Shahriar, T., & Matsuda, N. (2024). "I am confused! How to differentiate between…?" Adaptive Followup Questions Facilitate Tutor Learning with Effective Time-on-task. In Andrew, Irene & Zitao (Eds.), Proceedings of the International Conference on Artificial Intelligence in Education (pp. 17-30): Springer.

"I am confused! How to differentiate between…?" Adaptive Follow-up Questions Facilitate Tutor Learning with Effective Time-on-task

Tasmia Shahriar [0000-0003-0199-7757] and Noboru Matsuda [0000-0003-2344-1485]
North Carolina State University, Raleigh NC 27695, USA
(tshahri,noboru.matsuda)@ncsu.edu

Abstract. Within the learning-by-teaching paradigm, students, who we refer as *tutors*, often tend to dictate what they know or what to do rather than reflecting on their knowledge when assisting a teachable agent (TA). It is vital to explore more effective ways of fostering tutor reflection and enhancing the learning experience. While TAs can employ static follow-up questions, such as "Can you clarify or explain more in detail?" to encourage reflective thinking, the question arises: Can Large Language Models (LLMs) generate more adaptive and contextually-driven questions to deepen tutor engagement and facilitate their learning process? In this paper, we propose ExpectAdapt, a novel questioning framework for the TA using three stacked LLMs to promote reflective thinking in tutors, thereby, facilitating tutor learning. ExpectAdapt generates adaptive follow-up questions by directing tutors towards an expected response based on the tutor's contributions using conversation history as a contextual guide. Our empirical study with 42 middle-school students demonstrates that adaptive follow-up questions facilitated tutor learning by effectively increasing problem-solving accuracy in the learning-by-teaching environment when compared to tutors answering the static follow-up questions and no follow-up questions at all.

Keywords: Learning by teaching, conversational questions, large language model, in-context learning

1 Introduction

Students learn more by assisting a teachable agent (TA)—a synthetic peer they can iteratively teach—compared to solitary learning [1]. This phenomenon is known as *tutor learning* [2-4]. In our work, we address students who teach a TA as *tutors*. Empirical studies reported that tutors often tend to dictate what they know, instead of reflecting on their understanding and critical thinking that results in a limited benefits from learning-by-teaching [5, 6]. The TA can promote tutors' reflective thinking by persistently asking follow-up questions [7-10]. Yet, automatically generating such follow-up questions is challenging due to the expertise required to formulate such questions and varying levels of prior knowledge among tutors. Effective questions must be tailored to individual tutors' understanding, while simultaneously pushing their cognitive boundaries, maintaining discourse coherence, context relevance, and inextricably bound to the conceptual content of the subject matter [11, 12].

Can we instruct Large Language Models (LLMs) in such a way that it can generate

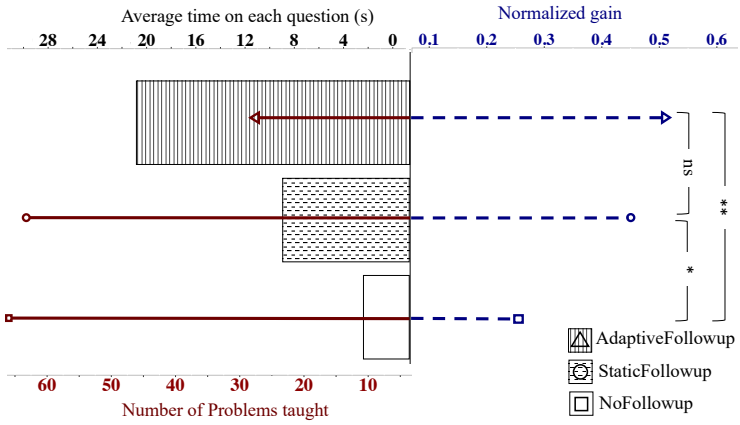


Fig 1: **ExpectAdapt wins in terms of more effective time-on-task**. Tutors spent more time on average to answer adaptive follow-up questions compared to static questions (shown in barplot) that helped them achieve the same gain (shown in dotted **blue line**) by teaching significantly fewer problems (shown in solid **darkred line**) to the teachable agent

questions to engage tutors in critical thinking in a learning-by-teaching environment?

In this paper, we propose ExpectAdapt, a novel follow-up questioning framework for the TA. ExpectAdapt consists of two LLMs. The first LLM generates an ideal tutor's response (to TA's question) that is reflective of tutor's critical thinking. The second LLM generates a follow-up question (relevant to the conversation history) if the student's response to TA's question is not satisfactory relative to the ideal response. **Fig 1** shows that tutors spent significantly more time on average to answer the expectation tailored adaptive (or ExpectAdapt for short) follow-up questions compared to static questions that only prompted tutors to explain more. Furthermore, spending more time on answering ExpectAdapt follow-up questions helped tutors *achieve the same learning gain by teaching fewer problems* to the TA compared to tutors who answered static questions. Additionally, tutors who engaged with ExpectAdapt follow-up questions achieved higher learning gains compared to those who did not answer any follow-up questions.

In this paper, we address following research questions. RQ1: Does answering ExpectAdapt follow-up questions help tutors learn? RQ2: Is ExpectAdapt follow-up questions more effective than the static follow-up questions?

Our main contributions are: (I) We propose ExpectAdapt that employs prompt engineering techniques to configure LLMs in a manner that enables them to generate contextually relevant follow-up questions. (II) We conduct an empirical evaluation study that showed the effectiveness of our proposed ExpectAdapt framework. (III) The ExpectAdapt questioning framework offers scalability, as it can be easily adapted to various problem-solving domains with minimal need for expert annotations.

2 APLUS: The Learning-By-Teaching Environment

Our study extends the traditional APLUS (**A**rtificial **P**eer **L**earning **U**sing **S**imStudent) where tutors assist SimStudent (the teachable agent) how to solve linear algebraic

equations [13, 14]. **Fig 2** displays the user interface of APLUS. Whenever tutor enters a linear equation to teach (**Fig 2-a**), SimStudent tries to solve one step at a time by consulting its knowledge base that consists of production rules once learned like, "if [*conditions*] hold then perform [*a solution step*]." In APLUS, the *solution step* allows four basic math operations: *add*, *subtract*, *multiply*, and *divide by* a term. If SimStudent has a production that can apply, it seeks feedback from the tutor. If the tutor agrees, it proceeds to the next step. If the tutor disagrees, it asks a focal question, "Why am I wrong?". The tutor is expected to provide their textual explanation in a chat box (**Fig 2-c**). If SimStudent does not have a production to apply, it requests the tutor to demonstrate the next step. After tutor demonstrates the solution step, it asks another focal question, "Why should we do it?". In the traditional APLUS, SimStudent does not ask follow-up questions after tutor's response to the focal questions.

Apart from teaching, tutor can quiz SimStudent anytime to evaluate how well SimStudent has learned thus far by observing the SimStudent's performance on the quiz. Quiz topics include one-step equations (level 1), two-step equations (level 2), equations with variables on both sides (level 3), and a final challenge that contains equations with variables on both sides (level 4) (**Fig 2-g)**. SimStudent works on a single quiz level at a time. Upon successfully passing a level, the subsequent level is unlocked.

Tutors may also review the resource tabs that include problem bank, unit overview, introduction video and worked out examples at any time (**Fig 2-d**). The teacher agent, Mr. Williams (**Fig 2-h**), provides on-demand, voluntary hints on how to teach. For example, if the tutor repeatedly teaches one-step equations, Mr. Williams might provide the hint, "SimStudent failed on the two-step equation. Teaching similar equations will help him pass that quiz item."

3 ExpectAdapt Framework

3.1 Motivation

In our past studies with APLUS, tutors often exhibited a tendency to neglect or

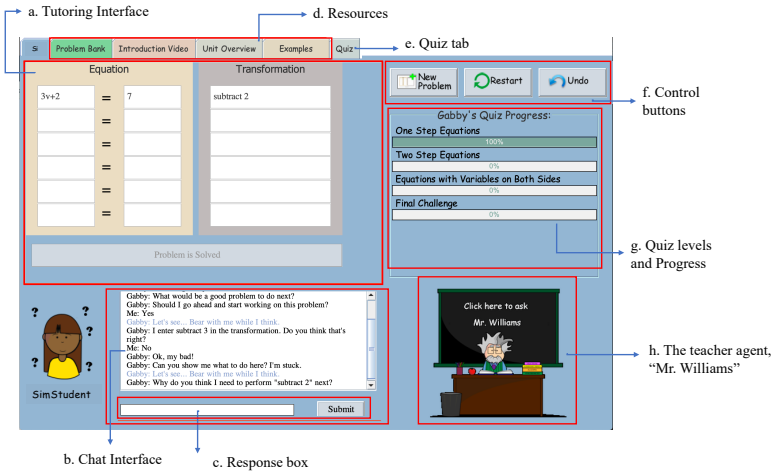


Fig 2: APLUS interface with SimStudent in the bottom left corner

inadequately respond to SimStudent's focal questions [6, 9]. We also found that tutors who could explain elaborately using conceptual terms learned significantly more than tutors who could not provide such responses irrespective of their prior knowledge [6]. Roscoe [8] linked tutors' inability to provide accurate, elaborated and sense-making response with their infrequent reflective behavior.

Building upon these insights, we design ExpectAdapt framework to generate questions directing tutors towards an elaborated response. ExpectAdapt generates follow-up questions tapping on the aspects of the elaborated response that tutor has not conveyed just yet throughout the current conversation history. Our work is closely aligned with AutoTutor's expectation & misconception-tailored dialogue (aka, *EMT* dialogue) [15]. AutoTutor is a computer-based tutor that attempts to simulate the dialogue moves of a human tutor to help students learn. However, our approach distinguishes itself by detecting a misalignment between tutors' response and the expected response to mimic a teachable agent's effort to bridge knowledge gaps when encountering unclear concepts. This process resembles the way a student clarifies ambiguities in a textbook, avoiding excessively corrective questioning. Additionally, AutoTutor relies heavily on scripted authoring tools, demanding significant human expertise to design various misconception cases and dialogue scenarios. In contrast, our LLM-based framework aims to reduce this substantial human effort, making it a more efficient and scalable solution.

ExpectAdapt consists of three modules: (1) Expected response generator, (2) Alignment detector, and (3) Expectation tailored follow-up question generator. All these three modules are implemented using OpenAI API key for the GPT-3.5-turbo model. **Fig 3** shows the ExpectAdapt framework with three stacked LLM.

We utilized various prompt engineering techniques such as few-shot demonstrations using chain-of-thought [16-19], role prompting [20], and adding extra context in the prompt [21, 22]. We intentionally avoided the term "teachable agent" in our prompts, opting instead for the term "student". This choice is grounded in the hypothesis that LLMs may encounter difficulties in assuming the role of a teachable agent, a scenario presumably less prevalent in their pretraining datasets—LLMs excel with more common terms from pretraining [23]. For researchers aiming to replicate our results, we recommend substituting "teachable agent" with "student" shown in the prompts.

3.2 Expected Response Generator

The Expected Response Generator (ERG) LLM outputs an accurate explanation that

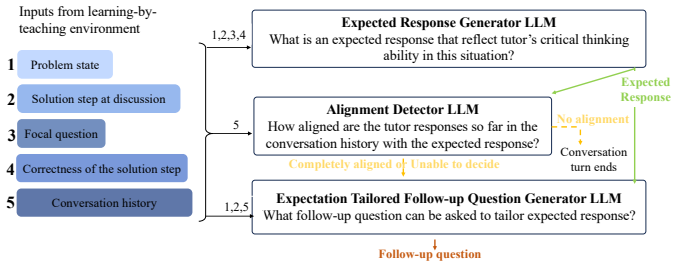


Fig 3: ExpectAdapt framework with three stacked LLM

reflects critical thinking by using domain relevant concepts to the question asked by the teachable agent. Zhang *et al.* [20] reported that providing the LLM with a specific role to play, such as a helpful assistant or a knowledgeable expert can be particularly effective in ensuring that the model's output align with the desired output. To ensure that the generated expected response is accurate we provide the ERG LLM with the role of a tutor who is expert in the domain in the task instruction part of the prompt. **Fig 4** shows the prompt used for the ERG LLM.

In general, LLMs learn to perform a new task by conditioning on a few *input*/*output* demonstrations, a phenomenon called *in-context* learning [16]. One of the key drivers of in-context learning is the distribution of the input text *specified by* the demonstrations [24]. We hypothesized that problem state, solution step at discussion, correctness of the solution step, and the focal question asked by the TA are necessary and sufficient components to capture the entire specification of *input*. We further hypothesize that this input specification facilitates accurate generation of expected response. We based our formulation of *output*, which is the expected responses, on the definition of reflective responses defined in [6] as "A reflective response is either descriptive or reparative in its intonation and elaborates in favor or disfavor of a solution step using relevant conceptual terms." We included eight demonstrations as few-shot examples, adhering to findings that LLM performance declines with more than eight examples [24, 25].

To enhance the LLM's ability to generate expected responses across problem states and solution steps that were not covered in the demonstrations, we incrementally integrated conceptual knowledge of algebraic domain addressing errors made by the LLM that we call *assertions* [22]. For instance, LLM generated an output, "dividing by 4 is wrong when the equation is $-4v = 6 + 3v$ since the coefficient is -4 not 4." Such error was prevented including an assertion like, "You cannot divide when an equation has two variable terms." Adding this assertion modified LLM output as, "there are two

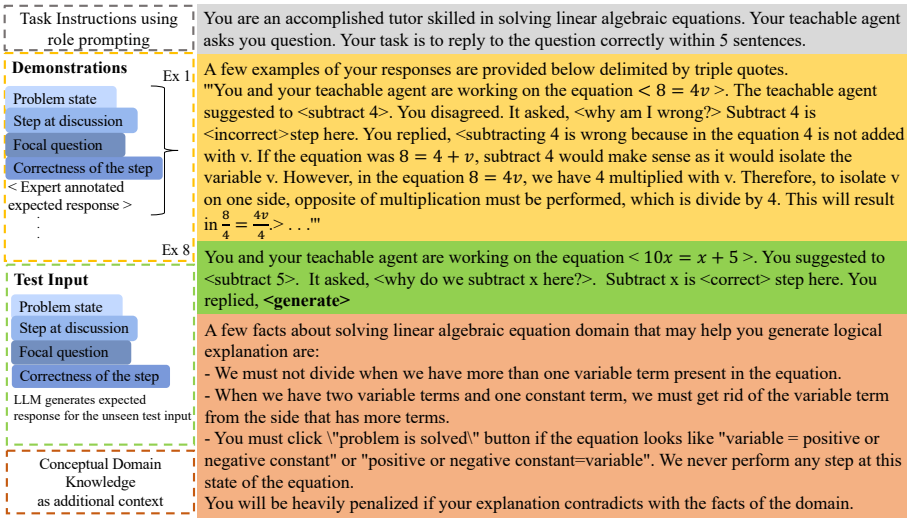


Fig 4: Overview of the prompt components and the resultant prompt used for the expected response generation.

variable terms in this equation. If the equation was $-4v = 6$, dividing by -4 would make sense." We used the greedy decoding strategy by setting the temperature to 0 in ERG LLM. Since the current work is primarily focused on the question generation, we encourage interested readers to refer to this paper [22].

We conducted a survey with 12 in-service middle school teachers to assess the quality of the generated expected responses. The survey data revealed that LLM generated responses are (1) relevant to the *input* specification, (2) elaborate optimal solution step, and (3) sound in terms of using concept terms in its reasoning. The survey further confirmed that including assertions in the prompt improved the accuracy of expected responses by 15% over solely relying on demonstrations [22].

3.3 Expectation Tailored Follow-up Question Generator

To generate a question, the goal is to find any missing stem from the expected response that was not conveyed by the tutor during the conversation and ask an open-ended question focusing on the missing stem. In the task instruction, the Expectation Tailored Follow-up Question Generator LLM was provided the role of the teachable agent to encourage curiosity driven questions. The task instruction also includes a set of rules to generate the question. The rules are as follows ("you" in the rules refer to the teachable agent role-playing LLM):

- To generate a question, you must find out a missing stem from the expected response that was not covered in the conversation history and generate a question.
- If you have previously asked question about a missing stem but tutor did not provide a relevant response, find another missing stem, and generate a question.
- If you have asked questions about every possible stem from the expected response, then say, *no question*.

To further mold the LLM output, we design eight few-shot demonstrations using the chain-of-thought prompting technique [25]. These demonstrations consist of *input* comprising the problem state, the solution step discussed, and the conversation history for that step. To create more realistic conversation histories, we drew from the data collected during past studies using APLUS [26, 27] that are available publicly in Datashop [28]. This allowed us to incorporate realistic instances containing grammatical or spelling errors in tutors' responses. The *output* consists of three parts: (1) the chain-of-thought to find the missing stem from the expected response, (2) relevant acknowledgement or summarization from the conversation history to maintain conversational context, and (3) the formulated question focusing on the missing stem. **Fig 5** shows the prompt used for question generation. The three parts of the output are marked using numbers in the figure.

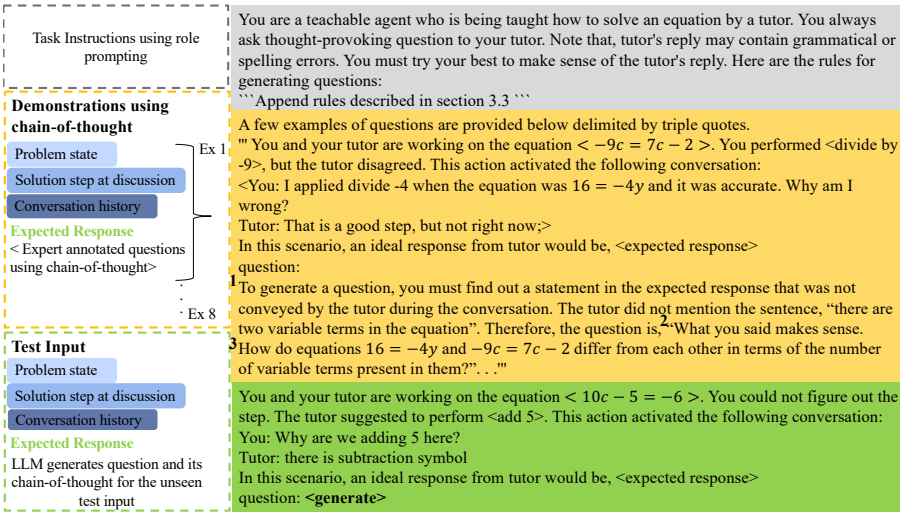


Fig 5: Overview of the prompt components and the resultant prompt used for the expectation tailored question generation. The output components are marked using numbers in the demonstration.

Finally, to encourage diverse reasoning pathways and prevent the TA from asking repetitive questions, we deliberately set the temperature of the LLM decoder to 0.5 instead of using a greedy decoding approach at 0.

3.4 Alignment Detector

Our initial observation revealed that questions generated by the question generator tend to be overly corrective or out of conversational context. This happened when (1) tutor disagreed with the TA's suggested step and had a different solution step in mind than the expected response, (2) tutor's response diverged completely from the expected response, and (3) expected response was erroneous due to hallucinations [29].
An instance of (1) is shown below:

A tutor and SimStudent are working on the equation $3 - 2x = 4$. SimStudent suggested to perform add 3, but the tutor disagreed. This action activated the following conversation:

SimStudent: Why am I wrong?
Tutor: You have to add 2 because we have -2 in the equation.
SimStudent: I understand you are suggesting to add 2. Why can't I perform subtract 3?

In this case, SimStudent asking "Why can't I perform subtract 3?" is overly corrective. This question was suggested by the question generator because the expected response in this scenario was: "Add 3 is not correct because 3 is added with $-2x$. The equation can also be written as $-2x + 3 = 4$. To undo the +3 and isolate -2x, we must perform the opposite of +3 which is subtract 3 on both sides. Subtract 3 will result in $-2x + 3 - 3 = 4 - 3$."

Johns [30] identified that correcting tutors' contribution is limiting for learning. These observations inspired us to incorporate the alignment detector that indicates if tutor's response and the expected response are (1) completely aligned, (2) not aligned,

Bloom's Taxonomy levels	Questions generated by ExpectAdapt
Remember	What is the coefficient in this equation?
Understand	I see that we have 7c and -9c in the equation. Can you explain what makes them like terms? How do we identify like terms in an equation?
Apply	After dividing both sides of the equation by 3," what will be the final equation? How does this step help us isolate v on its own?
Analyze	I am confused! How can we differentiate between the coefficient and the constant in an equation? Can you provide some examples to clarify this concept?
Evaluate	It appears we can cancel out the addition of 3 by subtracting 3 as well as adding -3. Can you explain why they are same?
Create	None found

Table 1: Examples of questions generated by ExpectAdapt with corresponding Bloom's Taxonomy levels. Labels are author-coded for demonstration only and do not reflect an assessment of question quality or effectiveness.

or (3) unable to detect. We designed eight few-shot demonstrations including chain-of-thoughts covering three scenarios to design the prompt for the alignment detector LLM.

We show sample questions generated by ExpectAdapt framework in **Table 1.**

4 Method

The central research questions are: (1) Does answering adaptive follow-up questions help tutors learn? (2) Are adaptive questions more effective than static follow-up questions?

To address these research questions, we conducted a semi-secondary data analysis study where two sets of empirical data were combined: (1) The data collected from a past study [6], and (2) the data that we collected from a new study that we conducted. For clarity, we call the past study **Study A** and the newly conducted study **Study B.**

Study A included two conditions of APLUS: **N**o**F**ollowup condition where tutors only answered the focal question and TA never asked follow-up questions, and **S**tatic**F**ollowup condition where tutors answered focal questions along with static "explain more" follow-up questions. There were 16 and 17 participants in the **N**o**F**ollowup and the **S**tatic**F**ollowup conditions respectively.

Study B, which we recently conducted, involved only one condition, **A**daptive**F**ollowup condition, where tutors answered focal questions along with adaptive follow-up questions generated by ExpectAdapt. Nine 6^{th} to 8^{th}-grade middle school students from local areas were recruited through a study flyer shared within the previous participants' network (aka purposive and snowball sampling). Participants received monetary compensation. The study was conducted online where APLUS was accessed through Zoom screen-sharing.

Consequently, the current analysis compares three conditions with the total of 42 middle school students involved in three conditions. Study B followed the same format and used the same measures as Study A. That is, participants took a pre-test for 30 minutes on the first day of the study. Immediately after taking the pre-test, all

participants watched a 10-minute tutorial video on how to use APLUS. Participants were informed in the video that their goal was to help their TA pass the quiz. Participants were free to use APLUS for three days for a total of 2 hours or to complete their goal (i.e., passing the quiz), whichever came first. Upon completion, participants took a 30-minute post-test.

The pre- and post-tests were isomorphic, and each consisted of 22 questions: 10 questions on solving the equation and 12 multiple-choice questions to measure the proficiency of algebra concepts. Details on the test items can be found in our previous paper [6]. We utilized a binary scoring system for each test items, i.e., answers were marked strictly as either correct or incorrect. Test scores are normalized as the ratio of participant's score to the maximum score.

One-way ANOVA with the normalized pre-test score and condition confirmed no condition difference; $M_{\text{NoFollowup}}$ = 0.63±0.24 vs. $M_{\text{StaticFollowup}}$ = 0.60±0.18 vs. $M_{\text{AdaptiveFollowup}}$ = 0.62±0.25; $F(2,39) = 0.06$, $p = 0.94$. We controlled the time on task. A one-way ANOVA confirmed no condition difference on the minutes participants spent on APLUS; $M_{\text{NoFollowup}}$=215 vs $M_{\text{StaticFollowup}}$=242 vs. $M_{\text{AdaptiveFollowup}}$=204; $F(2,39)=0.66$, $p=.52$. To maintain consistency with StaticFollowup condition in our previous study, we purposefully limited the ExpectAdapt framework to generate a maximum of three follow-up questions for each focal question.

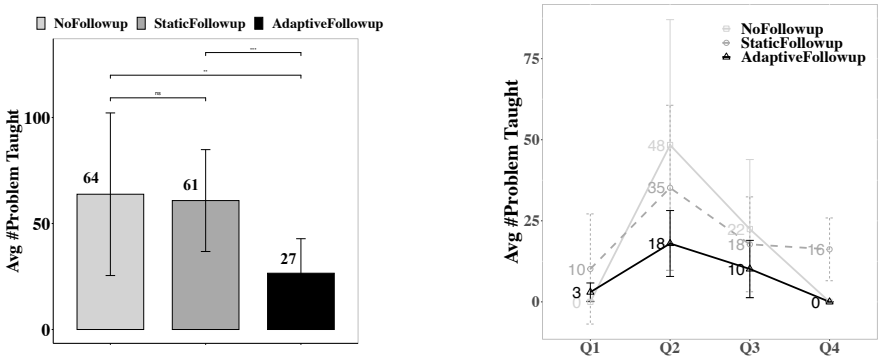
In the following analysis, we use the learning outcome data (normalized pre-and post-test scores) along with participant's interaction data collected by APLUS, interface actions, TA inquiries, and participants' responses.

5 Results

5.1 Tutors in follow-up question modes had higher test score improvement than NoFollowup mode, whereas AdaptiveFollowup tied with StaticFollowup.

We conducted a repeated-measures ANOVA with test score as a dependent variable, whereas test-time (pre- vs. post-test) as the within-subject and condition (**N**o**F**ollowup vs. **S**tatic**F**ollowup vs. **A**daptive**F**ollowup) as the between-subject independent variables. There was an interaction between test-time and condition; $F(2, 39)=3.05$, $p=0.05$. A simple main effect of condition (paired t-test with test-time as the independent variable) revealed that tutors in follow-up conditions showed a reliable increase from pre- to post-test (**S**tatic**F**ollowup: paired-$t(16)=3.86$, $p<0.05$ and **A**daptive**F**ollowup: paired-$t(8)=2.87$, $p<0.05$); but no reliable increase in the **N**o**F**ollowup condition (paired-$t(15)=1.2$, $p=0.24$). We further ran ANCOVA analysis with the normalized post-test as dependent variable and condition as the independent variable while controlling the normalized pre-test. No condition effect was found between **A**daptive**F**ollowup and **S**tatic**F**ollowup tutors; $F(1,23) = 0.03$, $p = .88$. However, there was a condition difference between **A**daptive**F**ollowup vs **N**o**F**ollowup; $F(1,22) = 3.90$, $p = .06$ and **S**tatic**F**ollowup vs **N**o**F**ollowup tutors; $F(1,30) = 5.20$, $p < .05$.

The results suggests that *tutors who answered any kind of follow-up questions (static or adaptive) ended up having a higher post-test improvement than the tutors who did not answer follow-up questions.* The result also suggests that *tutors who answered*



(a) Avg problems taught in APLUS (b) Avg problems taught to pass quiz levels

Fig 6: Average problems taught by tutors across conditions

adaptive follow-up questions ended up having the same post-test improvement as the tutors who answered static follow-up questions.

5.2 Tutors in AdaptiveFollowup passed quiz levels by teaching fewer problems compared to StaticFollowup and NoFollowup.

Tutors in **A**daptive**F**ollowup and **S**tatic**F**ollowup conditions tied on post-test. This outcome prompted an investigation into whether the time spent on APLUS was comparable across these conditions. Therefore, we first analyzed the number of problems taught by tutors across all conditions. A one-way ANOVA with number of problems taught as dependent variable and condition as independent variable revealed a main-effect of condition; $F(2,39) = 5.3$ $p < .01$ $M_{NoFollowup} = 64 \pm 38$ vs. $M_{StaticFollowup} = 61 \pm 24$ vs. $M_{AdaptiveFollowup} = 27 \pm 16$. We ran pairwise *T*-tests with Bonferroni correction as the post-hoc analysis. The results revealed that the average problems taught by AdaptiveFollowup tutors were statistically different compared to StaticFollowup ($t(22.0)=4.30$, $p<.05$) and NoFollowup ($t(22.0)=3.38$, $p<.05$) tutors, whereas, the average problems taught by StaticFollowup vs. NoFollowup tutors were not different ($t(25.0)=0.27$, $p=.27$). Therefore, the data suggests that *StaticFollowup and NoFollowup tutors taught equal number of problems on average, whereas AdaptiveFollowup tutors taught significantly fewer number of problems compared to the other two conditions*.

This finding led us to question whether this difference influenced their ability to pass quiz levels. When we ran one-way ANOVA with the maximum quiz level passed by tutors across conditions, we found no condition effect; $M_{NoFollowup} = 2$ vs. $M_{StaticFollowup} = 2$ vs. $M_{AdaptiveFollowup} = 2$ $F(2,39) = .50$ $p = .61$. We further visualized the number of problems taught before tutors could pass a quiz level across conditions, illustrated in **Fig 6.** As shown in the plot, *tutors in AdaptiveFollowup condition passed the second quiz level with much fewer problems (18) taught in average compared to StaticFollowup (35) and NoFollowup (48) tutors.*

5.3 AdaptiveFollowup tutors spent more time on average to answer TA questions that helped them teach problems accurately

Why did **A**daptive**F**ollowup tutors pass the quiz by teaching fewer problems? Our naïve hypothesis conjectures that tutors in the **A**daptive**F**ollowup condition spent more time on answering the questions that facilitated accurate problem-solving, thereby enabling them to pass the quiz with fewer problems taught.

To test this hypothesis, we began by calculating the average time spent by tutors on various activities while teaching each problem within APLUS. Activities include question answering (QA), teaching (T), reviewing resource tabs like quiz (Qu), example (Ex), unit overview (Uo), problem bank (Pb), and introduction video (Iv). We conducted separate mixed model analysis with each of the activity duration per problem as a dependent variable while condition as fixed factor and tutors as random factor (shown in **Table 2).** The data suggest that *AdaptiveFollowup tutors spent reliably more time on QA (21.4s on average) compared to Static (8.9s) and NoFollowup (2.8s) tutors per problem.*

Our next aim is to understand how time spent on these activities relates to problem-solving accuracy, measured as the percentage of correctness (*%correctness*). *%correctness* was calculated per problem based on the ratio of correctly demonstrated steps and feedback to the total number of steps and feedback provided for that problem. We employed a linear regression model with *%correctness* as the dependent variable and prior groupings based on pre-test scores as the first term, followed by significant activities found in our previous mixed model analysis (i.e. QA, Ex, and Qu), condition, and their interactions. The result revealed a significant interaction between time spent on QA and condition; $F_{Condition:QA}$ (2, 2279) = 9.5, p<.05 and Example tab and condition; $F_{Condition:Example}$ (2, 2283) = 5.5, p<.05. Other interaction terms were not main effects. The regression model suggests that *spending 1 minute more on adaptive follow-up questions results in 7.8% increase in problem-solving accuracy in APLUS, which is a notable correlation given that AdaptiveFollowup tutors spent 10 min on QA in average for all the problems taught in APLUS as shown in* ***Fig 6a.***

6 Discussion

Our first research question was, RQ1: Does answering expectation tailored adaptive follow-up questions help tutors learn? Our data revealed that the post-test improvement

	Condition						*F*	*p*
	NoFollowup		**S**tatic**F**ollowup		**A**daptive**F**ollowup			
	M	*SD*	*M*	*SD*	*M*	*SD*		
QA	2.8[a]	1.3	8.9[b]	5.1	21.4[c]	8.0	$[\frac{2}{32}]$ 26.82	< .05
Qu	13.1[a]	6.3	13.5[a]	9.1	49.2[c]	24.4	$[\frac{2}{29}]$ 55.35	< .05
Ex	6.9[a]	6.2	3.1[a]	2.8	13.9[c]	15.2	$[\frac{2}{35}]$ 6.52	< .05
Uo	0.8	0.9	1.7	1.7	3.3	6.1	$[\frac{2}{31}]$ 2.98	0.07
Pb	1.7	1.6	0.9	0.8	3.1	3.2	$[\frac{2}{43}]$ 1.83	0.17
Iv	1.2	2.0	0.6	0.7	2.3	3.2	$[\frac{2}{55}]$ 1.23	0.29
T	61.1	33.3	46.0	42.7	67.8	15.2	$[\frac{2}{31}]$ 1.94	0.15

Table 2: Average time spent (s) across different activities per problem.
Means that do not share superscripts differ significantly at $p < 0.05$ in the post-hoc analysis

is higher for tutors who answered any follow-up questions (both static and adaptive) compared to tutors who only answered the focal questions followed by no follow-up. Tutors who answered adaptive follow-up questions ended up having the same post-test improvement as the tutors who answered static follow-up questions.

Our next research question was, RQ2: Are adaptive follow-up questions more effective than static follow-up questions? Our data revealed that tutors spent more time answering the adaptive follow-up questions than static follow-up questions. We also found a strong correlation between time spent on answering adaptive follow-up questions and accuracy in solving problems. This observation suggests that *spending more time on adaptive questions helped tutors solve problems more accurately, which resulted in teaching fewer problems in APLUS for passing the quiz levels*.

This efficiency in learning was particularly evident in the early quiz levels as shown in **Fig 6b.** The relatively low number of problems taught before passing the first quiz level can be attributed to the APLUS design. APLUS allows tutors to pass the first level automatically if they quiz SimStudent after launching the app for the first time, as advised in the intro video. Similarly, teaching problems more accurately in the second and third levels increases the likelihood of SimStudent passing the fourth level without requiring additional problems taught. This is because fourth level involves equations with variables on both sides that have similar difficulties to the third level. The fewer number of problems taught at the second and third levels suggests that *AdaptiveFollowup tutors enhanced their problem-solving accuracy effectively by engaging with adaptive questions*. In contrast, StaticFollowup tutors, despite engaging with static questions, did not achieve the same level of problem-solving accuracy, leading to their need for teaching more problems to pass the quiz.

In this paper, we proposed ExpectAdapt, an expectation tailored follow-up question framework for teachable agents using large language models and showed that tutors learned efficiently by answering adaptive questions. In this work, we narrow our focus to learning outcomes to assess the efficacy of our framework. One of our future works includes delving into the cognitive level of the adaptive questions.

The observation that tutor learning outcomes were comparable between static and adaptive questions is intriguing. One possible explanation is that teaching many problems with "shallow" question answering could be as effective for tutor learning as teaching fewer problems with elaborated question answering. In other words, the quantity of problems tackled could be as crucial as their quality. Further research could explore the optimal balance between problem quantity and quality, as well as the differential impacts these factors have on learning gains in educational settings.

7 Conclusion

We proposed ExpectAdapt, a novel follow-up questioning framework for the teachable agent using large language models that generates follow-up questions adapting based on tutors' contributions to the conversation history. We found that adaptive follow-up questions facilitated tutor learning by ensuring productive use of instructional time. Our current data demonstrated that tutors interacting with ExpectAdapt's questions exhibited greater improvement from pre- to post-test than interacting with focal questions only. We also found that while tutors achieved equivalent learning outcomes

when responding to adaptive as opposed to static questions, the former demanded a higher level of engagement, as evidenced by extended question answering durations. This extended duration with adaptive question answering correlated with improved problem-solving abilities within the APLUS learning environment.

Our research provides strong evidence supporting the use of large language models to generate adaptive follow-up questions. Furthermore, in-context learning capabilities of these models provide an opportunity to incorporate expert knowledge. This results in more polished and insightful question generation with minimal efforts and easily scalable across educational domains.

Acknowledgment: This research was supported by the Institute of Education Sciences, U.S. Department of Education, through grant No. R305A180517 and National Science Foundation grant Nos. 2016966 and 2112635 to North Carolina State University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education and NSF.

8 References

1. Okita, S.Y., J. Bailenson, and D.L. Schwartz, *Mere Belief of Social Action Improves Complex Learning*, in *Proceedings of the International Conference for the Learning Sciences*, K.H. S. Barab, D. Hickey, Editor. 2008 in press, Lawrence Erlbaum: New Jersey.
2. Roscoe, R.D. and M.T.H. Chi, *Understanding tutor learning: Knowledge-building and knowledge-telling in peer tutors' explanations and questions.* Review of Educational Research, 2007. **77**(4): p. 534-574.
3. Chi, M.T.H., et al., *Learning from human tutoring.* Cognitive Science, 2001. **25**: p. 471-533.
4. Graesser, A.C., N.K. Person, and J.P. Magliano, *Collaborative dialogue patterns in naturalistic one-to-one tutoring.* Applied Cognitive Psychology, 1995. **9**(6): p. 495-522.
5. Roscoe, R.D., *Opportunities and barriers for tutor learning: Knowledge-building, metacognition, and motivation*. 2008, University of Pittsburgh.
6. Shahriar, T. and N. Matsuda. *What and How You Explain Matters: Inquisitive Teachable Agent Scaffolds Knowledge-Building for Tutor Learning*. 2023. Cham: Springer Nature Switzerland.
7. Roscoe, R.D. and M. Chi, *Tutor learning: the role of explaining and responding to questions.* Instructional Science, 2008. **36**(4): p. 321-350.
8. Roscoe, R.D., *Self-monitoring and knowledge-building in learning by teaching.* Instructional Science, 2014. **42**(3): p. 327-351.
9. Shahriar, T. and N. Matsuda. *"Can you clarify what you said?": Studying the impact of tutee agents' follow-up questions on tutors' learning*. in *International Conference on Artificial Intelligence in Education*. 2021. Springer.
10. Peterson, D.S. and B.M. Taylor, *Using higher order questioning to accelerate students' growth in reading.* The Reading Teacher, 2012. **65**(5): p. 295-304.
11. Otero, J. and A.C. Graesser, *PREG: Elements of a model of question asking.* Cognition and instruction, 2001. **19**(2): p. 143-175.
12. Azevedo, R. and J.G. Cromley, *Does Training on Self-Regulated Learning Facilitate Students' Learning With Hypermedia?* Journal of Educational Psychology, 2004. **96**(3): p. 523-535.

13. Matsuda, N., et al., *Learning by Teaching SimStudent – An Initial Classroom Baseline Study comparing with Cognitive Tutor*, in *Proceedings of the International Conference on Artificial Intelligence in Education*, G. Biswas and S. Bull, Editors. 2011, Springer: Berlin, Heidelberg. p. 213-221.
14. Li, N., et al., *Integrating representation learning and skill learning in a human-like intelligent agent.* Artificial Intelligence, 2015. **219**: p. 67-91.
15. Graesser, A.C., *Conversations with AutoTutor help students learn.* International Journal of Artificial Intelligence in Education, 2016. **26**(1): p. 124-132.
16. Brown, T., et al., *Language models are few-shot learners.* Advances in neural information processing systems, 2020. **33**: p. 1877-1901.
17. Radford, A., et al., *Language models are unsupervised multitask learners.* OpenAI blog, 2019. **1**(8): p. 9.
18. Devlin, J., et al., *Bert: Pre-training of deep bidirectional transformers for language understanding.* arXiv preprint arXiv:1810.04805, 2018.
19. Liu, J., et al., *What Makes Good In-Context Examples for GPT-$3 $?* arXiv preprint arXiv:2101.06804, 2021.
20. Zhang, Z., et al., *VISAR: A Human-AI Argumentative Writing Assistant with Visual Programming and Rapid Draft Prototyping.* arXiv preprint arXiv:2304.07810, 2023.
21. Liu, J., et al., *Generated knowledge prompting for commonsense reasoning.* arXiv preprint arXiv:2110.08387, 2021.
22. Shahriar, T., N. Matsuda, and K. Ramos, *Assertion Enhanced Few-Shot Learning: Instructive Technique for Large Language Models to Generate Educational Explanations.* arXiv preprint arXiv:2312.03122, 2023.
23. Razeghi, Y., et al. *Impact of pretraining term frequencies on few-shot numerical reasoning*. in *Findings of the Association for Computational Linguistics: EMNLP 2022*. 2022.
24. Min, S., et al., *Rethinking the role of demonstrations: What makes in-context learning work?* arXiv preprint arXiv:2202.12837, 2022.
25. Wei, J., et al., *Chain-of-thought prompting elicits reasoning in large language models.* Advances in Neural Information Processing Systems, 2022. **35**: p. 24824-24837.
26. Matsuda, N., D. Lv, and G. Zheng, *Teaching How to Teach Promotes Learning by Teaching.* International Journal of Artificial Intelligence in Education, 2022: p. 1-32.
27. Matsuda, N., et al., *Studying the Effect of Tutor Learning using a Teachable Agent that asks the Student Tutor for Explanations*, in *Proceedings of the International Conference on Digital Game and Intelligent Toy Enhanced Learning (DIGITEL 2012)*, M. Sugimoto, et al., Editors. 2012, IEEE Computer Society: Los Alamitos, CA. p. 25-32.
28. Koedinger, K.R., et al., *A Data Repository for the EDM community: The PSLC DataShop*, in *Handbook of Educational Data Mining*, C. Romero, et al., Editors. 2010, CRC Press: Boca Raton, FL.
29. Ling, C., et al., *Beyond One-Model-Fits-All: A Survey of Domain Specialization for Large Language Models.* arXiv preprint arXiv:2305.18703, 2023.
30. Johns, J.P., *The relationship between teacher behaviors and the incidence of thought-provoking questions by students in secondary schools.* The Journal of Educational Research, 1968. **62**(3): p. 117-122.