

Bringing Generative Artificial Intelligence (GenAI) to Education



Hang Li



Kaiqi Yang



Yucheng Chu



Jiliang Tang

Michigan State University

Tutorial Outline

Morning

Introduction
to GenAI
(40 Min)



Overview of GenAI
in Education
(40 Min)



Applications in
Education Part I
(60 Min)



Afternoon

Applications in
Education Part II
(60 Min)



Demo
Presentation
(40 Min)



Future Direction
and QA
(20 Min)

Tutorial Outline

Morning

Introduction
to GenAI
(40 Min)



Overview of GenAI
in Education
(40 Min)



Applications in
Education Part I
(60 Min)

Afternoon

Applications in
Education Part II
(60 Min)

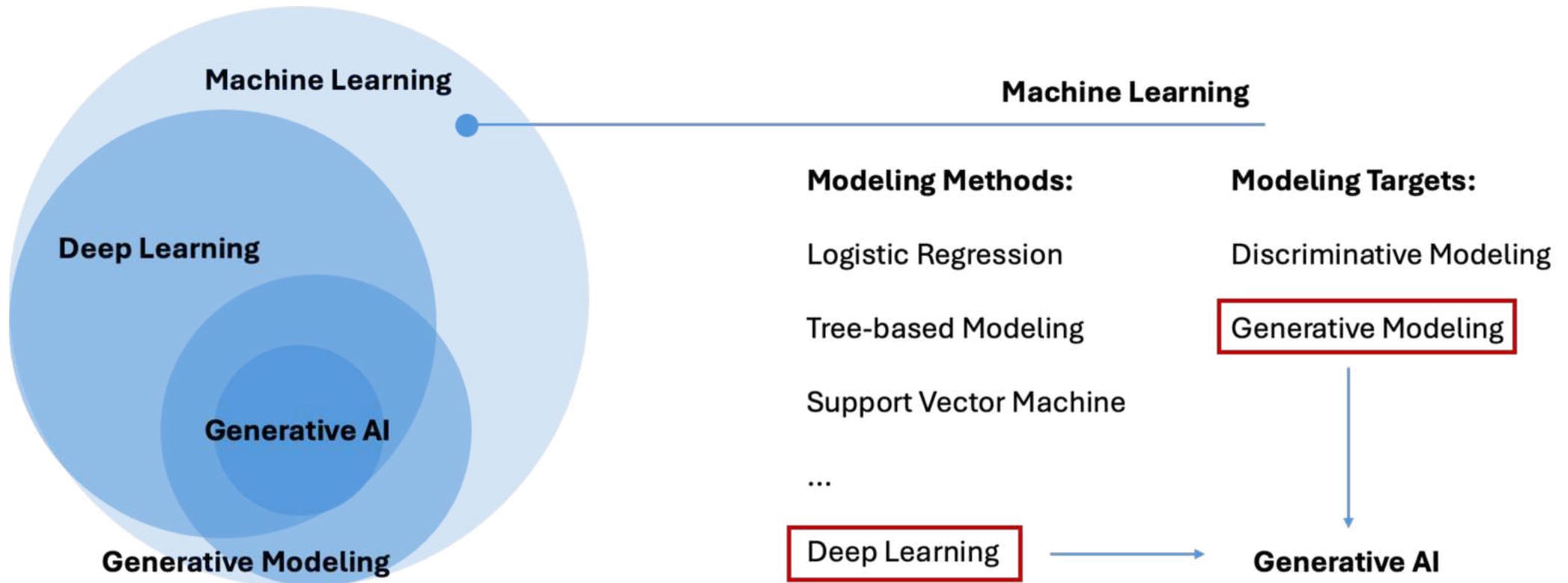


Demo
Presentation
(40 Min)

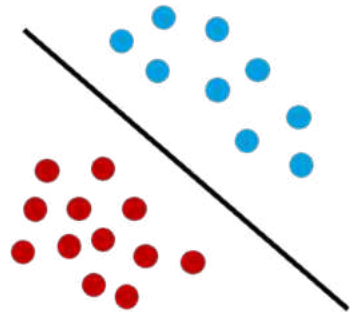


Future Direction
and QA
(20 Min)

What is Generative AI?



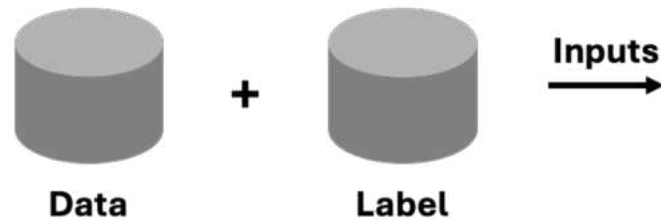
Discriminative v.s. Generative Modeling



Discriminative Modeling

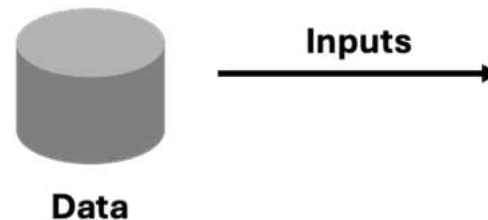


Generative Modeling



Predictive ML Model

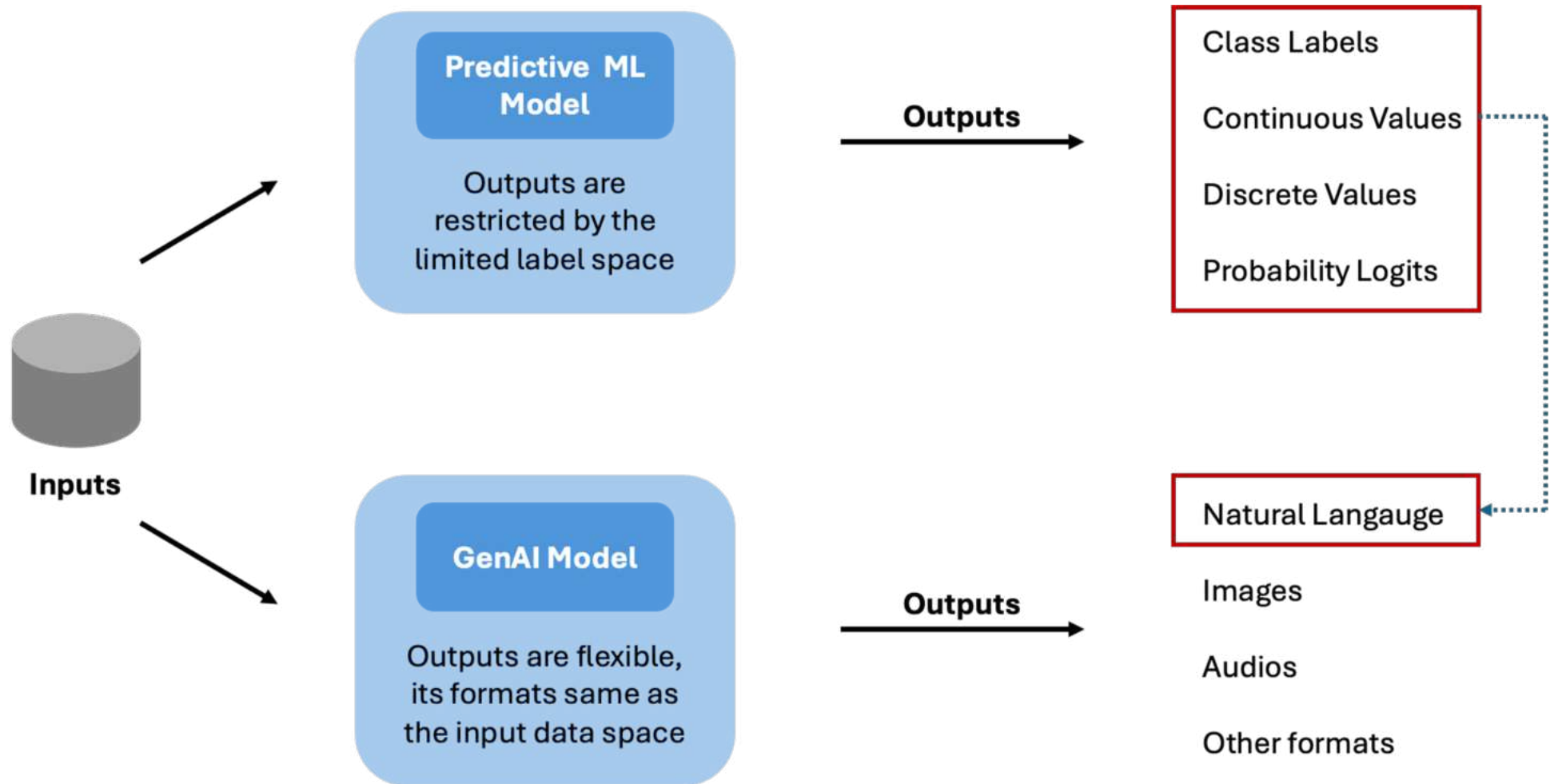
Learns the relationship between data and label



GenAI Model

Learns the pattern and distribution of the data

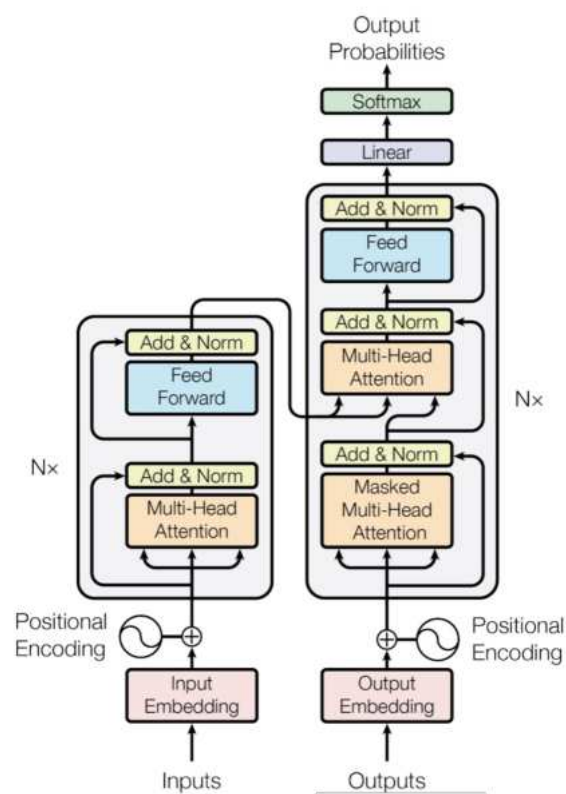
Discriminative v.s. Generative Modeling



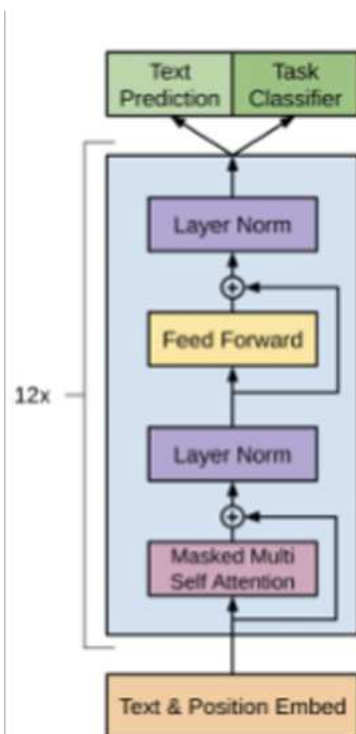
Popular Generative Frameworks

- Variational Autoencoders (VAEs)
- Generative Adversarial Networks (GANs)
- Flow-based Models
- Generative Diffusion Models (GDMs)
- **Large Language Models (LLMs)**

Generative Pre-Training (GPT)



Transformer



GPT(Decoder-only)

Text sample:

LLMs learn to predict one word at a time
LLMs learn to predict one word at a time
LLMs learn to predict one word at a time
LLMs learn to predict one word at a time
LLMs learn to predict one word at a time
LLMs learn to predict one word at a time
LLMs learn to predict one word at a time
LLMs learn to predict one word at a time
LLMs learn to predict one word at a time
LLMs learn to predict one word at a time

Input the LLM receives

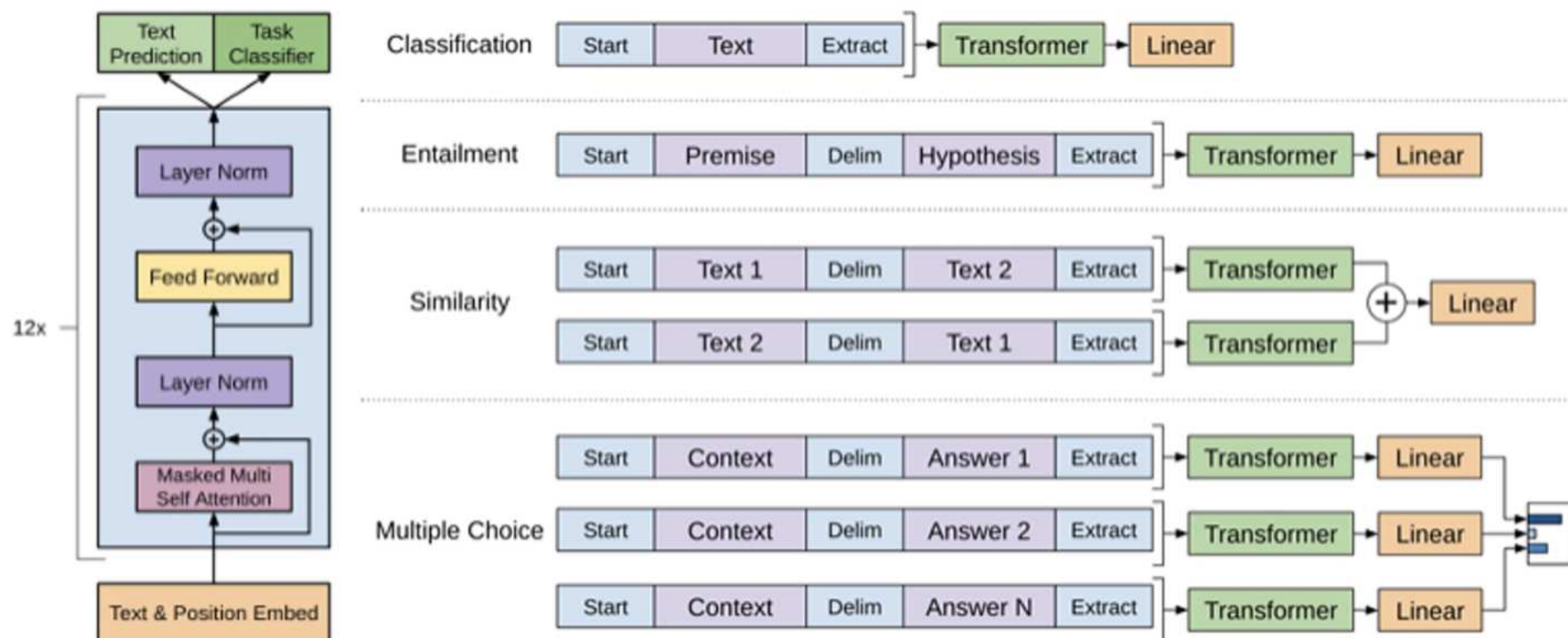
The LLM can't access words past the target

Target to predict

Next-word Prediction

Generative Pre-Training (GPT)

GPT was the first framework successfully unified different types of NLP tasks.



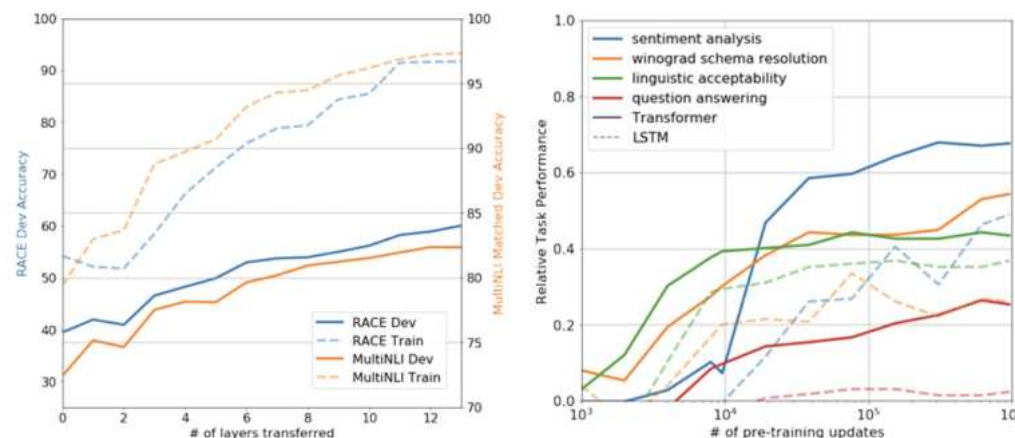
Results of GPT

Task	Datasets
Natural language inference	SNLI [5], MultiNLI [66], Question NLI [64], RTE [4], SciTail [25]
Question Answering	RACE [30], Story Cloze [40]
Sentence similarity	MSR Paraphrase Corpus [14], Quora Question Pairs [9], STS Benchmark [6]
Classification	Stanford Sentiment Treebank-2 [54], CoLA [65]

Method	MNLI-m	MNLI-mm	SNLI	SciTail	QNLI	RTE
ESIM + ELMo [44] (5x)	-	-	<u>89.3</u>	-	-	-
CAFE [58] (5x)	80.2	79.0	<u>89.3</u>	-	-	-
Stochastic Answer Network [35] (3x)	<u>80.6</u>	<u>80.1</u>	-	-	-	-
CAFE [58]	78.7	77.9	88.5	<u>83.3</u>	-	-
GenSen [64]	71.4	71.3	-	-	<u>82.3</u>	59.2
Multi-task BiLSTM + Attn [64]	<u>72.2</u>	<u>72.1</u>	-	-	<u>82.1</u>	61.7
Finetuned Transformer LM (ours)	82.1	81.4	89.9	88.3	88.1	56.0

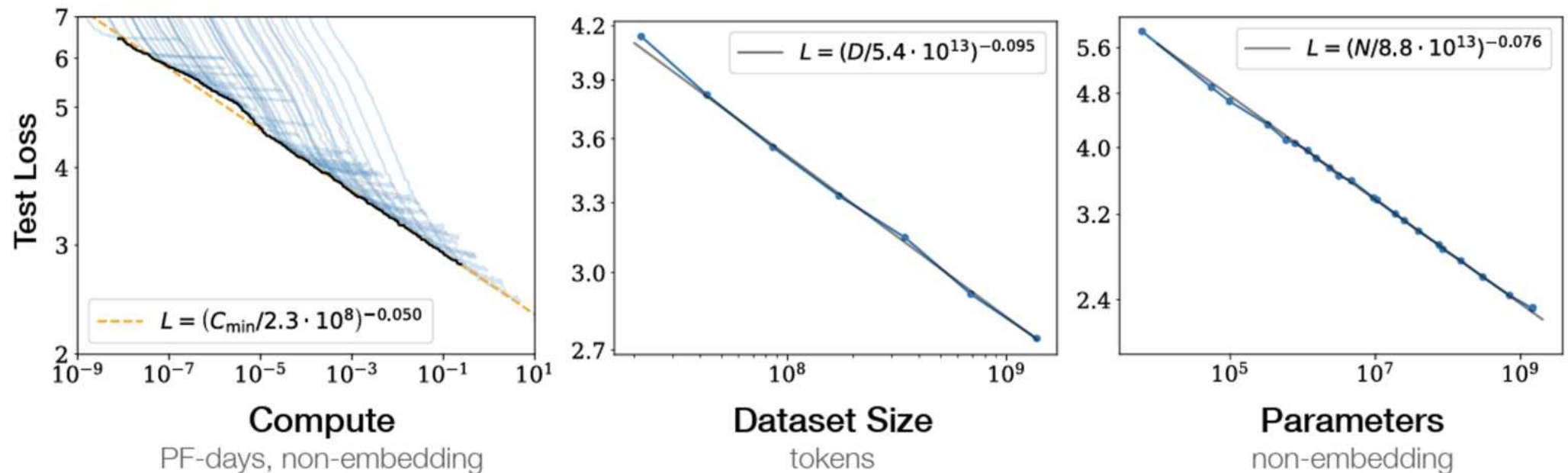
Method	Story Cloze	RACE-m	RACE-h	RACE
val-LS-skip [55]	76.5	-	-	-
Hidden Coherence Model [7]	<u>77.6</u>	-	-	-
Dynamic Fusion Net [67] (9x)	-	55.6	49.4	51.2
BiAttention MRU [59] (9x)	-	<u>60.2</u>	<u>50.3</u>	<u>53.3</u>
Finetuned Transformer LM (ours)	86.5	62.9	57.4	59.0

Method	Classification		Semantic Similarity			GLUE
	CoLA (mc)	SST2 (acc)	MRPC (F1)	STSB (pc)	QQP (F1)	
Sparse byte mLSTM [16]	-	93.2	-	-	-	-
TF-KLD [23]	-	-	86.0	-	-	-
ECNU (mixed ensemble) [60]	-	-	-	<u>81.0</u>	-	-
Single-task BiLSTM + ELMo + Attn [64]	<u>35.0</u>	90.2	80.2	55.5	<u>66.1</u>	64.8
Multi-task BiLSTM + ELMo + Attn [64]	18.9	91.6	83.5	72.8	<u>63.3</u>	<u>68.9</u>
Finetuned Transformer LM (ours)	45.4	91.3	82.3	82.0	70.3	72.8



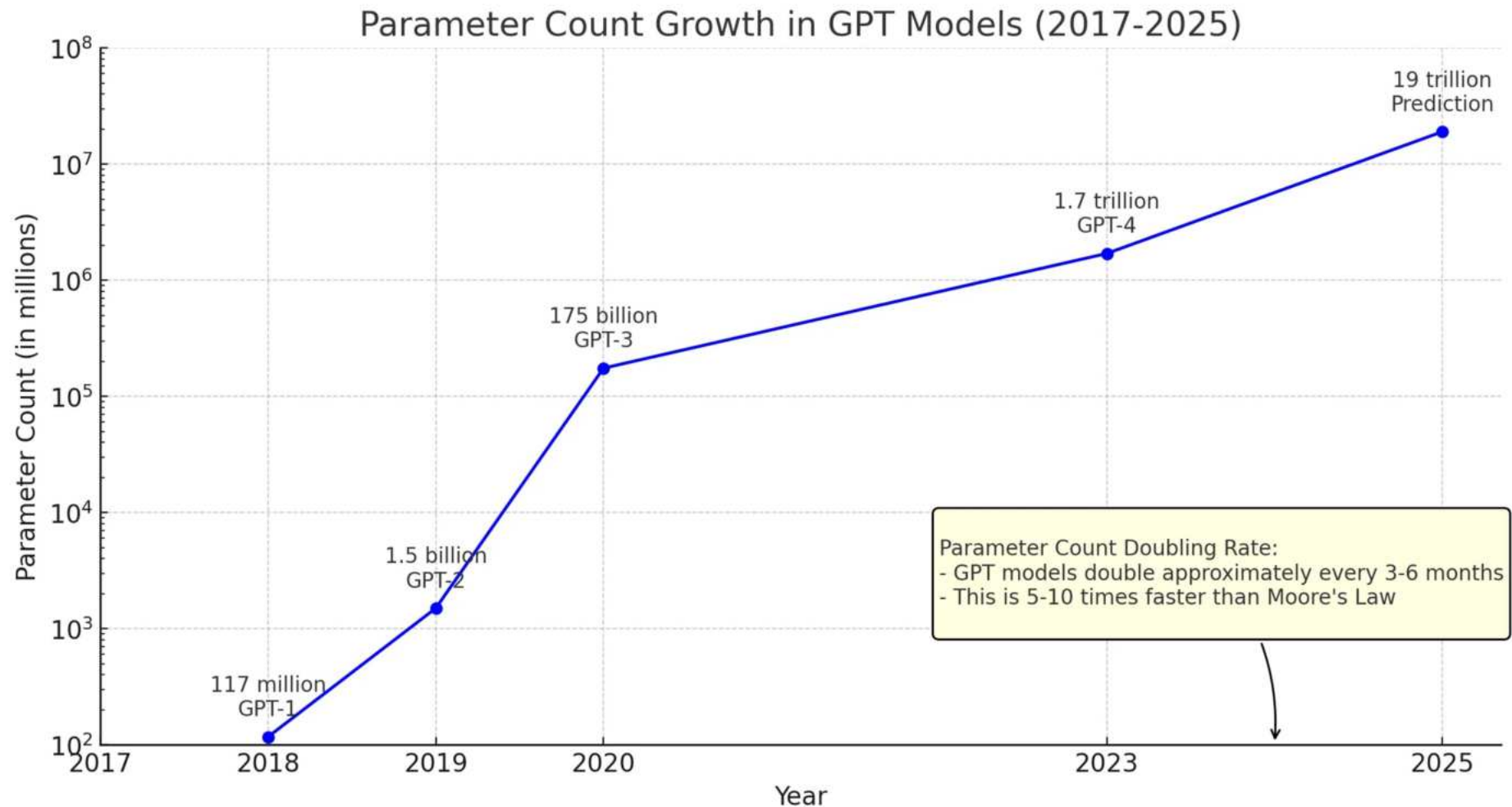
GPT achieved leading performance across various tasks. Besides, its performance kept improving with the increase in model parameter size.

Scaling Law for Language Models

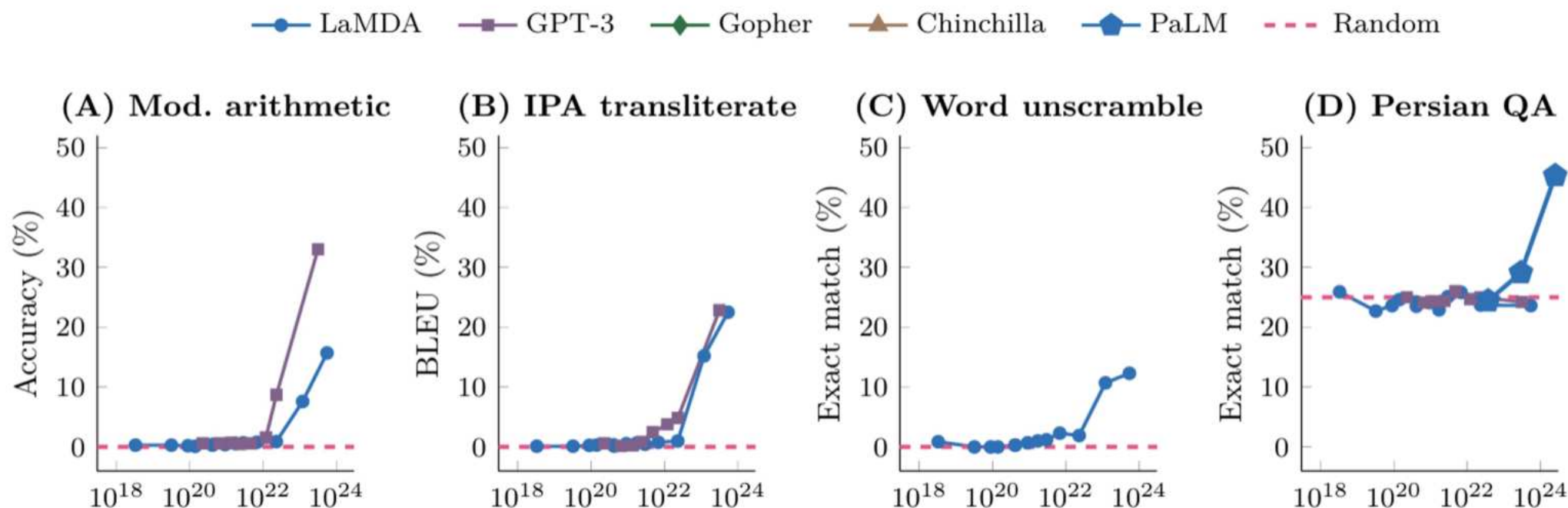


Scaling laws describe how three key factors in training, dataset size, model parameters, and training duration, influence the upper-bound capabilities of language models.

GPT Series Model Size Changes

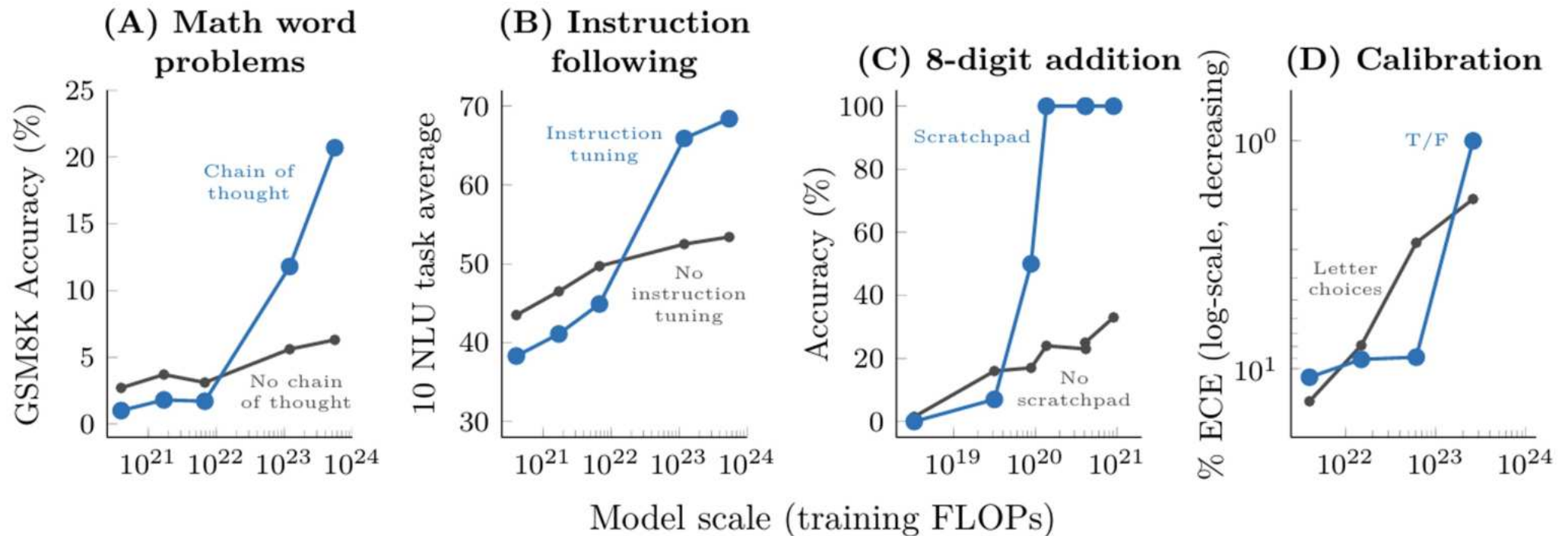


Emergent Abilities of Large Language Models



Emergent of few-shot learning capability of LLMs.

Emergent Abilities of Large Language Models



Specialized prompting or fine-tuning methods can be emergent

From GPT-3 to ChatGPT

- Differences between GPT-3 and ChatGPT?
 - GPT-3 needs manual prompt engineering or external fine-tuning to guide behavior.
 - ChatGPT let user to instruct models to finish flexible tasks via conversation directly.
- Benefits brought by ChatGPT?
 - The conversation based design allows the instruction become context informed.
 - Flexible instructions of different users helps model become more intelligent.

Challenges for Training ChatGPT

- Challenges:
 - Collect labeled samples for training models to finish the varied instructions of users is facing scalability issues .
- Solutions:
 - Use reinforcement learning to replace supervised learning to reduce the reliance of labeled data.
 - Using binary comparisons to generate reward signal to instruct the training of reinforcement learning process.

Bradley-Terry model

- Bradley-Terry model helps to estimate the score of item i β_i by maximum the likelihood of the following probability.

$$\Pr(i > j) = \frac{e^{\beta_i}}{e^{\beta_i} + e^{\beta_j}}.$$



$$\mathcal{L}_{BT} = -\log \left(\frac{e^{r(A)}}{e^{r(A)} + e^{r(B)}} \right)$$

- Human annotators provide preference labels, and a reward is trained to automatically estimate the rewards for new samples.

RLHF Framework

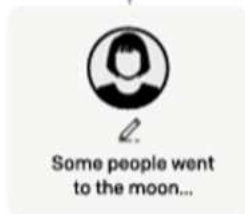
Step 1

Collect demonstration data, and train a supervised policy.

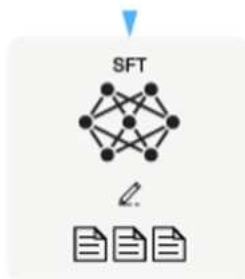
A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



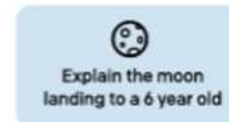
This data is used to fine-tune GPT-3 with supervised learning.



Step 2

Collect comparison data, and train a reward model.

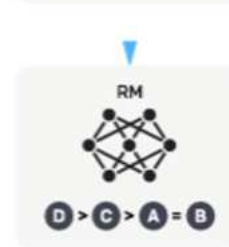
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



Step 3

Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.



The policy generates an output.

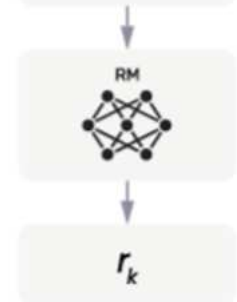


Once upon a time...

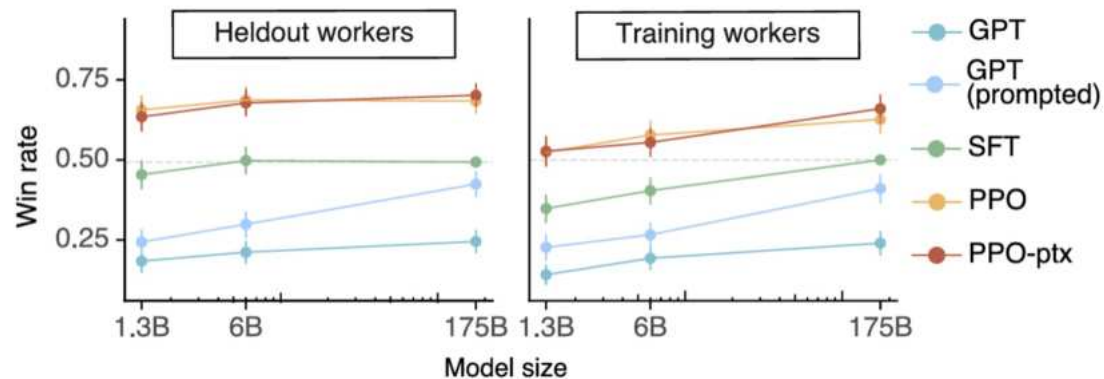
The reward model calculates a reward for the output.



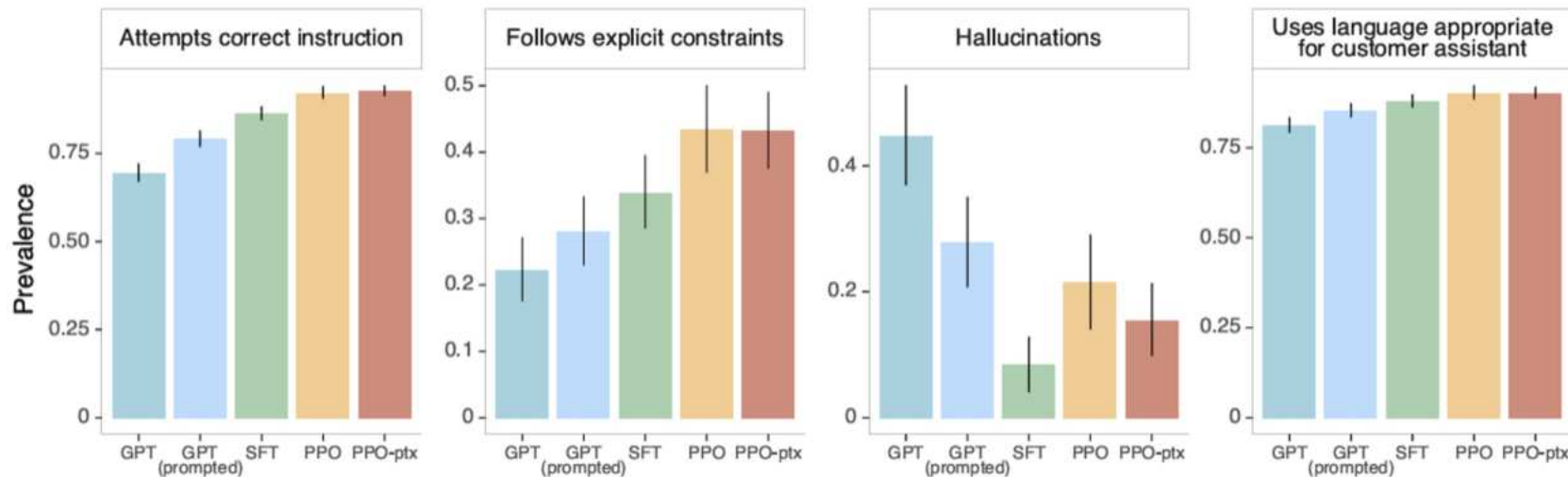
The reward is used to update the policy using PPO.



Improved Results by RLHF



Conclusions: RLHF enhances GPT-3's instruction-following capabilities and leads to consistently higher user preference compared to other models fine-tuned through different methods.



Recent Advancement in LLMs

- **Backbone Designs:**

- Mixture of Experts
- Attention Mechanism ...

- **Fine-tuning Methods:**

- Low-rank Parameter Efficient Fine-tuning (LoRa)
- Training Losse: DPO, GROP ...

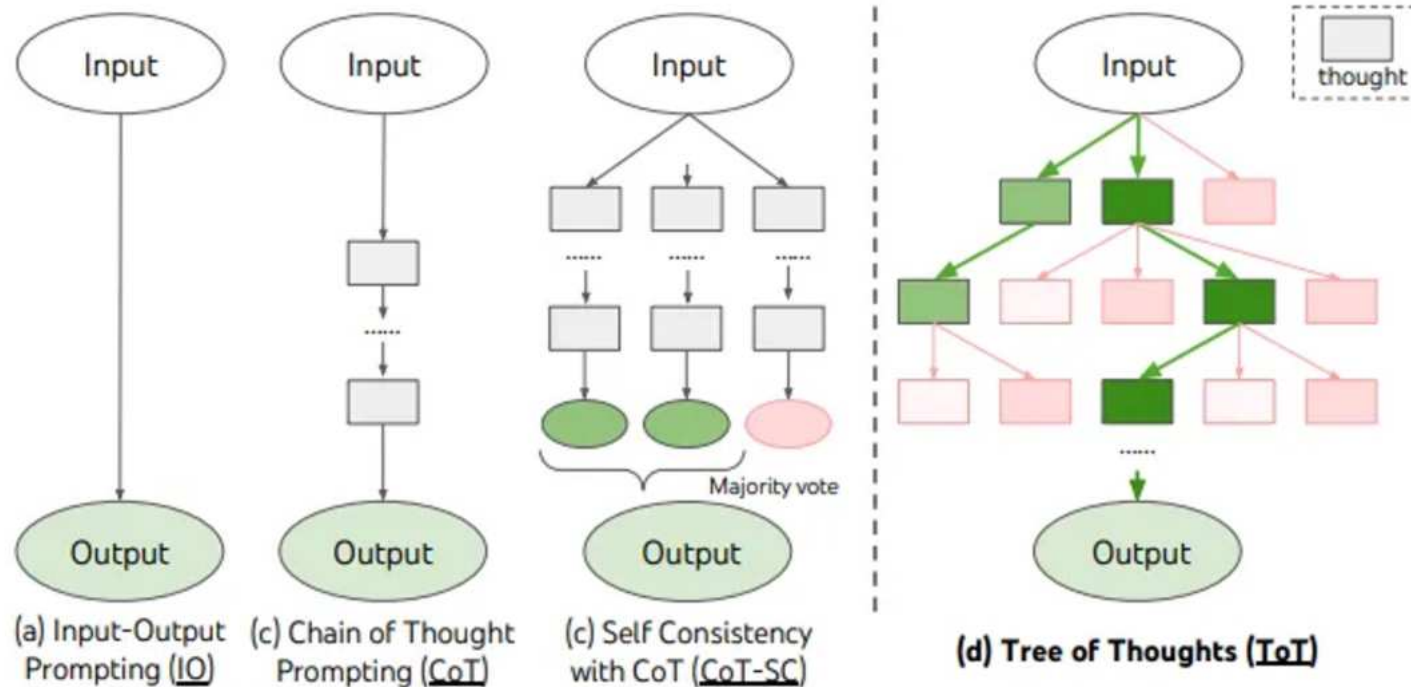
- **Enhancing Strategies:**

- Prompt Engineering
- In-content Learning
- Multi-Agent System
- Retrieval Augmented Learning

Prompt Engineering

- Idea:

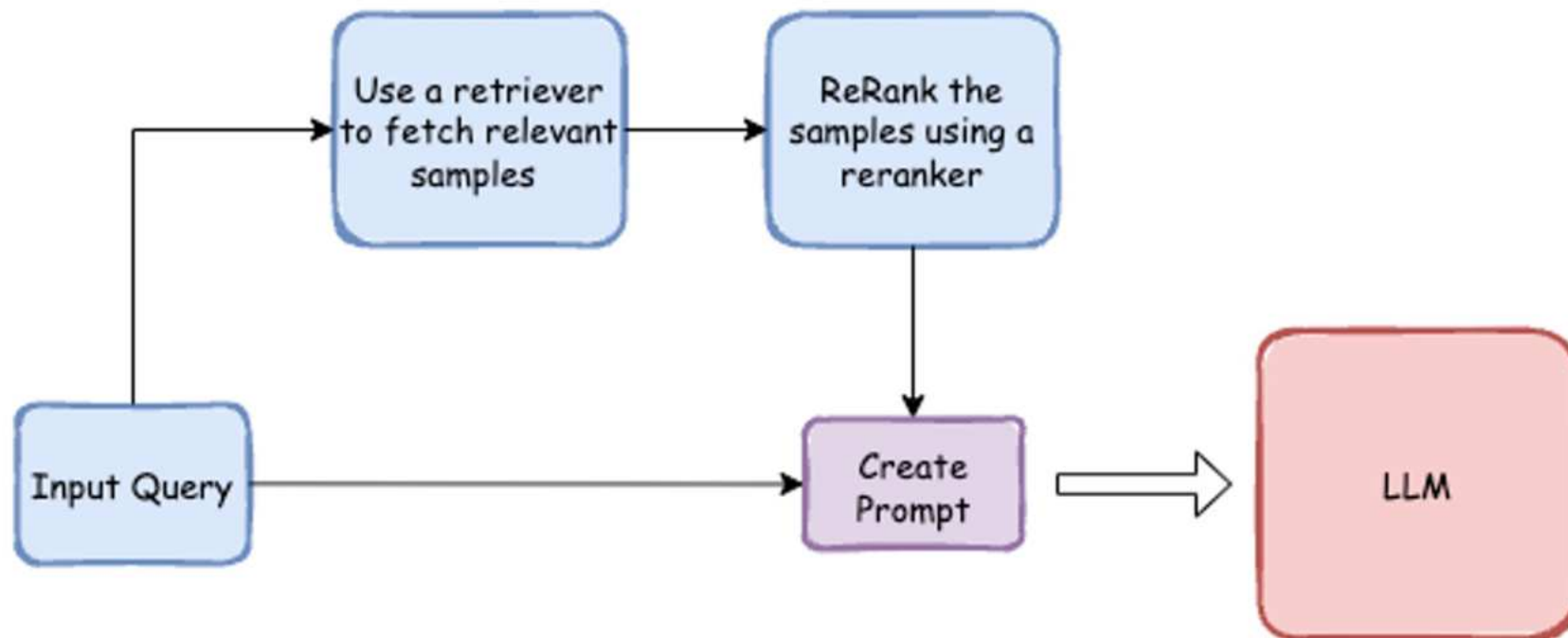
- Improve the LLMs' task performance by refining the instruction text and strategies.



In-context Learning

- **Idea:**

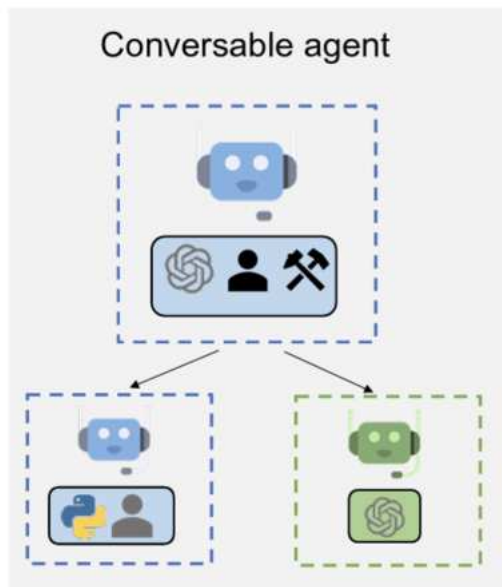
- Provide demonstration samples to help LLMs better understand the task and get the better task performance.



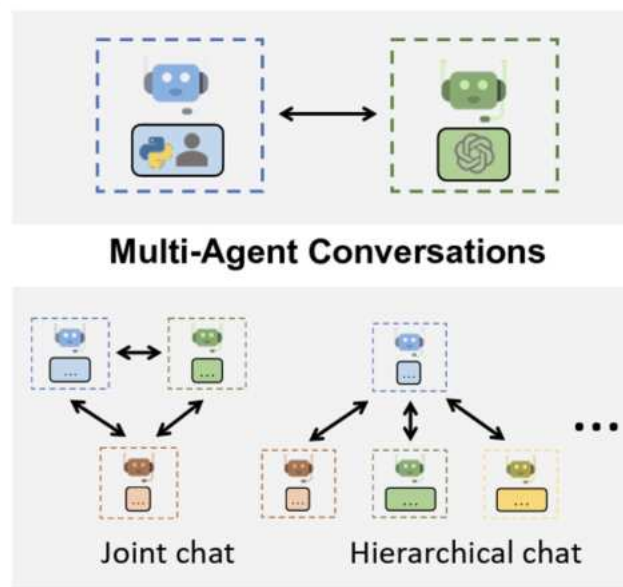
Multi-Agent System (MAS)

- **Idea:**

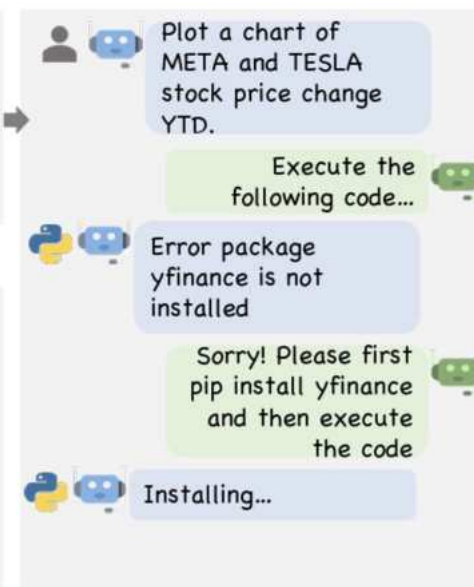
- Facilitate the task difficulties by decomposing into several sub-tasks.
LLMs can also improve its performance through communications.



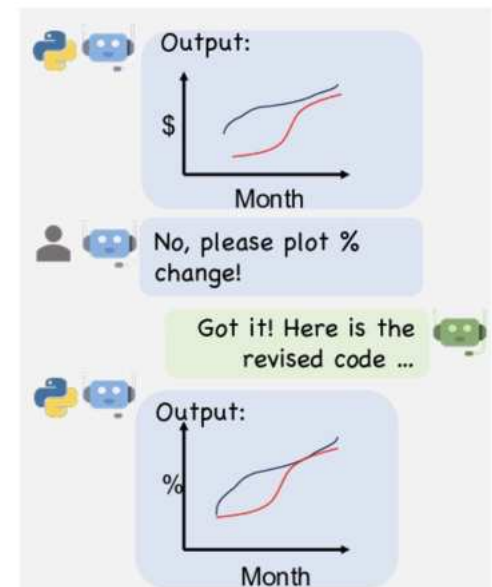
Agent Customization



Flexible Conversation Patterns



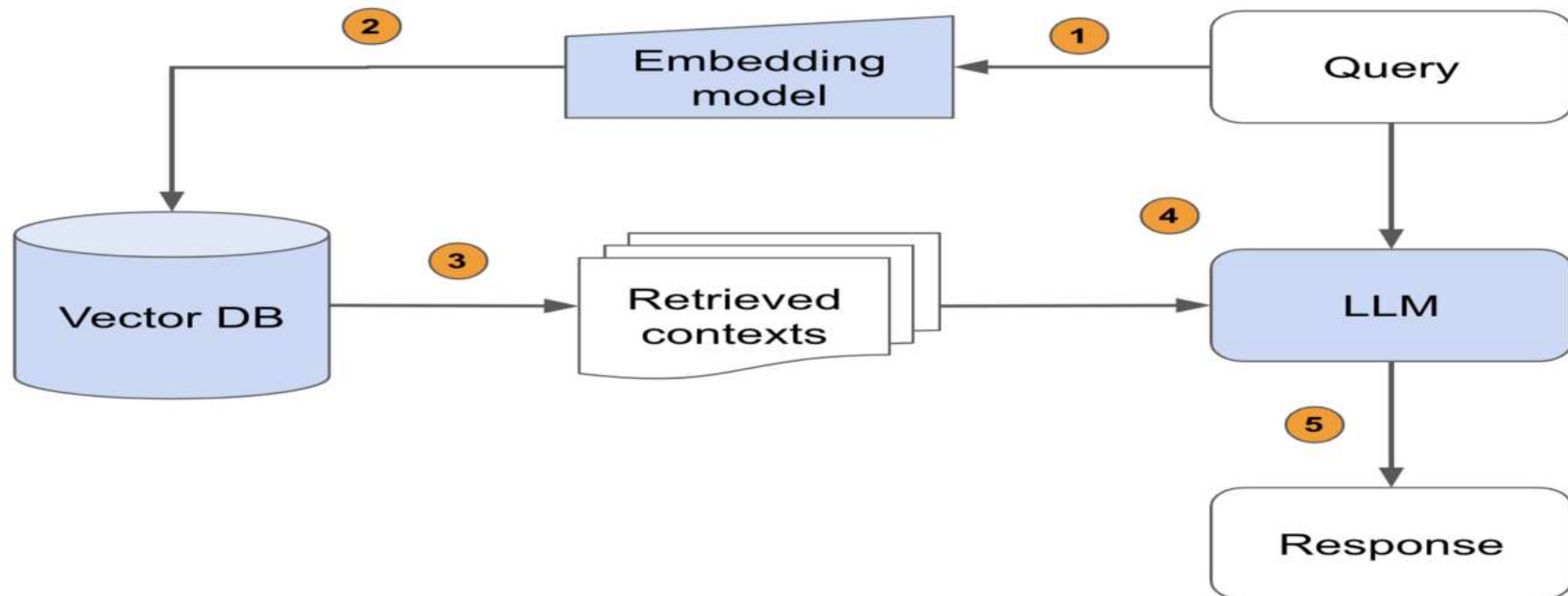
Example Agent Chat



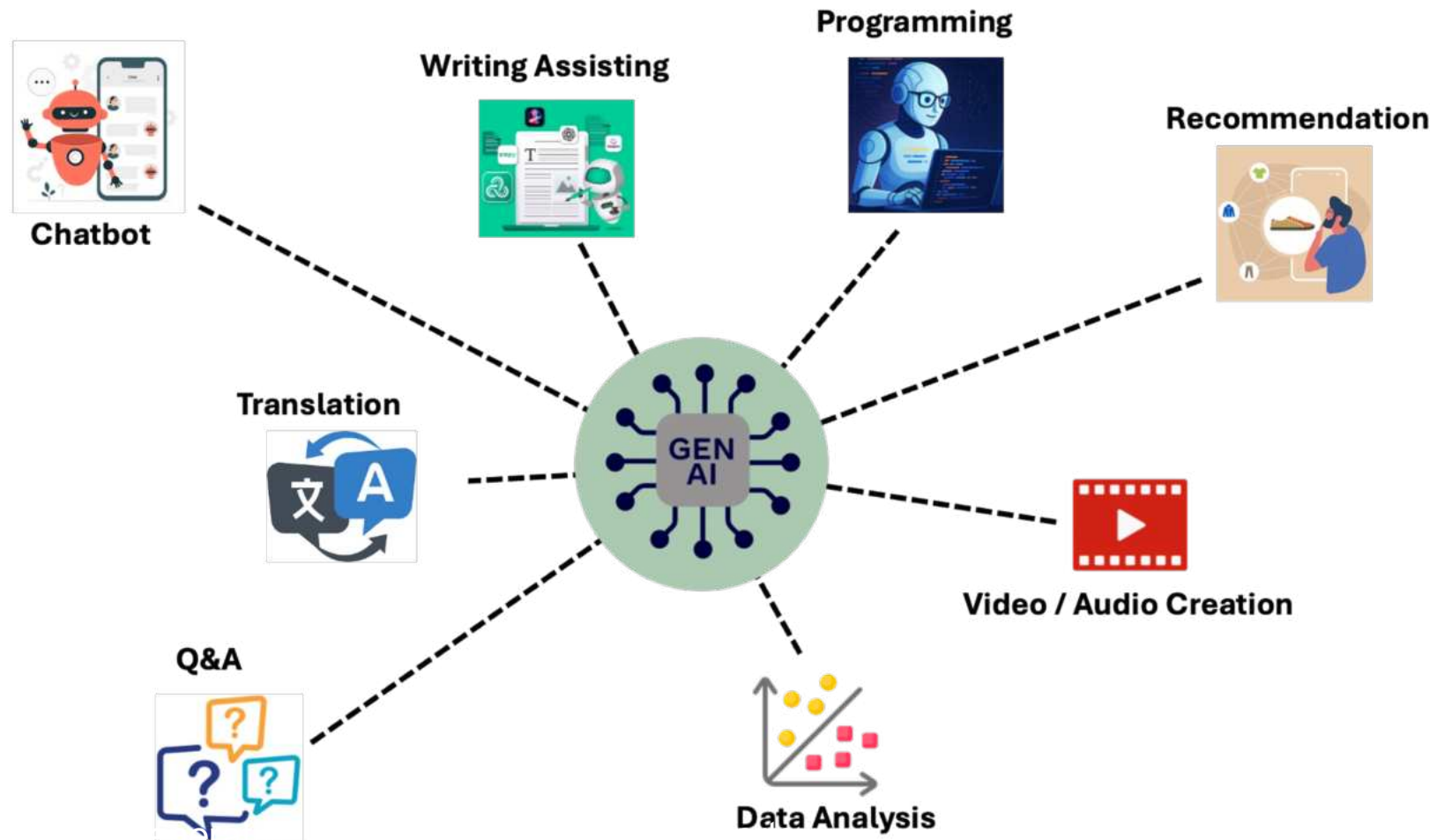
Retrieve Augmented Generation (RAG)

- **Idea:**

- Combines information retrieval with text generation to improve the accuracy and factuality of LLMs.



Applications of LLM



Tutorial Outline

Morning

Introduction
to GenAI
(40 Min)



Overview of GenAI
in Education
(40 Min)



Applications in
Education Part I
(40 Min)

Afternoon

Applications in
Education Part II
(40 Min)



Demo
Presentation
(40 Min)



Future Direction
and QA
(40 Min)

Role of AI in Education

Learning:

- Assigning tasks based on individual competence.
- Providing human-machine conversations.
- Analyzing student work for feedback.
- Increasing adaptability and interactivity in digital environments.

Assessment:

- Providing automatic marking.
- Providing dynamic and adaptive assessment questions for examination.
- Predicting students' performance.

Teaching:

- Providing adaptive teaching strategies.
- Enhancing teachers' ability to teach.
- Supporting teachers' professional development.

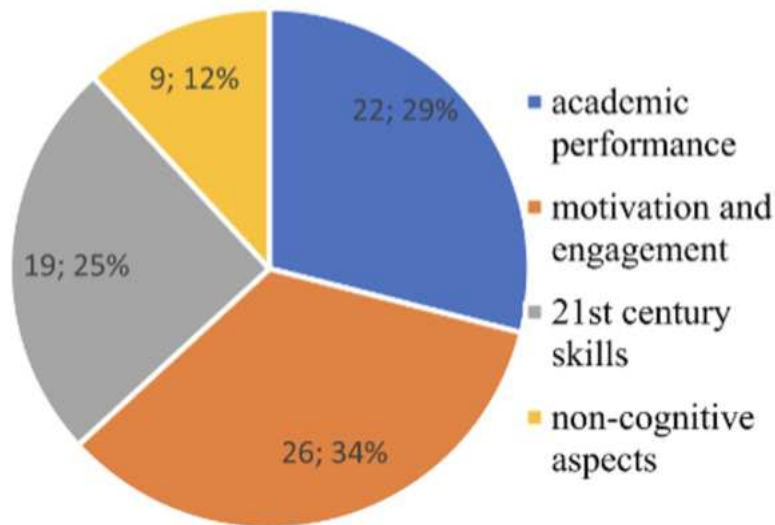
Administration:

- Improving the performance of management platforms.
- Providing convenient and personalized service.
- Supporting educational decision-making with evidence.

Outcomes of AI in Education

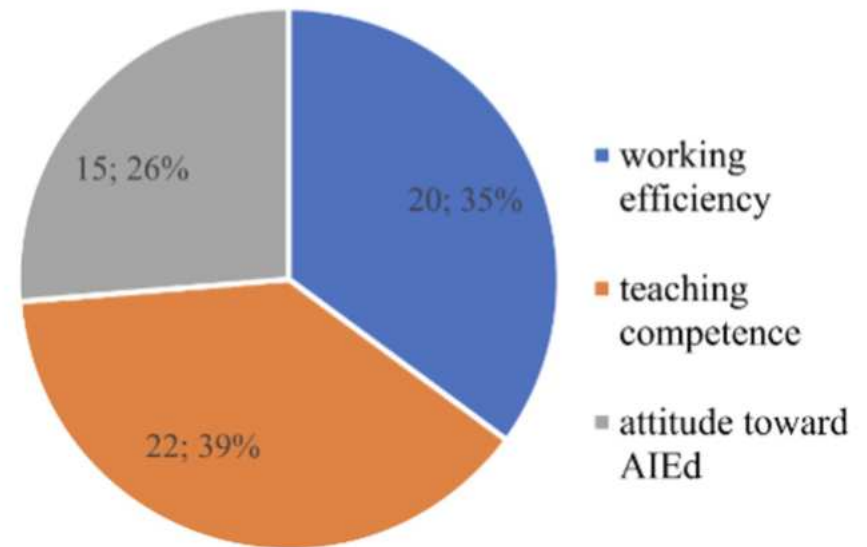
Student learning outcomes:

- Academic Performance.
- Motivation and Engagement.
- 21st Century Skills.
- Non-cognitive Aspects

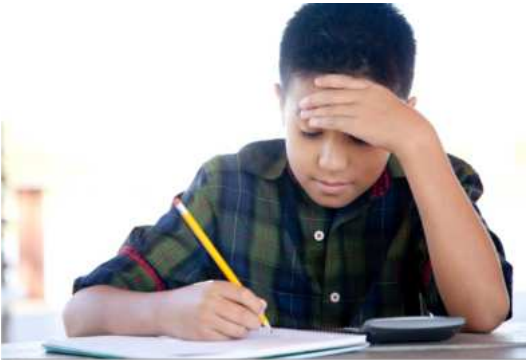


Teacher outcomes:

- Working efficiency.
- Teaching Competence.
- Attitude toward AIEd.



Overview of AI Applications in Education



● Learning

- Question Solving
- Confusion Helper
- Error Correction

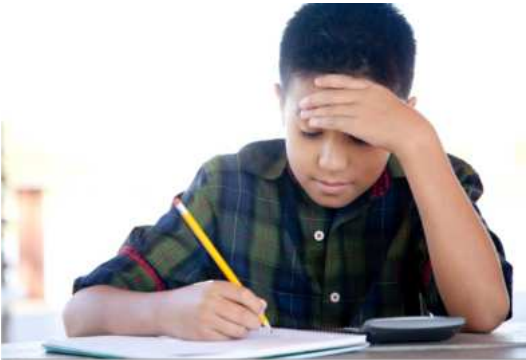
● Teaching

- Content Generation
- Automatic Grading
- Profession Development

● Assessment

- Knowledge Tracing
- Question Generation

Overview of AI Applications in Education



● Learning

- Question Solving
- Confusion Helper
- Error Correction

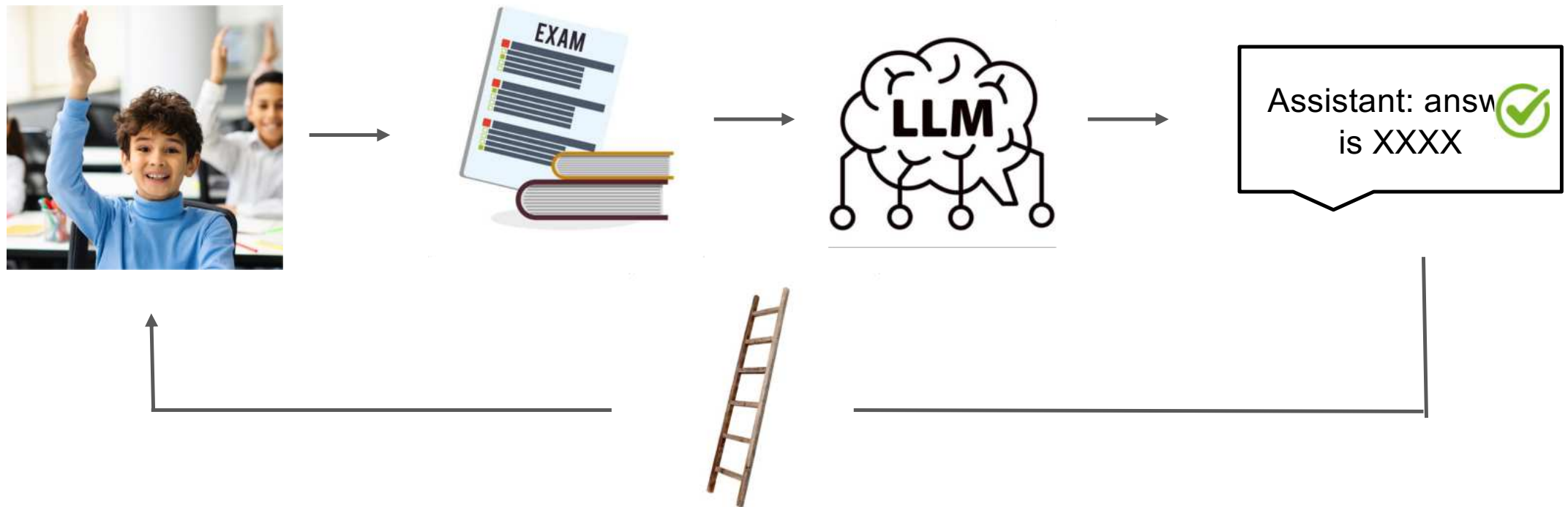
● Teaching

- Content Generation
- Automatic Grading
- Profession Development

● Assessment

- Knowledge Tracing
- Question Generation

Learning: Question Solving



LLM provides scaffolder to help student solve practice questions.

Example of scaffold for students

Show instructions

Study: Math Word Problem Exercise 1

James decides to run 3 sprints 3 times a week. He runs 60 meters each sprint. How many total meters does he run a week?

Try again. Let's break down the problem and answer step by step:

How many sprints does James run in a week?

Answer to the subquestion:
Partial answer here.

How many meters does James run in a week?

Answer to the subquestion:
Partial answer here.

Final answer (numerical format):
Write your answer here.

Continue



Break questions into simple sub-problems, guiding students to find the solution.

Summary of Question Solving

- **Goal**

- Help students solve the challenging questions.

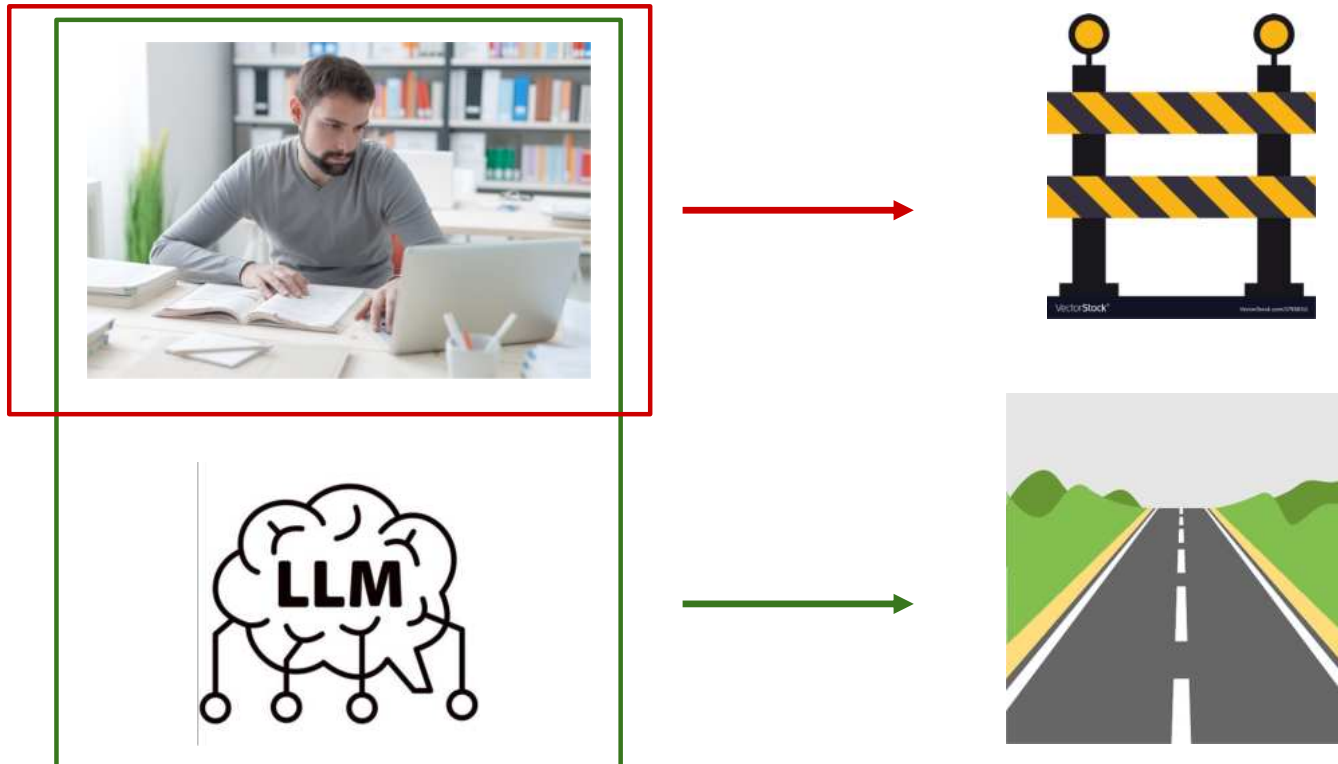
- **Advantages**

- The advancement of LLMs enable to solve most of questions across different subjects.
- LLMs is able to generate adaptive solution scaffolders.

- **Directions**

- Build the benchmark and evaluate the question solving result correctness.
- Generate valid and efficient guidance to help students solve challenging questions.

Learning: Confusion Helper



LLM is used to solve student's confusions to textbooks, course materials or any other contents during learning process.

Example of Confusion Helper

Definition



User: How do you find the first n terms of an arithmetic sequence?
EduChat: Are you asking for the formula for the sum of the first n terms of an arithmetic sequence, or how to apply the formula to find the sum of the first n terms?
User: I don't know the formula for the sum of the first n terms of an arithmetic sequence.
EduChat: Well, do you know what an arithmetic sequence is?
User: Not really.

EduChat: Okay, let's first understand what an arithmetic sequence is. An arithmetic sequence is a sequence in which the difference between any term and the one before it is constant. For example, 1, 3, 5, 7, 9 is an arithmetic sequence because the difference between each term and the one before it is 2.

Discussion



User: So how do you find the sum of the first N terms?
EduChat: We can use the formula for the sum of an arithmetic sequence to solve it. The formula for the sum of the first N terms of an arithmetic sequence is: $S_n = 1/2 * n * (a_1 + a_n)$, where a_1 is the first term of the arithmetic sequence, and a_n is the N th term.

User: But I don't know what the N th term is.
EduChat: Well, we can derive the N th term from known conditions. If we know the first term of an arithmetic sequence a_1 , the common difference d , and the N th term a_n , then the formula for the N th term is $a_n = a_1 + (N-1) * d$.

User: What do you mean by common difference?
EduChat: The common difference refers to the difference between any two terms in an arithmetic sequence. In an arithmetic sequence, the difference between each term and the one before it is the same, and this equal difference is the common difference.

Exempling



User: Can you give me an example?
EduChat: Of course. For example, if the first term of an arithmetic sequence is 3, the common difference is 2, what is the 6th term? According to the formula $a_n = a_1 + (N-1) * d$, we can get $a_n = 3 + (6-1) * 2$, which gives $a_n = 13$. Therefore, the first 6 terms of this arithmetic sequence are 3, 5, 7, 9, 11, 13.

User: I understand now, thank you.
EduChat: You're welcome, feel free to ask if you have any other questions.

Summary of Confusion Helper

- **Goal**

- Help to solve students' confusion to learning contents in their study process.

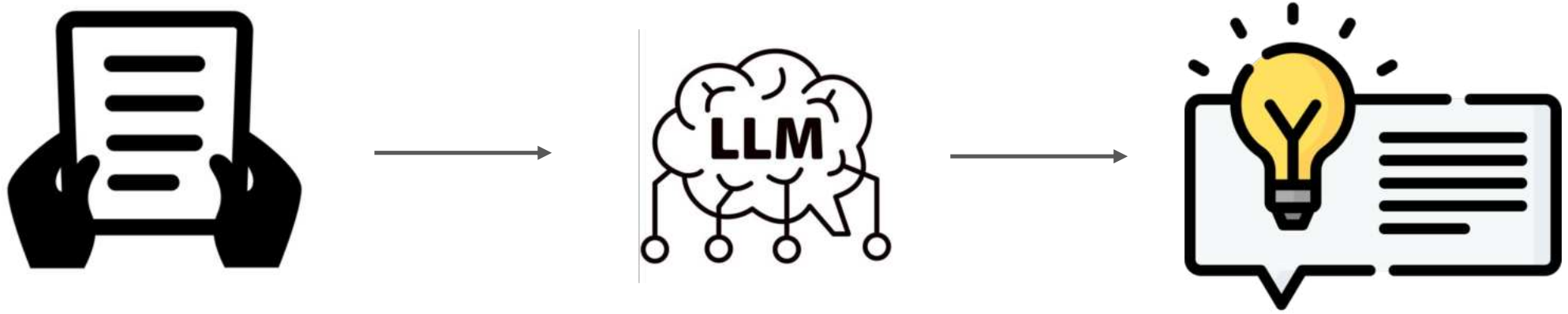
- **Advantages**

- The prior knowledge of LLMs make it easy to answer many K12 level questions.
- LLM provide flexible form of response to student's question is available, where student can choose the one which helps them the most.

- **Directions**

- Integrating education theories with LLMs and generate response meets pedagogical standards.
- Extend application to more domains, including stronger domain knowledges.

Learning: Error Correction



LLMs help students find the errors from exercises and then provide the modification suggestions.

Examples of Error Correction

< original_text
> modified_text
1 / 1 If Galton's results provide facts that we can trust, then the task of explaining mental imagery just got a little bit harder. For example, in addition to explaining how people mentally image things, we also need to explain how some people can do it very well and others can't do it at all. This is a good example of the increasing complexity that comes along with the research cycle: asking questions uncovers more facts that raise new questions requiring additional explanation. If Galton's results provide reliable facts, then explaining mental imagery becomes more challenging. Not only do we need to explain how individuals create mental images, but we also need to account for why some individuals can vividly create these images while others cannot. This illustrates the growing complexity that follows the research cycle. Every question answered leads to new facts, which in turn result in new questions requiring thorough explanation.

Error Correct for Essay Writing

Question & Answer	Question analysis	Student Draft	Student Answer
Q: Calculate the equation: $567 \times 89 + 345 =$. A: In operations without parentheses, when the expression includes multiplication and addition, perform the ...	Calculate according to the order of operations in mixed calculations.		50968
Cause of error	Error Feedback		
The student made an error in the multiplication process. Specifically, in the second line of the multiplication, where 567 should be multiplied by 9, the student mistakenly wrote 3 instead of 6 as the units digit of the product ...	When performing vertical multiplication, it is important to accurately sum each product result. For products that exceed one digit, carry over the tens digit (or higher) to the corresponding position in the next row...		

Error Correct for Image

Problem: The square root of t is greater than 2 and less than 3.5. How many integer values of t satisfy this condition?

Ground truth: 8

Question-level math correction:

Student Solution:

Step 1: Translate the condition into inequalities: $2 < \sqrt{t} < 3.5$.

Step 2: Square all parts of the inequality to eliminate the square root: $2^2 < t < 3.5^2$.

Step 3: Calculate the squares: $4 < t < 12.25$.

Step 4: Determine the integer values of t that satisfy the inequality: The integers that fall strictly between 4 and 12.25 are 5, 6, 7, 8, 9, 10, 11, 12. However, since t must be strictly less than 12.25, 12 is not included in the solution set because it is not strictly less than 12.25.

Step 5: Count the integer values: The integers that satisfy the condition are 5, 6, 7, 8, 9, 10, 11, which gives us a total of 8 integer values of t .

Extract match?  Final Answer: 8 $\longleftrightarrow$ Ground truth: 8

Step-level math correction:

Student Solution:

Step 1: Translate the condition into inequalities: $2 < \sqrt{t} < 3.5$. 

Step 2: Square all parts of the inequality to eliminate the square root: $2^2 < t < 3.5^2$. 

Step 3: Calculate the squares: $4 < t < 12.25$. 

Step 4: Determine the integer values of t that satisfy the inequality. The integers that fall strictly between 4 and 12.25 are 5, 6, 7, 8, 9, 10, 11, 12. However, since t must be strictly less than 12.25, 12 is not included in the solution set because it is not strictly less than 12.25.  12 should be included in the solution set.

Step 5: Count the integer values: The integers that satisfy the condition are 5, 6, 7, 8, 9, 10, 11, which gives us a total of 8 integer values of t .  Student missed a number but got the correct final answer 8. Twelve also satisfies the condition.

Step-wised Error Correct on Math

Summary for Error Correction

- **Goal:**

- Help students to find mistakes and improve from corrected solutions.

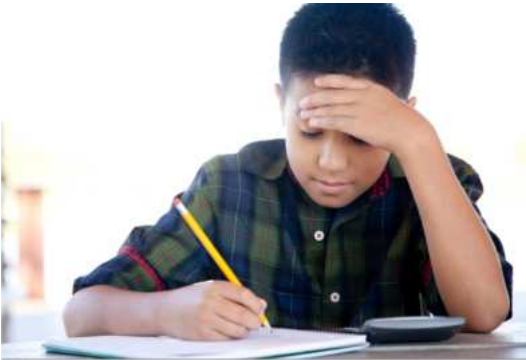
- **Advantage:**

- LLMs' prior knowledge makes it easy to find and correct errors in K12 subjects.
- LLMs' reasoning capability enables it to correct errors over the challenging STEM subjects.

- **Direction:**

- Improve the error detection accuracy in STEM subjects.
- Extend the application to scenarios including multi-modal cases.

Overview of AI Applications in Education



● Learning

- Question Solving
- Confusion Helper
- Error Correction

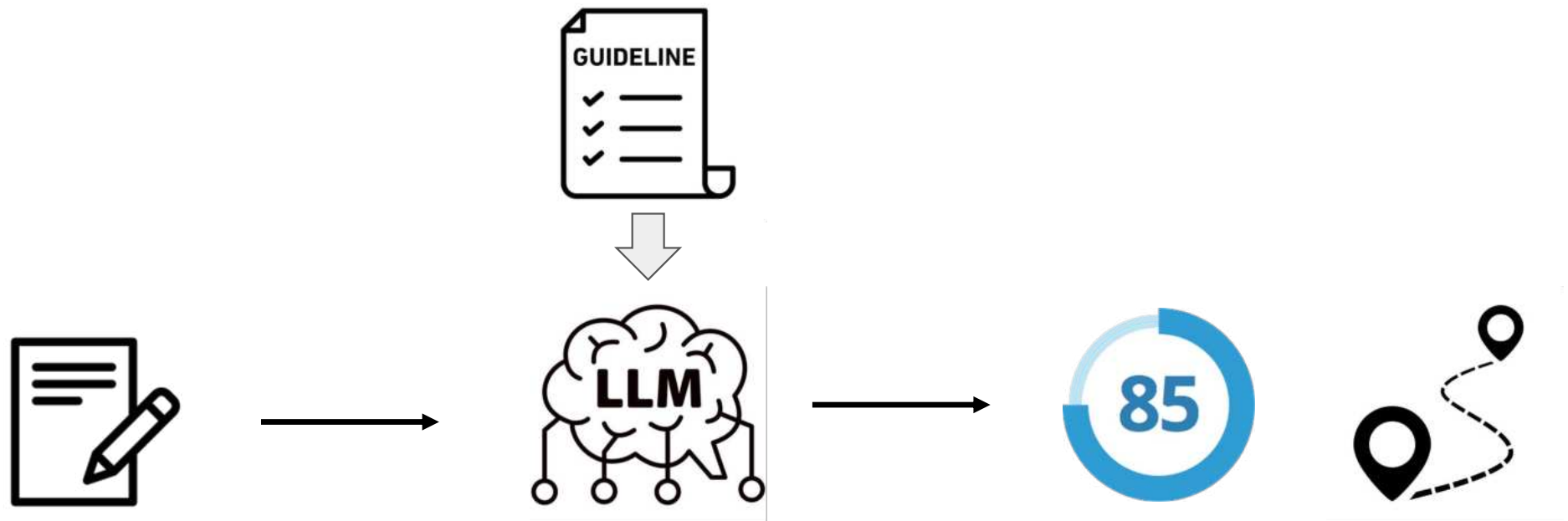
● Teaching

- Content Generation
- Automatic Grading
- Profession Development

● Assessment

- Knowledge Tracing
- Question Generation

Teaching: Automatic Grading



LLMs can grade directly according to rubrics and provide explanations for their decisions.

Summary for Automatic Grading

- **Goal**

- Help teacher to grade student answers correctly and reliably.

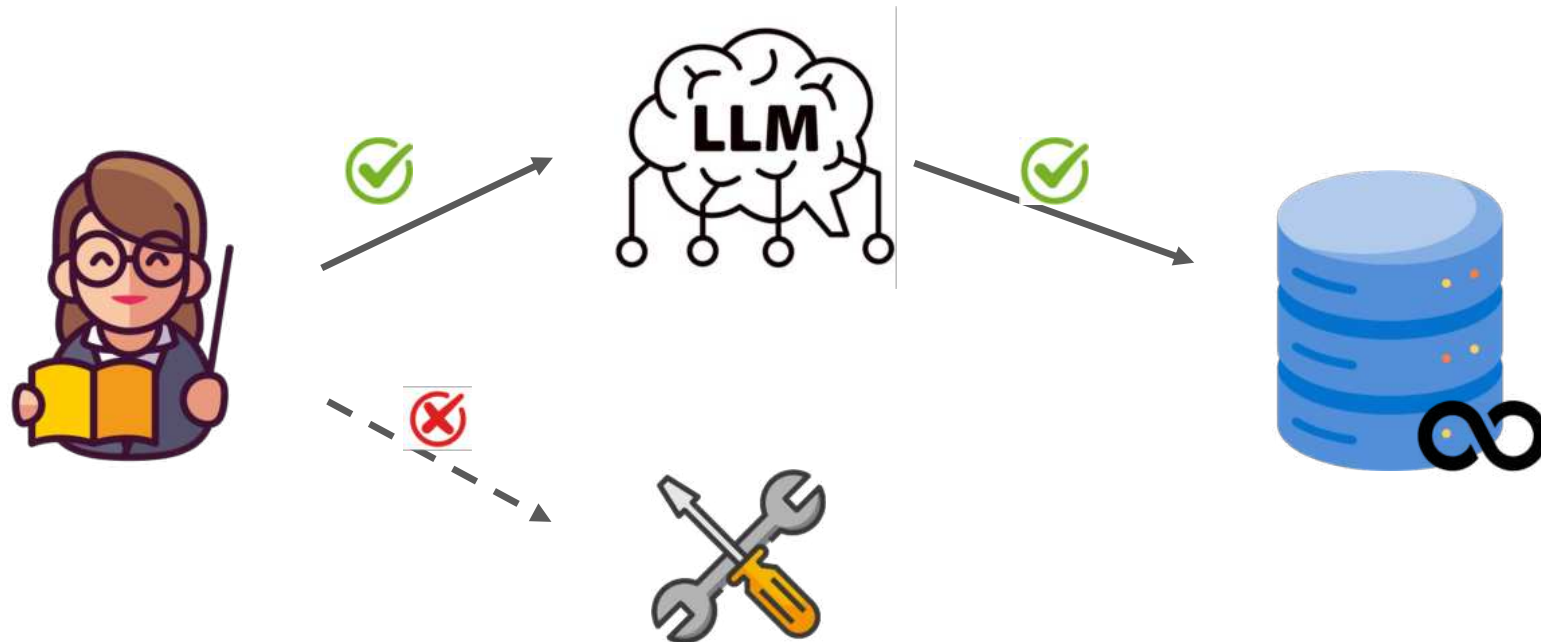
- **Advantages**

- LLMs' instruction following capability makes it avoid additional training for grading on tasks.
- The human readable outputs on rationals make the grading results interpretable to human.

- **Directions**

- Improve the general grading accuracy with the optimized rubric instructions.
- Improve the reliability of the grading results via human supervisions and uncertainty measurements.

Teaching: Material Generation



LLMs provide teachers with limitless teaching resources and reduce the need for complex content creation skills.

Example of Material Generation

Parallelogram

Student learning analysis

Considering that our student group is second-year junior high school students, they should have mastered basic geometry knowledge in the first year.....

Lesson objectives

1. Knowledge and Skills: Students should be able to understand and master the basic properties of parallelograms, including equal diagonals, 2. Process and Methods: Enable students to understand the properties of parallelograms intuitively, for example, through physical models 3. Emotional Attitudes and Values: Stimulate students' interest in learning about parallelograms, and enhance their enthusiasm and participation in learning..... 4. Comprehensive Application: Enable students to apply the knowledge of parallelograms they have learned to.....

Key and difficult points

1. Understanding the Properties of Parallelograms: The main content of this lesson is the properties of parallelograms, 2. Drawing Parallelograms: Students need to master how to

Materials

1. Teaching Aids: Ensure there are enough rulers and protractors for students to use. Prepare some 2. Textbook Preparation: In addition to being familiar with the content about parallelograms in the textbook, it is necessary to 3. Courseware Preparation: Design a PPT courseware of parallelograms that includes theoretical knowledge, examples, exercises, etc. Ensure that 4. Lesson Plan Preparation: A detailed lesson plan should be developed based on the student's learning situation and teaching objectives. The objectives of 5. Exercise Preparation: Design.....

Lesson procedure

1. Reviewing Old Knowledge: Teacher: Let's start by reviewing what we've learned about shapes. Can you tell me what parallel and perpendicular lines are? 2. Introducing New Knowledge: Teacher: Today we are going to learn about parallelograms. Before that, can you try to tell me the definition of a parallelogram and 3. Learning New Knowledge: Teacher: Now, I want you to try to draw a parallelogram with a ruler and protractor, and try to find out its properties. Students: (Students try to draw.) Activity Intent: Through hands-on practice, students can 4. Consolidating New Knowledge: Teacher: Okay, now let's do some exercises about parallelograms, which will help you better understand and master the properties of parallelograms. Students: Activity Intent: Through exercise training, consolidate students' new 5. Expanding New Knowledge: Teacher: In fact, parallelograms can be seen everywhere in our daily life. Now Activity Intent: By guiding students to discover applications in life, enhance students'

Toy Redis



Introduction

Welcome to this exciting project where you will learn to build a Redis clone using Python! Redis is a popular in-memory data store that is used by many organizations around the world. By building a Redis clone, you will gain a deep understanding of how it works and how it can be used to handle basic commands like PING, GET, and SET. As Robert Frost once said, "Two roads diverged in a wood, and I - I took the one less traveled by, and that has made all the difference." So, take the road less traveled and embark on this journey to build your very own Redis clone!

Environment Setup

Taking the Windows as an example, we will introduce the local development configuration of the project.

Preparation

(1) Clone the project

Click the Editor button to access the project's Git URL, then open the console and execute the following commands:

```
git clone https://GIT URL.git
```

(2) Set up a virtual environment

Learning Step 1

In this step, we will learn about the Redis protocol and how it's used to parse incoming requests and generate appropriate responses. We will also learn about asyncio library and how it can be used to build event-driven applications. We will not implement any functionality in this step.

Here's the sample code:

```
import asyncio
```

AI Scenario Dialogue

Hi John, I'm supposed to build a Redis clone. What is Redis anyway?

Well, Redis is like a superhero that can store and retrieve data really quickly. It's used as a database, cache, and message broker.

Pomodoro Plan

- ☐ Read about the Redis Protocol and how it is used to parse incoming requests and generate appropriate responses. (1-2 Pomodoros)
- ☐ Learn about the asyncio library and how it can be used to build event-driven applications. (2-3 Pomodoro s)

Questioning

- What is Redis, and why is it a popular in-memory data store?
- How does the Redis Protocol work, and how is it used to parse incoming requests and generate appropriate responses?
- What is an event loop, and how is it used in this

Summary

Congratulations on completing this project! By building a Redis clone, you have gained a deep understanding of the Redis Protocol, event loops, and how to implement basic key-value store functionality. As you continue on your journey to learn more about Python and software development, remember that the most important thing is to keep learning and experimenting. As Albert Einstein once said, "I have no special talent, I am only passionately curious." So, stay curious and keep on learning!

Summary for Material Generation

- **Goal**

- Allow teachers to create any form of desired learning content.

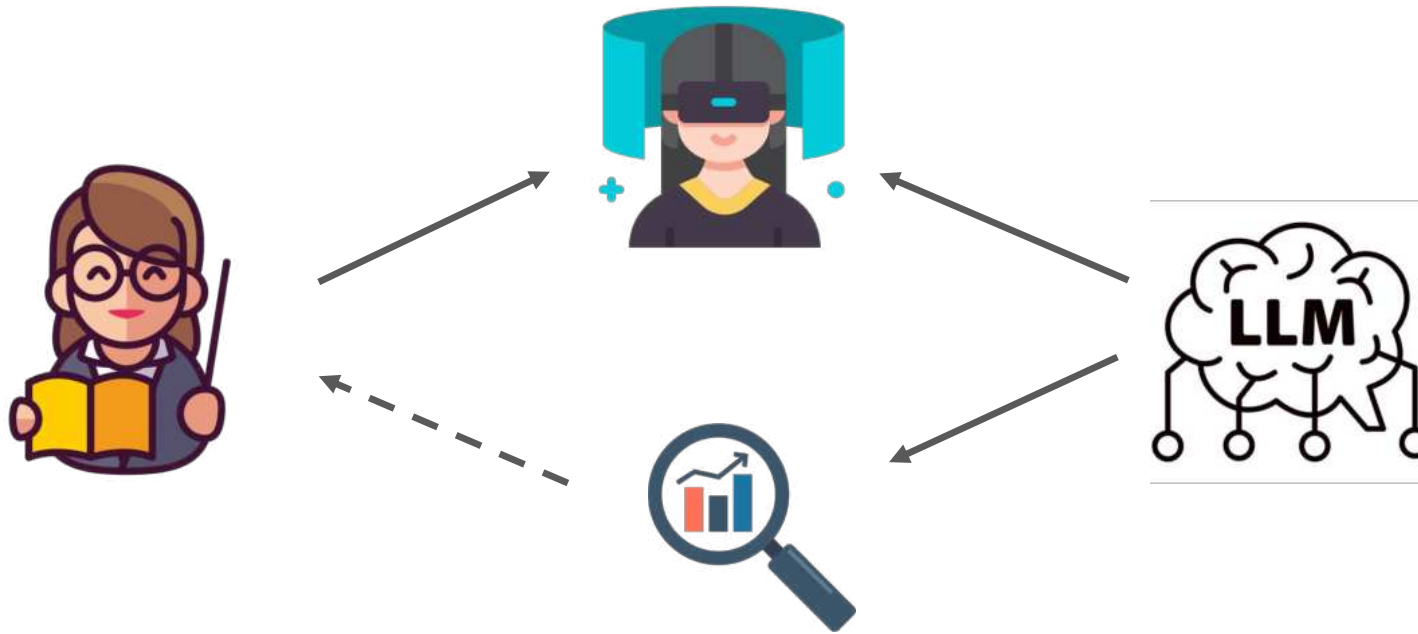
- **Advantages**

- LLMs can generate contents simply by instructions text, facilitate the workloads and requirements to teachers during course preparation.
- The generative nature allows LLMs to generate infinity contents and easily customized to the additional conditions.

- **Directions**

- Extend from the single modality to multimodality.
- Fine-grained controls to the contents and improve generation qualities.

Teaching: Profession Develop



LLMs can serve as simulators and analyst to help teachers practice their pedagogical skills.

Summary for Professional Development

- **Goal**

- Improve teacher's the pedagogical skills.

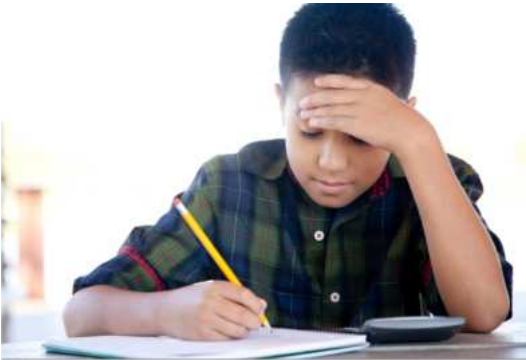
- **Advantages**

- The simulation enables teachers to practice their skills with minimal cost.
- By automate the analyzing process, teachers can always get intime suggestions and feedbacks anywhere anytime.

- **Directions**

- Improve the behavior of LLMs better aligned with true students.
- Provide suggestion algin better with human experts.

Overview of AI Applications in Education



● Learning

- Question Solving
- Confusion Helper
- Error Correction

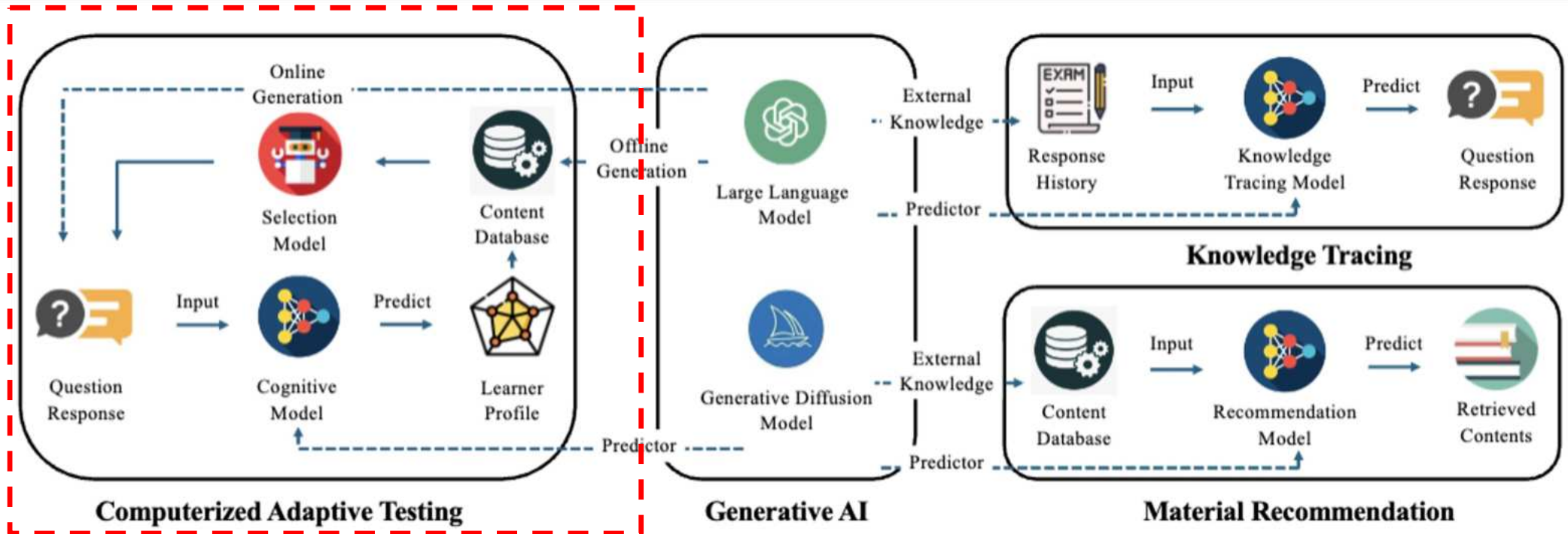
● Teaching

- Content Generation
- Automatic Grading
- Profession Development

● Assessment

- Knowledge Tracing
- Question Generation

Assessment: Question Generation



The generated questions avoids repetition in question bank, which also enables the dynamic assessment process.

Example of Question Generation

Math-Based MCQ

Question: In a simple linear regression model, the slope of the best-fit line for a training dataset was calculated to be 1.5. If the model predicts a value of 4.5 for an input of 2, what is the intercept of the model?

Options: A) 1.5 B) 2.0 C) 3.0 D) 4.5

Concept-Based MCQ

Question: In a Support Vector Machine, what is the primary purpose of using the kernel trick?

Options:
A) To reduce the dimensionality of the feature space
B) To linearly separate data in a higher-dimensional space
C) To increase the computational efficiency of large datasets
D) To implement online learning in real-time

Coding-Based MCQ

Question: Given the following Python code snippet for a binary tree node and a function 'find_max_depth', which is designed to find the maximum depth of the binary tree:

```
python
class TreeNode:
    def __init__(self, val=0, left=None, right=None):
        self.val = val
        self.left = left
        self.right = right

def find_max_depth(root):
    if root is None:
        return 0
    else:
        left_depth = find_max_depth(root.left)
        right_depth = find_max_depth(root.right)
        return max(left_depth, right_depth) + 1
```

What concept does the 'find_max_depth' function primarily demonstrate?

Options:
A) Breadth-first search (BFS)
B) Depth-first search (DFS)
C) Binary search
D) Dynamic programming

Requirements

Topics:

CEFR: ☐ A1 ☒ A2 ☐ B1 ☐ B2

Length:

Genre:

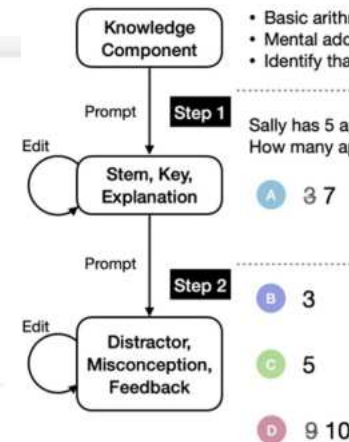
Based on the example? ☐ Yes ☒ No

Example:

Questions: ☐ 1 ☐ 2 ☐ 3 ☒ 4 ☐ 5

Choices: ☐ 3 ☒ 4 ☐ 5

Question Type:



Sally has 5 apples. She gives gets 2 apples to from her friend. How many apples does Sally have left?

The problem states that Sally gives gets 2 apples to from her friend, which means she is subtracting adding 2 from to the total number of apples she has. Therefore, the correct answer is 3 7.

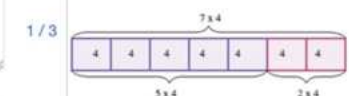
Incorrectly subtracting.

This answer suggests that Sally subtracts the 2 apples she received from her friend, which is incorrect. Remember, the problem asks for the total number of apples Sally has after receiving the additional 2 apples.

Complete the equations.

$7 \times 4 = 5 \text{ fours} + 2 \text{ fours} = ? + 8 = ?$

Let's use a tape diagram to break the 7×4 into two parts.



The tape diagram splits 7 fours into 5 fours and 2 fours. Why? Because $5+2=7$. You could also use $4+3$ or $1+6$.

$7 \times 4 = 5 \text{ fours} + 2 \text{ fours}$, $7 \times 4 = 20 + 8$, $7 \times 4 = 28$

```
<svg height="128" viewBox="0 0 320 128"
width="320" xmlns="http://www.w3.org/2000/svg">
  <path
d="M6.154 44v36H50.11V44H6.154M50.11
44v36h43.956V44H50.11M94.066
44v36h43.956V44H94.066M138.022
44v36h43.956V44H138.022
44v36h43.956V44H43.956M181.978
44v36h43.956V44H181.978
44v36h43.956V44H43.956"
fill="#7854ab" fill-opacity=".1"
stroke="#7854ab" stroke-width="2"></path>
  <path {{Pink Squares}}></path>
  <path {{Curly Brackets}}></path>
  <text fill="black"
font-size="12" x="30" y="65">4</text>
  {{Rest of the annotations}}
</svg>
```

Purple Squares

Summary for Question Generation

- **Goal**

- Generate adaptive questions for assessment usages.

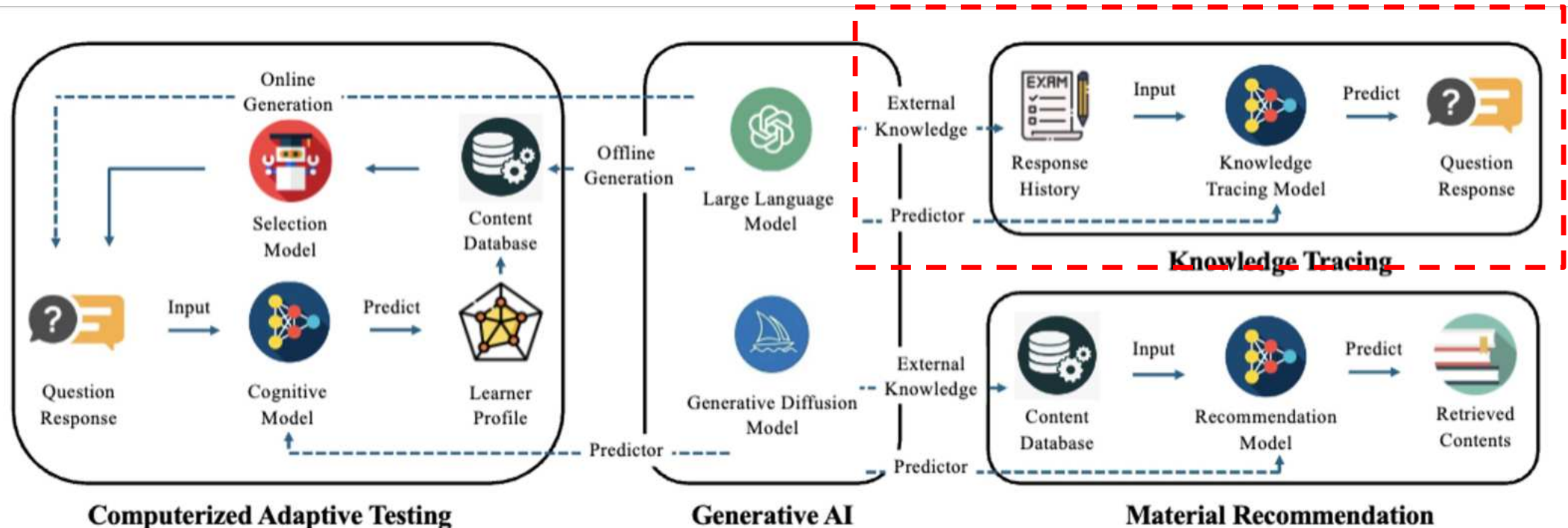
- **Advantages**

- LLMs' language capability make it easy to generate the questions while meeting various scenarios requests.
- The prior knowledge of LLM makes it support generate questions for various subjects without additional training.

- **Directions**

- Improve the controls over education specific conditions like difficulty.
- Extend the creation to multimodality outputs.

Assessment: Knowledge Tracing



LLMs can serve as data enhancer to help existing models, meanwhile it can also be a new form of predictor.

Summary for Knowledge Tracing

- **Goal**

- Provide accurate estimations to the student's ability during the learning process.

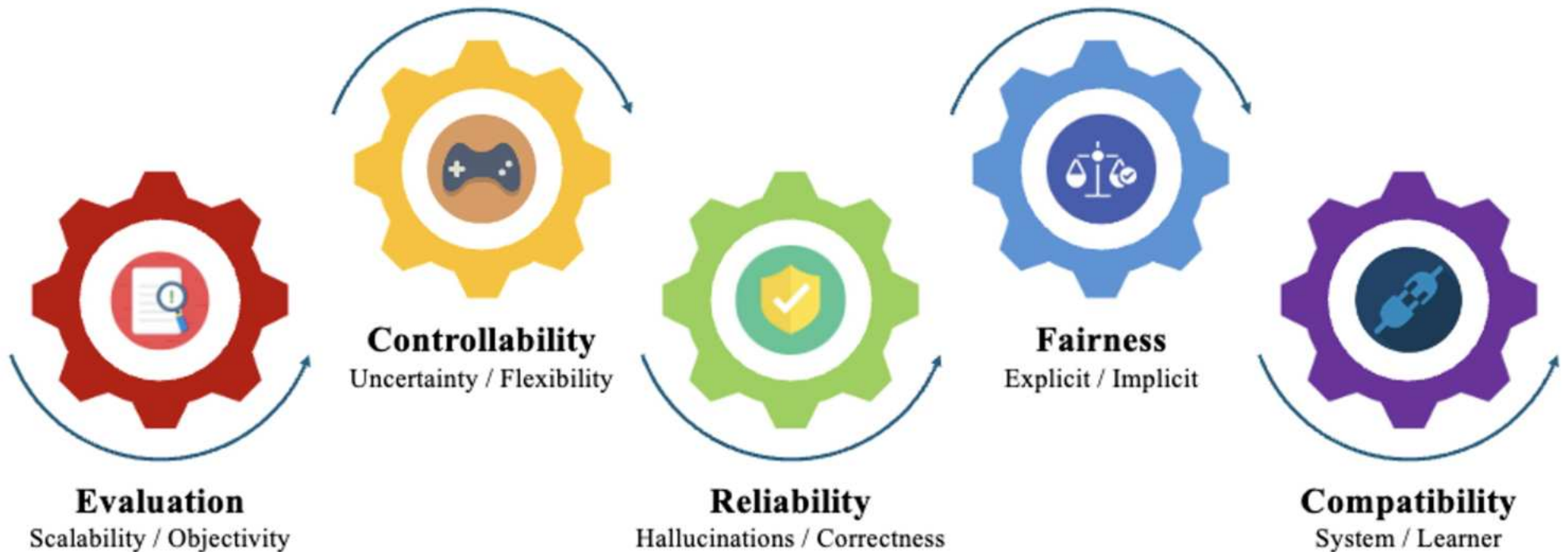
- **Advantages**

- LLM can process with more flexible testing data, and leverage additional external information if needed.
- The prior knowledge of LLM makes it support complementary feature for modeling, which is useful for solving the cold start problem in the traditional ML methods..

- **Directions**

- Explore the prediction potential of LLMs and improve accuracy.
- Expand the tracing scenarios with more diverse input forms.

Challenges and Opportunities



Challenge 1: Evaluation

Challenges:

- Individual differences among learners and the complexity of scaling experiments.
- Low feasibility for simple reference based evaluation.
- Objectively assessing personalized generated content is hard.

Opportunities:

- Develop reliable, automated evaluation methods tailored for educational applications.
- Create simulation algorithms that imitate human learners to enrich the diversity and quantity of testing samples.

Challenge 2: Controllability

Challenges:

- The generated content may deviate from expectations.
- Struggle to satisfy all requirements simultaneously.

Opportunities:

- Enhance the efficiency of human supervision during the generation process.
- Train specialized models to achieve precise, fine-grained control over generated outputs.

Challenge 3: Reliability

Challenges:

- Producing misleading information that can have severe consequences for learners.
- Incorporating human supervision is constrained by limited human resources.

Opportunities:

- Enhance the efficiency of human supervision during the process.
- Automated verification techniques to reduce the workload and reliance on human experts.

Challenge 4: Fairness

Challenges:

- Risk of generating biased content that reflects majority interests.
- Users with limited exposure to GenAI tools may face unfair disadvantages in fully leveraging its capabilities.

Opportunities:

- Models capable of filtering and correcting biased or unfair content.
- Policies and Laws aimed at enhancing fairness and ensuring equitable access and outcomes in the use of GenAI.

Challenge 5: Compatibility

Challenges:

- The operational costs and resource requirements of the most advanced GenAI models can be substantial.
- Compatibility with existing educational practices and prior experience.

Opportunities:

- Reduce the burden and costs associated with integrating LLMs into educational settings.
- Address the unique challenges and emerging issues introduced by the adoption of LLMs in education.

Tutorial Outline

Morning

Introduction
to GenAI
(40 Min)



Overview of GenAI
in Education
(40 Min)



Applications in
Education Part I
(40 Min)

Afternoon

Applications in
Education Part II
(40 Min)

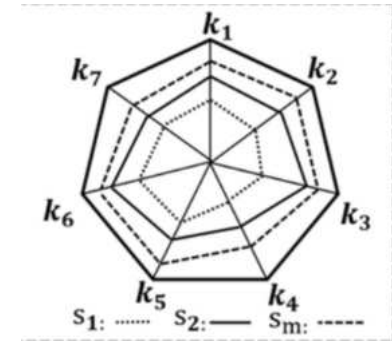


Demo
Presentation
(40 Min)



Future Direction
and QA
(40 Min)

Applications in Education Part I



- **Automatic Grading**

- Multi-Agent
- Human-in-the-Loop
- Retrieval Augmented Generation

- **Content Generation**

- Material Generation
- Question Generation

- **Knowledge Tracing**

- LLM as annotator
- LLM as predictor

Automatic Grading

- **Goal:**

- Follow instructions of teachers and predict score / grades to submissions to the problems.

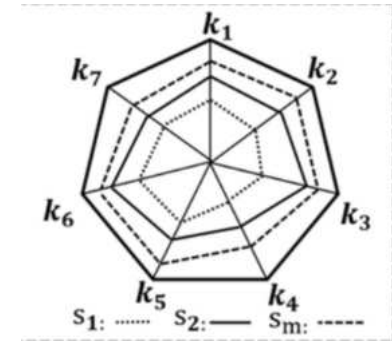
- **Problems:**

- Most automatic grading models are question specific.
- Teacher cannot provide additional supervisions except provide labels.
- Results are not directly interpretable to users.

- **Advances:**

- LLMs' prior knowledge enables it to work with multiple subjects easily.
- The instruction following and in-context learning capability enables the teacher to build new grading tasks easily.
- The output rationale enable teachers to understand the grading process.

Applications in Education Part I



- **Automatic Grading**

- Multi-Agent
- Human-in-the-Loop
- Retrieval Augmented Generation

- **Content Generation**

- Material Generation
- Question Generation

- **Knowledge Tracing**

- LLM as annotator
- LLM as predictor

Naive LLM-based Automatic Grading



Version 1: Single Agent

Teacher Response: "Use a graph where the cost values are better defined initially. I can illustrate the point (5, 30) on the graph but it would help if I could show 2 or more. I could also demonstrate slope on the graph if the **cost quantities clearly showed multiples of 6.**"

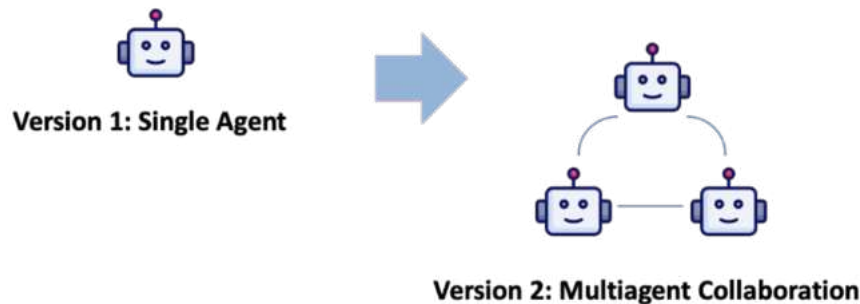
Human Code: **1**

AI Agent Code: **0 (wrong)**

AI Agent Reasoning:

- **Covariance in Ratio Table: No mention**
- **Invariance in Ratio Table: No mention.**
- **Covariance in Graph: No mention (wrong).**
- **Invariance in Graph: Explicit.**
- **Covariance and Invariance Shown Through Connection between Ratio Table and Graph: Explicit.**

Multiagent Collaboration



Reflector Agent → Analyze Errors

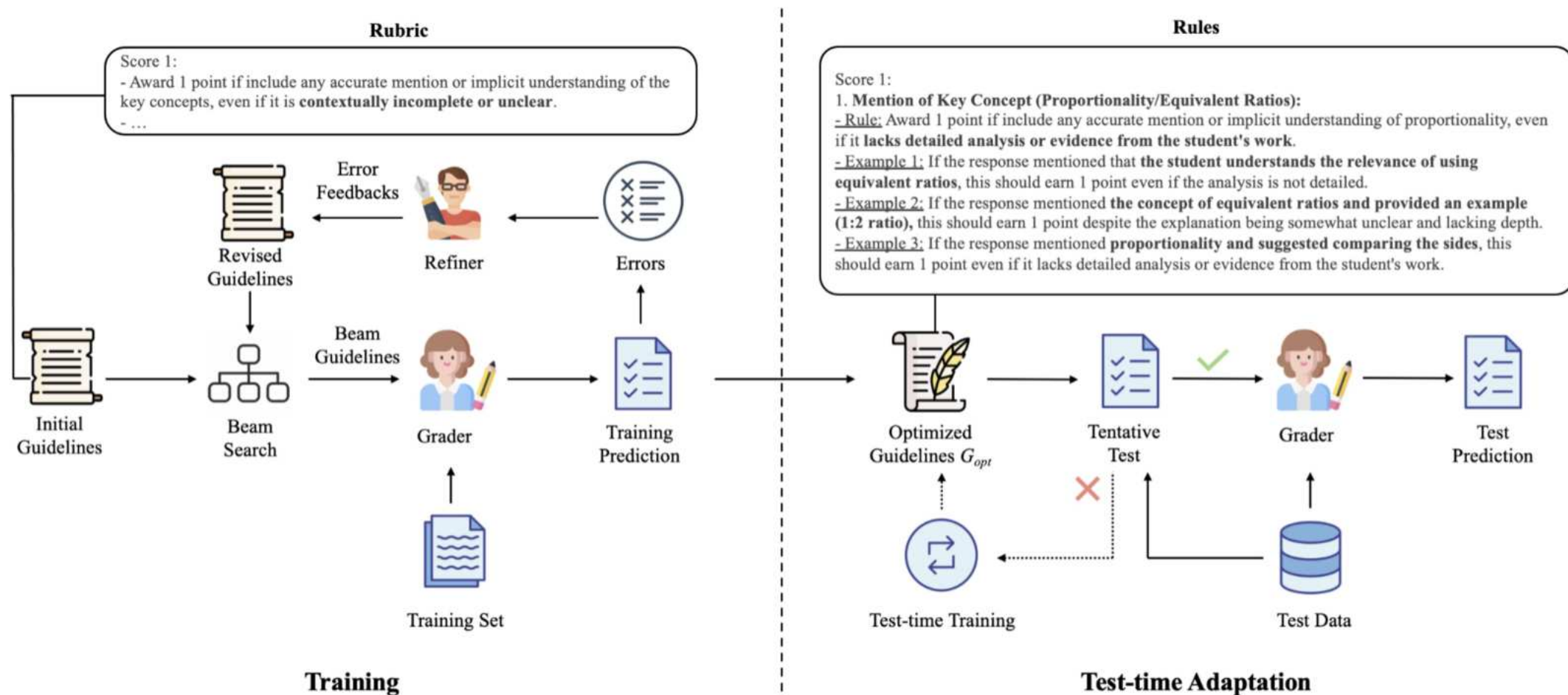
- The current rules might have *failed* to recognize the explicit mention of **invariance** in the graph and the **connection** between the table and graph.
- The *lack* of explicit or implicit mentions of **covariance** in both the table and graph might have led to a lower score prediction.

Refiner Agent → Modify Rules

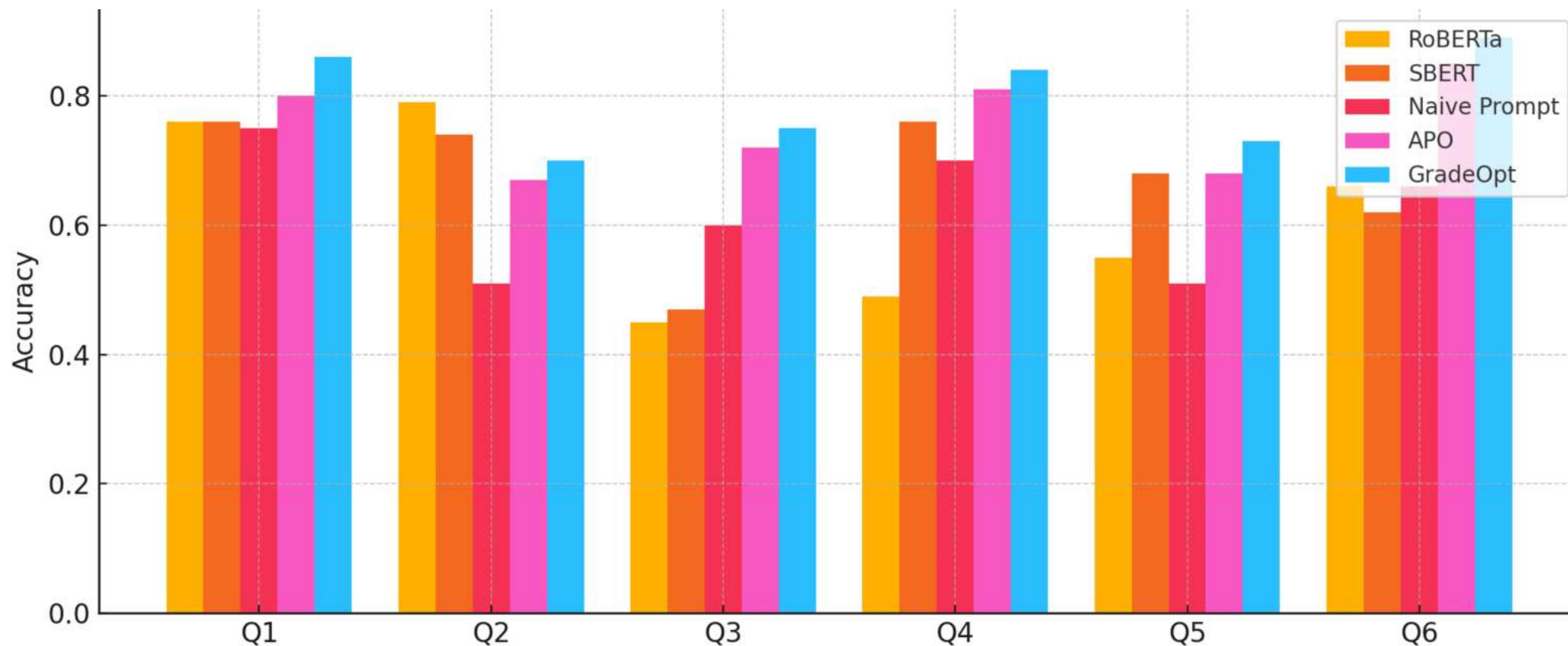
Covariance in Graph:

- Implicit Mention:
 - If the response shows the **change in the graph without explicitly stating that values change together**, consider it implicit.
 - Example: "slope on the graph if the **cost quantities clearly showed multiples of 6**."

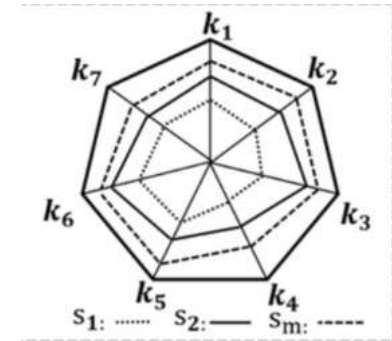
Multi-Agent Framework



Results and Findings



Applications in Education Part I



- **Automatic Grading**

- Multi-Agent
- **Human-in-the-Loop**
- Retrieval Augmented Generation

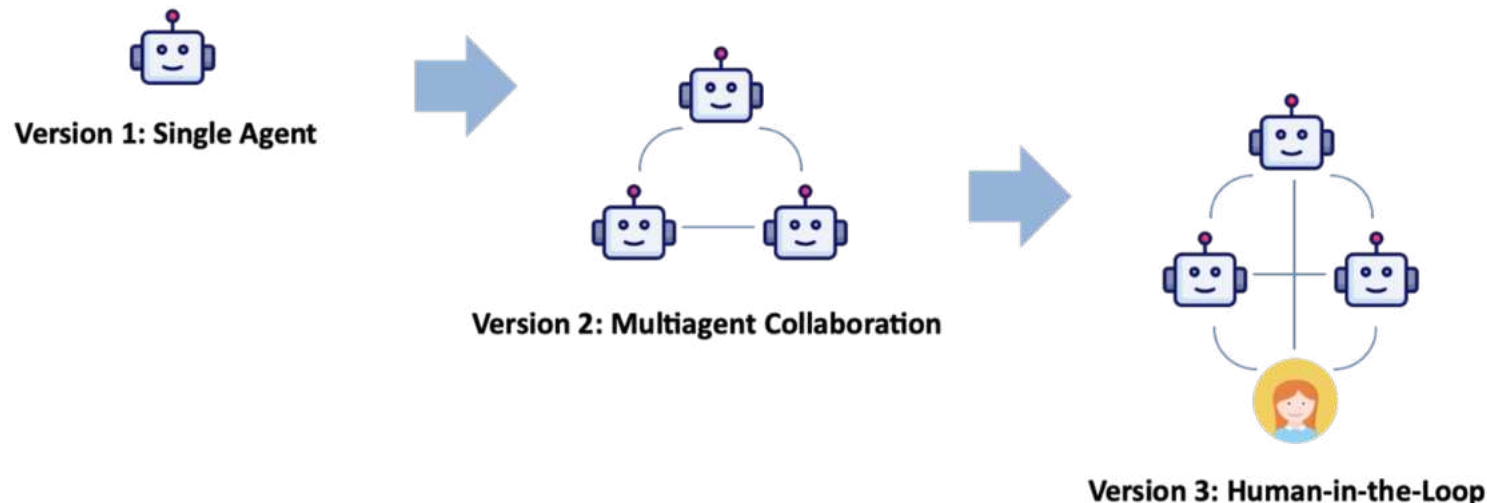
- **Content Generation**

- Material Generation
- Question Generation

- **Knowledge Tracing**

- LLM as annotator
- LLM as predictor

Human-in-the-Loop in Automatic Grading



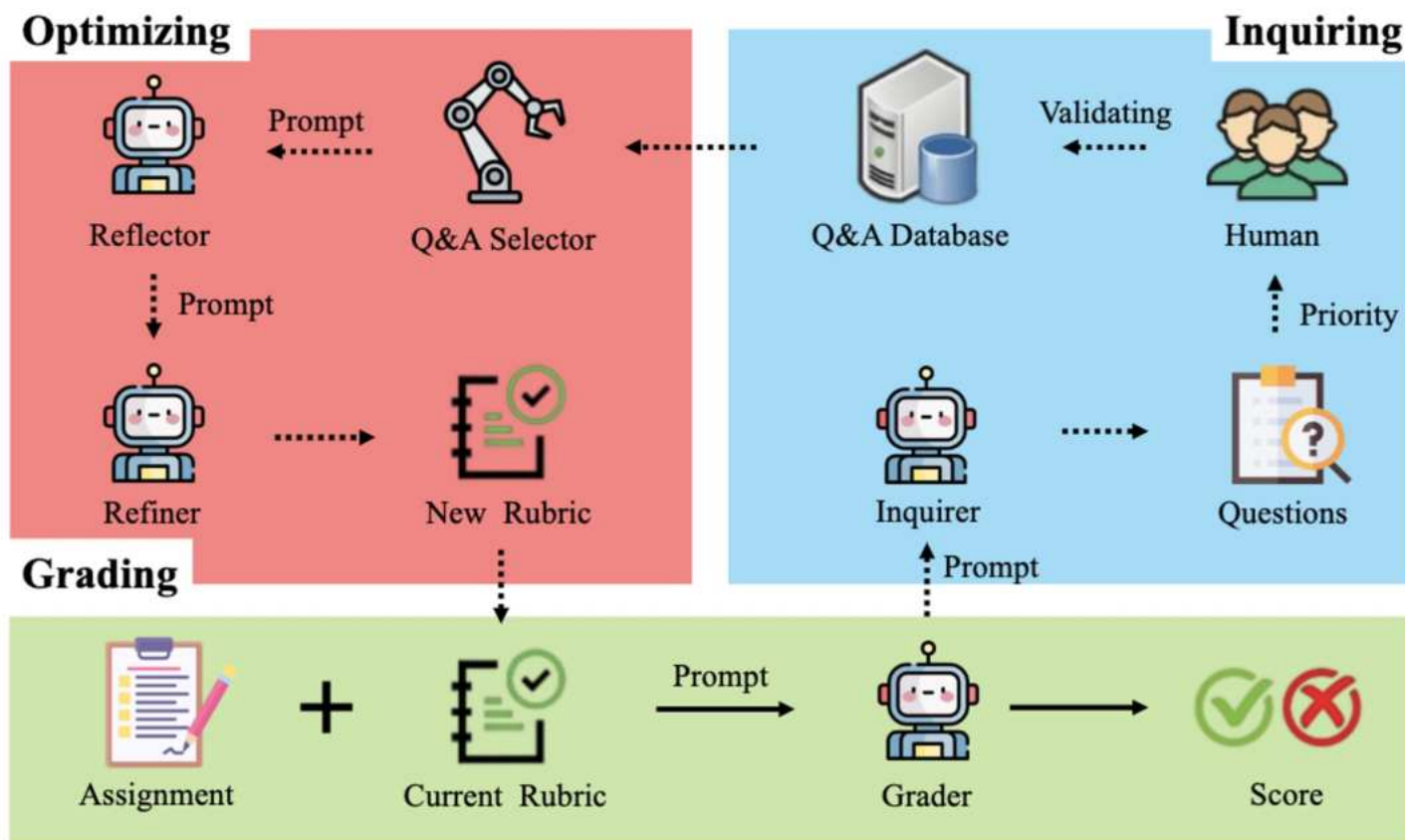
AI Question:

For **covariance in graph**, if a teacher says they want to *"illustrate the point (5, 30)"* and add *"2 or more" points*, is that considered an **implicit** way of showing that values change together, or does that still fail the **covariance** criterion because it lacks discussion of change or joint variation?

Human:

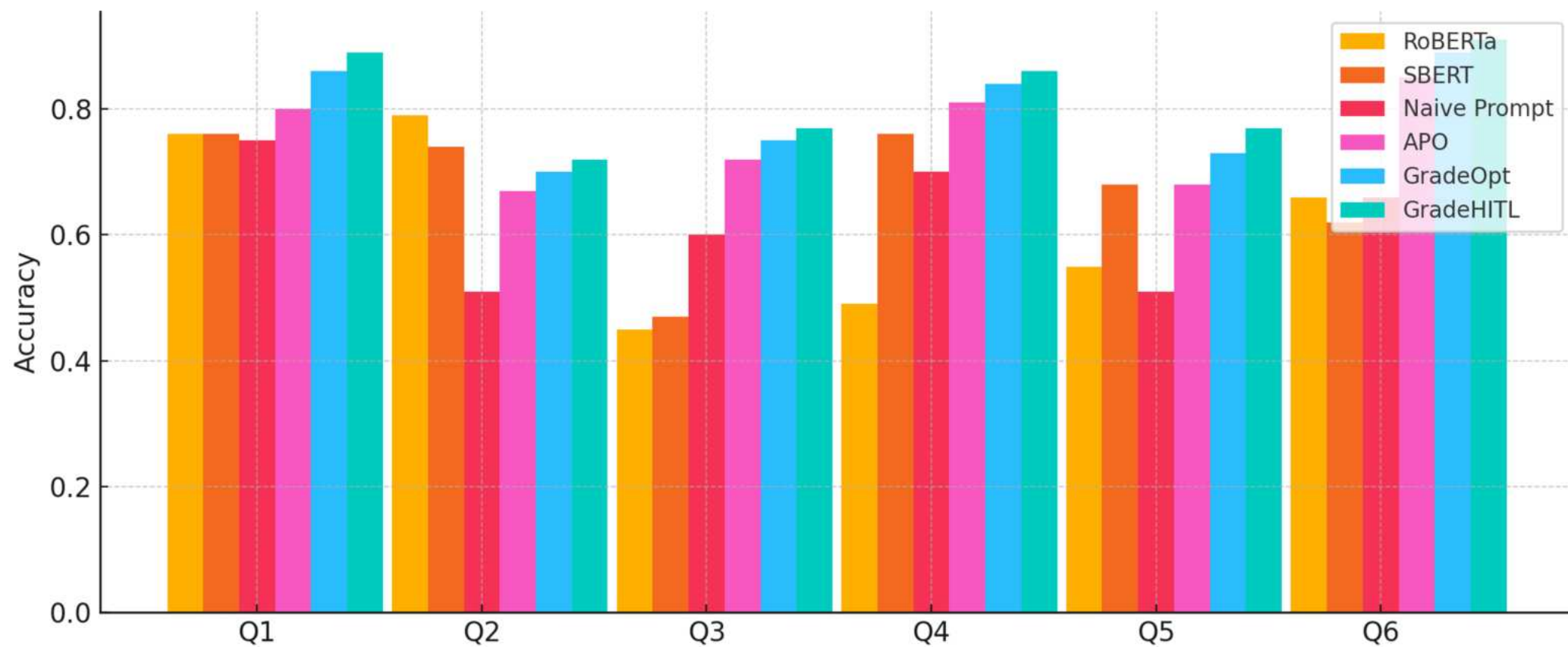
Yes, because plotting multiple points implies an awareness of the relationship between changing x and y values, even without explicitly stating that the quantities change together.

GradeHITL Framework

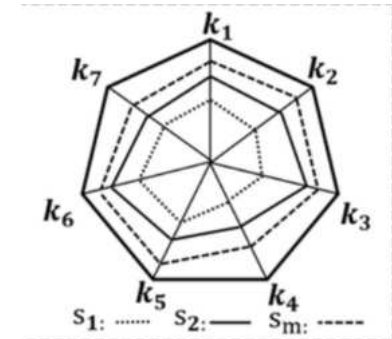


- **Grading:**
 - Find the defects of the current rubric from failure samples.
- **Inquiring:**
 - Generate question for human to answer.
- **Optimizing:**
 - Improve the performance with RL.

Results and Findings



Applications in Education Part I



- **Automatic Grading**

- Multi-Agent
- Human-in-the-Loop
- Retrieval Augmented Generation

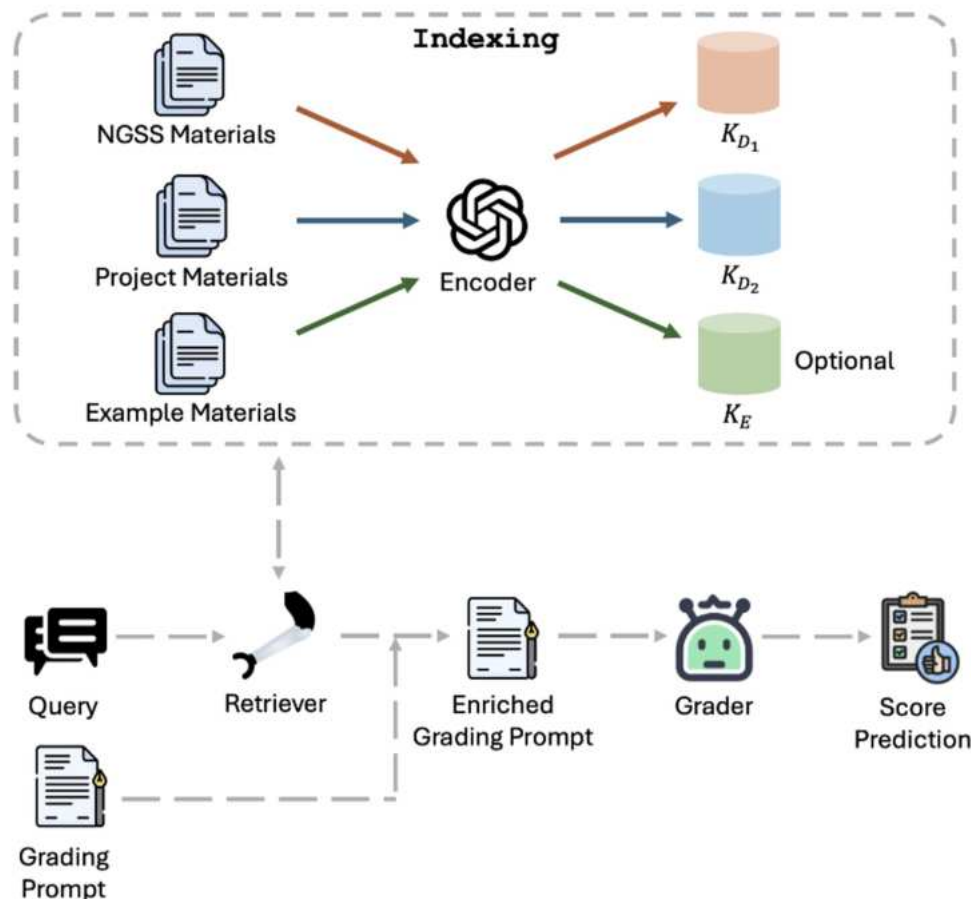
- **Content Generation**

- Material Generation
- Question Generation

- **Knowledge Tracing**

- LLM as annotator
- LLM as predictor

Grading and Feedback



- **RAG-Enhanced Accuracy:**
 - Combine LLMs with RAG
- **Semantic Search + Expert Rationales:**
 - Semantic search over educational materials
- **Efficient and Reliable:**
 - Consistent with reliable educational database

Grading and Feedback



Query

- **Scoring Criteria:**

Learning Goals for SEP: Relational [Level 2] and Descriptive explanation [Level 1].

- Level 2: Relational explanations should include a claim with scientific ideas (e.g., using the properties of substances to identify whether a chemical reaction occurs) to account how one event relates to another in a specific phenomenon that using multiple sources of evidence, without distinguishing the correlational or casual relationships in the phenomenon.

- Level 1: Descriptive explanations should include a claim with scientific ideas (e.g., properties can be used to identify substances) to account what happened in a specific phenomenon that using the observed evidence, without attempting to provide any relationships in the phenomenon.

- **Instruction**

Examine following example graded answers using the scoring rationales:

Student's Short Answer: When coconut oil is mixed with lye, a chemical reaction occurs because soap and glycerol are new substances. From the table, I found that the odor, density, solubility in water, and melting point are different from each other. They are properties that can be used to identify substances and whether a chemical reaction occurs.

Grading and Feedback



“Statement: Descriptive explanations should include a claim to account what happened in a specific phenomenon that using the observed evidence, without attempting to provide any relationships in the phenomenon.”

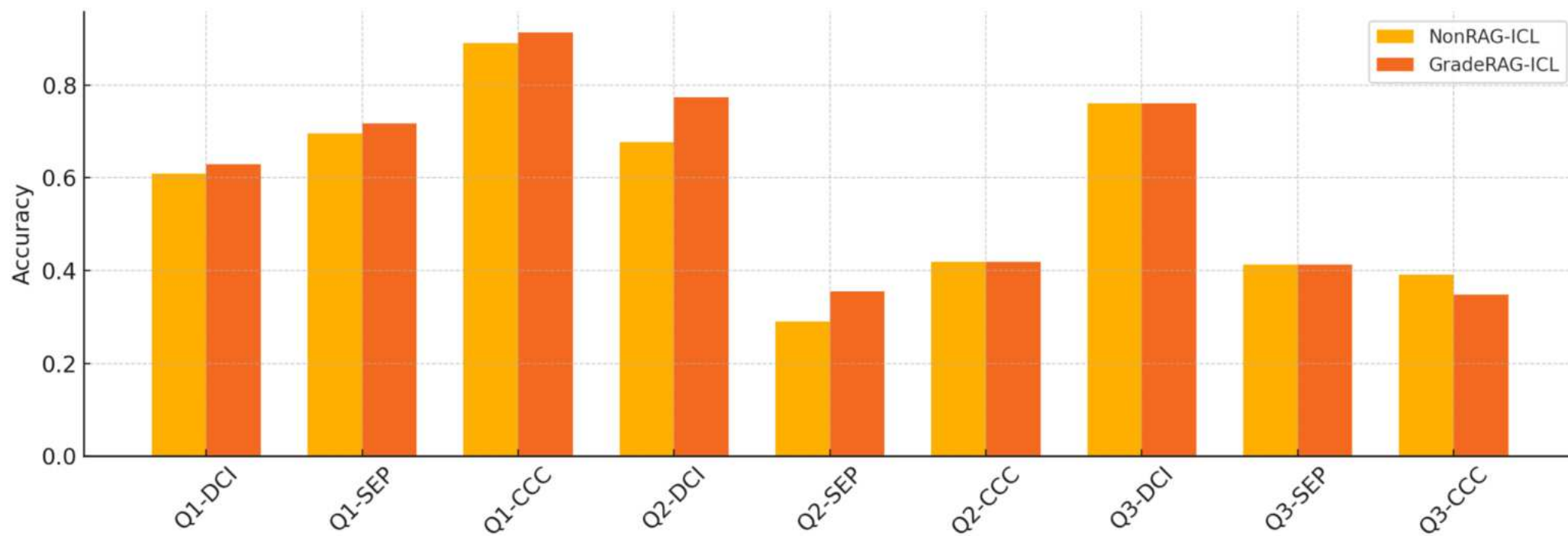
“Substances react chemically in characteristic ways. In a chemical process, the atoms that make up the original substances are regrouped into different molecules, and these new substances have different properties from those of the reactants. The total number of each type of atom is conserved, and thus the mass does not change. Some chemical reactions release energy, others store energy.”



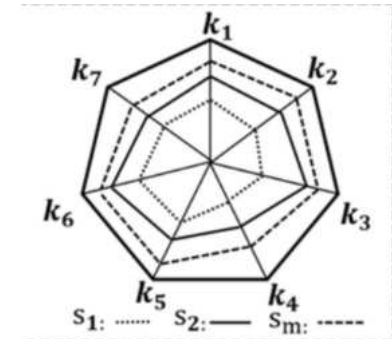
Score: SEP-1

Scoring Rationale: “When coconut oil is mixed with lye, a chemical reaction occurs because soap and glycerol are new substances.” – this part meets the partial SEP criteria that mentioned a descriptive explanation, including a claim of a chemical reaction occurs and observed evidence of the data before and after the process. However, the response did not use the evidence to connect to the phenomenon – new substance produced and a chemical reaction occurred.

Results and Findings



Applications in Education Part I



- **Automatic Grading**

- Multi-Agent
- Human-in-the-Loop
- Retrieval Augmented Generation

- **Content Generation**

- Material Generation
- Question Generation

- **Knowledge Tracing**

- LLM as Annotator
- LLM as Predictor

Content Creation

- **Goal:**

- Provide instructional content for teaching assistants or generate questions for use in assessment systems.

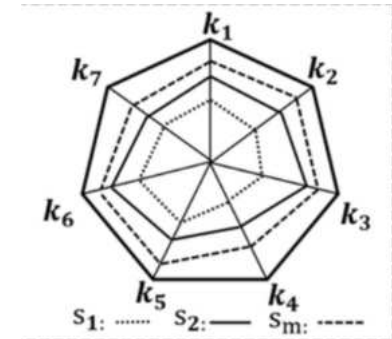
- **Problems:**

- Integrating diverse, personalized content needs to enhance learner engagement.
- Ensuring the delivery of high-quality questions to support a robust and effective assessment process.

- **Advances:**

- LLMs are capable of generating novel content, which helps mitigate concerns about question leakage during assessments.
- Their extensive prior knowledge allows them to produce rich, creative, and high-quality educational materials.

Applications in Education Part I



- **Automatic Grading**

- Multi-Agent
- Human-in-the-Loop
- Retrieval Augmented Generation

- **Content Generation**

- **Material Generation**
- Question Generation

- **Knowledge Tracing**

- LLM as Annotator
- LLM as Predictor

Material Generation

- **Goal:**

- Generate instructional content that aligns with teachers' specific requirements for use as course materials.

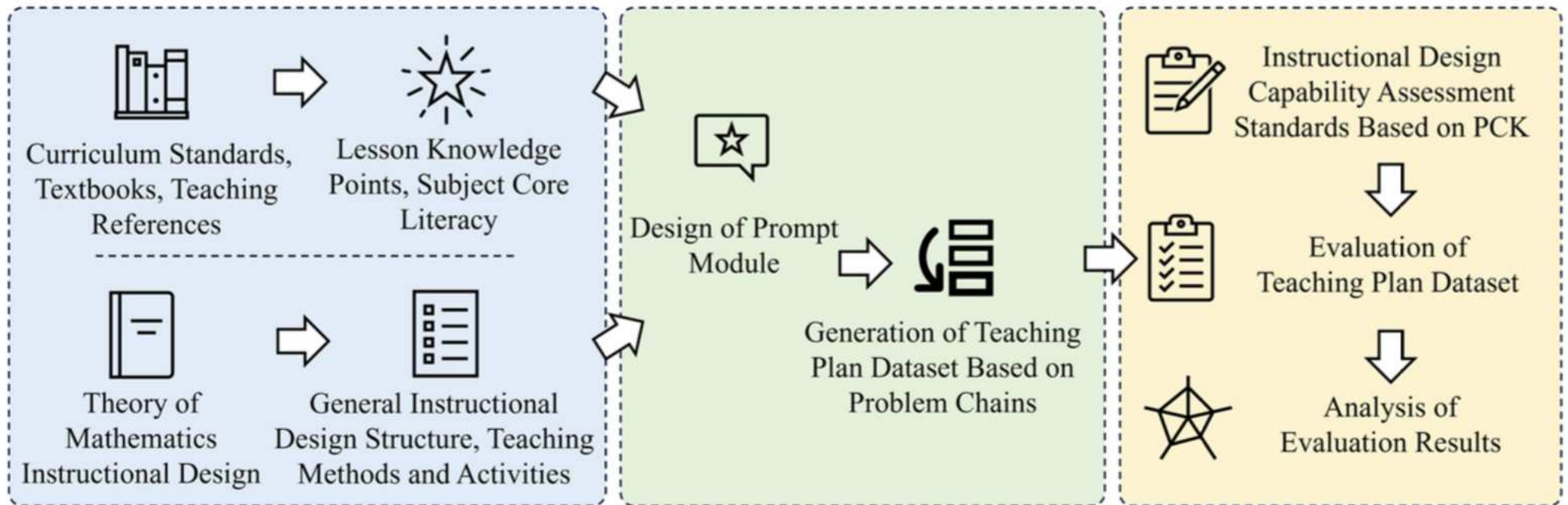
- **Problems:**

- Producing high-quality content involves meeting multiple criteria, often requiring significant teaching experience and time.
- While individual students have different learning needs, limited teacher capacity often results in standardized, sub-optimal materials being used for all.

- **Advances:**

- LLMs can generate expert-level content while accounting for multiple instructional requirements.
- LLMs offer a scalable solution that enables teachers to create personalized materials from simple oral or written instructions.

Teaching Plan Generation with LLM



Generate teaching plan with the problem chains.

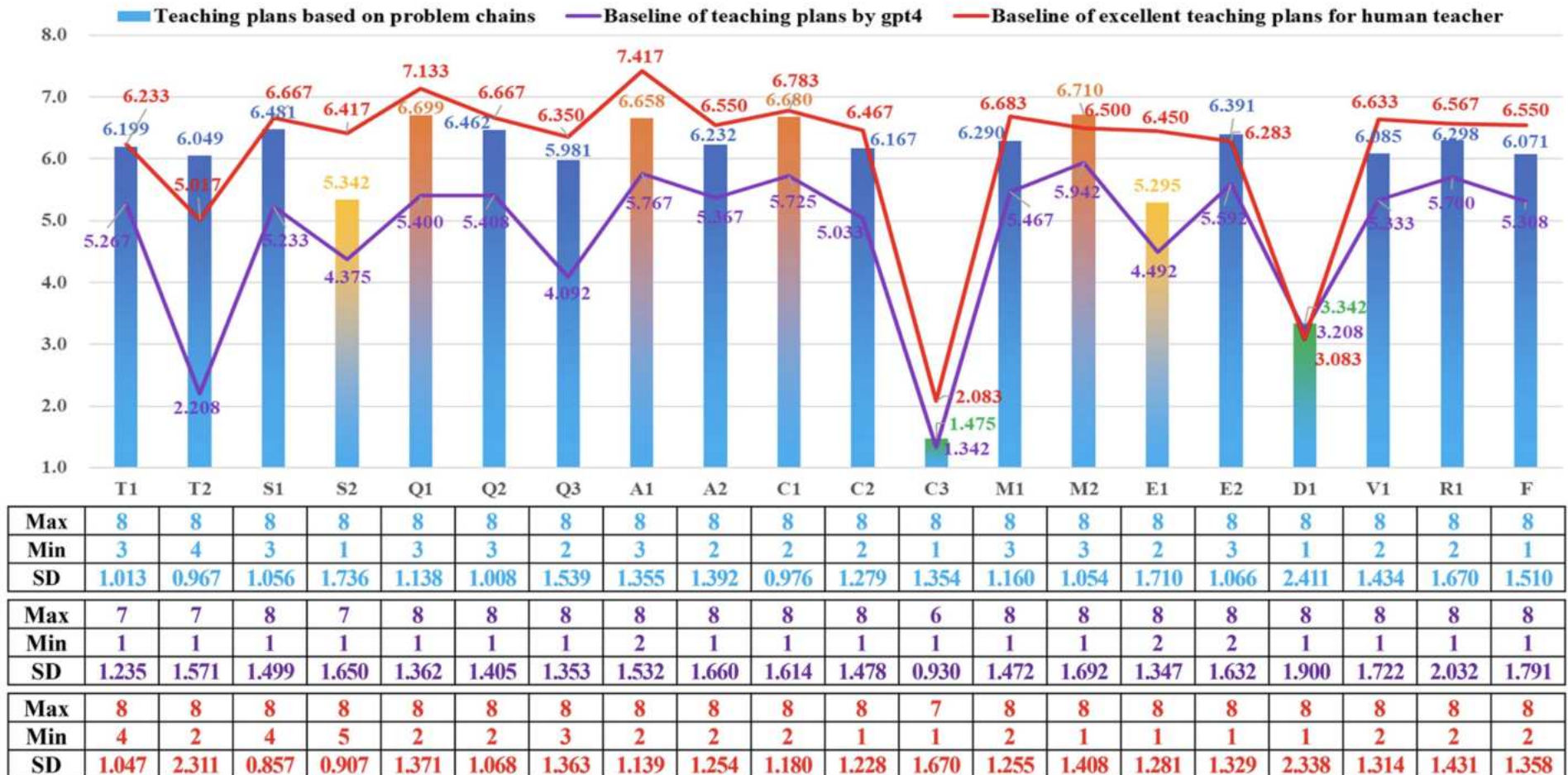
Evaluate Results by Pedagogical Factors

Category	Number	Dimension	Sources
Instructional Objectives	T1	Ability to design instructional objectives that align with the class content and are logically formulated	[70] [28]
	T2	Ability to design instructional objectives that reflect the core literacy of the subject	[70] [28]
Teaching Situation and Textbook Analysis	S1	Recognize the emphasis points of the teaching content	[28]
	S2	Identify common misunderstandings and learning difficulties of students	[71]
Problem Chain	Q1	Design and introduce rich, effective, and scientific problem scenarios	[72]
	Q2	Design challenging learning problems and tasks that connect and promote step by step	[72]
	Q3	Design suitable variant exercises to consolidate learned knowledge	[72]
Teaching Activities	A1	Design content that promotes teacher—student and student—student interactive discussions	[71]
	A2	Select teaching activities that align with the teaching objectives and content	[28]

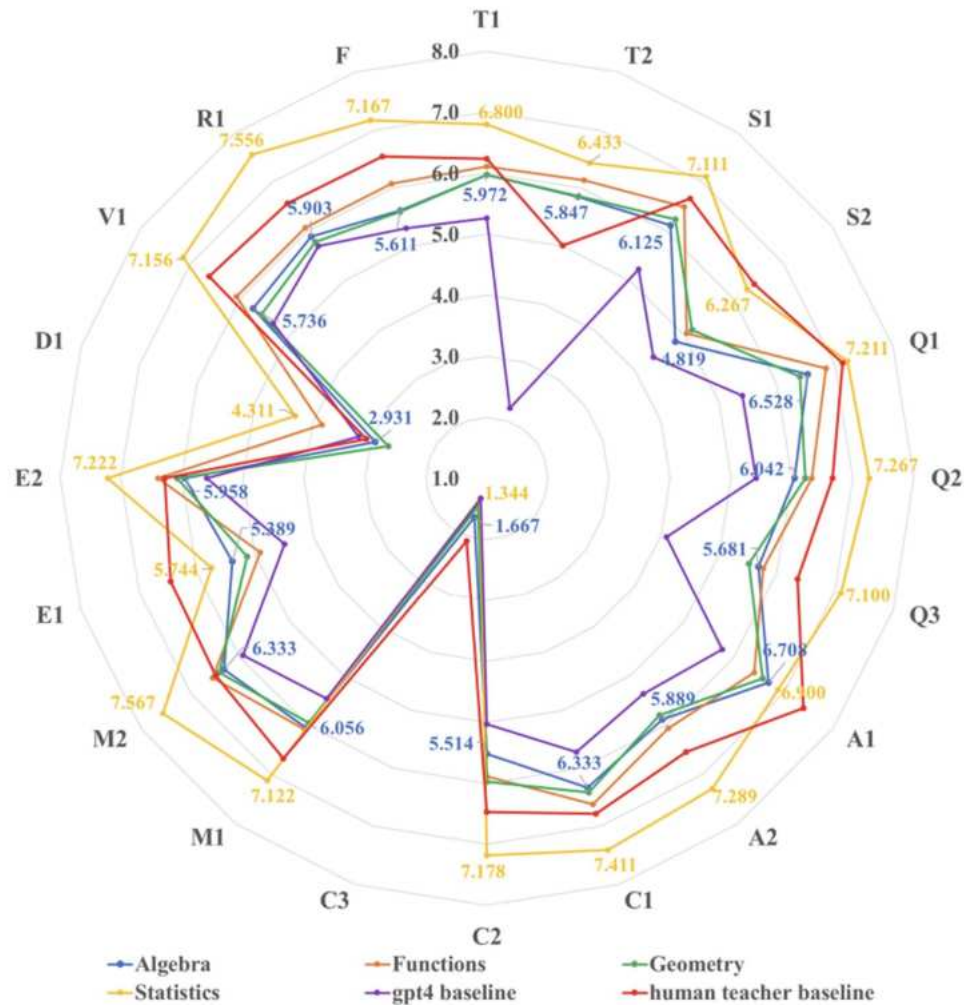
Evaluate Results by Pedagogical Factors

Knowledge Content	C1	Correctly and reasonably explain the basic theories and concepts of the subject	[28] [72]
	C2	Extract and summarize the subject content of the teaching period	[28]
	C3	Introduce the historical development of important theories in the subject and the culture of mathematics	[28]
Teaching Methods and Strategies	M1	Choose suitable teaching methods and strategies	[71]
	M2	Apply a variety of teaching methods and strategies	[28]
Teaching Evaluation	E1	Recognize the individual differences and learning differences among students	[73] [74]
	E2	Use a variety of ways to assess students' learning situation	[73] [28]
Interdisciplinary	D1	Establish appropriate connections with other disciplines and design activities that develop students' interdisciplinary abilities	[74] [71]
Usable Value	V1	Can be used for secondary modification and to provide reference value for teaching plans	\
Scope	R1	Ability to design the content within the scope of the curriculum standard	[73]
Overall Score	F	Overall evaluation of the teaching plan	\

Evaluate Results by Pedagogical Factors

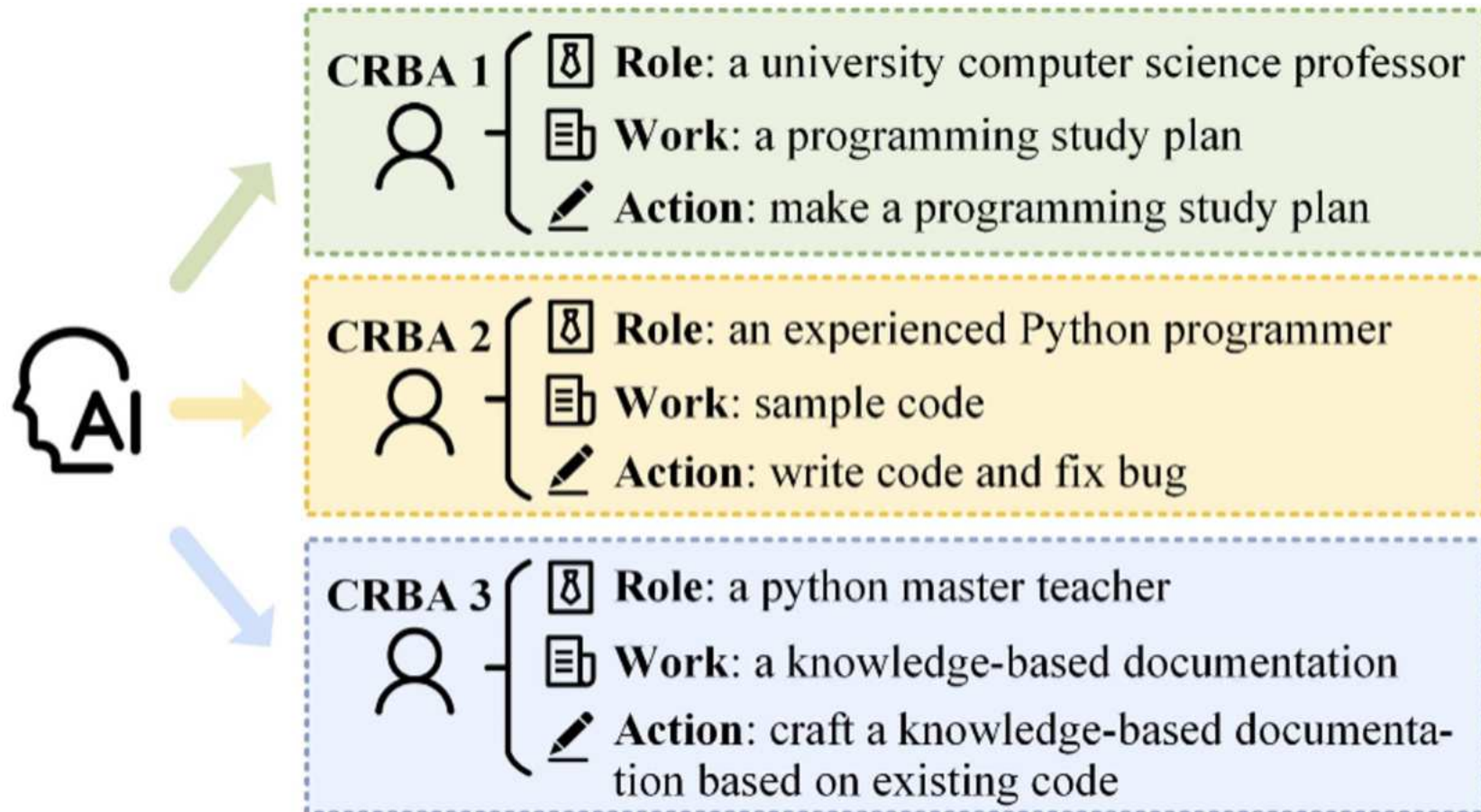


Evaluate Results by Modules

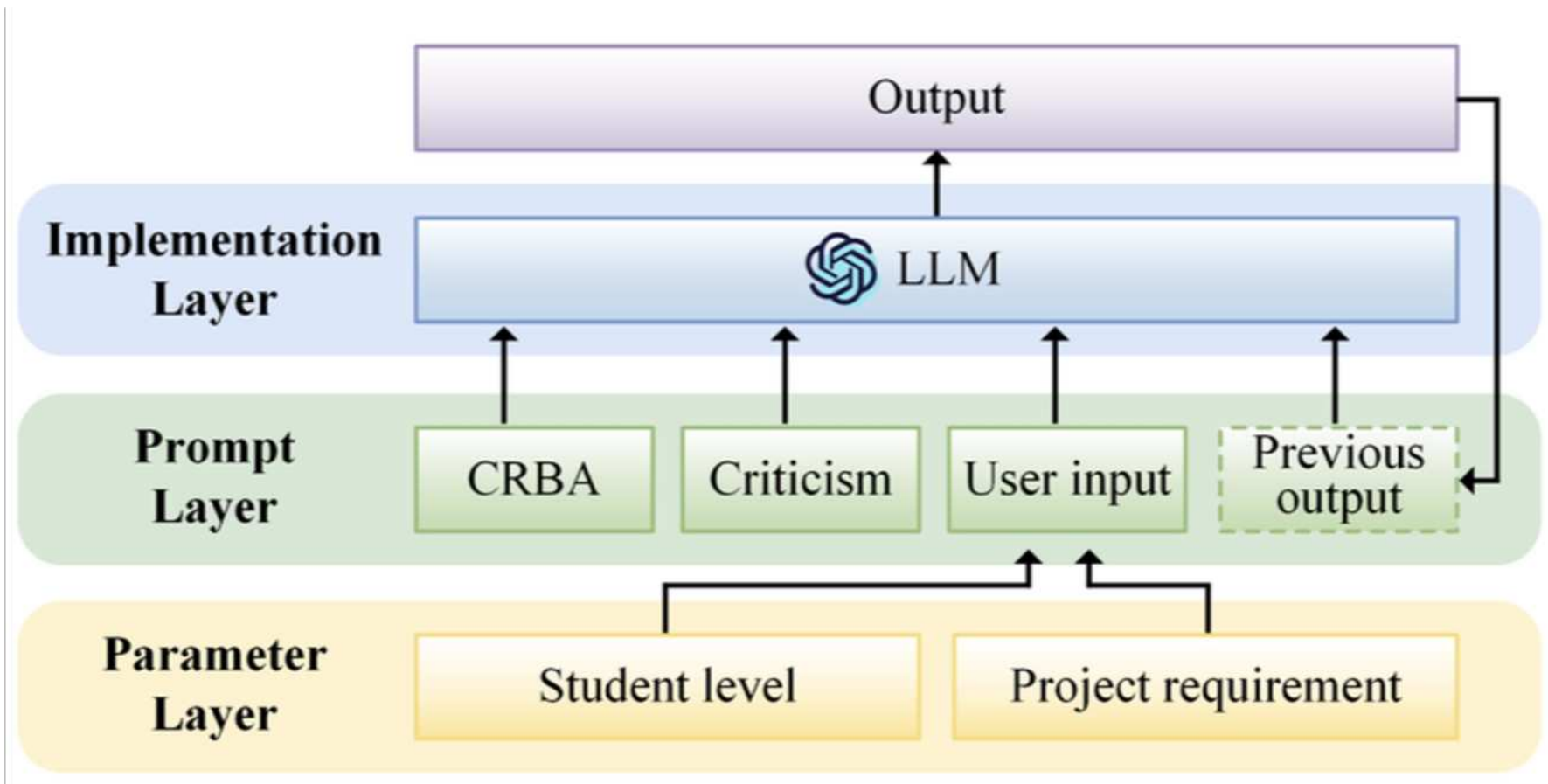


Based on plot, we can see the plan is overly good to human annotators.

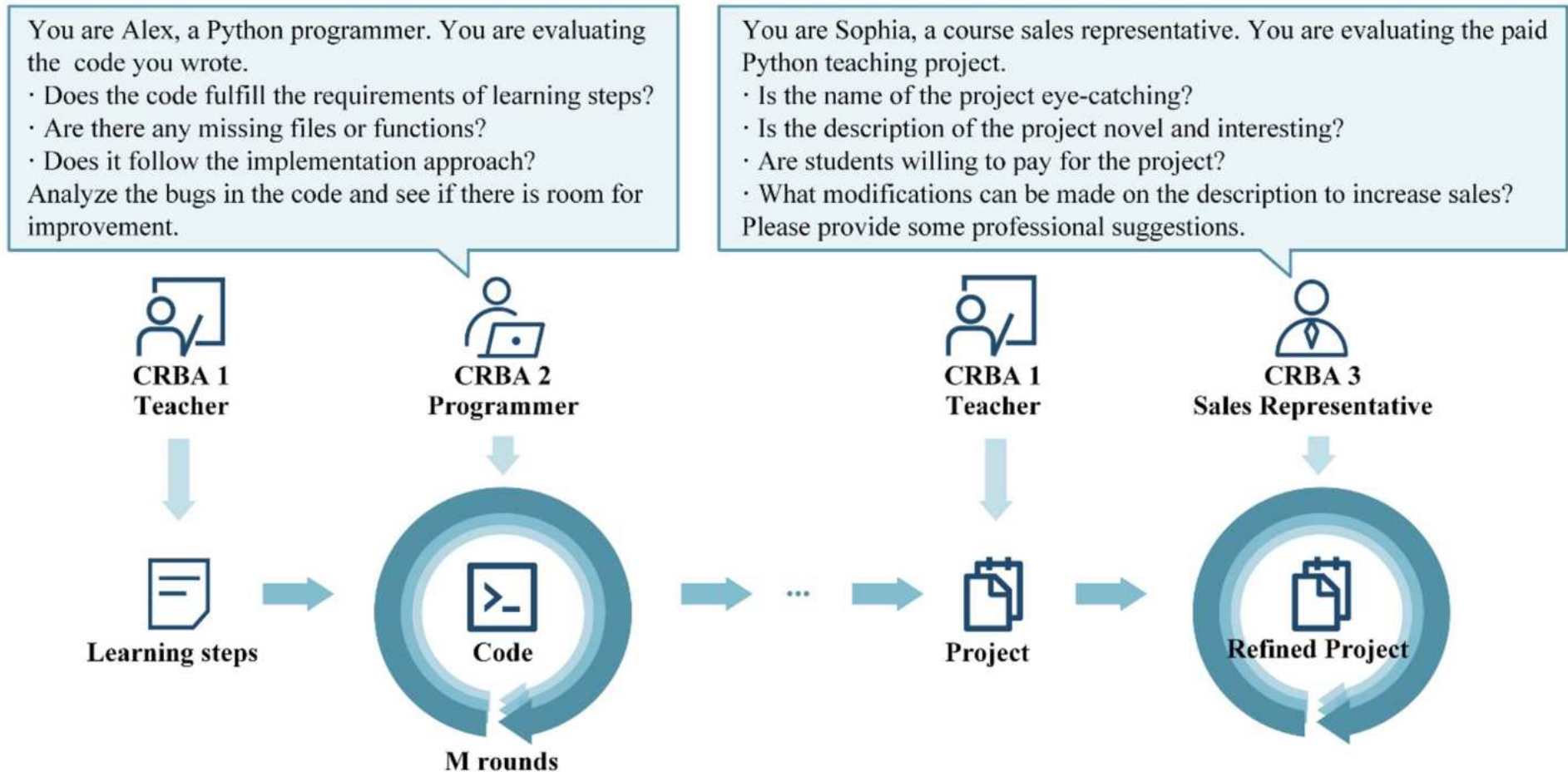
Generate Programming Project by Agent



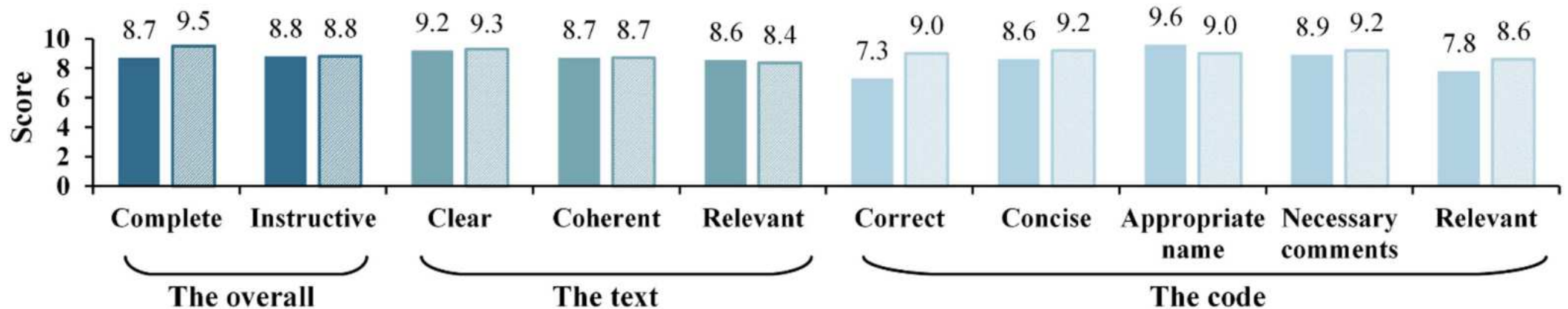
Using Criticism to Improve the Plans



Improve with Agent-based Iterative Refining



Evaluate the Quality of Generated Projects

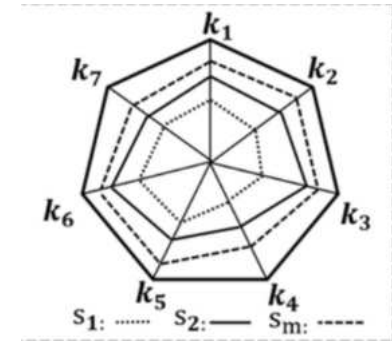


Based on plot, we can see the plan is overly good to human annotators.

Evaluate Stats Difference in Student's Responses

Question	Project	Group	Mean	Std	Shapiro–Wilk			Mann–Whitney U	Wilcoxon W	Asymp. Sig. (2-sided)
					Stat	dif	Sig.			
1	Redis	Man-made	3.200	1.095	0.916	30	0.021	329.500	794.500	0.072
		LLM	3.933	0.907	0.839	30	<0.001			
	Match 3 game	Man-made	3.767	0.817	0.861	30	0.001	442.500	907.500	0.911
		LLM	4.000	0.947	0.846	30	<0.001			
2	Redis	Man-made	3.367	0.964	0.916	30	0.002	342.000	807.000	0.082
		LLM	3.733	0.907	0.904	30	0.010			
	Match 3 game	Man-made	3.200	0.925	0.863	30	0.001	468.500	933.500	0.770
		LLM	3.533	0.819	0.830	30	<0.001			
3	Redis	Man-made	2.933	0.980	0.905	30	0.011	364.000	829.000	0.181
		LLM	3.267	0.828	0.772	30	<0.001			
	Match 3 game	Man-made	3.067	0.868	0.894	30	0.006	359.000	824.000	0.148
		LLM	2.900	1.029	0.801	30	<0.001			
4	Redis	Man-made	5.065	2.407	0.958	30	0.283	283.000	748.000	0.010
		LLM	6.448	2.131	0.935	30	0.066			
	Match 3 game	Man-made	5.484	2.234	0.954	30	0.022	378.500	843.500	0.263
		LLM	5.600	2.762	0.917	30	0.222			

Applications in Education Part I



- **Automatic Grading**

- Multi-Agent
- Human-in-the-Loop
- Retrieval Augmented Generation

- **Content Generation**

- Material Generation
- Question Generation

- **Knowledge Tracing**

- LLM as Annotator
- LLM as Predictor

Question Generation

- **Goal:**

- Generate questions that effectively reveal gaps in students' knowledge.

- **Problems:**

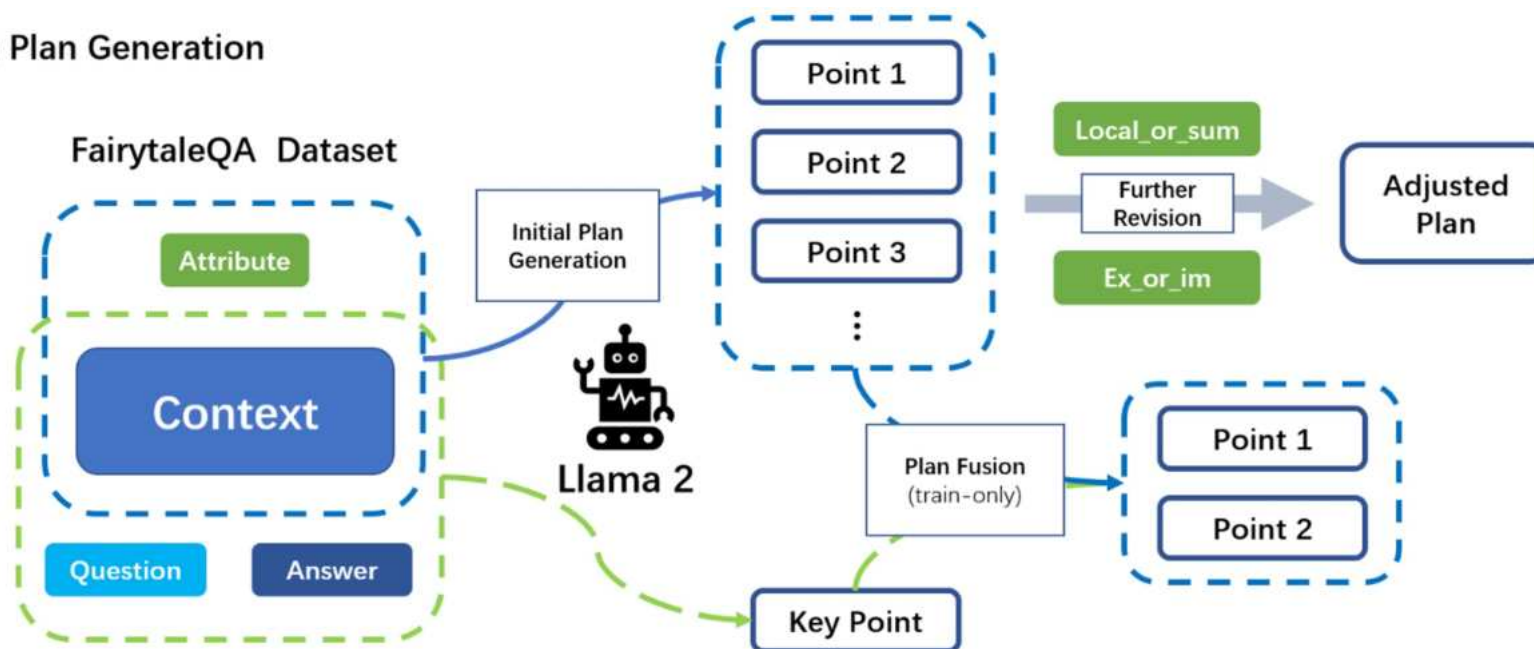
- Well-designed questions go beyond simple prompts, they require the integration of multiple areas of expertise.
- High-quality questions should provide teachers with the most direct and meaningful insights into students' learning challenges.

- **Advances:**

- LLMs can incorporate domain-specific knowledge during generation using techniques such as Retrieval-Augmented Generation (RAG).
- As flexible language models, LLMs are capable of understanding and emulating human question design strategies.

Question Generation via Explicit Plans

(a) Plan Generation



(b) Question Generation



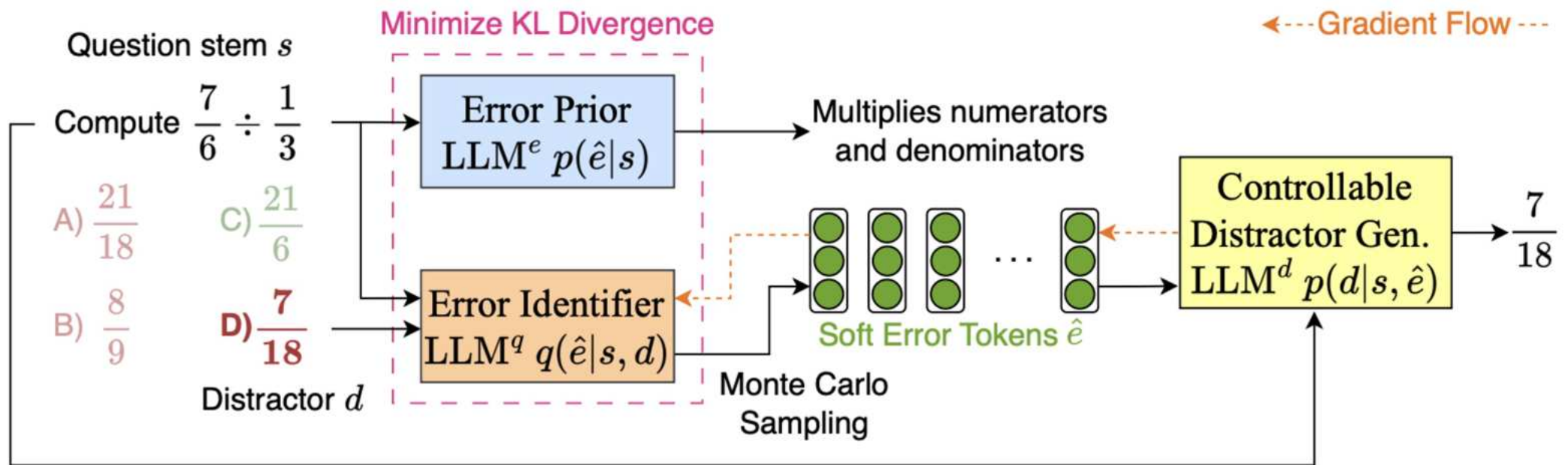
Question Generation

Method	MAP@N (Rouge-L F1)				MAP@N (BERTScore F1)			
	Top 10	Top 5	Top 3	Top 1	Top 10	Top 5	Top 3	Top 1
FQAG (Yao et al., 2022)	0.440/0.435	0.375/0.374	0.333/0.324	0.238/0.228	0.9077/0.9077	0.8990/0.8997	0.8929/0.8922	0.8768/0.8776
SQG (Dugan et al., 2022)	0.460/0.455	0.392/0.388	0.344/0.337	0.234/0.242	0.9056/0.9062	0.8953/0.8955	0.8876/0.8878	0.8707/0.8723
DQAG (Eo et al., 2023)	0.500/0.503	0.426/0.429	0.369/0.372	0.247/0.254	0.9156/ 0.9178	0.9046/0.9068	0.8956/0.8977	0.8752/0.8783
BART-large (Lewis et al., 2020)	0.375/0.353	0.354/0.332	0.337/0.314	0.298/0.276	0.8911/0.8900	0.8878/0.8866	0.8851/0.8839	0.8794/0.8784
+Prompt,Plan(PFQS, <i>ours</i>)	0.569/0.547	0.535/0.510	0.506/0.487	0.431/0.413	0.9198/0.9173	0.9144/0.9121	0.9099/0.9082	0.8988/0.8965

Ablation Study

Model	MAP@N (Rouge-L F1)					MAP@N (BERTScore F1)				
	SLMQ	Top 10	Top 5	Top 3	Top 1	SLMQ	Top 10	Top 5	Top 3	Top 1
BART-large	0.306	0.353	0.332	0.314	0.276	0.8794	0.8900	0.8866	0.8839	0.8784
+Prompt	0.385	0.446	0.431	0.420	0.394	0.8908	0.9020	0.8996	0.8980	0.8942
+Plan	0.408	0.538	0.496	0.456	0.381	0.8945	0.9157	0.9096	0.9037	0.8920
+Prompt,Plan	0.418	0.542	0.507	0.477	0.401	0.8963	0.9166	0.9113	0.9069	0.8951
+Plan(fused)	0.419	0.545	0.506	0.466	0.389	0.8968	0.9165	0.9111	0.9050	0.8932
+Prompt,Plan(fused)	0.443	0.547	0.510	0.487	0.413	0.9001	0.9173	0.9121	0.9082	0.8965

Control Generation via Latent Variable



$$p(d|s) = \sum_{e \in \mathcal{E}} p(e|s)p(d|s, e), \quad \longrightarrow \quad \begin{aligned} \log p_{\theta}(d|s) &\geq \text{ELBO}(d|s) = \mathcal{L}(\theta, \phi) \\ &= \mathbb{E}_{q_{\phi}(e|s, d)} [\log p_{\theta_d}(d|s, e)] \\ &\quad - \beta D_{\text{KL}}(q_{\phi}(e|s, d) \parallel p_{\theta_e}(e|s)), \end{aligned}$$

Control Generation via Latent Variable

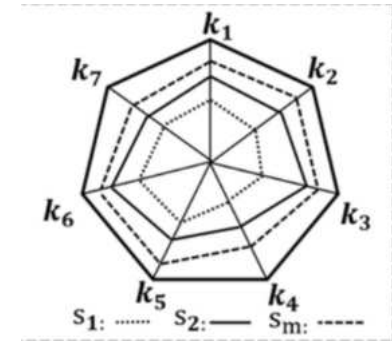
Automatic Performance Metrics

Model	K=3			K=10		
	Exact@3	Partial@3	Prop@3	Exact@10	Partial@10	Prop@10
Proprietary Base LLM, GPT-4o						
GPT-4o Zero-shot CoT (Feng et al., 2024)	6.22 $\pm$ 2.18	69.14 $\pm$ 3.97	35.27 $\pm$ 1.74	19.47 $\pm$ 3.30	78.66 $\pm$ 3.71	50.00 $\pm$ 1.96
GPT-4o kNN (Feng et al., 2024)	21.28 $\pm$ 3.43	78.42 $\pm$ 5.81	49.63 $\pm$ 3.88	33.47 $\pm$ 3.48	85.14 $\pm$ 4.77	60.19 $\pm$ 3.87
Open-source Base LLM, MetaMath-Mistral 7B						
DisSearch-D	13.74 $\pm$ 2.86	74.13 $\pm$ 4.52	41.76 $\pm$ 0.50	34.12 $\pm$ 2.33	86.17 $\pm$ 4.91	61.41 $\pm$ 2.68
DisSearch-ED CoT	<u>14.11</u> $\pm$ 1.22	73.18 $\pm$ 3.81	42.14 $\pm$ 1.52	<u>36.21</u> $\pm$ 1.21	<u>86.77</u> $\pm$ 3.80	<u>62.83</u> $\pm$ 2.52
DisSearch-ED CoT Pipeline	13.53 $\pm$ 1.74	73.63 $\pm$ 4.05	41.26 $\pm$ 1.87	32.97 $\pm$ 3.48	86.23 $\pm$ 3.88	60.42 $\pm$ 2.80
DiVERT (ours)	13.37 $\pm$ 1.70	<u>76.33</u> $\pm$ 4.33	<u>42.87</u> $\pm$ 2.45	37.00 $\pm$ 3.29	87.26 $\pm$ 4.29	63.24 $\pm$ 3.37

	Human	DiVERT	GPT-4o
Rating	3.23 $\pm$ 1.28	3.07 $\pm$ 1.39	2.56 $\pm$ 1.25

Manual Judging Metrics

Applications in Education Part I



- **Automatic Grading**

- Multi-Agent
- Human-in-the-Loop
- Retrieval Augmented Generation

- **Content Generation**

- Material Generation
- Question Generation

- **Knowledge Tracing**

- LLM as Annotator
- LLM as Predictor

Knowledge Tracing (KT)

- **Goal:**

- Evaluate the learning status of student by their answer correctness history, describe their capabilities on each knowledge concepts.

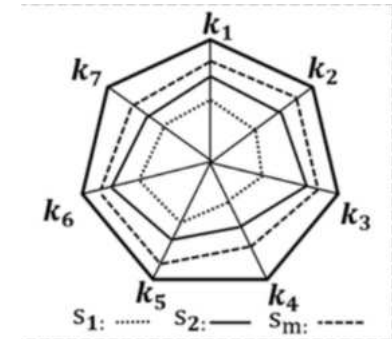
- **Problems:**

- Features of KT models are inconsistent, the usage of KT model is limited .
- Cold-start problem causes the inarracuate results for new students.

- **Advances:**

- The prior knowledge of LLM over the common knowledges can help model to get rid of the features limitation.
- The analyzing and reasoning capability of LLMs will make it a competitive predictor model.

Applications in Education Part I



- **Automatic Grading**

- Multi-Agent
- Human-in-the-Loop
- Retrieval Augmented Generation

- **Content Generation**

- Material Generation
- Question Generation

- **Knowledge Tracing**

- LLM as Annotator
- LLM as Predictor

LLM as Annotator

- **Goal:**

- Reduce the reliance on manual labor in creating features such as knowledge concepts and skill relationships, empowering the existing KT methods.

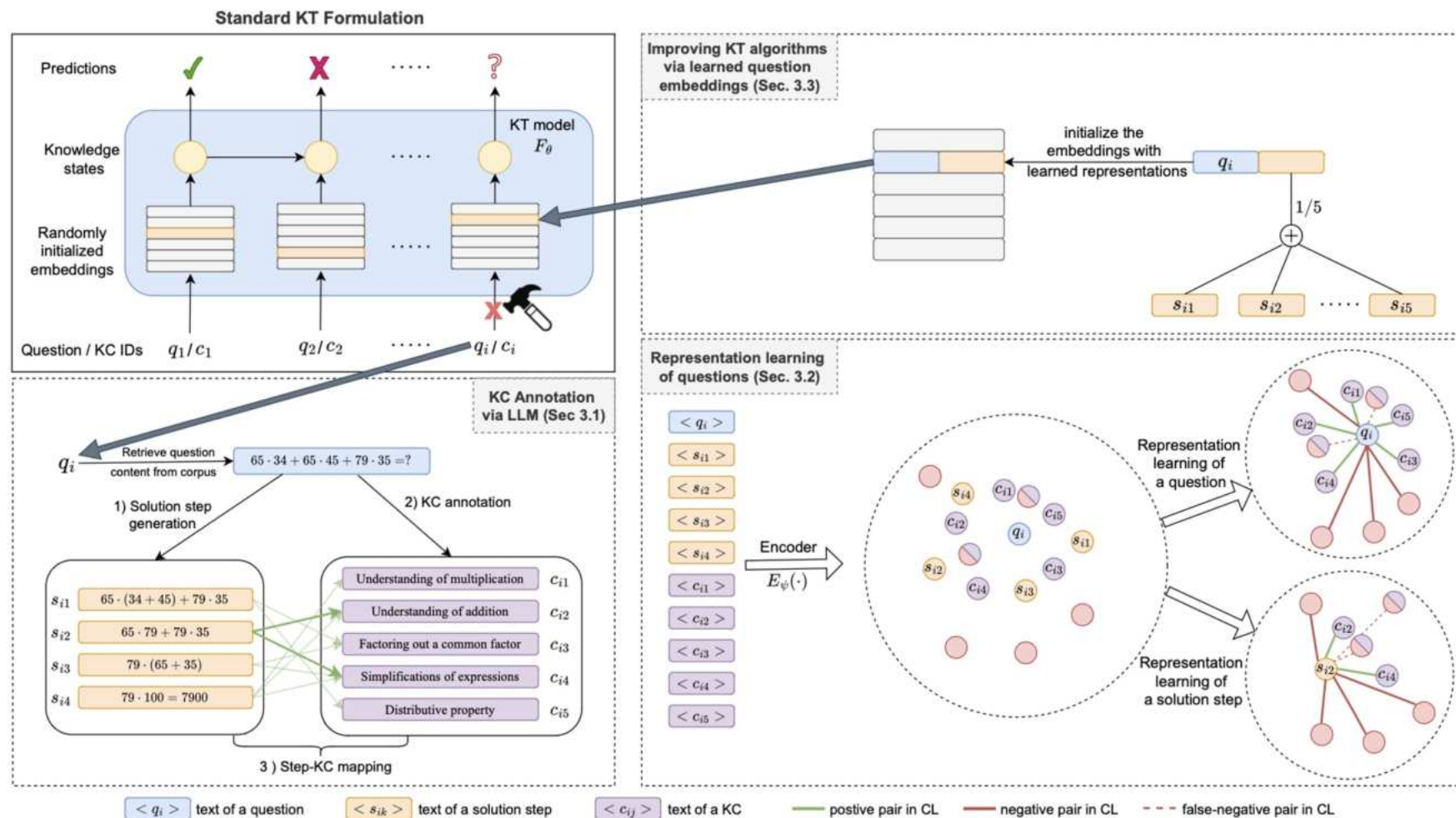
- **Problems:**

- Existing KT methods depend heavily on external annotations, limiting their practical applicability.
- existing features are often represented by simple ID values, which overlook important semantic information.

- **Advances:**

- LLMs are capable of producing human-level knowledge annotations.
- They can also comprehend and enrich language-based features directly, adding depth and meaning to the data representation.

LLM-Assisted Annotation for KT



Result by Integrate LLM into KT Methods

Model	XES3G5M			Eedi		
	Default	w/ KCQRL	Imp. (%)	Default	w/ KCQRL	Imp. (%)
DKT	78.33 $\pm$ 0.06	82.13 $\pm$ 0.02	+4.85	73.59 $\pm$ 0.01	74.97 $\pm$ 0.03	+1.88
DKT+	78.57 $\pm$ 0.05	82.34 $\pm$ 0.04	+4.80	73.79 $\pm$ 0.03	75.32 $\pm$ 0.04	+2.07

Criteria	XES3G5M						Eedi					
	Original	KCQRL w/o sol. steps	Original	KCQRL	KCQRL w/o sol. steps	KCQRL	Original	KCQRL w/o sol. steps	Original	KCQRL	KCQRL w/o sol. steps	KCQRL
Correctness	33.9	66.1	6.8	93.2	15.9	84.1	44.2	55.8	25.9	74.1	27.0	73
Coverage	41.9	58.1	13.5	86.5	13.3	86.7	25.9	74.1	7.7	92.3	22.5	77
Specificity	33.5	66.5	25.5	74.5	36.0	64.0	37.0	63.0	39.2	60.8	55.8	44
Ability of Integration	40.3	59.7	12.7	87.3	12.5	87.5	34.7	65.3	20.6	79.4	25.0	75
Overall	38.6	61.4	7.8	92.2	13.1	86.9	36.7	63.3	21.2	78.8	24.1	75

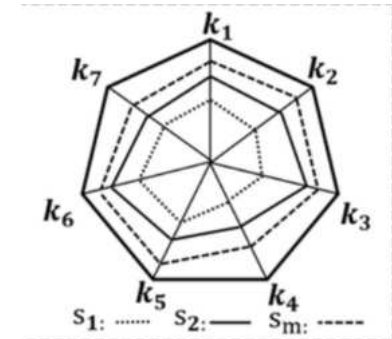
SAKT	75.90 $\pm$ 0.05	81.64 $\pm$ 0.03	+7.56	71.60 $\pm$ 0.03	74.77 $\pm$ 0.02	+4.43
SAINT	79.65 $\pm$ 0.02	81.50 $\pm$ 0.07	+2.32	73.96 $\pm$ 0.02	75.20 $\pm$ 0.04	+1.68
AKT	81.67 $\pm$ 0.03	83.04 $\pm$ 0.05	+1.68	74.27 $\pm$ 0.03	75.49 $\pm$ 0.03	+1.64
simpleKT	81.05 $\pm$ 0.06	82.92 $\pm$ 0.04	+2.31	73.90 $\pm$ 0.04	75.46 $\pm$ 0.02	+2.11
sparseKT	79.65 $\pm$ 0.11	82.95 $\pm$ 0.09	+4.14	74.98 $\pm$ 0.09	78.96 $\pm$ 0.08	+5.31

Best values are in bold. The shading in green shows the magnitude of the performance gain.

Result by Integrate LLM into KT Methods

Criteria	XES3G5M					
	Original	KCQRL w/o sol. steps	Original	KCQRL	KCQRL w/o sol. steps	KCQRL
Correctness	33.9	66.1	6.8	93.2	15.9	84.1
Coverage	41.9	58.1	13.5	86.5	13.3	86.7
Specificity	33.5	66.5	25.5	74.5	36.0	64.0
Ability of Integration	40.3	59.7	12.7	87.3	12.5	87.5
Overall	38.6	61.4	7.8	92.2	13.1	86.9
Pairwise Win Rate	Eedi					
	Original	KCQRL w/o sol. steps	Original	KCQRL	KCQRL w/o sol. steps	KCQRL
	44.2	55.8	25.9	74.1	27.0	73.0
	25.9	74.1	7.7	92.3	22.5	77.5
	37.0	63.0	39.2	60.8	55.8	44.2
	34.7	65.3	20.6	79.4	25.0	75.0
	36.7	63.3	21.2	78.8	24.1	75.9

Applications in Education Part I



- **Automatic Grading**

- Multi-Agent
- Human-in-the-Loop
- Retrieval Augmented Generation

- **Content Generation**

- Material Generation
- Question Generation

- **Knowledge Tracing**

- LLM as Annotator
- LLM as Predictor

LLM as Predictor

- **Goal:**

- Leverage LLMs to replace traditional knowledge tracing (KT) methods and address limitations such as the cold-start problem.

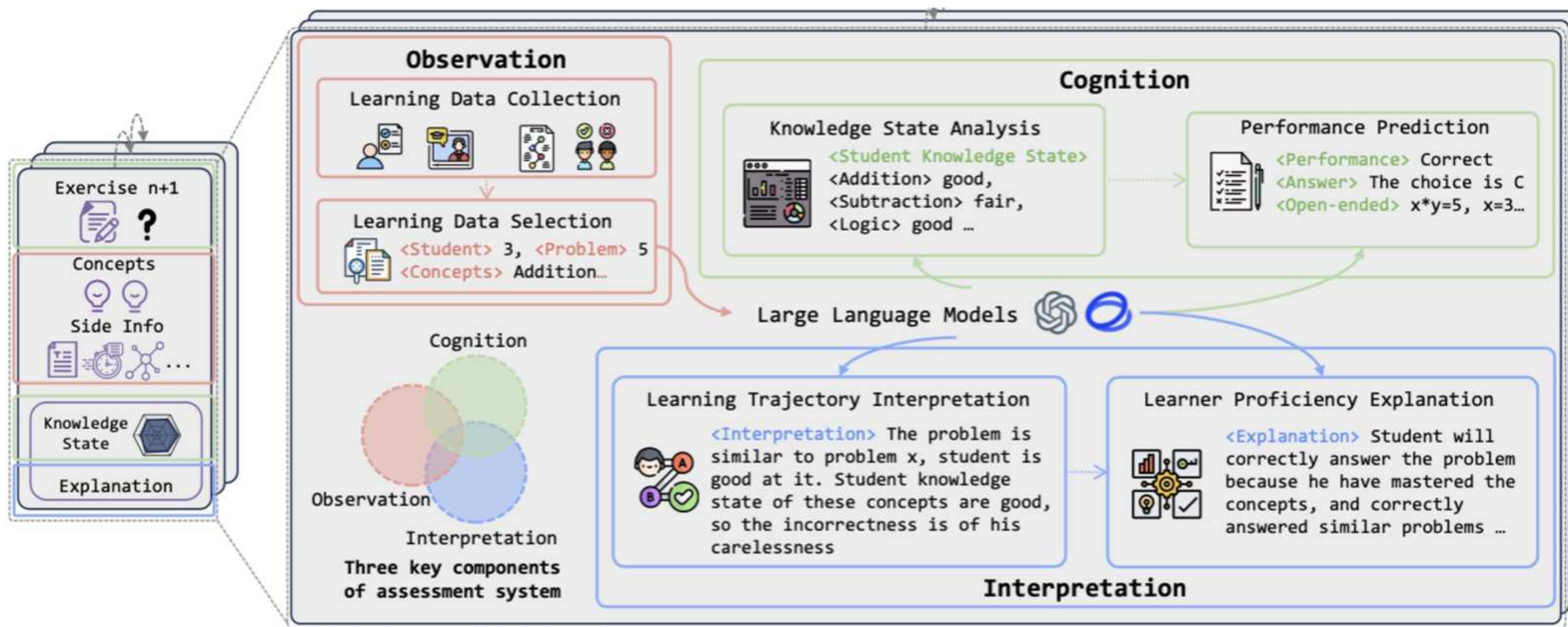
- **Problems:**

- Conventional KT methods struggle with students with limited learning history.
- Additional tools or techniques are required to adapt the most KT framework to open ended questions and interpret prediction results effectively.

- **Advances:**

- LLMs, with their few-shot learning and contextual reasoning capabilities, can generate accurate predictions even with limited data.
- Moreover, their ability to produce human-readable reasoning steps provides a transparent and interpretable prediction process.

Explainable Few-shot Knowledge Tracing



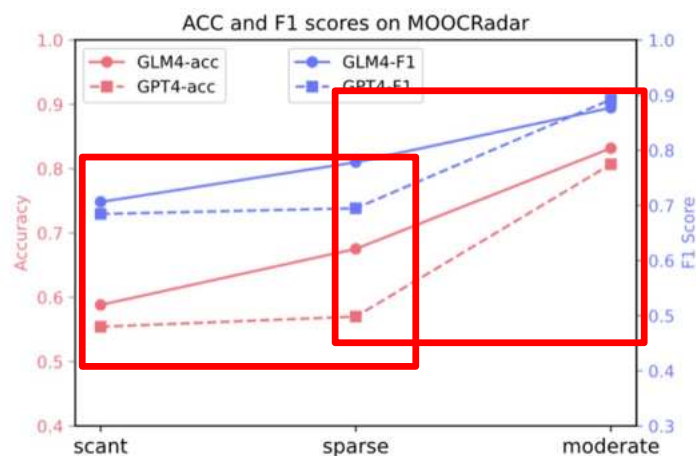
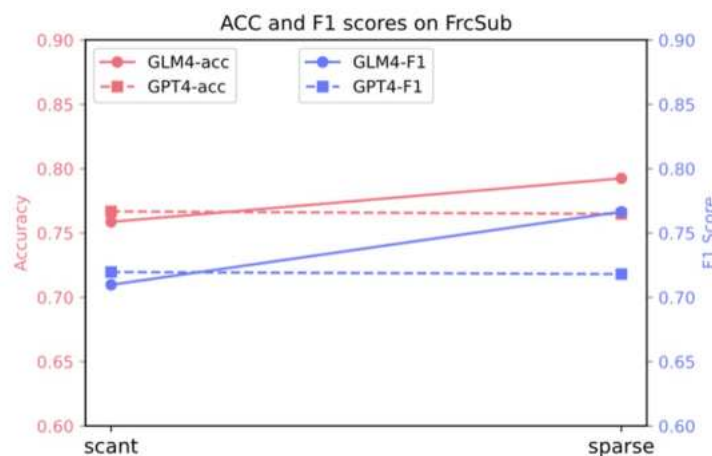
Explainable Few-shot Knowledge Tracing

Input scale	Baselines	Dataset					
		FrcSub		MOOCRadar		XES3G5M	
		Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score
full-set	DKT	0.7481	0.7514	0.8210	0.8882	0.8355	0.9017
	DKVMN	0.7909	<u>0.8077</u>	0.8147	0.8836	0.8372	<i>0.9037</i>
	GKT*	0.5480	0.3043	0.7991	0.8772	0.8169	0.8923
	AKT	0.7747	0.7869	0.8194	0.8870	<u>0.8435</u>	0.9063
	SAKT	0.7476	0.7389	0.7956	0.8706	0.8298	0.8990
	SAINT	0.8061	0.8201	<i>0.8241</i>	<i>0.8904</i>	<i>0.8399</i>	<u>0.9044</u>
few-shots	GLM3-6b	0.6571	0.6496	0.5378	0.6753	0.5434	0.6580
	GLM4	<i>0.7939</i>	<i>0.7889</i>	0.8489	0.9052	0.8491	0.8978
	GPT-4	<u>0.7968</u>	0.7471	<u>0.8246</u>	<u>0.9029</u>	0.8176	0.8714

Explainable Few-shot Knowledge Tracing

GLM4	FrcSub-sparse		XES3G5M-sparse		XES3G5M-moderate	
	Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score
4	0.7192 ^{+0.0}	0.7086 ^{+0.0}	0.4399 ^{+0.0}	0.4707 ^{+0.0}	0.6672 ^{+0.0}	0.7592 ^{+0.0}
8	0.7771 ^{+8.1}	0.7568 ^{+6.8}	0.7057 ^{+60.4}	0.7752 ^{+64.7}	0.7928 ^{+18.8}	0.8623 ^{+13.6}
16	0.7939 ^{+10.4}	0.7889 ^{+11.3}	0.7542 ^{+71.4}	0.8395 ^{+78.4}	0.8491 ^{+27.3}	0.8978 ^{+18.3}

Student_id: 8087	Scant
Exercise_id: 20	
is_correct: right	Sparse
Knowledge concept ids: ["105", "39", "106"]	
+Knowledge concepts: ['proposition', 'interrogative sentence', 'propositional form']	Moderate
++Exercise content: Statements can be expressed by propositions...	



Explainable Few-shot Knowledge Tracing

GLM4	GPT4
<pre>Exercise_id: 20 ... Exercise_id: 24 Exercise content: The propositions whose truth values can be determined independently by logic itself are: ... Knowledge concepts: ['truth value of proposition', 'compound proposition', 'logic'] is_correct: right Explanation: 1. The new exercise contains the knowledge points: ['proposition', ...]. 2. The new question is similar to question 2 in the student's record of work, as it also involves the knowledge points 'proposition' and 'interrogative sentence'. 3. Student's Knowledge state: - 'proposition', good (from previous exercise 2) ... 4. The student gets it right, possibly because correct mastery. Exercise_id: 29 ... <Exercise to Predict> Exercise content: Which of the following is not a partition? 'A': ... Knowledge concepts: ["predicate", "basic proposition", "partition"] <Output Predicted is_correct> Explanation: 1. The new exercise contains the knowledge points: ['predicate', ...]. 2. Similar to question 24, the student has encountered the knowledge points 'basic proposition' and 'partition' before. 3. Student's Knowledge state: - 'proposition', good (from previous exercise 24) - 'interrogative sentence', fair (from previous exercise 24) ... 4. The student gets it right, possibly because correct mastery.</pre>	<pre>Exercise_id: 20 ... Exercise_id: 24 Exercise content: The propositions whose truth values can be determined independently by logic itself are: ... Knowledge concepts: ['truth value of proposition', 'compound proposition', 'logic'] is_correct: right Explanation: 1. The new exercise contains 'proposition', ... 2. It's a new question, there is some kind of connection between the previous knowledge points and questions. The knowledge point 'proposition' was also present in the previous question. 3. Student's Knowledge state: 'proposition', fair ... 4. The student gets it wrong, possibly due to incorrect mastery. The student may not have fully understood the knowledge points... Exercise_id: 29 ... <Exercise to Predict> Exercise content: Which of the following is not a partition? 'A': ... Knowledge concepts: ["predicate", "basic proposition", "partition"] <Output Predicted is_correct> Explanation: 1. The new exercise contains 'predicate',... 2. It's a new question, there is some kind of connection between the previous knowledge points and questions. The knowledge point 'basic proposition' was also present in the previous question. 3. Student's Knowledge state: 'proposition', fair, 'interrogative sentence', fair, ... 4. The student gets it right, possibly due to correct mastery. The student may have understood the knowledge points involved in the question, leading to the correct answer.</pre>

Tutorial Outline

Morning

Introduction
to GenAI
(40 Min)



Overview of GenAI
in Education
(40 Min)



Applications in
Education Part I
(40 Min)



Afternoon

Applications in
Education Part II
(40 Min)



Demo
Presentation
(40 Min)



Future Direction
and QA
(40 Min)

Applications in Education Part II



- **Education Assistant**

- Question-Answer
- Assessment
- Tutoring

- **Role-play Simulation**

- Student Simulation
- Classroom Simulation
- Role-play Interaction

- **Profession Develop**

- Exam Evaluation
- Studying Assistant
- Outcome Study

Education Assistant

- **Goals:**

- Support teachers, students, and other human participants (parents, TAs, etc.)
- Helping with learning, tutoring, and administrative support

- **Problems:**

- Lack of personalization in traditional tools.
- Students have limited access to timely help outside classroom hours.

- **Advances**

- Adapt to individual learner.
- Rich tutoring experiences with interactive dialogues.
- Scale efficiently to support large numbers of students.

Applications in Education Part II



- **Education Assistant**

- Question-Answer
- Assessment
- Tutoring

- **Role-play Simulation**

- Student Simulation
- Classroom Simulation
- Role-play Interaction

- **Profession Development**

- Exam Evaluation
- Studying Assistant
- Outcome Study

Question-Answer Assistant

- **Goals:**

- Answer questions based on input text and knowledge
- Involve understanding, retrieving, and reasoning

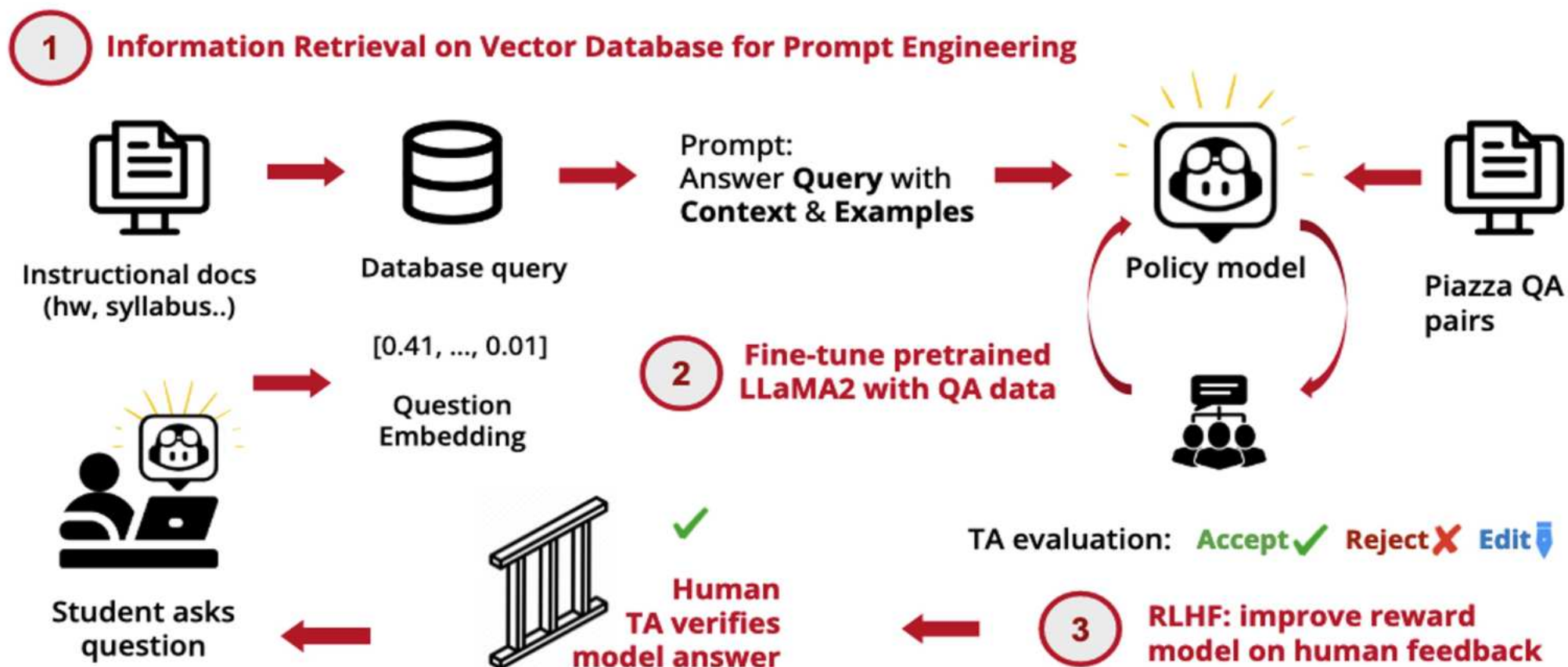
- **Problems**

- Manual responses: time-consuming and not scalable
- Forum-based Q&A: lacks quality, speed, and engagement
- General LLMs: may produce hallucinations or irrelevant answers

- **Advances**

- For student: accurate, timely, context-aware, and engaging
- For teachers: pedagogically meaningful, aligned with instructional goals

Architecture of Question-Answer Assistant



Example: Logistical question answering

- Student Question:
 - *“Do we get penalties when the code is quite long?”*
- Human Answer:
 - *“Best is to try to simplify as much as possible. If it is too long, it is likely not efficient and can be improved upon.”*
- LLMs Answer:
 - *“**Hello!** About your concern with penalties for long codes—I’m happy to clarify that there are no penalties in this course. That said, long code can be harder to read and maintain, so it’s a good idea to aim for concise, clear solutions. This helps keep your code manageable and lowers the chance of bugs or mistakes. So, focus on writing clean, efficient, and well-structured code! If you have any other questions or concerns, feel free to ask—I’m always here to help.”*

Conclusions of Q-A Assistant

- LLMs answers are generally helpful
 - Over 81% usefulness and 72% accuracy at best
 - Sometimes surpass human with comprehensiveness and positive sentiment
- Depending on Base LLMs
 - GPT-4 has the best performance yet privacy concerns
 - Open-source models are secure, while more augmenting techniques (RAG, SFT, DPO, etc.)

Applications in Education Part II



- **Education Assistant**

- Question-Answer
- **Assessment**
- Tutoring

- **Role-play Simulation**

- Student Simulation
- Classroom Simulation
- Role-play Interaction

- **Profession Development**

- Exam Evaluation
- Studying Assistant
- Outcome Study

Assessment Assistant

- **Goals:**

- Evaluate student learning by automatically generated responses.
- Support both formative and summative assessment types.
- Provide personalized and actionable feedback to guide learning.
- Help teachers track progress and make instructional decisions.

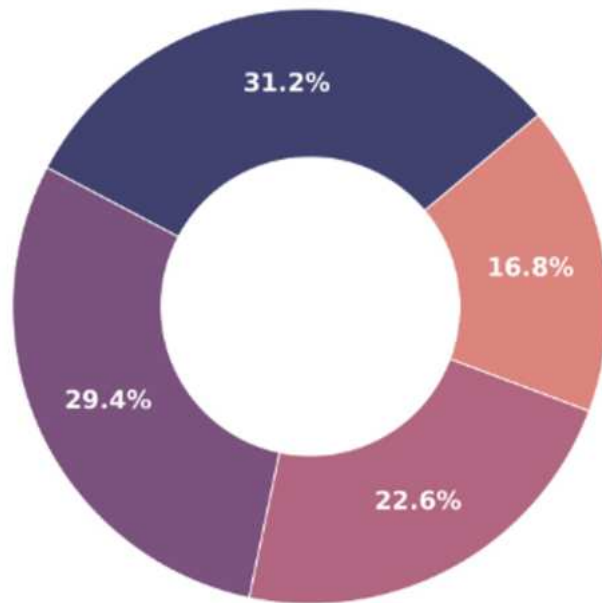
- **Problems:**

- Relies heavily on human effort, limiting scalability and efficiency.
- Manual feedback is slow and untimely.
- Lack personalization and consistent quality.

- **Advances:**

- Generate high-quality, adaptive feedback with minimal effort.
- Ensure consistent, scalable, and fair student evaluation.

Building Assessment Assistant



■ Essay Assessment
■ Emotional Support
■ Retrieval-Augmented Open QA
■ Socratic Teaching

Data:

Textbooks

Fundamental
Instruction

Educational
Instruction



Tools

- Web Search
- Calculator
- Self-Check



Skill

- Psychology
- Pedagogical skills
- Socrates teaching



Assessment
Assistant

Example: Multi-grained Essay Assessment

User:

First Encounter with Fireflies

Today, my mother said she would take me to see the fireflies at ...

This firefly-watching experience made my night wonderful ...

Please kindly rate the composition above.

Assessment:

- **Content Evaluation:** revolves around the theme ...
- **Expression Evaluation:** The quality of this article is high, ...
- **Highlight Sentence:** *Cicadas, frogs, and cuckoos formed a choir of summer ...*
- **Overall Evaluation:**
 - **(Summary)** The essay revolves around ...
 - **(Suggestion)** In writing, more attention should be paid to ...

Fine-grained,
Metric-based
Assessment

Task-specific
Assessment

Holistic, Text-based
Assessment

Findings

- Psychology and education theories are helpful in assistant building
 - Socratic Teaching, Rational Emotive Behavior Therapy (REBT), ABC theory (Ellis, 1991), etc.
- LLM-based assessment is competent in STEM as well as Social Science, Humanities and other professional majors.

Applications in Education Part II



- **Education Assistant**

- Question-Answer
- Assessment
- **Tutoring**

- **Role-play Simulation**

- Student Simulation
- Classroom Simulation
- Role-play Interaction

- **Profession Develop**

- Exam Evaluation
- Studying Assistant
- Outcome Study

Tutoring Assistant

- **Goals:**

- Guide students with adaptive hints, explanations, and examples.
- Provide personalized, instructionally aligned support for learning.

- **Problems:**

- Human tutors are limited in availability, consistency, and scalability.
- Existing systems offer static, non-personalized, and impersonal feedback.

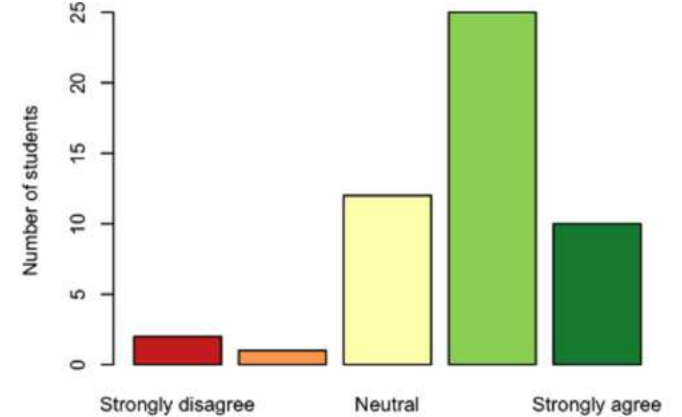
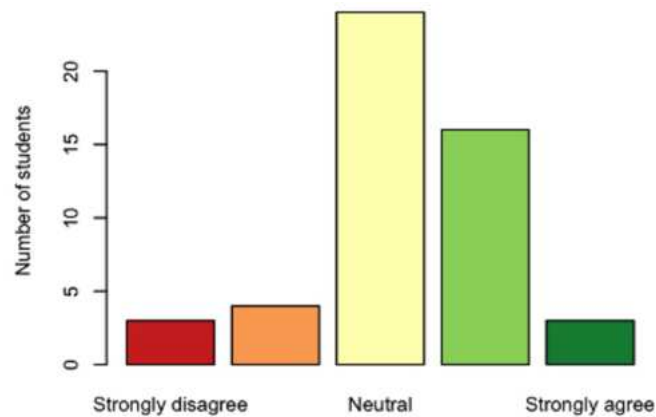
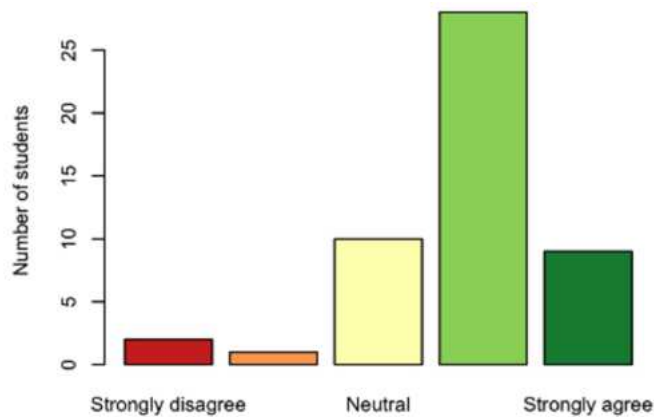
- **Advances:**

- LLMs enable dynamic, personalized tutoring through dialogue.
- Scales high-quality tutoring support across diverse learners.

Build Tutoring Assistant by Prompts

- Explaining solutions
 - You should help them understand the concept and solve the problem using explanations, examples, and analogies...
 - When you provide an equation, explain why it works and what it means ...
- Offering hints
 - You should ask them to first identify all values in the equation...
 - You should ask questions whenever you explain a new concept to check understanding...
- Providing iterative teaching
 - You should end your responses with questions to check their understanding...

Performance and Satisfaction of *NewtBot*



⇐ Strongly Disagree Disagree Neutral Agree Strongly Agree ⇒

Accomplish
academic tasks
more quickly?

Increase chances
of getting better
grades?

I would use
NewtBot in the
future

Applications in Education Part II



- **Education Assistant**

- Question-Answer
- Assessment
- Tutoring

- **Role-play Simulation**

- Student Simulation
- Classroom Simulation
- Role-play Interaction

- **Profession Development**

- Exam Evaluation
- Studying Assistant
- Outcome Study

Role-play simulation

- **Goals:**

- Simulate human-like agents with beliefs, emotions, and goals.
- Model dynamic student-agent interactions to explore teaching strategies.
- Practice or predict the potential scenarios

- **Problems:**

- Manual role-play is labor-intensive and hard to scale
- Inconsistent pedagogical quality across simulations
- Ethical Concerns: unproven or potentially harmful teaching settings on human students

- **Advances**

- Enable safe, scalable testing of diverse interventions and behaviors.
- Create engaging, personalized role-based learning experiences.

Applications in Education Part II



- **Education Assistant**

- Question-Answer
- Assessment
- Tutoring

- **Role-play Simulation**

- Student Simulation
- Classroom Simulation
- Role-play Interaction

- **Profession Develop**

- Exam Evaluation
- Studying Assistant
- Outcome Study

TeachTune: Simulation of Students

- **Goals:**

- Simulate dynamic interactions between pedagogical agents and diverse student personas
- Explore and compare pedagogical strategies under varied scenarios;

- **Problems:**

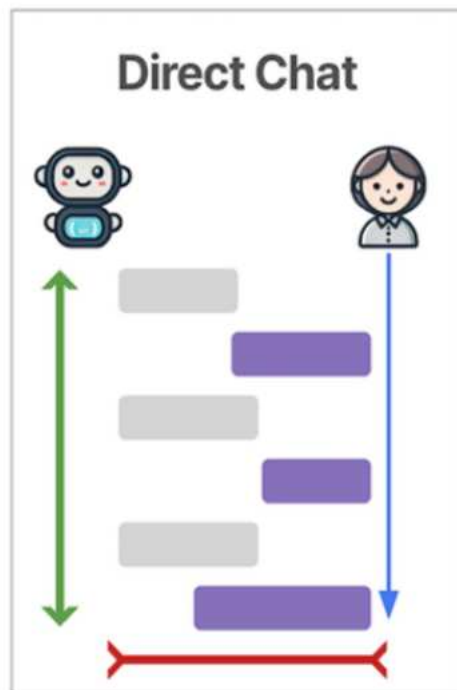
- Manual role-play is labor-intensive and hard to scale
- Inconsistent pedagogical quality across simulations

- **Advances:**

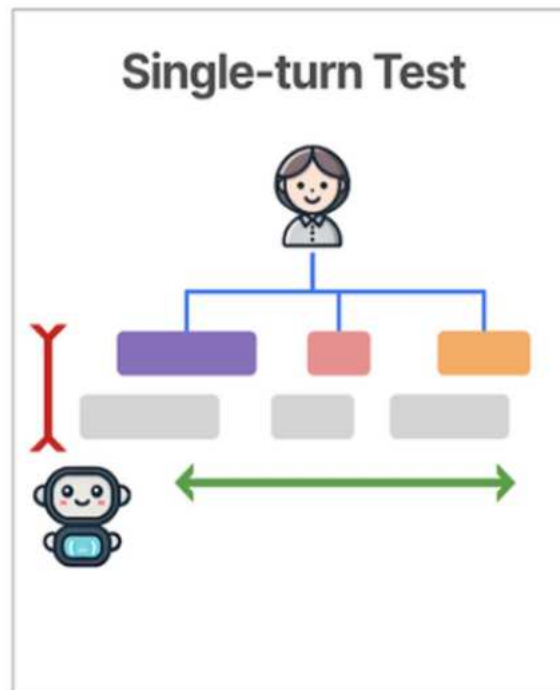
- Injects student profiles and materials into LLMs for realism.
- Enables both broad and detailed assessments of teaching quality.
- Helps teachers evaluate interaction effectiveness across strategies.

Types of Simulated Interactions

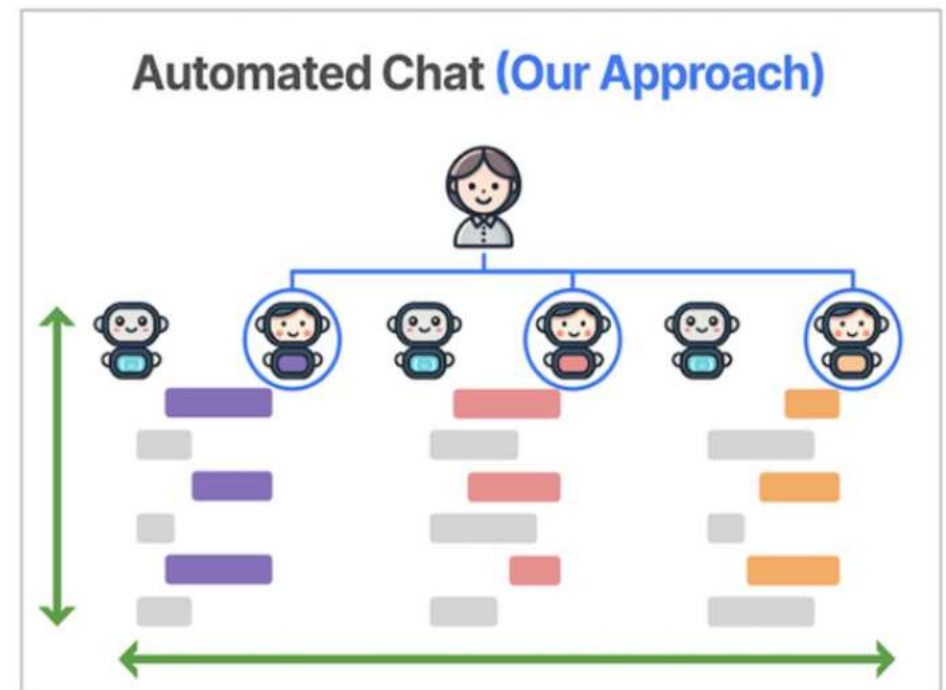
 Instructor  Pedagogical Conversational Agent  LLM-based Simulated Student



Depth ○
Breadth ✗



Depth ✗
Breadth ○



Depth ○
Breadth ○

Predict Student Behaviors with TeachTune

 : Hi! Do you have any questions about phase transition?

 : Hi, I didn't get what phase transition is

 : No problem! Phase transition is when a substance changes from one state of matter to another.

There are four main states of matter... that's a phase transition.

Phase transitions are caused by ... Does it help clarify things a bit?

 : Do you have examples for phase transitions?

 : ...

Conclusions

- *TeachTune* provides a structured way to evaluate pedagogical conversational strategies;
- Simulated Students are aligned to the human counterparts
 - ~ 5% error in knowledge behavior alignment
 - ~ 10% error in trait behavior alignment
- Helpfulness to Teachers
 - Lower task load both physically and temporally
 - Predicting a wider range of students
 - Helping in finding edge cases and improving robustness of pedagogical strategies

Applications in Education Part II



- **Education Assistant**

- Question-Answer
- Assessment
- Tutoring

- **Role-play Simulation**

- Student Simulation
- Classroom Simulation
- Role-play Interaction

- **Profession Develop**

- Exam Evaluation
- Studying Assistant
- Outcome Study

SimClass: Simulation of Classroom

- **Goals:**

- Simulate dynamic classroom interactions using multi-agent LLM
- Model collective learning behaviors in classroom communities

- **Problems:**

- Human-based simulations: time-consuming and hard to replicate
- Difficult to explore rare or edge-case scenarios in controlled settings
- Ethical concerns: risks to student well-being and classroom equity from experimenting with social manipulation

- **Advances:**

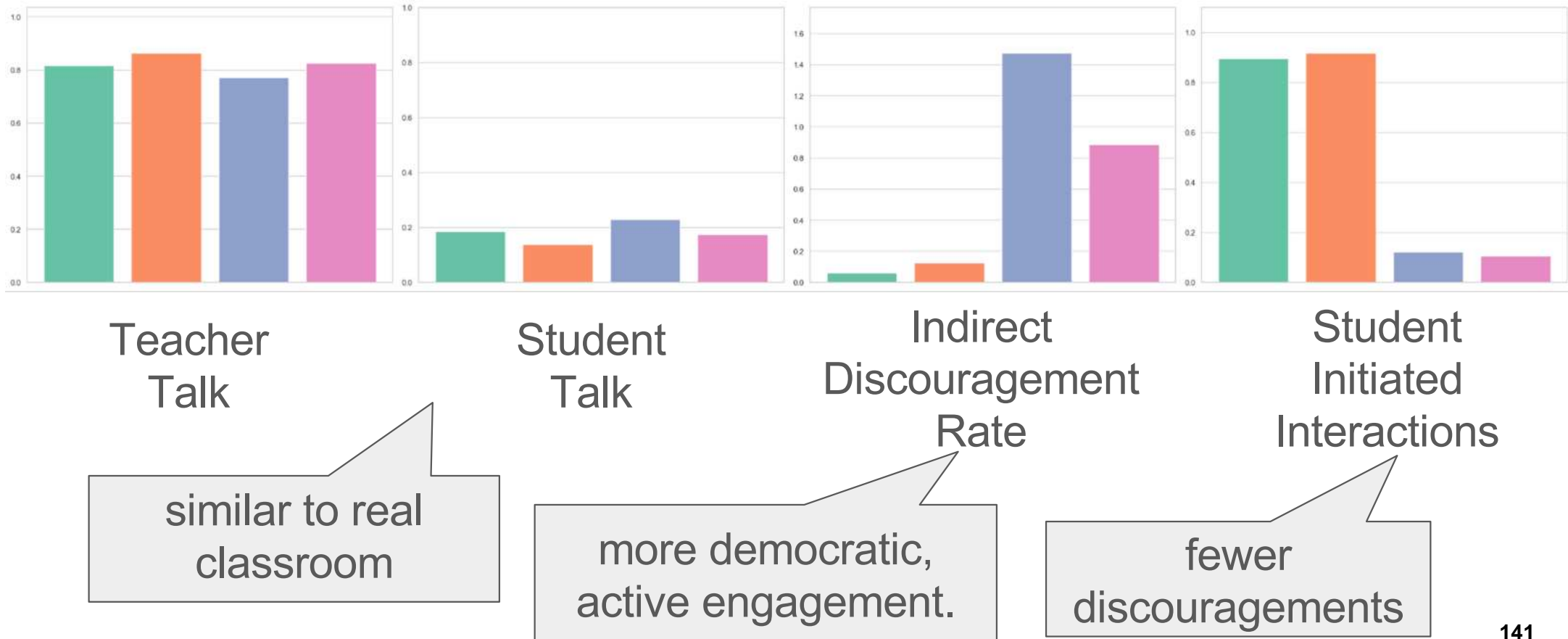
- Scalable, repeatable simulations of diverse classroom settings
- Safe testing of interventions without ethical risks.

Roles in the Virtual Classroom of SimClass

Teacher 	present material; answer questions based on discussions.
Assistant 	supplement teaching information; participate in discussion; maintain discipline.
Class Clown 	initiate ideas; helps teachers to steer class flow.
Deep Thinker 	promote deep thinking; raise challenging topics.
Note Taker 	summarize notes of lessons; help to organize ideas.
Inquisitive Mind 	pose questions about lectures; stimulate others' thinking and discussion.

Performance of *SimClass* Simulation

'SimClass-TAGI' 'SimClass-HSU' v.s. 'Expert Teacher' 'Novice Teacher'



Applications in Education Part II



- **Education Assistant**

- Question-Answer
- Assessment
- Tutoring

- **Role-play Simulation**

- Student Simulation
- Classroom Simulation
- **Role-play Interaction**

- **Profession Develop**

- Exam Evaluation
- Studying Assistant
- Outcome Study

Role-Play Interaction

- **Goals:**

- Simulates realistic patient interactions for clinical students.
- Maintain dialogue and provide automated assessments.

- **Problems:**

- Human patients: costly, labor-intensive, and limited in diversity
- In-person simulations: lack scalability and consistent feedback
- Ethical concerns:
 - Cause unintended harm to patients
 - Undermine student confidence

- **Advances:**

- Repeated, low-cost practice of real-world scenarios safely
- Build communication, decision-making, and empathy together

Role-based Interaction of CureFun

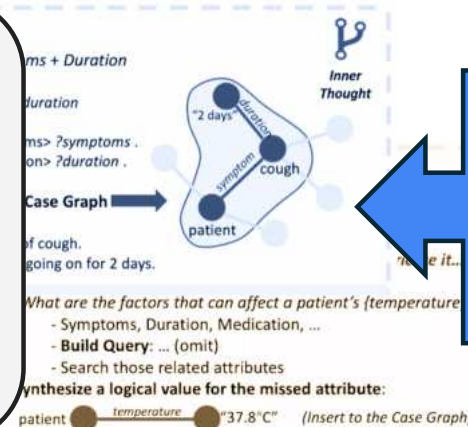
So, what kind of symptoms have you been having and how long have you been experiencing them?



Human
Student

Inner Thoughts:

- Patient profile and demographics
- Medical facts and attributes
- Conceptual relationships
- Dialogue and interaction history...



Thinking with
background
+ knowledge

I've been coughing for two days, and it's really uncomfortable. 🥱



LLM
Patient

CureFun: Similar Scenarios with Better Experience

Model	Information Density	Emotional Tendency	Response Length	Turn Number	Overall Score
Llama2-70B	0.02	0.93	2996.17	5.39	0.34
Mixtral-8x7B	0.11	0.69	647.50	5.88	0.33
BianQue-2 (6B)	0.14	0.69	178.00	4.25	0.25
DISC-MedLLM (13B)	0.15	0.66	361.87	4.57	0.43
ERNIE-4-Bot	0.13	0.69	662.09	5.70	0.37
ChatGPT (3.5-turbo)	0.15	0.70	371.74	7.74	0.51
Human (Non-medical background)	0.15	0.71	125.00	8.45	0.45
Human (Clinical medicine student)	0.19	0.48	212.99	23.11	0.72
Human (Expert)	0.27	0.56	135.75	13.38	0.78

- Friendliness: LLMs consistently perform well in emotional tendency
- Details: LLMs tend to generate longer responses
- Efficiency: LLMs tend to end the conversation in fewer rounds

Applications in Education Part II



- **Education Assistant**

- Question-Answer
- Assessment
- Tutoring

- **Role-play Simulation**

- Student Simulation
- Classroom Simulation
- Role-play Interaction

- **Profession Development**

- Exam Evaluation
- Studying Assistant
- Outcome Study

Profession Develop

- **Goals**

- Assist learners in developing skills and preparing them for real-world practice.

- **Problems**

- Professional education emphasizes practical experience, yet students often rely solely on textbooks for learning.
- Challenging to understand the content and develop effective learning tools.
- Relying exclusively on textbooks limits the adult learners' opportunities for discussion and inquiry

- **Advances**

- Easy adopt domain-specific knowledges via various algorithms.
- Support practical training with simulated the capability.
- Engaging Interactive experiences for learners.

Applications in Education Part II



- **Education Assistant**

- Question-Answer
- Assessment
- Tutoring

- **Role-play Simulation**

- Student Simulation
- Classroom Simulation
- Role-play Interaction

- **Profession Develop**

- Exam Evaluation
- Studying Assistant
- Outcome Study

Exam Evaluation

- **Goals:**

- Evaluate the model's familiarity with domain-specific content.

- **Problems:**

- The evaluation must be objective and comprehensive.
- The results should serve as a valid reference for future use within the domain.

- **Advances:**

- LLMs can comprehend and respond to domain-specific questions in a human-like manner, supporting the need for tailored designs.
- Their ability to generate human-readable answers and explain reasoning allows users to thoroughly assess the model's domain-specific capabilities.

Methods in Examining LLMs

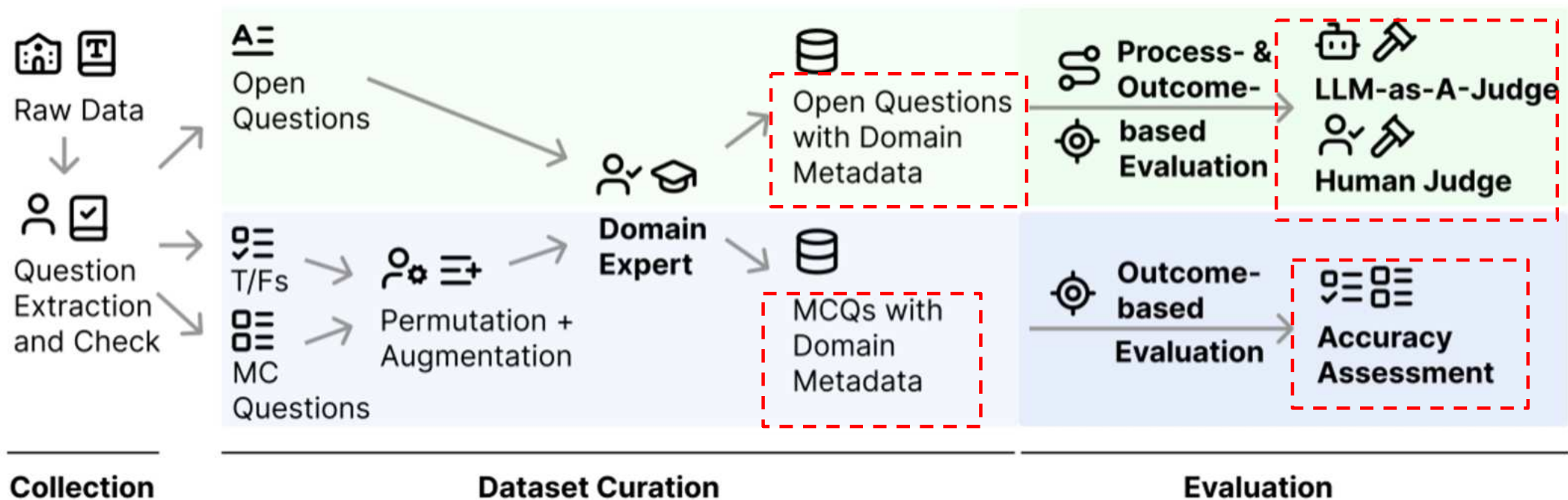
- **Past:**

- Type: multiple choice questions, true/false question.
- Target: knowledge concept
- Origin: selected questions from single textbook / human exam
- Metric: answer correctness

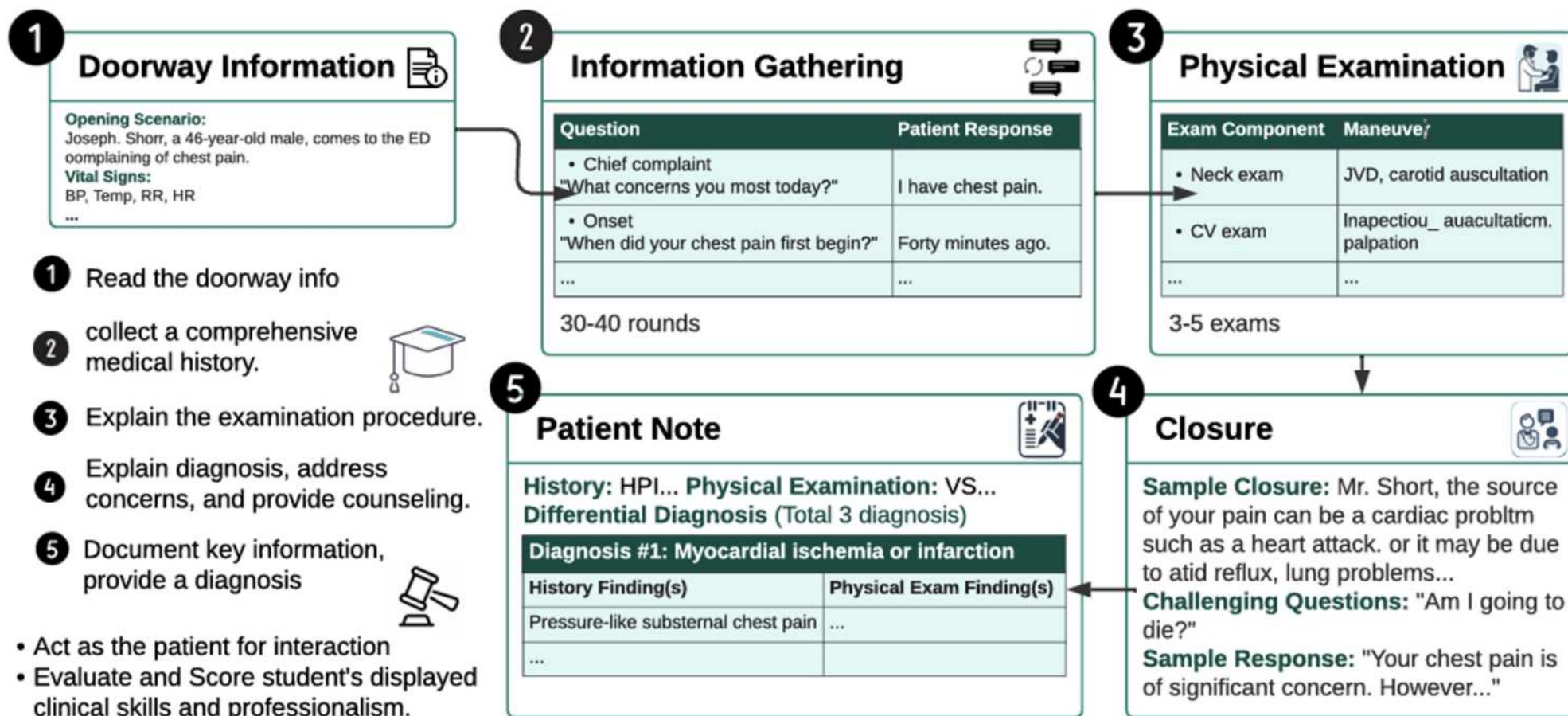
- **Now:**

- Type: open-ended questions.
- Target: real-world problem in practice
- Origin: any related contents (books,
- Metric: all the answering (intermediate steps, final conclusions)

Example of Dataset Preparation



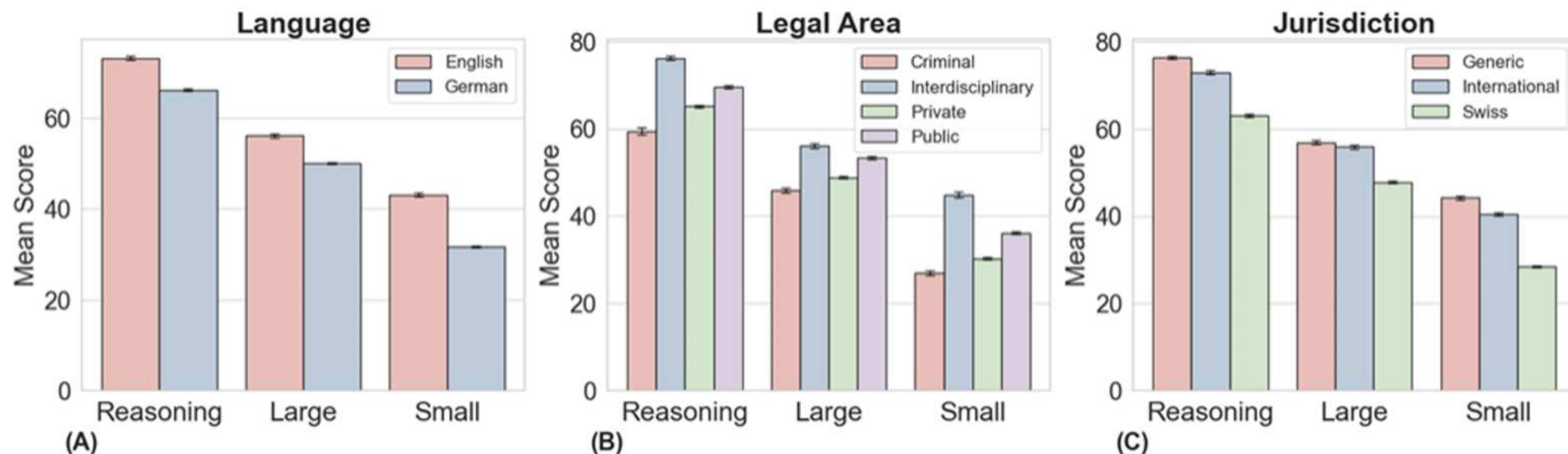
Example of Create Practical Questions



Example of Multilingual and Multisource

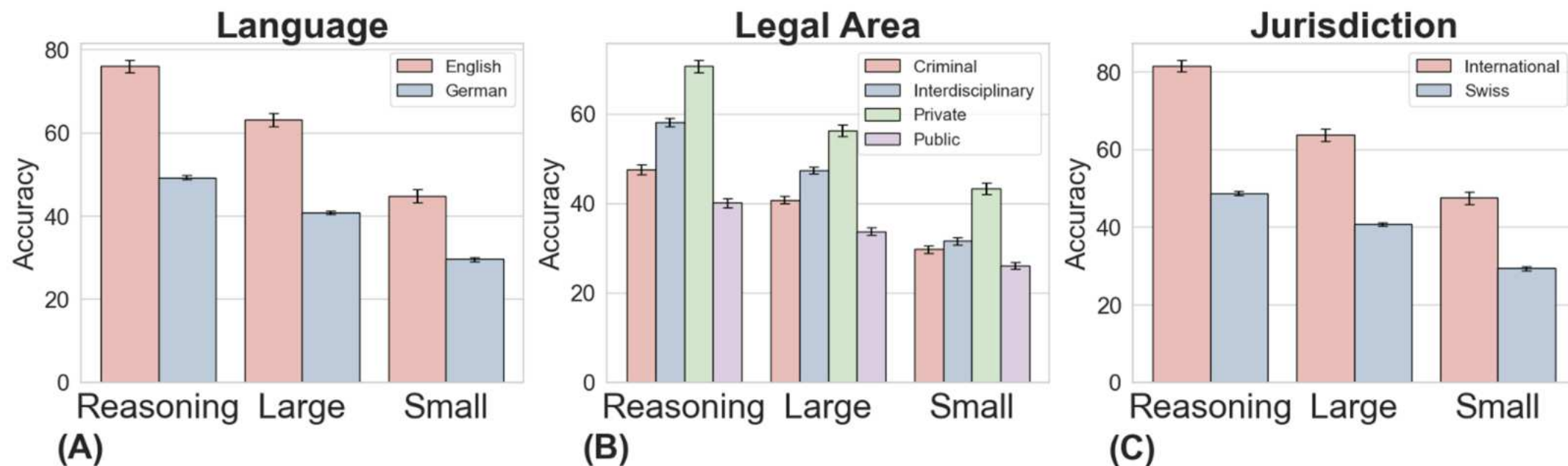
Benchmark	License	Jurisdiction	Languages	Legal Domain
Housing Statutes [88]	Unknown	US	English	Statutory Housing Law
Sara [28]	Unknown	US	English	Tax Liability
Brazilian Bar Exams [15]	Unknown	Brazil	Portuguese	17 law areas
COLIEE [31]	CC BY 4.0	Japan, Canada	Japanese, English	Japanese Civil Law, Canadian Case Law
GLOBALCIT [78]	CC BY 4.0	International	English	International Citizenship Law
JecQA [89]	CC BY-NC-ND	China	Chinese	Chinese Law
MMLU (Legal Subset) [27]	CC BY 4.0	Primarily US	English	General Legal Knowledge
Multistate Bar Exam [88]	Unknown	US	English	Bar Exam questions
PrivacyQA [65]	MIT	Global	English	Privacy Law & Data Protection
LEXAM (ours)	CC BY 4.0	CH, EU & Int.	German, English	78 Subdomains

Findings of Open Ended Questions



The ability of LLMs to solve domain-specific questions varies across different languages and subject areas.

Findings of Multiple Choice Questions



Similar observations can be found with results on MCQs.

Findings of Multiple Choice Questions

Model	4 Choices	8 Choices	16 Choices	32 Choices
Gemini-2.5-Pro [72]	68.6 ($\pm$ 2.37)	51.6 ($\pm$ 2.55)	45.2 ($\pm$ 2.43)	35.6 ($\pm$ 2.30)
Claude-3.7-Sonnet [3]	60.9 ($\pm$ 2.48)	48.6 ($\pm$ 2.44)	40.4 ($\pm$ 2.43)	33.0 ($\pm$ 2.31)
DeepSeek-R1 [13]	57.5 ($\pm$ 2.49)	44.1 ($\pm$ 2.45)	36.9 ($\pm$ 2.36)	24.9 ($\pm$ 2.21)
GPT-4.1 [58]	58.0 ($\pm$ 2.53)	42.3 ($\pm$ 2.51)	33.2 ($\pm$ 2.38)	26.3 ($\pm$ 2.13)
GPT-4o [55]	53.7 ($\pm$ 2.56)	36.4 ($\pm$ 2.37)	22.6 ($\pm$ 2.09)	21.8 ($\pm$ 2.18)
DeepSeek-V3 [14]	58.6 ($\pm$ 2.57)	36.1 ($\pm$ 2.40)	28.9 ($\pm$ 2.28)	16.0 ($\pm$ 1.88)
o3-mini [59]	50.0 ($\pm$ 2.55)	33.5 ($\pm$ 2.34)	24.5 ($\pm$ 2.16)	17.0 ($\pm$ 1.89)

Question designs significantly influences conclusions
about model performance.

Applications in Education Part II



- **Education Assistant**

- Question-Answer
- Assessment
- Tutoring

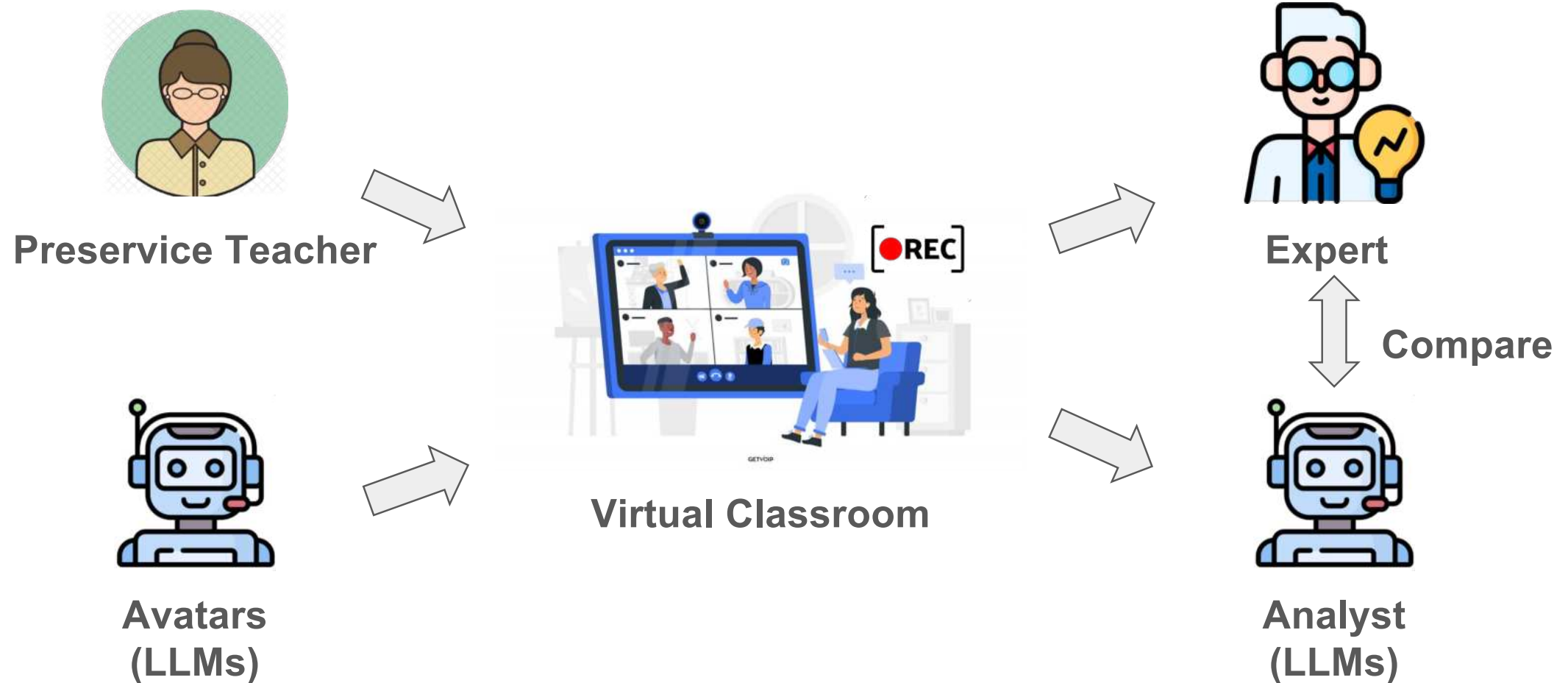
- **Role-play Simulation**

- Student Simulation
- Classroom Simulation
- Role-play Interaction

- **Profession Development**

- Exam Evaluation
- **Studying Assistant**
- Outcome Study

Function as Simulators to Support Skill Practice



Using LLMs as Avatars



Avatars	Personality traits
Ava	Ava predominantly displays disinterest in the subject and often attempts to divert the discussion away from the planned topics. That is a way to redirect the preservice teachers' focus. She presents challenges by resorting to sarcasm, particularly directed towards Ethan
Dev	Dev is profoundly inclined towards facts and knowledge acquisition, often pursuing independent learning. He engages with the topic avidly and occasionally surpasses the understanding of others
Ethan	Ethan exhibits intermittent engagement with the lesson content, often succumbing to distractions from his mobile phone. However, when he does actively participate and display attentiveness, he tends to dominate the discussion. He challenges his peers and the preservice teachers through sarcasm directed at Ava, adopting a provocative demeanour and strategically guiding the discourse by emphasising Dev's knowledge
Jasmine	Jasmine possesses diminished self-confidence and exhibits insecurity. Her sense of self-worth relies heavily on recognition from Ava or assistance from Dev. She rarely does respond to inquiries posed to her. She requires an interactive and practical approach from her peers to comprehend the expectations placed upon her
Savannah	Savannah appreciates collaborative learning and welcomes intellectual challenges within the group. Frequently, she seeks clarification from the preservice teacher or her peers. Generally, she maintains a diplomatic stance concerning the subject matter, yet occasionally, her fervent opinions become evident

Using LLM as Analyst

Raters	Burnings of the Quran			Conspiracy theories		
	CK	PCK	GPk	CK	PCK	GPk
Expert 1	42	34	38	34	19	16
Expert 2	43	31	32	36	23	40
ChatGPT-4	25	24	17	24	30	22
ChatGPT-4o	25	32	17	22	32	15
Claude 3.5 Sonnet	45	62	30	19	32	17

Raters	Burnings of the Quran			Conspiracy theories		
	CK	PCK	GPk	CK	PCK	GPk
Experts	82%	84%	87%	85%	85%	26%
GPT models	25%	21%	21%	10%	11%	23%
Expert 2 vs GPT-4	14%	7%	18%	10%	11%	27%
Expert 2 vs GPT-4o	5%	14%	29%	25%	5%	12%
Expert 2 vs Claude 3.5 Sonnet	3%	36%	−16%	0%	42%	10%

Career Development - Learning Assistant

The preservice teacher doesn't make it easier for the pupils by confusing them by mixing words, concepts, opinions, and values (Expert 2)

The preservice teacher creates a dialogue around an abstract concept and relates it to the pupil's personal experiences. (ChatGPT-4o)

Function as Tutor to Support Skill Learning

 I-VIP: Intelligent Virtual Interactive Program

ProgressLogoutOpen CalculatorTake NotesOpen Whiteboard

7 feet by 10 feet17 feet by 20 feet
27 feet by 30 feet37 feet by 40 feet

Professional
Development Materials

In this activity, we will focus on reasoning about a mathematical situation.

The dimensions of four rectangles are given. Which rectangle looks most like a square? Explain your thinking using complete sentences.

Dialogue



The dimensions of four rectangles are given. Which rectangle looks most like a square? Explain your thinking using complete sentences.

👍👍

Interactive
Learning Area

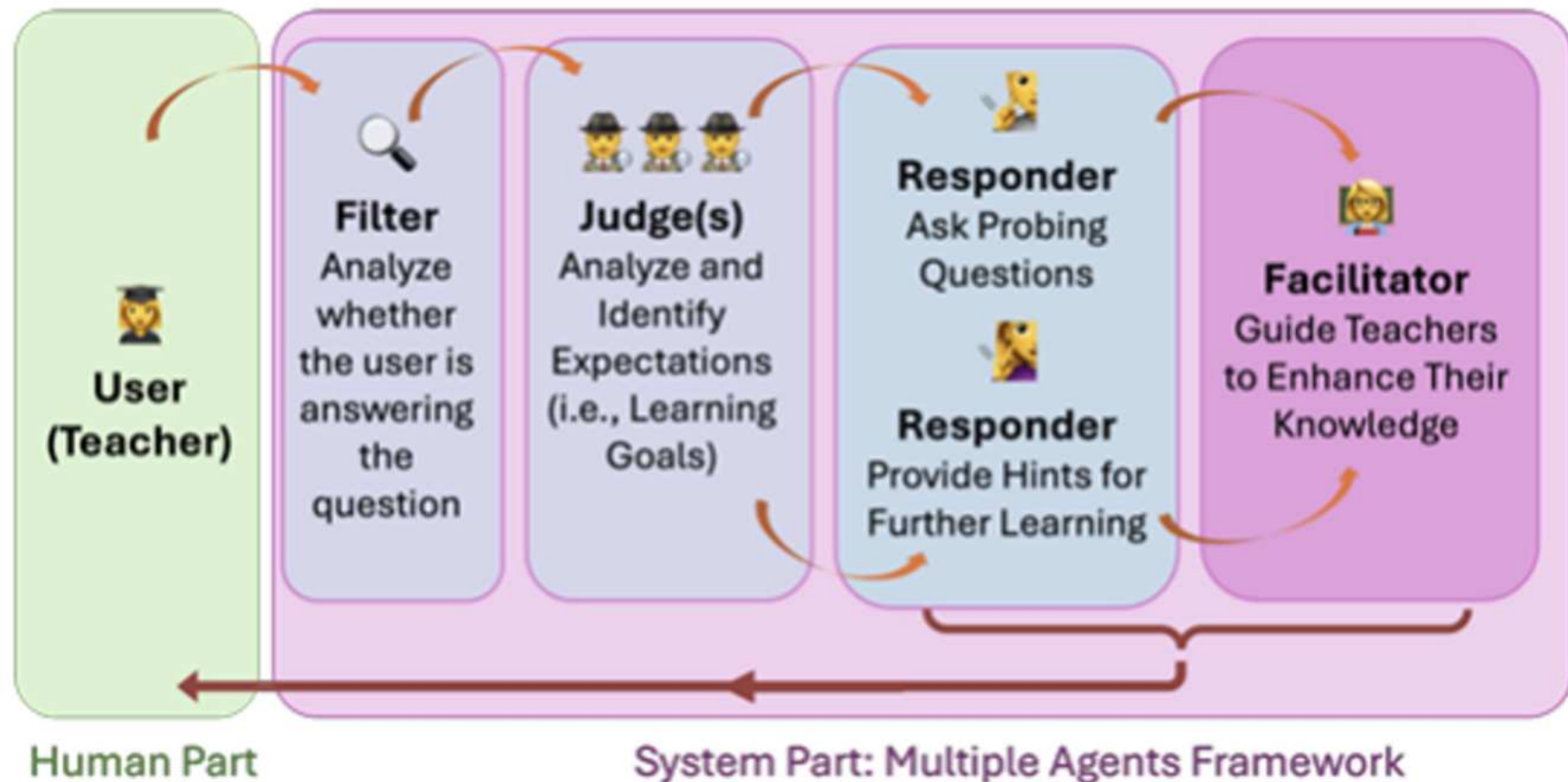
The 7 feet by 10 feet looks most like a square



Type your message here...

1

Multi-Agent Backend Framework



Career Development - Learning Assistant

Component	# Response	# Positive	# Negative	% Positive
Filter	1,900	1,791	109	94.26
Judger	1,538	1,419	119	92.26
Responder	7,135	6,999	136	98.09
Facilitator	1,538	1,538	0	100.0
Tools	2,380	2,380	0	100.0
Total	14,491	14,127	364	97.49

Method	Rubric-Opt	Few-Shot	Both
Improve	34.83	73.03	74.16

Applications in Education Part II



- **Education Assistant**

- Question-Answer
- Assessment
- Tutoring

- **Role-play Simulation**

- Student Simulation
- Classroom Simulation
- Role-play Interaction

- **Professional Development**

- Exam Evaluation
- Studying Assistant
- Outcome Study

Outcome Study in Engineering Education I

1 Conversation History

CONVERSATIONS + NEW

- Help with assignments
- Answers to LeetCode
- Teaching Assistant in CS P...
- Python Answers for CS Co...

2 Main Conversation

Python Answers for CS Course

Hi!

Hello! How can I assist you today?

What is Python?

Python is a high-level programming language that is easy to learn and widely used for a variety of applications. It is known for its readability and simplicity, making it a popular choice for beginners and experienced programmers alike.

Send a message

3 Conversation-level feedback

On a scale of 1 (Strongly Disagree) to 5 (Strongly Agree), please rate the following statement:

CodeTutor accurately understood what I intended to ask. ----- **Comprehension**

CodeTutor challenged me to think critically. ----- **Critical Thinking**

CodeTutor improved my understanding of programming syntax. ----- **Syntax Mastery**

CodeTutor helped me develop my own understanding of the subject. ----- **Independent Learning**

I feel like I still need to ask human TA the questions I asked to CodeTutor. ----- **TA Replacement**

Any specific feedback you would like to provide about this session with CodeTutor?

SUBMIT

4 Message-level feedback

Provide additional feedback

What was the issue with the response? How could it be improved?

Writing code in that style is beyond my skill level.

It does not provide the correct answer.

It gives me an irrelevant answer.

SUBMIT

Conversation-level feedback mode triggers when

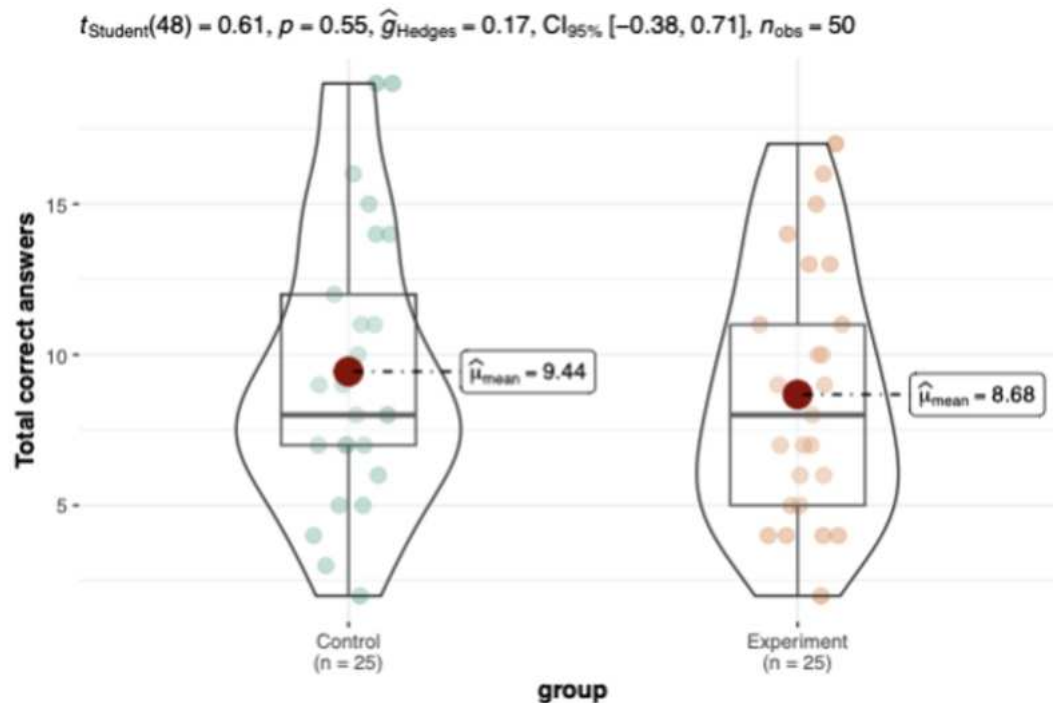
- 1) users are inactive for 10 minutes, or
- 2) users end the conversation; or
- 3) users click on the providing feedback button

Message-level feedback mode triggers when

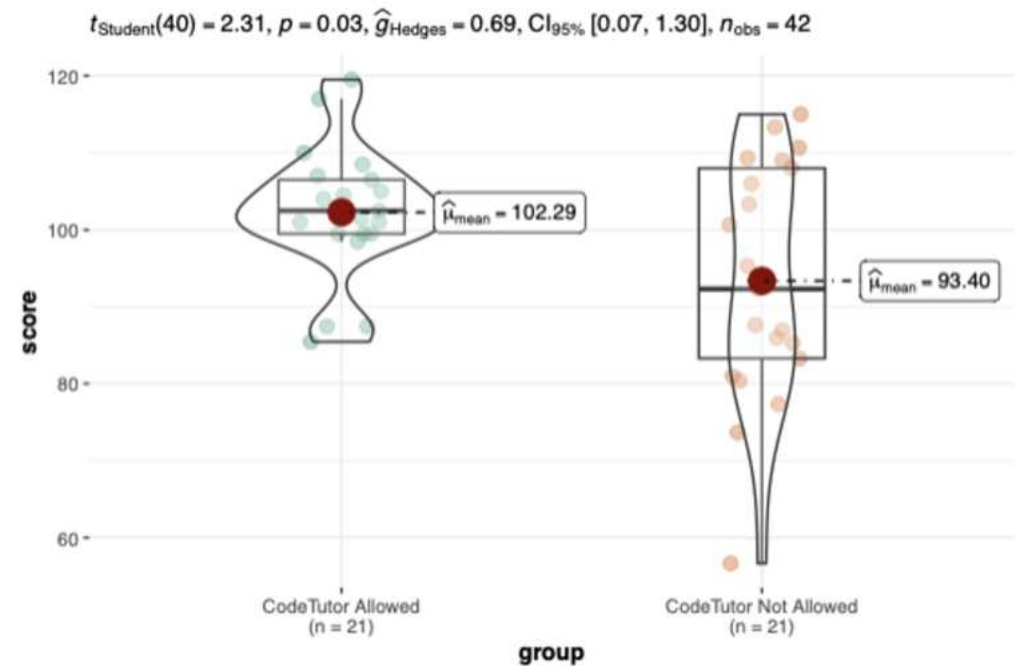
users click on the upvote or downvote

Characteristics	Options	Number of participants
Gender	Woman	22
	Man	25
	Non-binary	1
	Prefer not to say	2
Major	Computer Science	18
	Data Science	9
	Biology	5
	Mathematics	4
	Economics	3
	Others	10
	Not reported	1
Year of Study	Freshman	37
	Sophomore	5
	Junior	6
	Senior	1
	Not reported	1
Race	African American or Black	1
	Asian	17
	Multiracial	3
	White	26
	Not reported	3
Ethnicity	Latino/Hispanic	3
Prior Experience with LLM tools	Only ChatGPT	28
	ChatGPT and other tools	11
	Never used	11

Performance Comparison Between Pre- and Post-Tests

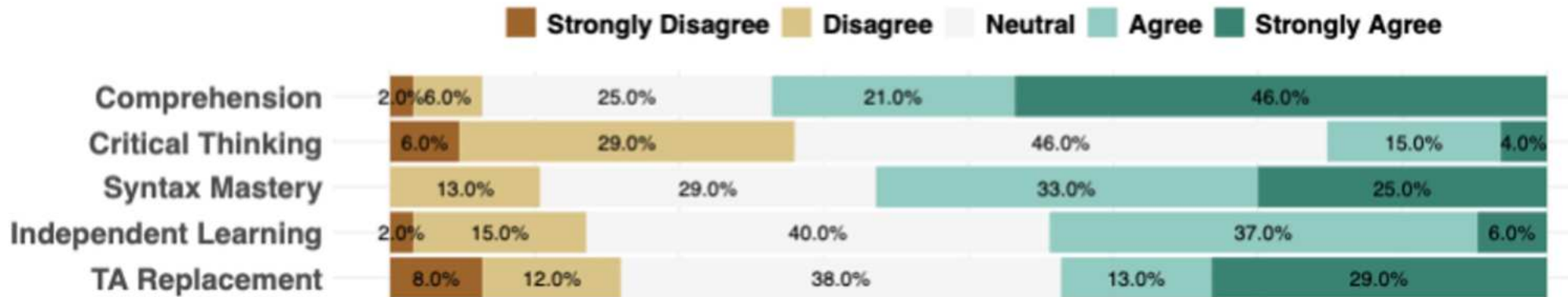


Pre-test performance comparisons



Post-test performance comparisons

Extended Analysis of User Attitudes

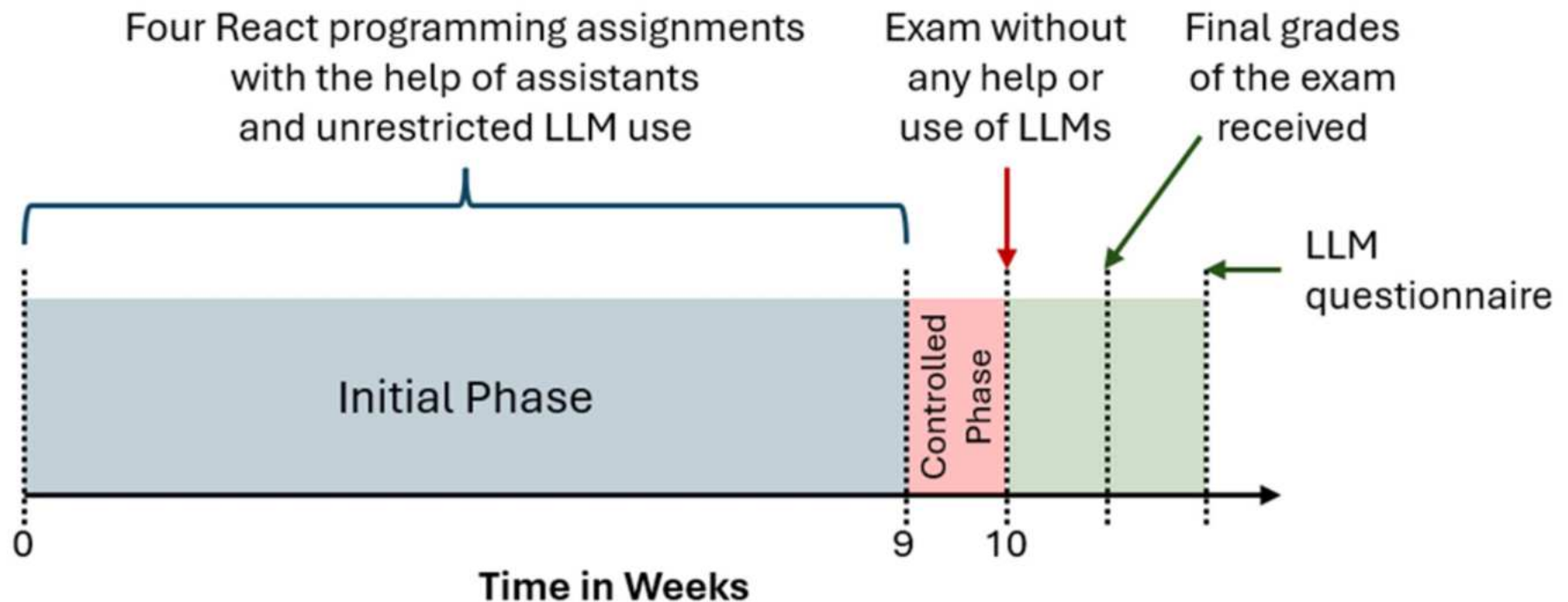


Perceptions of LLM Functions in Learning

	Comprehension β (Std. Error)	Critical Thinking β (Std. Error)	Syntax Mastery β (Std. Error)	Independent Learning β (Std. Error)	TA Replacement β (Std. Error)
Const	4.700(0.297)***	2.690(0.247)***	3.760(0.262)***	3.044(0.218)***	3.964(0.330)***
Time	-0.114(0.039)**	0.040(0.037)	-0.018(0.041)	0.054(0.036)	-0.099(0.051) [†]

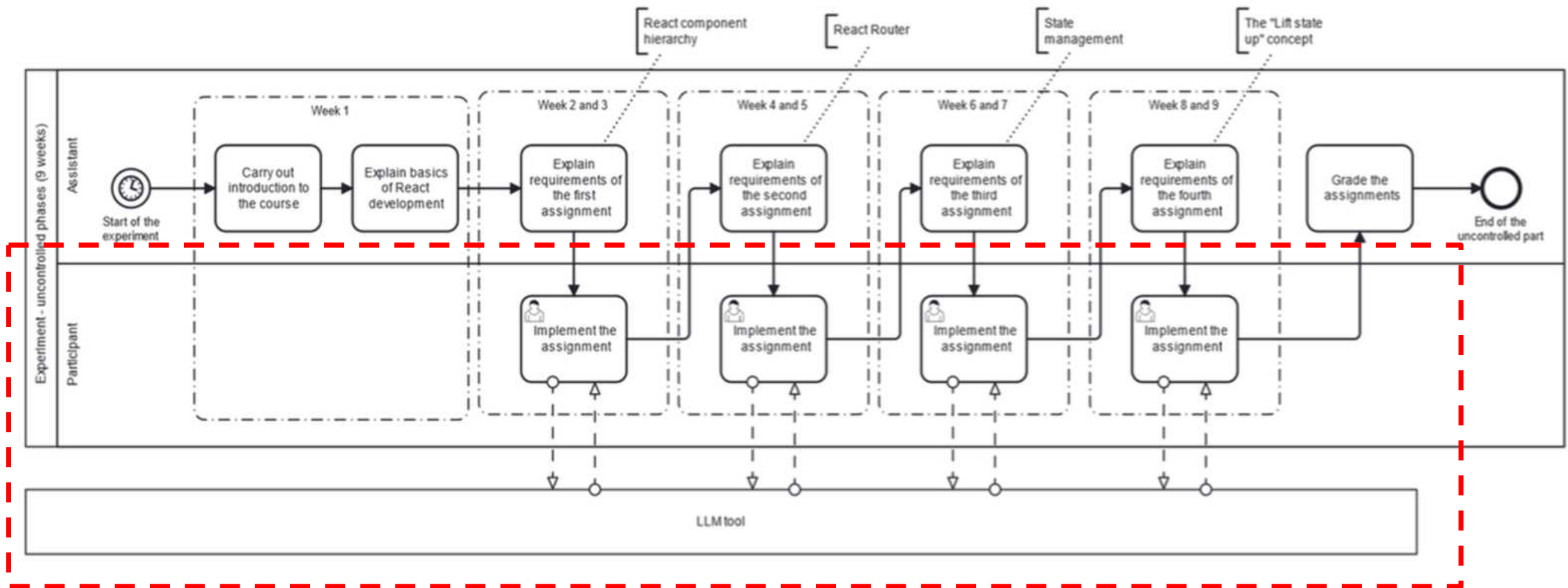
Shifts in User Attitudes Over Time

Outcome Study in Engineering Education II

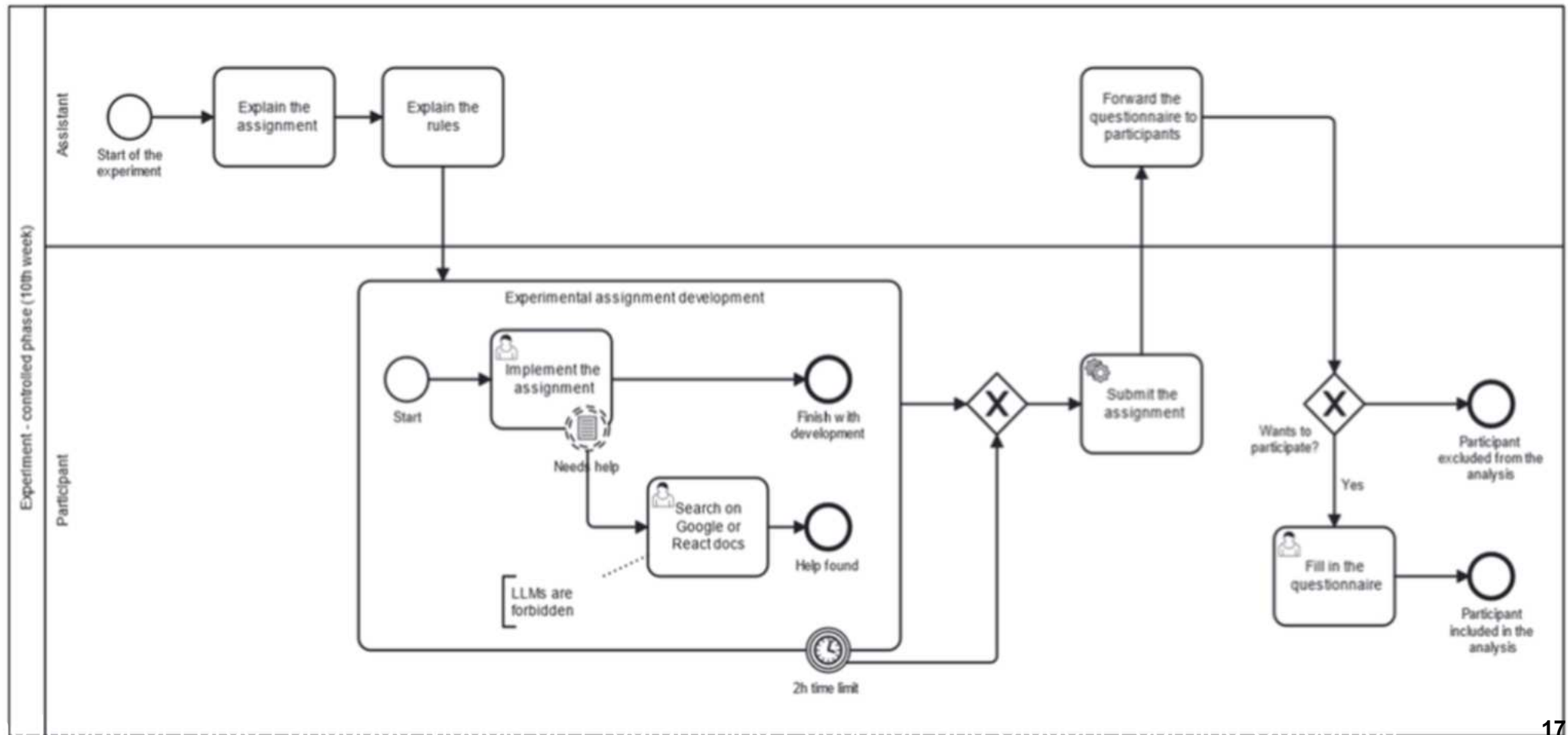


Specify the Uses of LLMs in Assisting with Programming Assignments

Designs of Initial Phase



Designs of Control Phase



Career Development - Outcome Study

		Mean	Median	SD	Min	Max
LLM use	Generating code	2.59	2.50	1.10	1.00	5.00
	Additional explanations	3.75	4.00	1.24	1.00	5.00
	Debugging	3.78	4.00	1.16	1.00	5.00
	Average	3.38	3.67	0.94	1.00	5.00
	Final grade	6.72	8.00	3.10	0.50	10.00

LLM Use	Spearman's Rho	95% Bootstrap CI	<i>p</i>
Generating code	−0.305	(−0.595, −0.058)	0.045
Additional explanations	−0.201	(−0.523, 0.220)	0.135
Debugging	−0.360	(−0.628, −0.011)	0.021
Average	−0.347	(−0.626, −0.044)	0.026

Tutorial Outline

Morning

Introduction
to GenAI
(40 Min)



Overview of GenAI
in Education
(40 Min)



Applications in
Education Part I
(40 Min)



Afternoon

Applications in
Education Part II
(40 Min)



Demo
Presentation
(40 Min)



Future Direction
and QA
(40 Min)

Live Demo for I-VIP System

Live Demo for Automatic Grading System

Tutorial Outline

Morning

Introduction
to GenAI
(40 Min)



Overview of GenAI
in Education
(40 Min)



Applications in
Education Part I
(60 Min)



Afternoon

Applications in
Education Part II
(60 Min)



Demo
Presentation
(40 Min)



Future Direction
and QA
(20 Min)

Future Directions

