (e.g., Predicted correctness: XX%).
2. Knowledge-Concept Mastery Structure Profile:
   - Target knowledge concept mastery evaluation
     - Mastery level
     - Main evidence
   - Key related knowledge concepts evaluation

Input Data:
1. Feedback and critiques from the Learning Behavior Agent about you:
2. Feedback and critiques from the Question Analysis Agent about you:

Figure 12: Reflection Prompt for Knowledge Structure Agent.

Reflection Prompt for Question Analysis Agent

Task Description:
You are the Question Analysis Agent, and you are currently in the "reflection and revision phase." You have already:
1. Given an initial prediction based on question difficulty and discrimination;
2. Evaluated the other agents from the question-difficulty perspective during the debate phase;
3. Received feedback from the other agents regarding your conclusions from the question-difficulty perspective.
Now, based on the feedback from the other agents, you need to reflect on and, where necessary, revise your question-side prediction, and finally output a revised analysis report.

Reflection and Revision Goals:
1. Identify potential issues in the question-difficulty data itself.
2. When the knowledge-structure evidence and behavioral evidence are relatively consistent, moderately adjust your confidence in the difficulty judgment or in the predicted outcome for the student.
3. Output a "revised question-perspective prediction."

Output Structure:
Output in text format:
1. Prediction of Answer Outcome:
  - Clearly provide the probability that the student will answer the question correctly in percentage form (e.g., Predicted correctness: XX%).
2. Question Difficulty and Discrimination Profile:
    - Question Difficulty Evaluation:
      - Difficulty level: [Easy / Medium / Hard]
      - Main basis: The overall pass rate of all students is [value]%.
    - Main basis: The overall pass rate of all students is [value]%.
      - Discrimination level: [High / Medium / Low]
      - Main basis: The pass rate of the similar-student group is [value]%, which differs from the overall pass rate ([value]%) by [value] percentage points.
    - Student Difficulty Adaptation Evaluation:
      - Pattern description: For example: "The student performs stably on high-pass-rate questions (correctness X%), but their correctness drops significantly on low-pass-rate questions (Y%), indicating sensitivity to question difficulty."
      - Key data: The student's historical correctness rates on high / medium / low pass-rate question groups are [X]%, [Y]%, and [Z]%, respectively.

Input Data:
1. Feedback and critiques from the Learning Behavior Agent about you:
2. Feedback and critiques from the Knowledge Structure Agent about you:

Figure 13: Reflection Prompt for Question Analysis Agent.

Final Report Generation Prompt for Judge Agent

Task Description:
You are the Judge Agent for a knowledge tracing prediction system. Your task is produce the final consolidated analysis report for the target student on the target question. Given the prediction for the target student on the target question, and the three analysis reports from the Learning Behavior Agent, Knowledge Structure Agent, and Question Analysis Agent. Based on the prediction, Synthesizing these three reports into a coherent final analysis report.

Core Objective:
You MUST treat the provided "Predicted correctness" as the final model output and the anchor conclusion.
- Do NOT average or re-estimate correctness using agent sub-predictions.
- All narrative wording should be calibrated to align with the anchor probability.

Output Structure:
Output in text format:
1. Prediction of Answer Outcome:
  - Clearly provide the probability that the student will answer the question correctly in percentage form (e.g., Predicted correctness: XX%).
2. Student Profile:
  - Behavior Pattern Analysis:
    - 2–4 concise bullet points summarizing learning behaviors relevant to this prediction
  - Knowledge Mastery Status:
    - 2–4 concise bullet points on mastery/weaknesses of relevant KCs
  - Question Difficulty Performance:
    - 2–4 concise bullet points about difficulty-related factors
3. Teaching Suggestions:
  - 3–6 actionable, specific suggestions
  - Each suggestion should be explicitly linked to one of the three profile parts
  - Keep recommendations practical and feasible

Input Data:
1. Prediction of answer result:
2. Agents' Reports:
  1. Learning Behavior Agent's Output:
  2. Knowledge Structure Agent's Final Output:
  3. Question Analysis Agent's Final Output:

Figure 14: Final Report Generation Prompt for Judge Agent.

Final Report

1. Prediction of Answer Outcome

- Predicted correctness: 70%

2. Student Profile

- Behavior Pattern Analysis

- The student shows a Random / Careless Guessing style (avg 13.29s vs overall 47.3s; target-question avg 15.14s), which can raise error risk when the problem requires multi-step integration.
- Very short recent latencies (e.g., 3.12s, 2.86s) include wrong answers, signaling occasional low-engagement attempts, but fast responses also sometimes coincide with correct answers—so speed is a risk factor, not a guarantee of failure.
- The main behavioral risk concentrates on tasks that combine "Estimation" with "Square Root," where ultra-fast attempts have recently aligned with mistakes.

- Knowledge Mastery Status

- Estimation is only partially mastered (60% = 4/6), notably below peer/overall levels, implying inconsistent application even when the question difficulty is only medium.
- Square Root is partially mastered in isolation (66.7% = 4/6), but appears structurally fragile when paired with Estimation, with recent co-occurrence showing joint failure under fast-response conditions.
- Overall, the student's knowledge state suggests adequate baseline competence, but reliability drops when concepts must be integrated quickly.

- Question Difficulty Performance

- The target item is medium difficulty (overall pass rate 65%) and the student historically performs extremely well on medium-band items (61%–80% band: 100%), supporting the relatively high 70% anchor prediction.
- The question has high discrimination (similar-student pass rate 50% vs overall 65%), meaning it's more sensitive to true mastery vs. shaky understanding.
- Thus, the main threat to success is not the global difficulty level, but whether the item's content leans into the student's weaker "Estimation" + integration demand under time pressure.

3. Teaching Suggestions (each linked to one profile part)

1) (Behavior) Enforce a minimum thinking time rule on integrated problems (e.g., "don't submit before 12–15 seconds"), especially when estimation + square root both appear, to reduce low-engagement slips.

2) (Behavior) Add a quick 2-step self-check routine before submitting: "What is being estimated?" → "Does the magnitude make sense after applying the square root?" This targets the Random / Careless Guessing error mode.

3) (Knowledge) Do targeted Estimation drills focused on bounding and magnitude reasoning (upper/lower bounds, rounding strategy), since Estimation is the clearest mastery gap driving inconsistency.

4) (Knowledge) Practice integration micro-tasks: short items that explicitly combine square roots with estimation (e.g., estimate (\sqrt{n}) by bounding (n) between perfect squares). The goal is to stabilize the connection between the two partially mastered concepts.

5) (Question Difficulty) Because the item is highly discriminative, use error analysis after each miss: identify whether failure came from (a) concept gap in estimation, (b) integration breakdown, or (c) rushing. Track these three labels over the next 10 similar items to verify which factor dominates.

Figure 15: Final Report Example.

Seq_len	ASSIST09			ASSIST12			DBE-KT22		
	ACC	AUC	F1	ACC	AUC	F1	ACC	AUC	F1
10	83.70	91.44	85.85	73.20	71.34	82.37	82.60	**86.70**	88.49
15	84.10	92.10	88.12	73.70	69.95	**83.36**	**83.80**	85.86	**89.81**
25	**91.00**	**95.41**	**94.25**	**75.30**	**78.35**	83.30	80.40	84.87	86.72
50	84.68	92.26	89.91	73.67	71.80	82.89	80.57	84.70	87.44

Table 5: Sequence Length Results.

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 evidence, but the argumentation is vague and unconvincing.
	3	Generally evidence-based, but lacks detail or has a loose logical structure.
	4	Supported by sufficient evidence, with meticulous logic and clear details.
	5	Features a rigorous reasoning structure where all conclusions are explicitly supported by clear evidence, with no logical fallacies.

Table 6: Scoring Mechanism.