ConceptKT: A Benchmark for Concept-Level Deficiency Prediction in Knowledge Tracing

Yu-Chen Kang, Yu-Chien Tang, An-Zi Yen
Department of Computer Science, National Yang Ming Chiao Tung University, Taiwan
{connie.cs12,tommytyc.cs10}@nycu.edu.tw, azyen@nycu.edu.tw

Abstract

Knowledge Tracing (KT) is a critical technique for modeling student knowledge to support personalized learning. However, most KT systems focus on binary correctness prediction and cannot diagnose the underlying conceptual misunderstandings that lead to errors. Such fine-grained diagnostic feedback is essential for designing targeted instruction and effective remediation. In this work, we introduce the task of concept-level deficiency prediction, which extends traditional KT by identifying the specific concepts a student is likely to struggle with on future problems. We present ConceptKT, a dataset annotated with labels that capture both the concepts required to solve each question and the missing concepts underlying incorrect responses. We investigate in-context learning approaches to KT and evaluate the diagnostic capabilities of various Large Language Models (LLMs) and Large Reasoning Models (LRMs). Different strategies for selecting informative historical records are explored. Experimental results demonstrate that selecting response histories based on conceptual alignment and semantic similarity leads to improved performance on both correctness prediction and concept-level deficiency identification.

Keywords: Concept-Level Deficiency Prediction, Historical Response Selection, Knowledge Tracing

1. Introduction

With the proliferation of online education platforms becoming mainstream learning environments, knowledge tracing (KT) (Anderson et al., 1990) has emerged as a foundational technique for monitoring learners' evolving knowledge states and supporting personalized learning interventions. Specifically, KT seeks to predict a student's latent knowledge state by modeling their historical response records, thereby enabling the dynamic adaptation of instructional strategies and learning resources. Prior work (Corbett and Anderson, 1994; Piech et al., 2015; Pandey and Karypis, 2019; Ghosh et al., 2020; Pandey and Srivastava, 2020) in KT has primarily focused on leveraging deep sequential models and attention mechanisms to capture students' evolving knowledge states, with the goal of predicting their overall answer correctness. Memory-augmented models (Zhang et al., 2017; Abdelrahman and Wang, 2019; Liu et al., 2019; Wang et al., 2020a) have been proposed to incorporate exercise content and concept-level information via memory-augmented architectures to infer a student-concept mastery matrix. Other approaches (Zhang et al., 2024; Cui et al., 2023; Luo et al., 2024; Qin et al., 2025) enhance interpretability through context-aware attention, relational modeling, hierarchical concept structuring, or difficulty-aware mechanisms. However, most works focus on binary correctness prediction, with limited attention to the prediction of concept-level deficiencies.

In practical educational scenarios, the utility of correctness prediction is enhanced when accompanied by insight into the specific concepts in which the student is likely to struggle. Such fine-grained diagnostic feedback is essential for informing personalized instruction, designing targeted remedial materials, and implementing effective scaffolding strategies. Taking Table 1 as an example, a student is required to first convert volume from liters to cubic meters, then compute the surface area of a cube, and finally estimate the number of metal sheets needed based on their area. The student's solution directly divides the volume in cubic meters by the area of a metal sheet, indicating a lack of understanding in both volume and surface area calculations, which ultimately leads to an incor-

Question: A welder received an order to make a 1 million liter cube - shaped tank . If he has only 4×2 meter sheets of metal that can be cut , how many metal sheets will be required for this order r ? (1 cubic meter = 1000 liters)
Student Process:

$$1000000L = 1000m^3$$

$$1000 \div 8 = 25$$

Correctness: Wrong

Associated Concepts: Volume Calculation, Area Calculation, Unit Conversion
Error Type: Wrong Mathematical Operation/Concept
Error Equation:

$$1000 \div 8 = 25$$

Missing Concepts: Volume Calculation, Area Calculation
Teacher Feedback: The concept is incorrect. The metal sheets are only used for the six surfaces of the tank. Therefore, you should calculate the area of each surface and then calculate the number of metal sheets required, instead of using the total volume. The area of each surface is $1000 = 10^3$, $10 \times 10 = 100$. Therefore, the number of metal sheets required is $100 \div 8 \times 6 = 75$.

Table 1: Example of a ConceptKT Instance Used for Answer Correctness and Concept-Level Deficiency Prediction based on Student Solution Processes.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 333–343
11–16 May 2026.

rect answer. Beyond predicting whether a student will answer correctly, the ability to anticipate specific concept-level deficiencies can significantly enhance the design of diagnostic assessments and adaptive instruction. By anticipating concept-level deficiencies, instructional systems can adaptively select assessment items that align more closely with each student's learning needs. This strategy enhances the effectiveness of adaptive testing by targeting concepts where the student is likely to struggle, thereby improving diagnostic precision and instructional relevance. For instance, if a system predicts that a student is likely to struggle with surface area calculation, it can provide preparatory exercises or modify the task flow accordingly.

In this work, we construct ConceptKT by extending the MathEDU dataset (Hsu et al., 2025) with concept-level annotations provided by three experts in mathematics education. MathEDU consists of 4,048 solution process records from 6 students with diverse academic backgrounds, including Applied Mathematics, Finance, Japanese, Information Management, Mathematics Education, and Physics. Each record includes rich error-related annotations, such as error type (e.g., incorrect operations or conceptual misunderstandings), error equation (the specific erroneous step), and teacher feedback that addresses the student's misconceptions. As shown in Table 1, we further label two types of annotations (1) the **Associated Concepts** required to solve the problem, and (2) the **Missing Concepts** that the student failed to demonstrate mastery of when answering incorrectly.

Recently, large language models (LLMs) and large reasoning models (LRMs) have demonstrated remarkable capabilities in task understanding and reasoning. An increasing number of studies show that LMs have strong potential in educational contexts, including math problem solving (Yang et al., 2023; Didolkar et al., 2024), student feedback generation (Hsu et al., 2025; Baral et al., 2024), and even knowledge tracing (Cho et al., 2024). Therefore, this study examines the capability of LMs to support concept-level diagnosis by extending KT beyond correctness prediction to include the identification of likely conceptual deficiencies through in-context learning. Note that KT often requires models to process and integrate long sequences of student responses to accurately capture the evolution of their knowledge states. Incorporating extensive historical data can introduce information redundancy or contextual noise, which may hinder the model's ability to effectively represent and reason about the student's current knowledge. This raises two research questions:

RQ1: Should the entire history be fed into the model, or is it more effective to select only a subset of responses?

RQ2: If selection is required, which records are most informative for predicting correctness and concept-level deficiencies?

To answer these questions, we explore strategies for selecting prior responses according to the concepts required by the target question, in order to predict both the correctness of a student's response and the specific concept deficiency underlying an incorrect answer. In sum, our contributions are threefold: (1) To support fine-grained instructional decisions in KT, we introduce an augmented task, i.e., concept-level deficiency prediction, that extends beyond correctness prediction by identifying the concept a student is likely to struggle with. (2) We present a novel dataset, ConceptKT,[1] with expert-annotated concept-level labels that capture both the required concepts for each problem and the concepts in which students failed to apply correctly in their erroneous solutions. (3) We evaluate various LMs within an in-context learning paradigm for knowledge tracing, and analyze response history selection strategies. The experiments demonstrate promising effects of selecting prior responses by conceptual relevance to the target question.

2. Related Work

2.1. LLM-Enhanced and Open-Ended Knowledge Tracing

Previous studies on knowledge tracing (KT) have primarily focused on deep sequential and attention-based architectures, such as DKT (Piech et al., 2015), SAKT (Pandey and Karypis, 2019), and AKT (Ghosh et al., 2020), which model students' response histories to capture temporal dependencies. Memory-augmented approaches, including DKVMN (Zhang et al., 2017) and EKT (Liu et al., 2019), extend these models to track concept-level mastery with external memory components. Graph-based extensions such as GKT (Nakagawa et al., 2019), AGKT (Long et al., 2022) and DyGKT (Cheng et al., 2024) further incorporate structural relationships among concepts for enhanced reasoning. While these models achieve strong correctness prediction, they provide limited diagnostic interpretability at the concept level.

Recently, LLMs have been applied to educational tasks involving knowledge tracing. LLM-KT (Wang et al., 2025) uses a pre-trained language model (e.g., BERT (Devlin et al., 2019) or LLaMA (Touvron et al., 2023)) to convert questions and concepts into embeddings, and employs traditional sequence learning models to encode question IDs and concept IDs. These embeddings are then injected into

[1] https://github.com/NYCU-NLP-Lab/ConceptKT

the prompt of an LLM to capture both students' sequential behavioral patterns and the semantic features of the question text. LLMKT (Scarlatos et al., 2025) is applied to tutor-student dialogue scenarios. It utilizes prompt-based techniques to identify the knowledge components involved in student responses and provides real-time assessments of their mastery levels. DDKT (Cen et al., 2025) enables LLMs to perform step-by-step reasoning over questions and generate solution processes in order to estimate difficulty, it also calculates statistical difficulty based on students' historical correctness rates. By combining both the difficulty estimated by the LLMs and the statistical difficulty derived from performance data, DDKT (Cen et al., 2025) enhances the predictive accuracy in KT tasks. KnowTrace (Li et al., 2025) and SINKT (Fu et al., 2024) focus on knowledge structure and graph-based modeling. In addition, several studies have begun to explore the application of knowledge tracing to open-ended problems, such as OKT (Liu et al., 2022) and ECKT (Yu et al., 2024).

2.2. Mathematics Education Datasets

Several datasets used in mathematics education research have been constructed. The ASSISTments datasets[2] are derived from the online mathematics tutoring platform. It includes multiple versions of student interaction records, covering secondary school mathematics curricula in the United States. The Junyi dataset (Pojen et al., 2020) comes from the learning platform Junyi Academy. The data was collected after 2016 and covers mathematical concepts and problem types from elementary to junior high school. Eedi2020 (Wang et al., 2020b) is a multiple-choice dataset released by the educational platform Eedi. The questions are designed for students in Grades 7-9. The Algebra datasets, presented at the KDD Cup 2010 Educational Data Mining Challenge (Stamper et al., 2010), target secondary school students (Grades 8–10) studying algebra. EdNet (Choi et al., 2020) contains student interaction data from 2018 to 2020, mainly focusing on secondary school mathematics. DBE-KT22 (Abdelrahman et al., 2022) covers responses from junior high school students.

DrawEduMath (Baral et al., 2025) is collected around 2022 from the DrawEdu platform. It contains data from students in Grades 7-12 solving interactive geometry drawing problems. Erickson et al. (2020) constructed a dataset comprising student responses to open-ended mathematics questions, sourced from open educational resources such as EngageNY, Illustrative Mathematics, and Utah Math. The questions were manually graded by teachers, and each student response is labeled as "Correct," "Partially Correct," or "Incorrect." Previous datasets used for KT have included annotations of the corresponding knowledge concepts for each question. However, most of these datasets only record the final responses of students, lacking detailed traces of the reasoning steps involved in their problem-solving processes. MathEDU is a dataset derived from the MathQA dataset (Amini et al., 2019), focused on word-based math problems at the secondary school level. Students from diverse backgrounds were invited to write detailed reasoning and solutions. The dataset includes students' step-by-step responses, correctness annotations, and error type labels, such as conceptual misunderstandings and calculation errors. Thus, we extend the MathEDU dataset into ConceptKT by incorporating concept-level annotations.

[2] https://www.etrialstestbed.org/resources/featured-studies/dataset-papers

	#Ans.	Correct Answers	Wrong Answers	
			Careless Mistake	Concept-Level Deficiencies
S1	683	70.57%	18.01%	11.42%
S2	685	87.59%	7.45%	4.96%
S3	678	71.09%	13.28%	15.63%
S4	660	75.30%	15.00%	9.70%
S5	682	67.01%	17.30%	15.69%
S6	660	80.61%	8.18%	11.21%
Overall	4,048	75.35%	13.21%	11.44%

Table 2: Distribution of Student Responses.

3. Dataset Construction and Analysis

3.1. From MathEDU to ConceptKT

MathEDU includes annotated error types related to students' problem-solving processes:

Wrong mathematical operation/concept: Student applies an incorrect mathematical operation or uses an inappropriate mathematical concept to solve a problem.

Lack of necessary mathematical concepts: Errors in answering caused by a lack of essential mathematical knowledge or techniques.

Calculation Error: Mistakes in calculations, such as errors in solving equations, arithmetic mistakes, and incorrect unit conversions.

Incomplete Answer: Student used a correct formula or procedure but did not complete it.

Careless Error: Errors caused by students' carelessness in answering, including number substitution errors and missing digits.

We consider "Wrong Mathematical Operation/Concept" and "Lack of Necessary Mathematical Concepts" as indicators of **conceptual deficiency**, while the remaining three error types are regarded as **careless mistakes**. Thus, we focus

on the two concept-related error types and annotate the corresponding data with concept labels.

To characterize students' problem-solving performance, Table 2 reports the proportions of correct and incorrect problem-solving results for each individual. Students exhibit different error patterns. While most make more careless mistakes, some (e.g., Students 3 and 6) show greater concept-level deficiencies, indicating deeper conceptual gaps. These findings highlight that modeling students' learning states requires not only predicting correctness but also distinguishing error types to reveal their underlying causes.

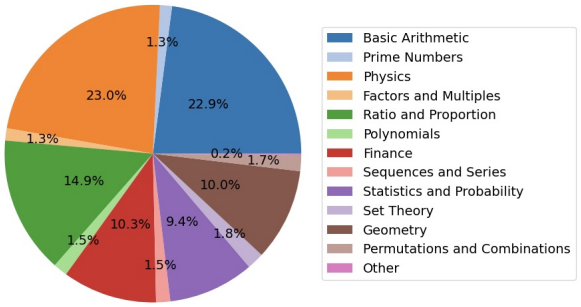

Figure 1: Distribution of Questions Across 13 Categories.

3.2. Data Annotation

Based on the U.S. Common Core State Standards for mathematics[3] for K-12 students and the questions in MathEDU, we define 55 core mathematical concepts. Three experts were invited in mathematics education to review each student's problem-solving process and annotate both the associated concepts assessed by the problem and the missing concepts demonstrated by the student's errors. Before annotation, they were instructed to carefully review the guideline. Both associated concepts and missing concepts were annotated using a multi-label format, as each problem may involve multiple relevant concepts and a student's error may arise from multiple conceptual deficiencies.

To assess annotation quality, we computed Fleiss' κ to evaluate inter-annotator agreement among the three experts. The κ score was 0.6799 for associated concepts and 0.6328 for missing concepts, indicating substantial agreement according to standard interpretation thresholds. In cases of annotation disagreement, we applied a majority voting scheme. Each label was determined by the majority of annotators. When majority voting failed to resolve uncertainty, the annotators engaged in collaborative discussion until consensus was reached. As a result, a total of 4,048 records were annotated. Each question was labeled with an average of 1.2441 associated concepts. Among these, 463 records involved incorrect responses requiring the annotation of missing concepts, with an average of 1.112 missing concepts per question.

3.3. Dataset Statistics and Analysis

3.3.1. Concept Statistics

In consultation with the experts, we further grouped the 55 defined concepts into 13 high-level categories, including "Basic Arithmetic," "Prime Numbers," "Factors and Multiples," "Physics," "Ratio and Proportion," "Finance," "Statistics and Probability," "Polynomials," "Sequences and Series," "Geometry," "Set Theory," "Permutations and Combinations," and "Other". The 13 categories contain the following 55 concepts, each associated with its respective ID:

Basic Arithmetic: 1. Basic Math Operations, 2. Expression and Operations with Symbols, 3. Unit Conversion, 4. Range of Values, 5. Modular Arithmetic, 6. Factorial

Prime Numbers: 7. Prime, 8. Composite Number, 9. Prime Factorization

Factors and Multiples: 10. Number of Factors, 11. Greatest Common Divisor, 12. Least Common Multiple

Physics: 13. Distance-Time-Speed, 14. Relative Speed, 15. Workload-Time-Speed, 16. Mixed Solution Concentration

Ratio and Proportion: 17. Direct Proportion and Inverse Proportion, 18. Ratio Calculation

Finance: 19. Simple Interest, 20. Compound Interest, 21. Effective Annual Interest Rate, 22. Profit, 23. Loss, 24. Discount Problem, 25. Price Calculation

Statistics and Probability: 26. Arithmetic Mean, 27. Mode, 28. Median, 29. Standard Deviation, 30. Normal Distribution, 31. Probability

Polynomials: 32. Square of the Sum, 33. Difference of Squares, 34. Sum of Squares, 35. Factorization of Polynomials, 36. Function, 37. Roots and Coefficients, 38. Quadratic Equation, 39. Absolute Value Equation

Sequences and Series: 40. Arithmetic Sequence/Series, 41. Geometric Sequence/Series

Geometry: 42. Perimeter Calculation, 43. Area Calculation, 44. Volume Calculation, 45. Graphs of Cartesian Coordinates and Linear Equations, 46. Pythagorean Theorem, 47. Solid Geometry, 48. Plane Geometry

Set Theory: 49. Inclusion-Exclusion Principle, 50. Fundamental Counting Principles

Permutations and Combinations: 51. Permutations and Combinations, 52. Pigeonhole Principle

Other: 53. Exact Value, 54. Local Value, 55. Face Value

[3] https://corestandards.org/mathematics-standards/

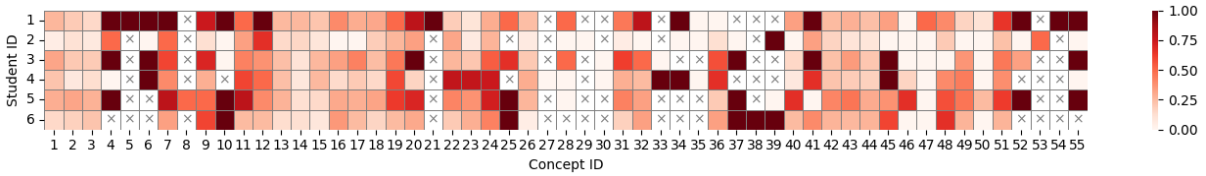

Figure 2: Concept-level error rates across students, where darker shades indicate higher error proportions and "×" marks denote concepts that were not covered in the student's problem-solving history.

Figure 1 shows the distribution of questions across these categories. Among them, "Basic Arithmetic", "Physics", and "Ratio and Proportion" constitute the largest proportions, suggesting that these categories encompass a broad range of problem types and are frequently encountered in ConceptKT. In contrast, categories such as "Prime Numbers" and "Factors and Multiples" each account for less than 2% of the dataset.

3.3.2. Student Concept Mastery Analysis

To better understand each student's concept mastery, we present a heatmap in Figure 2. where the vertical axis corresponds to the six student IDs and the horizontal axis indicates the concept IDs. Each cell in the heatmap represents the error rate of a student on a specific concept. Darker colors indicate higher error rates, corresponding to lower levels of conceptual understanding. Cells marked with "×" indicate that the student did not encounter any problems involving the corresponding concept.

Students 1, 3, and 5 exhibit high error rates in Concept IDs 1–6 (i.e., "Basic Arithmetic"), suggesting a weaker mathematical foundation that may have negatively impacted their performance across other concepts as well. Their errors span a wide range of concepts, indicating generally lower mastery across multiple areas. In contrast, Student 2 demonstrates relatively strong performance across most concepts. Student 4 shows lower performance in Concept IDs 19–25 (i.e., "Finance") and Concept IDs 32–39 (i.e., "Polynomials"). Student 6 also struggle with Polynomials.

4. Methodology

4.1. Task Formulation

This study predicts a student's performance on new math questions and identifies potential concept deficiencies based on their problem-solving history. We adopt an in-context learning framework with LLMs to perform this extended KT task. To guide the model's reasoning process, we incorporate a prompt-based Chain-of-Thought (CoT) strategy, in which structured instructions are provided to elicit

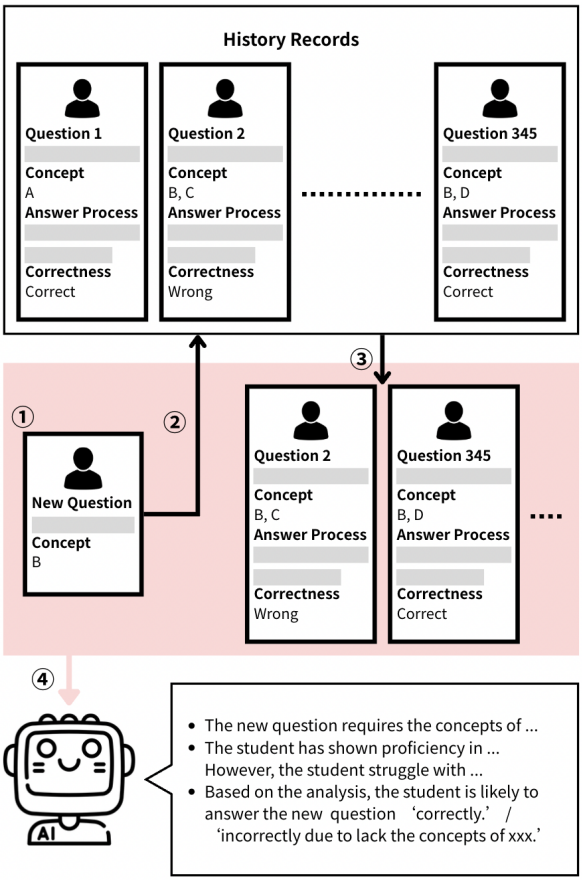

Figure 3: Overview of Knowledge Tracing.

step-by-step conceptual analysis. Figure 3 shows the overview of our task.

Specifically, let q_i denote the i-th question, r_i denote the student's solution process for q_i, and $c_i \in \{correct, wrong\}$ indicate the corresponding correctness label. For the target question q, the student's response history $H = \{(q_1, r_1, c_1), \ldots, (q_n, r_n, c_n)\}$ is presented to the model $\mathcal{M}$ as few-shot examples. The model is instructed to perform three steps: (1) **Associated Concept Identification:** infer the associated concepts required to solve the target question q, with the resulting analysis denoted as α; (2) **Student Modeling:** analyze the student's response history H to assess their concept-level mastery, with the resulting analysis denoted as β; (3) **Answer Correctness and Concept-level Deficiency Prediction:** predict the student's response

	Total	Training Set	Test Set
Student 1	683	615	68
Student 2	685	616	69
Student 3	678	610	68
Student 4	660	594	66
Student 5	682	614	68
Student 6	660	594	66
Total	4,048	3,643	405

Table 3: Number of Questions in the Training and Test Sets.

correctness c for the target question q, and identify the set of concept-level deficiencies γ that may contribute to an incorrect response. The complete process can be expressed as $(\alpha, \beta, c, \gamma) = \mathcal{M}(H, q)$.

However, not all prior problem-solving records are equally relevant to the target question. Therefore, it is necessary to selectively sample the most informative ones for predicting a student's future performance. Formally, given a response history H consisting of n interactions, the objective is to select a subset of m records ($m < n$) to support effective correctness and Concept-level Deficiency Prediction. In our experiments, models capable of encoding up to 110K tokens are sufficient to fully accommodate a student's entire response history. We use models with sufficient context length (up to 1 million tokens), allowing the entire student history to be included. We investigate whether selective record inclusion remains beneficial.

4.2. Response Selection Strategies

The main principle behind the selection strategies is that examples involving similar concepts provide more relevant context for reasoning about a student's conceptual mastery. Accordingly, we select prior responses that cover one or more of the same concepts required by the target question q. The following selection strategies are tested:

(1) All Responses: All responses of the student are used without any filtering or selection.

(2) Same-Concept Only: All responses that involve the same concepts as q are selected.

(3) Conceptual Semantic Selection: This strategy prioritizes prior responses that are both conceptually and semantically relevant to q. We first identify the subset of records in the student's history H that involve one or more of the same concepts as q, yielding k candidate responses. If $k \leq m$, all available responses are used. If $k > m$, we compute the semantic similarity between each candidate question and q using BERTScore (Zhang et al., 2019), and select the top m responses with the highest similarity. We limit the number of responses to m. If fewer than m responses are available, only those records are used.

You will be provided with several sets of historical records, ending with a new question.
Task Instructions: 1. Identify the concepts required to answer the new question (Do not solve the new question). 2. Compare these concepts with the student's past performance.
3. Determine whether the student can answer the new question correctly.
- If the student is likely to answer correctly, output: Correct. No lacking concepts.
- If the student is likely to answer incorrectly, output the missing concepts using this strict format: Wrong. Lack of concept1 && concept2
- Replace concept1, concept2, etc., with actual missing concepts from the **Concept List**.
- If multiple concepts are missing, separate them using " && " (no extra spaces, punctuation, or explanations).
- Never use placeholders like concept1 && concept2—only real concepts from the list.
Concept List (Strict Selection Only):
{Concept list as specified in Section 3.3.1}
Historical Records:
Question 1: {Question}
Student Response: {Response}
Student Response Correctness: {Yes or No}

Question 2: {Question}
Student Response: {Response}
Student Response Correctness: {Yes or No}

...

New Question: {Target Question}

Table 4: KT Inference Prompt.

5. Experiments

5.1. Experimental Setup

We experiment with various LLMs and LRMs on our task, with parameter sizes ranging from 7B to 671B. All models are configured with a fixed temperature of 0 to ensure reproducibility. A unified prompt template is used for the KT task across all models. Adaptive prompt tuning tailored to each model is left for future work.

Table 3 shows the number of questions in the training and test sets. The training set is used for selecting historical responses to represent student knowledge mastery, while the test set provides the target questions used for evaluation. Table 4 presents the prompt used to instruct the model to perform the KT task.

We chronologically order each student's problem-solving records and use the first 90% as the response history H and the remaining 10% as the test set. Formally, for a student with N total responses, we define the first $n = \lfloor 0.9N + 0.5 \rfloor$ records as $H = (q_1, r_1, c_1), \ldots, (q_n, r_n, c_n)$, which

Selection Strategy	Model	Correctness	Missing Concept		
		Accuracy	Macro-F1	Macro-Precision	Macro-Recall
N/A	DKT	64.80%	-	-	-
N/A	DKVMN	63.10%	-	-	-
N/A	GKT	63.20%	-	-	-
N/A	SAKT	66.46%	-	-	-
N/A	OKT	69.25%	1.87%	2.07%	1.92%
All Responses	Gemini-2.0-Flash	48.01%	14.08%	12.12%	29.38%
	Llama-4	68.81%	5.43%	5.66%	5.44%
	o3-mini	70.35%	2.16%	2.67%	2.11%
	DeepSeek-R1	67.70%	14.82%	13.85%	19.06%
Same-Concept Only	Gemini-2.0-Flash	**60.83%**	**15.99%**	**13.63%**	**29.66%**
	Llama-4	73.67%	10.13%	12.07%	9.86%
	o3-mini	70.35%	10.90%	11.36%	11.81%
	DeepSeek-R1	**71.02%**	**17.40%**	17.00%	**23.46%**
Conceptual Semantic Selection	Gemini-2.0-Flash	60.62%	15.55%	13.23%	26.57%
	Llama-4	**73.89%**	**13.83%**	**14.46%**	**14.48%**
	o3-mini	**71.02%**	**12.77%**	**14.39%**	**15.61%**
	DeepSeek-R1	70.12%	16.87%	**18.63%**	20.56%

Table 5: Results of Answer Correctness and Concept-Level Deficiency Prediction.

are provided to the model $\mathcal{M}$ as few-shot examples. The model is then tasked with predicting the answer correctness c_{n+j} and concept-level deficiencies γ_{n+j} for each target question q_{n+j}, where $j = 1, 2, \ldots, N - n$.

We set m to 30 in our experiments, and further investigate the effect of varying m values in the subsequent analysis. Gemini-2.0-Flash,[4] Llama-4,[5] o3-mini,[6] DeepSeek-R1 (Guo et al., 2025) are employed. Additionally, we train DKT, DKVMN, GKT, SAKT, and OKT as baseline models, among which only OKT can predict concept deficiencies.

5.2. Experimental Results

We evaluate two tasks in this study: answer correctness prediction and concept-level deficiency prediction. Accuracy is used as the evaluation metric for the answer correctness prediction. Macro-F1 is adopted for the concept-level deficiency prediction, as it is a multi-label classification problem.

Table 5 presents the performance of various LLMs and LRMs under different historical response selection strategies. The best-performing result for each model across all selection strategies is highlighted in bold, while the second-best is underlined. Comparisons are made within each individual model (i.e., within-row grouping) to isolate the impact of different selection strategies.

OKT demonstrates strong accuracy in predicting answer correctness. However, it performs poorly in predicting concept-level deficiencies. Comparing different strategies, using all responses without filtering (*All Responses*) leads to a significant drop in concept-level deficiency prediction performance. For example, Llama-4 and o3-mini achieve Macro-F1 scores of only 5.43% and 2.16%, respectively. In contrast, restricting the input to responses involving the same concepts as the target question (Same-Concept Selection) leads to substantial performance gains. DeepSeek-R1 achieves the highest Macro-F1 score (17.40%) for concept-level deficiency prediction among all models in this setting, along with strong performance on answer correctness prediction (71.02%). These results indicate that even when a model is capable of ingesting the full student history, concept-aware input filtering remains critical for guiding model reasoning and maintaining diagnostic precision.

The results of "Conceptual Semantic Selection" show that this strategy further improves performance for certain models, confirming its effectiveness in selecting informative responses. Using this strategy, o3-mini exhibits a remarkable improvement in concept-level deficiency prediction, rising from 2.16% to 12.77%, with a statistically significant difference (t-test, $p < 0.01$). Hence, for models with reasoning capacity, carefully selecting a small set of conceptually relevant examples can be more beneficial than simply increasing the quantity of input history.

In summary, these findings suggest that simply increasing the number of input responses does not necessarily enhance model performance. Instead, the quality and conceptual relevance of the selected responses are key determinants of success in both answer correctness prediction and concept-level deficiency identification. Selection strategies that

[4] https://deepmind.google/technologies/gemini/flash-thinking/

[5] https://huggingface.co/meta-llama/Llama-4-Maverick-17B-128E-Instruct

[6] https://openai.com/index/openai-o3-mini/

jointly consider conceptual alignment and semantic similarity enable models to more effectively reason about students' conceptual understanding. Appropriate selection strategies enable small-scale LLMs to achieve promising performance even when historical responses are limited, as often seen in real-world settings.

We conducted a manual analysis of the model's predictive behavior and identified the primary source of difficulty. A key challenge is the variability of student performance. Student performance on target questions often diverges markedly from their prior responses. This unpredictability is especially pronounced among students with moderate mastery levels, whose learning trajectories tend to be unstable. Furthermore, a high frequency of non-conceptual errors (e.g., careless mistakes) in the historical responses can mislead the model into falsely inferring underlying conceptual deficiencies.

6. Discussion

Due to the limited availability of response records for some concepts (as shown in Figure 1), the conceptual semantic selection strategy may sometimes yield an insufficient number of historical responses for a given target question. This lack of same-concept context can hinder the model's ability to accurately assess the student's conceptual understanding. To investigate this challenge, we examine whether supplementing the input with additional responses involving different concepts can mitigate the sparsity of same-concept records and assist in maintaining prediction quality. Alternatively, such supplementation may introduce semantic noise and negatively affect model performance.

We restrict this analysis to cases where the number of same concept responses is fewer than 30, in order to assess the impact of supplementing the input with different-concept responses under data-sparse conditions. A total of 106 target questions are included in this experiment. The following three augmentation methods are evaluated:

Conceptual Semantic Selection (No Augmentation): This setting follows the method described in Section 4.2, without any additional augmentation from other concepts.

Random Augmentation: After applying the conceptual semantic selection strategy, additional responses are appended by randomly sampling from the remaining responses involving different concepts, until the total reaches 30.

Similarity-Based Augmentation: Similar to Random Augmentation, but instead of random sampling, responses from different concepts are selected based on their semantic similarity to the target question, measured using BERTScore.

We select Llama-4, which achieves the best performance under the conceptual semantic selection strategy in both selection-required and selection-optional settings. As shown in Table 6, Llama-4 performs better when using only same-concept responses (i.e., no augmentation), even if the total number of responses is less than 30. Supplementing with responses from different concepts, whether randomly or based on semantic similarity, tends to introduce contextual inconsistency, thereby reducing both correctness prediction accuracy and concept-level deficiency identification. These findings suggest that conceptual alignment is more critical than input quantity in supporting model reasoning and prediction reliability.

Augmentation	Correctness	Missing Concept
No	**68.87%**	**9.15%**
Random	66.98%	5.08%
Similarity-Based	66.04%	7.66%

Table 6: Results of Llama-4 under Different History Augmentation Strategies.

7. Conclusion

Knowledge tracing plays a crucial role in enabling personalized learning and adaptive instruction. However, existing research focuses on answer correctness prediction, with limited support for diagnosing concept-level deficiencies. Moreover, most datasets used in Mathematics KT studies lack students' authentic problem-solving processes. To address these limitations, we introduce a new task formulation that jointly predicts answer correctness and concept-level deficiencies, and systematically evaluate the capabilities of various LLMs and LRMs under an in-context learning setting. We construct ConceptKT, the first expert-annotated dataset with concept-level labels, and propose selection strategies for student response histories. Our findings show that strategies based on conceptual alignment and semantic similarity significantly improve performance under limited context length. At this stage, we have only explored a limited set of selection strategies. Identifying more effective strategies is left for future work. Additionally, predicting students' concept-level deficiencies remains challenging and still has significant room for improvement.

8. Limitations

This study adopts a single fixed prompt format and does not explore model-specific or diverse prompting strategies. Relying on a single format may limit the model's potential to fully capture the complexity of students' learning processes. The concept-level annotation in this study adopts a binary scheme,

indicating whether a student demonstrated a deficiency in a particular concept, without capturing varying degrees of understanding. However, in real-world educational scenarios, students often exhibit partial or context-dependent mastery rather than complete understanding or total deficiency. Such coarse-grained labeling may oversimplify students' conceptual states and constrain the model's capacity to reason about nuanced patterns of conceptual understanding.

9. Ethics Statement

Our dataset did not contain any personal information, and all annotators were fully informed about the research purposes of the data prior to the annotation process.

Experts were informed that they could provide feedback if the predefined concept set was insufficient to capture the observed reasoning. No such cases were reported.

Acknowledgments

This research was partially supported by National Science and Technology Council, Taiwan, under grant NSTC 114-2221-E-A49-057-MY3 and NSTC 114-2639-E-A49-001-ASP.

10. Bibliographical 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*.

Ghodai Abdelrahman and Qing Wang. 2019. Knowledge tracing with sequential key-value memory networks. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pages 175–184.

Aida Amini, Saadia Gabriel, Peter Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. Mathqa: Towards interpretable math word problem solving with operation-based formalisms. *arXiv preprint arXiv:1905.13319*.

John R Anderson, C Franklin Boyle, Albert T Corbett, and Matthew W Lewis. 1990. Cognitive modeling and intelligent tutoring. *Artificial intelligence*, 42(1):7–49.

Sami Baral, Li Lucy, Ryan Knight, Alice Ng, Luca Soldaini, Neil T Heffernan, and Kyle Lo. 2025. Drawedumath: Evaluating vision language models with expert-annotated students' hand-drawn math images. *arXiv preprint arXiv:2501.14877*.

Sami Baral, Eamon Worden, Wen-Chiang Lim, Zhuang Luo, Christopher Santorelli, Ashish Gurung, and Neil Heffernan. 2024. Automated feedback in math education: A comparative analysis of llms for open-ended responses. *arXiv preprint arXiv:2411.08910*.

Jiahui Cen, Jianghao Lin, Weixuan Zhong, Dong Zhou, Jin Chen, Aimin Yang, and Yongmei Zhou. 2025. Llm-driven effective knowledge tracing by integrating dual-channel difficulty. *arXiv preprint arXiv:2502.19915*.

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

Yongwan Cho, Rabia Emhamed AlMamlook, and Tasnim Gharaibeh. 2024. A systematic review of knowledge tracing and large language models in education: Opportunities, issues, and future research. *arXiv preprint arXiv:2412.09248*.

Youngduck Choi, Youngnam Lee, Dongmin Shin, Junghyun Cho, Seoyon Park, Seewoo Lee, Jineon Baek, Chan Bae, Byungsoo Kim, and Jaewe Heo. 2020. Ednet: A large-scale hierarchical dataset in education. In *International conference on artificial intelligence in education*, pages 69–73. Springer.

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

Jiajun Cui, Zeyuan Chen, Aimin Zhou, Jianyong Wang, and Wei Zhang. 2023. Fine-grained interaction modeling with multi-relational transformer for knowledge tracing. *ACM Transactions on Information Systems*, 41(4):1–26.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Aniket Didolkar, Anirudh Goyal, Nan Rosemary Ke, Siyuan Guo, Michal Valko, Timothy Lillicrap, Danilo Jimenez Rezende, Yoshua Bengio, Michael C Mozer, and Sanjeev Arora. 2024.

Metacognitive capabilities of llms: An exploration in mathematical problem solving. *Advances in Neural Information Processing Systems*, 37:19783–19812.

John A Erickson, Anthony F Botelho, Steven McAteer, Ashvini Varatharaj, and Neil T Heffernan. 2020. The automated grading of student open responses in mathematics. In *Proceedings of the tenth international conference on learning analytics & knowledge*, pages 615–624.

Lingyue Fu, Hao Guan, Kounianhua Du, Jianghao Lin, Wei Xia, Weinan Zhang, Ruiming Tang, Yasheng Wang, and Yong Yu. 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.

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.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Wei-Ling Hsu, Yu-Chien Tang, and An-Zi Yen. 2025. Mathedu: Towards adaptive feedback for student mathematical problem-solving. *arXiv preprint arXiv:2505.18056*.

Rui Li, Quanyu Dai, Zeyu Zhang, Xu Chen, Zhenhua Dong, and Ji-Rong Wen. 2025. Knowtrace: Explicit knowledge tracing for structured retrieval-augmented generation.

Naiming Liu, Zichao Wang, Richard Baraniuk, and Andrew Lan. 2022. Open-ended knowledge tracing for computer science education. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

Qi Liu, Zhenya Huang, Yu Yin, Enhong Chen, Hui Xiong, Yu Su, and Guoping Hu. 2019. Ekt: Exercise-aware knowledge tracing for student performance prediction. *IEEE Transactions on Knowledge and Data Engineering*, 33(1):100–115.

Ting Long, Yunfei Liu, Weinan Zhang, Wei Xia, Zhicheng He, Ruiming Tang, and Yong Yu. 2022. Automatical graph-based knowledge tracing. In *EDM*.

Huazheng Luo, Zhichang Zhang, Lingyun Cui, Ziqin Zhang, and Yali Liang. 2024. An efficient state-aware coarse-fine-grained model for knowledge tracing. *Knowledge-Based Systems*, 302:112375.

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.

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

Shalini Pandey and Jaideep Srivastava. 2020. Rkt: relation-aware self-attention for knowledge tracing. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 1205–1214.

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

Chen Pojen, Hsieh Mingen, and Tsai Tzuyang. 2020. Junyi academy online learning activity dataset: A large-scale public online learning activity dataset from elementary to senior high school students. *Dataset available from https://www. kaggle. com/junyiacademy/learning-activity-public-dataset-byjunyi-academy*.

Yang Qin, Xinning Zhu, Xiaosheng Tang, Chunhong Zhang, Kunbao Wu, Fengjie Chang, Jianzhou Diao, and Zheng Hu. 2025. Interpretable knowledge tracing with difficulty-aware attention and selective state space model. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 316–325.

Alexander Scarlatos, Ryan S Baker, and Andrew Lan. 2025. Exploring knowledge tracing in tutor-student dialogues using llms. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, pages 249–259.

John Stamper, Alexandru Niculescu-Mizil, Steve Ritter, GJ Gordon, and Kenneth R Koedinger. 2010. Challenge data set from kdd cup 2010 educational data mining challenge. *Find it at http://pslcdatashop. web. cmu. edu/KDDCup/downloads. jsp*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava,

Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Fei Wang, Qi Liu, Enhong Chen, Zhenya Huang, Yuying Chen, Yu Yin, Zai Huang, and Shijin Wang. 2020a. Neural cognitive diagnosis for intelligent education systems. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 6153–6161.

Zichao Wang, Angus Lamb, Evgeny Saveliev, Pashmina Cameron, Yordan Zaykov, José Miguel Hernández-Lobato, Richard E Turner, Richard G Baraniuk, Craig Barton, Simon Peyton Jones, et al. 2020b. Instructions and guide for diagnostic questions: The neurips 2020 education challenge. *arXiv preprint arXiv:2007.12061*.

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

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. 2023. Large language models as optimizers. In *The Twelfth International Conference on Learning Representations*.

Yang Yu, Yingbo Zhou, Yaokang Zhu, Yutong Ye, Liangyu Chen, and Mingsong Chen. 2024. Eckt: Enhancing code knowledge tracing via large language models. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.

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.

Suojuan Zhang, Jie Pu, Jing Cui, Shuanghong Shen, Weiwei Chen, Kun Hu, and Enhong Chen. 2024. Mlc-dkt: A multi-layer context-aware deep knowledge tracing model. *Knowledge-Based Systems*, 303:112384.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.