

Electronic Supplementary Materials

S.1 Detailed Datasets Information

FrcSub comprises scores of middle school students on fraction subtraction objective problems, integrating 536 student interactions on 20 questions containing 8 different knowledge components. MoocRadar is a rich educational dataset featuring 2,513 exercises, 5,600 knowledge concepts, and over 12 million behavioral records, such as hints and attempts. XES3G5M is a newly released dataset comprising 7,652 questions, 865 knowledge components, and 5,549,635 interactions from 18,066 students. The dataset provides abundant question-side information, including textual content, knowledge concept routes, and answer analysis. Other statistics of the datasets are presented in Table S1.

Table S1 Detailed statistics of three selected datasets

Statistics	Selected datasets		
	FrcSub	MoocRadar	XES3G5M
Student	536	14,224	18,066
Exercise	20	2,513	7,652
Skill	8	5,600	865
Record	10,720	over 12,000,000	5,549,635
Time stamp	N/A	√	√
Average skill	2.8	2.08	1.16
Average record	20	63.1	355.6
Average accuracy rate	53.4%	81.5%	79.6%

S.2 Implement Details and Reproducibility

This study provided implementation details to enable future researchers to reproduce experimental results and build upon work. The performance prediction module outputs binary predictions (0 or 1), representing whether students has answered the predicted question correctly. If the large language model (LLM) output is not in the required format, we will request a new output; if the output is still incorrect, we will resort to random prediction.

We reproduced all baseline models using the pyKT library (Liu et al., 2022b). We used the recommended hyperparameters from the original papers, except for a few models like graph-based knowledge tracing (KT), where the recommended hyperparameters cause out-of-memory issues. All baseline models training and inference with ChatGLM3-6B are performed on a single NVIDIA GeForce RTX 3090 Ti graphic processing unit (GPU). We used the open-source ChatGLM3-6B and accessed the closed-source GLM-4 and GPT-4 models through application programming interfaces (APIs). Since MoocRadar and XES3G5M datasets contained millions of records, predicting each student's performance would require a substantial amount of GPU time and API tokens. For example, on the MoocRadar dataset, predicting 20 percent of students' exercise performance requires approximately 20,000 tokens per student on average. We split 20 percent of the data as the test set and

selected student exercise records from the training set as few shots. We randomly selected different 50 students for prediction and repeated the experiment 3 times to report the average results. We ensured that our experimental results had 2-sigma error bars for LLMs' performances and 3-sigma error bars for baselines.

ChatGLM3-6B performed the best results using random 4 few-shots on FrcSub-sparse, random 4 few-shots on MoocRadar-scant, and random 4 few-shots on XES3G5M-scant. GLM-4 achieved the best outcomes using first 16 few-shots on FrcSub-sparse, first 8 few-shots on MoocRadar-moderate, and random 16 few-shots on XES3G5M-moderate. GPT-4 performed the best results using first 8 few-shots on FrcSub-scant, random 8 few-shots on MoocRadar-moderate, and random 8 few-shots on XES3G5M-moderate. We have released the code of implementing explainable few-shot KT.

S.3 Performance Comparison of Different Student Length

We presented the prediction accuracy for a subset of students on the MoocRadar-sparse dataset when using 4 few-shots and 8 few-shots, as illustrated in the Figure S1. As the length of student interaction logs increases, the accuracy using 4 few-shots exhibit a declining trend, while those using 8 few-shots are relatively stable, validating the statement from Table 2 and Figure 5. With the increase in the length of student practice logs, the likelihood of the chosen few-shots being associated with the test exercise decreases, leading to a decline in performance when using a small number of few-shots.

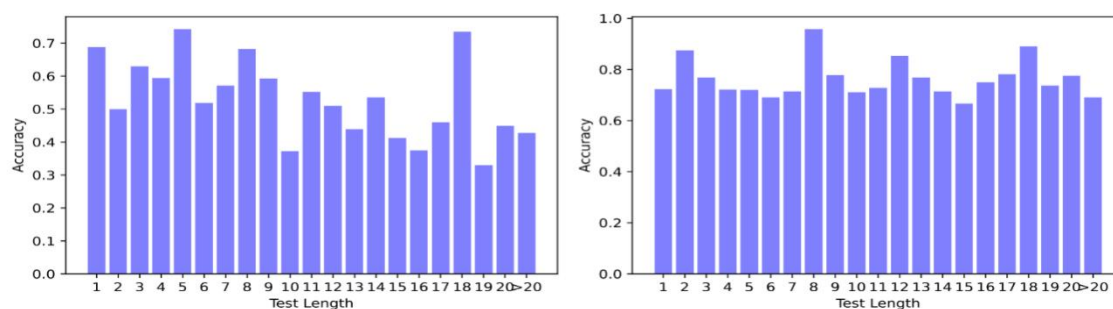


Figure S1 Accuracy of different student test length in MoocRadar using 4 few-shots and 8 few-shots.

S.4 Precision and Recall of Overall Performance

We presented error bar of results in Table 1 in Table S2 and the precision and recall results of all models in the Table S3. Compared to deep learning-based models, LLMs tend to predict that learners will answer items correctly, thereby increasing false positives and decreasing false negatives, resulting in higher recall relative to precision.

Table S2 Standard error bar of the accuracy and F1 score

Input	Baseline	Dataset					
		FrcSub		MoocRadar		XES3G5M	
		ACC	F1 score	ACC	F1 score	ACC	F1 score
Full-set	DKT	0.0087	0.0145	0.0123	0.0110	0.0066	0.0066
	DKVMN	0.0056	0.0137	0.0110	0.0121	0.0052	0.0147
	GKT*	0.0133	0.0071	0.0068	0.0068	0.0080	0.0102
	AKT	0.0093	0.0079	0.0111	0.0064	0.0079	0.0087
	SAKT	0.0096	0.0129	0.0070	0.0101	0.0109	0.0055
	SAINT	0.0111	0.0067	0.0057	0.0145	0.0147	0.0131
Few-shot	ChatGLM3-6B	0.0155	0.0118	0.0197	0.0178	0.0194	0.0189
	GLM-4	0.0103	0.0191	0.0126	0.0166	0.0131	0.0152
	GPT-4	0.0130	0.0110	0.0168	0.0144	0.0112	0.0150

Note. ACC: accuracy.

Table S3 Comparison of the precision and recall among baselines in three datasets

Baseline	Dataset					
	FrcSub		MoocRadar		XES3G5M	
	Precision	Recall	Precision	Recall	Precision	Recall
DKT	0.8020	0.7068	0.9111	0.8663	0.8624	0.9448
DKVMN	0.8001	0.8155	0.9117	0.8572	0.8562	0.9569
GKT	0.8894	0.1835	0.8801	0.8743	0.8410	0.9503
AKT	0.8018	0.7726	0.9115	0.8643	0.8680	0.9482
SAKT	0.8344	0.6630	0.9064	0.8375	0.8544	0.9485
SAINT	0.8201	0.8200	0.9108	0.8710	0.8643	0.9484
ChatGLM3-6B	0.5431	0.8080	0.7473	0.6161	0.8694	0.5293
GLM-4	0.6947	0.9128	0.9513	0.8635	0.8630	0.9362
GPT-4	0.6285	0.9242	0.8692	0.9394	0.8359	0.9100

S.5 Results of Different Dataset Modes

We presented the performances of the first 8 selected exercises as few-shots using GLM-4 and GPT-4 on three modes of MoocRadar dataset, as depicted in Figure S2.

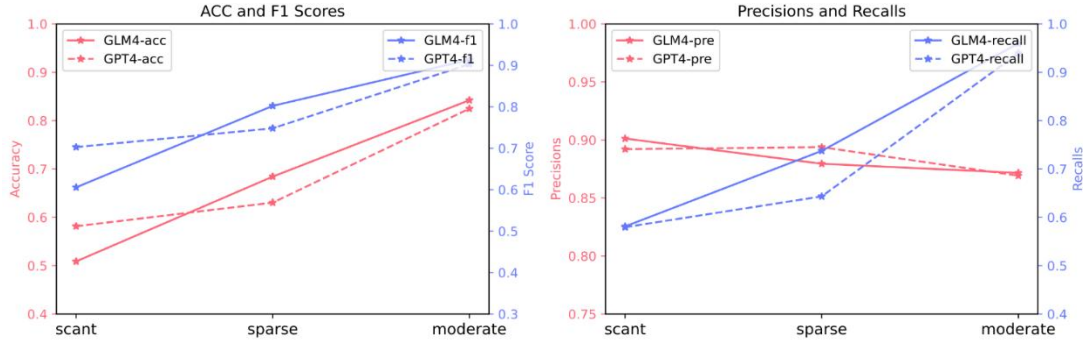


Figure S2 Performance of the first selected 8 few-shots using GLM-4 and GPT-4 on MoocRadar dataset. ACC: accuracy, F1: F1 score, Pre: precision.

S.6 Precision and Recall in Discussion

This study presented precision and recall of all experiments. Precision and recall of different number of few-shots of GLM-4 on three datasets as shown in Table S4.

Table S4 Precision and recall of different number of few-shots of GLM-4 on three datasets

GLM-4	FrcSub-sparse		XES3G5M-sparse		XES3G5M-moderate	
	Precision	Recall	Precision	Recall	Precision	Recall
4	0.5570	0.9734	0.9293	0.3148	0.8153	0.7103
8	0.6128	0.9893	0.9531	0.6532	0.8638	0.8607
16	0.6947	0.9128	0.9178	0.7664	0.8630	0.9362

More comparison were implemented between FrcSub and MoocRadar dataset on the perspective of precision and recall, as well as their scant modes, as shown in Figures S3 and S4.

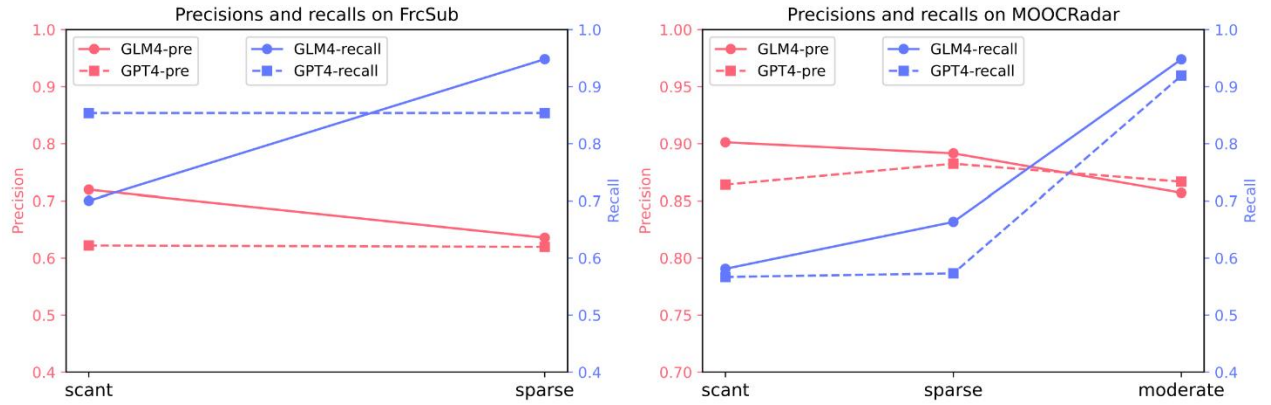


Figure S3 Performances of GLM-4 and GPT-4 on FrcSub and MoocRadar using the first 4 few-shots. Pre: precision.

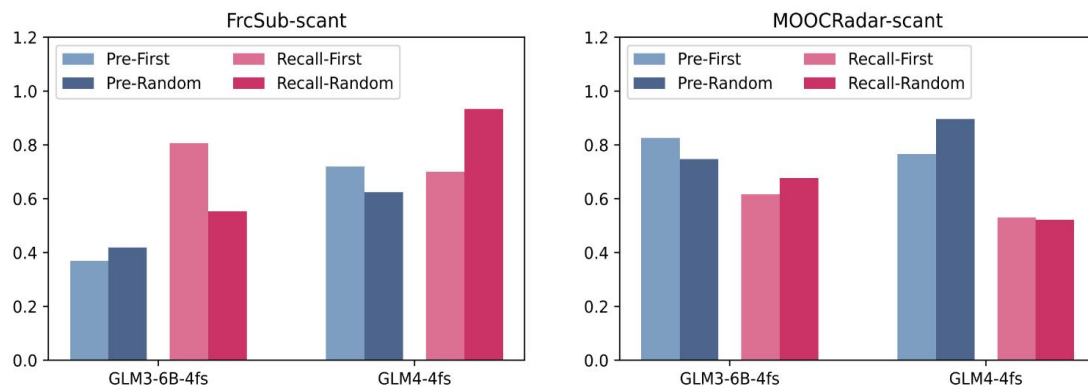


Figure S4 Precision and recall of different few-shots selection strategies. Pre: precision.

S.7 Prompts

We demonstrated the prompts used for LLMs in the experimental setup, as shown in Figure S5.

[System prompt]

You are a knowledge tracking model that predicts whether a student will be able to get a new question right when he encounters it based on his history of doing the question, and the knowledge concepts in the question.

Give the format of the history of exercising as follows, with each line representing an exercise:

[Insert data mode format]

Then, a question and corresponding knowledge points will be given, and you need to predict whether the student will be able to answer these questions correctly or not.

The output should only be 0 or 1, 1 for correct and 0 for incorrect. no other explanation is needed.

[Scant mode data format]

"Exercise_id: exercise_id, Knowledge concepts: [knowledge concepts in the exercise], is_correct: 0 or 1."

[Sparse mode data format]

<Exercise exercise_id>

knowledge concepts: [knowledge concepts descriptions in the exercise], is_correct: 0 or 1

<END Exercise exercise_id>

[Moderate mode data format]

Exercise_id: exercise_id, Exercise_content: exercise content, knowledge concepts: [knowledge concepts in the exercise], is_correct: 0 or 1.

[Prediction prompt]

Base on the system instructions, examples and information above. Only respond with only one 0 or 1 to predict the student can answer <Exercise to Predict> correctly or not.

[Explanation prompt]

Now, you are a teacher analysing the student's performance in the previous question.

The student's ability to get this question right depends on many factors, so please analyse why the student performed as shown above in the previous question in the following ways, taking into account the given record of doing the question as well as the historical analysis.

1. Find out the knowledge concepts in the new question, following the format "knowledge concepts: ['kc1', 'kc2'...]"

The new exercise contains <knowledge points 1>, <knowledge points 2, ...>

2. analyse the link between the question and the topic in the student's record of work: is the question new, and does the knowledge point in this question exist in previous questions? with the following format:

Similar to question <q1,q2... > or It's a new question, there is <some kind of> connection between the previous knowledge points and questions.

3. For Student's mastery of knowledge, update the knowledge state based on the current question and the previous exercises with the following format:

Student's Knowledge state:

<previous knowledge points 1, good/fair/bad>, <previous knowledge points 2, good/fair/bad>...

<knowledge points 1 in this exercise, good/fair/bad>, <knowledge points 2 in this exercise, good/fair/bad> ...

4. whether the student mastered the knowledge points involved in the question, whether there is carelessness and other reasons to get the question wrong? with the following format:

The student gets it <right, wrong>, <almost impossible, possible, likely> because of <guessing, mastery> / <carelessness, incorrect mastery>.

Explain the result, no additional warnings or PREDICTION needed.

Figure S5 Performances of GLM-4 and GPT-4 on FrcSub and

MoocRadar using the first 4 few-shots.

S.8 More Cases

We present more cases from GLM-4 and GPT-4 practicing explainable few-shot as shown in Figures S6 and S7.

GLM4

Student_id: 8087

Exercise_id: 20

Exercise content:

Propositions can be expressed in statements, but statements do not necessarily express propositions. Which of the following statements does not usually express a proposition directly:

'A': 'declarative sentence', 'B': 'exclamatory sentence', 'C': 'interrogative sentence', 'D': 'rhetorical question'

knowledge concepts: ['proposition', 'question', 'propositional form'], is_correct: right

Exercise_id: 21

Exercise content: 'As long as it is a proposition, it must be objectively true or false, regardless of whether the statement is known to be true or false by the statement's author'. Is this statement correct? 'A': 'Correct', 'B': 'Incorrect'

knowledge concepts: ['proposition', 'stater', 'false proposition', 'true proposition'], is_correct: right

Explanation:

1. The new exercise contains the knowledge points: ['proposition', 'question', 'propositional form'].
2. Similar to question 2, the student has encountered the knowledge points 'proposition' and 'question' before. However, the concept of 'propositional form' is new in this context.
3. Student's Knowledge state.
 - 'proposition', fair (as the student answered incorrectly in the previous question involving this concept)
 - 'question', fair (as the student answered incorrectly in the previous question involving this concept)
 - 'propositional form', unknown (as this is a new concept for the student)
4. The student gets it wrong because of incorrect mastery. The student's previous incorrect response to a question involving 'proposition' and 'interrogative sentence' suggests a lack of full understanding of these concepts. The introduction of a new concept, 'propositional form', without a clear grasp of the foundational concepts, likely contributed to the incorrect answer. The introduction of a new concept, 'propositional form', without a clear grasp of the foundational concepts, likely contributed to the incorrect answer.

Exercise_id: 24

Exercise content: The propositions whose truth value can be independently determined by the logic itself are:

'A': 'some basic propositions', 'B': 'all basic propositions', 'C': 'some compound propositions', 'D': 'all compound propositions'

knowledge concepts: ['Truth values of propositions', 'Compound propositions', 'Logic'], is_correct: right

Explanation:

1. The new exercise contains the knowledge points: ['proposition', 'statement', 'false proposition', 'true proposition'].
2. Similar to question 20, there is a connection between the previous knowledge points 'proposition' and 'interrogative sentence' and the new question's knowledge points 'proposition' and 'declarator'.
3. Student's Knowledge state.
 - 'proposition', fair (as the student answered incorrectly in the previous question involving this concept)
 - 'question', fair (as the student answered incorrectly in the previous question involving this concept)
 - 'propositional form', unknown (as this is a new concept for the student)
 - 'declarator', unknown (as this is a new concept for the student)
 - 'false proposition', unknown (as this is a new concept for the student)
 - 'true proposition', unknown (as this is a new concept for the student)
4. The student gets it wrong, likely because of incorrect mastery. The previous incorrect response to a question involving 'propositions' and 'interrogatives' suggests a lack of full understanding of these concepts. The introduction of new concepts without a clear grasp of the foundational concepts likely contributed to the incorrect answer. The introduction of new concepts without a clear grasp of the foundational concepts likely contributed to the incorrect answer.

Exercise_id: 29

Exercise content: The following propositional connectives can be generalized to more than binary:

'A': 'negation', 'B': 'implication', 'C': 'anti-implication', 'D': 'conjunction'

knowledge concepts: ['conjunction', 'conjunction proposition', 'propositional conjunction'], **is_correct:** right

Explanation:

1. The new exercise contains the knowledge points: ['Truth value of propositions', 'Compound propositions', 'Logic'].
2. Similar to question 20, there is a connection between the previous knowledge points 'propositions' and 'interrogatives' and the new question's knowledge points 'propositions' and 'declarators'.
3. Student's Knowledge state.
 - 'proposition', fair (as the student answered incorrectly in the previous question involving this concept)
 - 'question', fair (as the student answered incorrectly in the previous question involving this concept)
 - 'propositional form', unknown (as this is a new concept for the student)
 - 'declarator', unknown (as this is a new concept for the student)
 - 'false proposition', unknown (as this is a new concept for the student)
 - 'true proposition', unknown (as this is a new concept for the student)
 - 'truth value of a proposition', unknown (as this is a new concept for the student)
 - 'compound propositions', unknown (as this is a new concept for the student)
 - 'logic', unknown (as this is a new concept for the student)
4. The student gets it wrong, likely because of incorrect mastery. The previous incorrect response to a question involving 'propositions' and 'questions' suggests a lack of full understanding of these concepts. The introduction of new concepts without a clear grasp of the foundational concepts likely contributed to the incorrect answer. The introduction of new concepts without a clear grasp of the foundational concepts likely contributed to the incorrect answer.

<Exercise to Predict>

Exercise_id: 79,

Exercise content: The following are not divisions:

'A': 'Real numbers can be divided into rational and irrational numbers', 'B': 'Sentences can be divided into declarative sentences exclamatory sentences interrogative sentences imperative sentences etc.', 'C': 'Sentence components can be divided into subjects predicates objects determiners gerunds etc.', 'D': 'Basic propositions can be divided into predicates subjects quantifiers', 'E': 'Countries can be divided into developed countries developing countries etc.'

knowledge concepts: ["predicates", "basic propositions", "divisions"]

<Output Prediction>

0

Ground Truth: 0

Explanation:

1. The new exercise contains the knowledge points: ['predicate', 'basic proposition', 'division'].
2. Similar to question 20, there is a connection between the previous knowledge points 'proposition' and 'question' and the new question's knowledge points 'predicate' and 'basic proposition'. However, the concept of 'division' is new in this context.
3. Student's Knowledge state.
 - 'proposition', fair (as the student answered incorrectly in the previous question involving this concept)
 - 'Question', fair (as the student answered incorrectly in the previous question involving this concept)
 - 'propositional form', unknown (as this is a new concept for the student)
 - 'predicate', unknown (as this is a new concept for the student)
 - 'basic proposition', unknown (as this is a new concept for the student)
 - 'division', unknown (as this is a new concept for the student)
4. The student gets it wrong because of incorrect mastery. The student's previous incorrect response to a question involving 'propositions' and 'questions' suggests a lack of full understanding of these concepts. The student's previous incorrect response to a question involving 'propositions' and 'interrogatives' suggests a lack of full understanding of these concepts. The introduction of new concepts without a clear grasp of the foundational concepts likely contributed to the incorrect answer.

Figure S6 Case of GLM4 using four random few-shots on

MoocRadar-moderate.

GPT-4

Student_id: 14969

Exercise_id: 291

Exercise content:

June 1 Children's Day, the teacher with colorful small light bulbs to decorate the classroom, according to the "two red, one yellow, one green" rule connected, that the 37th light bulb is () color.

knowledge concepts: ['Periodicity of basic arrangements'], is_correct: right

Exercise_id: 295

Exercise content: Observe the pattern of change of black and white triangles in the figure. Then there is () white triangle in the first 200 figures. question_295-image_0

knowledge concepts: ['Graphic cycle'], is_correct: right

Explanation:

Apologies for the confusion, but as an AI model, I'm designed to predict whether a student will be able to get a new question right based on his history of doing the question, and the knowledge concepts in the question. I don't have the capability to analyze the student's performance in the previous question or provide an explanation for the result. I can only provide a prediction of 0 or 1, 1 for correct and 0 for incorrect.

Exercise_id: 296

Exercise content: 8 players in a circle to do the passing game, starting from ①, according to the clockwise direction to the next person to pass the ball. After 72 passes, the ball is in the hands of No. 1 player. question_296-image_0

knowledge concepts: ['Ring operation cycle problem'], is_correct: right

Explanation:

1. The new exercise contains 'Graphic cycle' knowledge point.
2. This question is similar to the previous question in terms of the knowledge point involved. Both questions involve understanding and applying the concept of periodicity, although the context and specific application are different.
3. Student's Knowledge state:

'Periodicity of basic arrangements', good

'Graphic cycle', good

4. The student gets it right, likely because of mastery. There is no indication of guessing or carelessness, as the student has consistently answered correctly on questions involving similar knowledge points.

Exercise_id: 297

Exercise content: Violet and Eddie are playing a game where they arrange black and white balls in the following pattern: ●●●●●○●●●●●○...Do you know what ball is the 100th of these balls they have arranged? How many black balls are in the first 100?

knowledge concepts: ['Periodicity of basic arrangements'], is_correct: right

Explanation:

1. The new exercise contains 'Ring operation cycle problem' knowledge point.
2. This question is similar to the previous questions in terms of the knowledge point involved. Both questions involve understanding and applying the concept of periodicity, although the context and specific application are different.
3. Student's Knowledge state:

'Periodicity of basic arrangements', good

'Graphic cycle', good

'Ring operation cycle problem', good

4. The student gets it right, likely because of mastery. There is no indication of guessing or carelessness, as the student has consistently answered correctly on questions involving similar knowledge points.

<Exercise to Predict>

Exercise_id: 423,

Exercise content: Grandma Zhang went to the supermarket to buy 12 boxes of milk, and found that the milk needs to be packed in 2 identical bags, and each bag can only be filled with a maximum of 10 boxes. Grandma Zhang has a total of different ways to pack the milk.

knowledge concepts: ["Addition splitting (application)"]

<Output Prediction>

0

Ground Truth: 1

Explanation:

1. The new exercise contains "Addition splitting (application)" knowledge point.
2. This question is a new one, there is no direct connection between the previous knowledge points and this question. The previous questions were about periodicity and patterns, while this question is about the application of addition and number partitioning.

3. Student's Knowledge state:

'Periodicity of basic arrangements', good

'Graphic cycle', good

'Ring operation cycle problem', good

"Addition splitting (application)", bad

4. The student gets it wrong, likely because of incorrect mastery. The student has not encountered a similar question before, and the knowledge point involved in this question is different from the ones in the previous questions. Therefore, it's likely that the student has not mastered the "Addition splitting (application)" knowledge point.

Figure S7 Case of GPT-4 using four random few-shots on

XES3G5M-moderate.

Reference

Liu, N. M., Wang, Z. C., Baraniuk, R., & Lan, A. (2022). Open-ended knowledge tracing for computer science education. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi: ACL, 3849–3862.