

# Towards Just-in-Time Adaptive Feedback: Enhancing Student Learning via Knowledge-Grounded LLM

Younghun Lee<sup>†</sup>, Amir Bralin<sup>‡</sup>, Nobel Sanjay Rebello<sup>†§</sup>, Dan Goldwasser<sup>†</sup>

<sup>†</sup>Department of Computer Science

<sup>‡</sup>Department of Physics and Astronomy

<sup>§</sup>College of Education

Purdue University

{younghun, abralin, rebellos, dgoldwas}@purdue.edu

## Abstract

Educational interventions are effective tools for enhancing student learning. While Large Language Models (LLMs) allow for generating adaptive feedback at scale, current studies lack clear methodologies for providing Just-in-Time (JiT) feedback in authentic instructional settings. In this paper, we present a framework that provides adaptive feedback by grounding LLMs with domain-specific expert knowledge. Our approach collects written reasoning logic (strategy essays) from students, analyzes potential error types based on the content of that reasoning, and delivers non-intrusive feedback designed to clarify missing or incorrect concepts. We deploy this framework in a large-scale university course ( $N > 1,000$ ), where it improved student performance by over 80% compared to previous semesters. Lastly, we validate the framework’s pedagogical utility by analyzing the learning trajectories; we demonstrate how iterative conversations with LLM facilitate shifting one’s misconception to correct understanding.

## 1 Introduction

In STEM education, students are often observed to engage in “Recursive Plug-and-Chug”—also known as “Formula Hunting”—where the goal is simply to fill a slot in a formula rather than to make sense of the problem (Chi et al., 1981; Tuminaro and Redish, 2007). Existing studies indicate that while experts analyze problems based on their deep logical structure and principles, novices are often primed to surface features, leading to unsuccessful problem-solving approaches. (Mestre et al., 1993).

Strategy writing has emerged as a powerful intervention to tackle this problem. Research has shown that asking students to articulate their strategies for solving problems can improve their problem-solving skills (Leonard et al., 1996). With a combination of argumentation and appropriate prompts,

students focus more on deep structure rather than the surface features of the problem and starts solving the problem with conceptual analysis, avoiding novice, unproductive strategies (Dufresne et al., 1992; Mestre et al., 1993; Docktor et al., 2010; Rebello and Piedrahita Uruena, 2019).

Despite the benefits, strategy writing is difficult to implement at scale. The primary impediment is the feedback bottleneck. Feedback is most effective when it is integrated into the learning process through formative assessment (Hattie and Timperley, 2007), and provided prior to completion (Henderson et al., 2021). However, providing real-time, formative feedback to thousands of students is prohibitively time-consuming for human instructors. Thus, the true potential of strategy writing to enhance student learning remains largely unrealized. The advance of generative Large Language Models (LLMs) offers a promising pathway toward overcoming the scalability barrier. Recent work proposes adaptive scaffolding frameworks that utilizes LLMs to help students understand course materials (Taneja et al., 2024; Kweon et al., 2025; Kestin et al., 2025; Zhang et al., 2025). However, they lack exploration of intervention-oriented interactions as well as grounding the models with domain-specific knowledge.

In this paper, we address these challenges by exploring dimensions of adaptive feedback and implementing a framework that provides Just-in-Time interventions using knowledge-grounded LLMs. First, we collect strategy essays from students over a couple of semesters and obtain historical data that connect the essays to the types of errors students had made. Human experts then annotate feedback on a small portion of the data considering the content of the essays, as well as the error types. Using these annotated instances as few-shot examples, the LLM identifies misconceptions in the strategy essays and provide real-time, adaptive feedback. Rather than providing a ‘shortcut’ to the correct

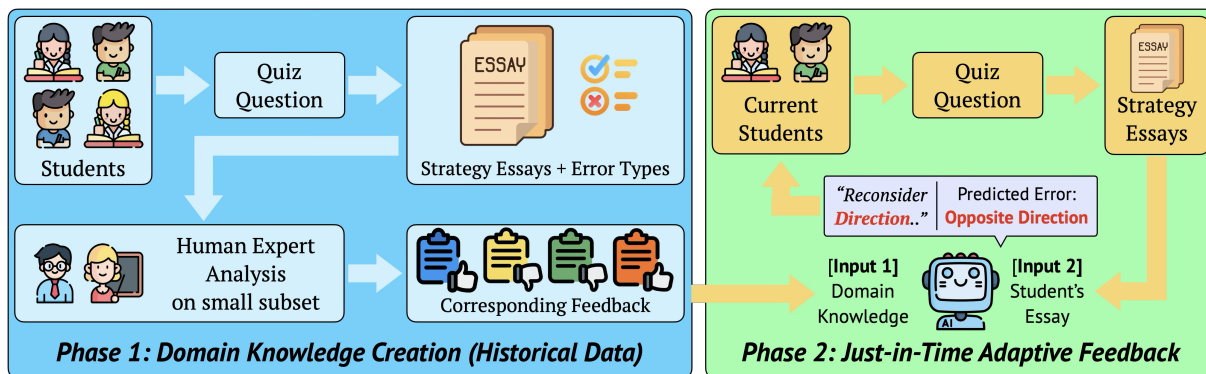


Figure 1: **Overall framework** of our Just-in-Time adaptive feedback LLM. In the first phase, we obtain domain knowledge from human experts regarding the appropriate feedback for students’ strategy essays and the types of errors they made. At deployment, we ground the LLM with this knowledge and provide adaptive feedback to the students based on their written strategy essays.

answer, the feedback directs the student’s attention to the underlying concepts and their reasoning logic (see Figure 4 and Figure 7 for examples). This ensures the intervention disrupts the ‘formula-hunting’ heuristic by letting students reflect on the problem’s deep structure. Figure 1 illustrates the overall framework we propose.

We experiment our framework on a large-scale university physics course ( $N > 1,000$ ). Experimental results show that the overall student performance is improved when the feedback framework is deployed, by over 80% compared to previous semesters. Based on further analyses on self-reported survey and conversational instances, we argue that the framework effectively facilitates a shift from initial misconceptions to the correct understanding.

## 2 Related Work

Educational research has shown that feedback is one of the important drivers of learning (Wisniewski et al., 2020; Foster, 2024). Feedback can facilitate improvements in learners’ understanding and skills (Henderson et al., 2021) by narrowing the gap between actual and desired performance (Burgess et al., 2020). The effectiveness of feedback increases with the information that it contains. Previous research (Kluger and DeNisi, 1996) has also shown that moderators such as timing, specificity, and task complexity affect how learners receive and use feedback (Hattie and Clarke, 2018; Brooks et al., 2019).

Feedback is effective if it is sufficiently detailed (Price et al., 2010), usable (Winstone et al., 2017), and facilitates change (Ryan et al., 2016), such

that learners can test their new understandings (Pitt and Norton, 2017). In asynchronous and isolated online settings (Orlando, 2016), interactive dialogues can be especially useful (Wolsey, 2008) as students cannot easily interact with their peers (Furnborough and Truman, 2009) which put significant weight on the feedback comments they receive (Ortiz-Rodríguez et al., 2005).

Recent research on educational LLM applications has shifted from general assistance to specialized interventions. For instance, Phung et al. (2024) used GPT-4 as a teacher to provide non-intrusive hints but relied on LLM-simulated student agents using a weaker LLM (GPT-3.5), rather than applying it to real classroom settings. In contrast, Dai et al. (2023) and Jia et al. (2024) designed LLM feedback to apply to students, but their evaluation metrics focused on student-reported helpfulness survey and qualitative analysis of feedback content, rather than measuring actual improvements in student performance. More recent studies have shown LLM-generated feedback can help improve student performance in class. Hashmi and Rebello (2025) used an LLM-based Socratic chatbot to scaffold expert-like reasoning. Zhang et al. (2025) implemented a persona-driven multi-agent dialogue system, showing the student’s performance in problem-solving improves as they interact more with LLMs. Kestin et al. (2025) integrated an AI tutor in a college Physics course and compared student’s performance on a quiz between an AI tutor and in-class lessons. The results show that when the model is applied to a class size of 200, the AI tutor almost doubled the learning gains compared to in-class lessons.

In this paper, we introduce three novel contri-

butions. Unlike reactive chatbots that are susceptible to formula-hunting behaviors of the students, our system proactively analyzes the reasoning logic within the strategy essays that guides students to focus on deep structures in solving quiz problems. Additionally, we ground LLMs with domain-specific knowledge that is carefully annotated by human experts, which brings significant benefits compared to persona-based prompt engineering. Lastly, we provide a robust, automated intervention that remains pedagogically sound at a massive scale ( $N > 1,000$ ).

### 3 Problem Formulation

#### 3.1 Instructional Setting

Our research context is a large enrollment calculus-based physics course for engineers and physical scientists at a large U.S. Midwestern land grant university. The focus of the course is the development of problem-solving skills focused on applying key physics principles pertaining to mechanics across three units: Newton’s Laws and linear momentum, work and energy, and angular momentum. The annual enrollment of the course is around 3,300 students (1,500 in fall, 1,800 in spring), about 25% women, 10% underrepresented minorities, and 8% international students.

Each week, students complete an online quiz administered via the Learning Management System, along with an online proctoring system to maintain integrity. Students have 40 minutes to complete each online quiz. In this work, we focus on the quiz data from four consecutive semesters, from the Fall 2024 to the Spring 2026 semester.

#### 3.2 Strategy Essays

Strategy essays refer to the student’s written reasoning logic in solving the quiz problem (Leonard et al., 1996). For one of the quiz questions, students are instructed to write a strategy essay that is at least 50 words long and does not include any numbers, symbols, or formulae. This was designed to evaluate the consistency between a student’s written reasoning logic and their performance on the quizzes. To maintain the quality of the essays, the instructors offer extra credit<sup>1</sup> for writing a strategy essay. In the Fall 2024 semester, for example, we gathered 11,948 essays from 1,418 students from a total of 11 quizzes.

<sup>1</sup>Extra credits equal to 10% of the total points on the quiz

#### 3.3 Desired Properties of LLM Feedback

LLM feedback refers to the LLM-generated text that is given to each student. The primary purpose of LLM feedback is to help students better understand the course materials related to the quiz problem and solve it correctly.

One of the most critical constraints in providing feedback is **to facilitate the student’s own problem-solving process**. The LLM should not disclose a direct solution or correct answers to the quiz problems when generating feedback to students. Another aspect we consider in designing the feedback is **adaptiveness**; rather than generating universal feedback, it is conditioned on student-specific features such as their in-class performance (e.g. pre-semester assessment scores, midterm exam scores, other quiz scores, etc.) and their strategy essays. Lastly, we focus on **grounding LLMs with domain-specific knowledge**. General-purpose LLMs often exhibit unreliable performance in understanding college-level STEM courses that require multi-step logic (Arora et al., 2023; Pang et al., 2025). To minimize the knowledge gap, we distill domain-specific knowledge from experts (i.e. instructors) and ground LLM generations with it.

#### 3.4 Post-Hoc vs. Just-in-Time (JiT)

LLM feedback can be presented to students in two different ways: after students finish solving the quiz (post-hoc) and in real-time while they solve it (JiT). These two methods serve different benefits. In this paper, we utilize LLM post-hoc feedback to mainly survey the student’s preference for the feedback style. Students’ preference patterns ultimately help identify how LLM feedback needs to be generated. We consider LLM JiT feedback as a means of real-time interventions which help students solve the quiz problem.

### 4 Designing Adaptive Feedback

#### 4.1 Student Preference Survey

Existing studies suggest the necessity of adaptive guidance that aligns with the student’s level of knowledge; unlike the way novice students perform better with detailed explanations, more experienced learners find it redundant and even show negative consequences (Kalyuga, 2007; Walker et al., 2012; Albacete et al., 2019). Since there is limited work that verifies this finding with LLM-based feedback, we conduct a series of surveys to find out whether

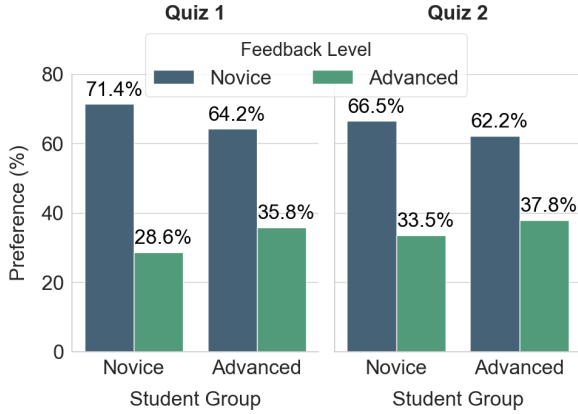


Figure 2: **Survey results regarding preference** between novice and advanced feedback in two quizzes. Regardless of their level of knowledge, students prefer LLM feedback that is targeted to the novice group.

students with different knowledge levels show a contrasting preference for the complexity of LLM feedback.

We provide post-hoc LLM feedback to the students for two quizzes in the Fall 2025 semester. Students are asked to write a short strategy essay that they employ to solve the problems. After having strategy essays from the students, we prompt LLMs to generate two different versions of the feedback, one for the novice-level and the other for the advanced-level. LLMs are instructed to differentiate not only linguistic styles such as tone and vocabulary, but also how feedback is framed with respect to the focus and goal<sup>2</sup>. Students are instructed to participate in the survey on a separate webpage showing two feedback texts, one targeting the novice-level and the other for the advanced-level, and are asked to indicate which version they prefer. The presentation order of the feedback was randomized to mitigate potential order effects and minimize response bias. Figure 10 shows the survey webpage (see Appendix A).

To find the relevance between the student’s level of knowledge and the complexity of the LLM feedback, we categorize students into two buckets, novice and advanced. Students were categorized as novice if they performed below average for all relevant assessments; this includes a pre-semester assessment<sup>3</sup>, mid-term exam, and relevant previous quizzes. In contrast, students who consistently exceeded the average score for all relevant assess-

<sup>2</sup>We used GPT-5.1 to generate post-hoc feedback. Detailed settings for prompting are described in Appendix A

<sup>3</sup>Energy and Momentum Conceptual Survey (Singh and Rosengrant, 2003)

**Quiz Problem**

**Question**

What is the horizontal component of the force that top block exerts on the bottom block?

**Answer Options**

A. $-m_2 \frac{\vec{F}}{m_1 + m_2}$	B. $m_2 \frac{\vec{F}}{m_1 + m_2}$
C. $-m_1 \frac{\vec{F}}{m_1 + m_2}$	D. $m_1 \frac{\vec{F}}{m_1 + m_2}$

Figure 3: **Example quiz problem.** While A is correct, other options map to common errors: B represents a direction error (misunderstanding force directions); C represents a position error (incorrect mass/location consideration); and D represents confusion involving both concepts ( position-direction error).

ments were categorized as advanced.

Figure 2 shows the results. Although the portion of students who favor advanced-level feedback is larger among students with advanced knowledge, both novice and advanced groups prefer feedback that is targeted to the novice group. This implies that LLMs are inadequate to generate advanced feedback that is pedagogically effective, suggesting that simply instructing LLMs to increase vocabulary complexity and reduce scaffolding may hinder rather than help the learning process.

## 4.2 Just-in-Time Intervention

Using insights from the survey results, we focus on a different aspect of adaptive LLM feedback for JiT intervention: **error types**.

All incorrect multiple-choice options in the quizzes are designed to elicit specific error types. Consider an example quiz in Figure 3. The question asks to find the force that the top block exerts on the bottom block, given a certain amount of horizontal force,  $\vec{F}$ , is applied to the bottom block and the two blocks move together. The correct answer is multiplying the mass of the top block by acceleration which is the opposite direction to the  $\vec{F}$ . Common errors can be made by not considering that the direction has to be opposite (B. direction

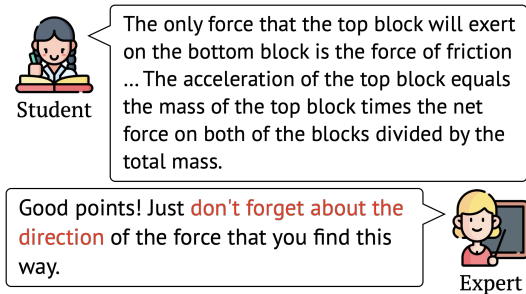


Figure 4: Example of a strategy essay and annotated feedback. Although the student in this example got the correct answer, the feedback mentions the importance of considering direction as it was missing in the essay.

error), or by multiplying the mass of another object (C. position error), or by having confusion involving both concepts (D. position-direction error). We posit that LLMs offer effective interventions by predicting a student’s likely error type and providing adaptive feedback that guides them to reconsider those aspects in their final answer.

### 4.3 Grounding LLM Feedback with Domain Knowledge

As stated in 3.3, LLMs need to be grounded with domain-specific knowledge to provide more effective feedback with a better understanding of the concepts. We integrate human-expert knowledge into LLMs’ feedback generation process.

Using historical data from previous semesters, we collected students’ strategy essays and their corresponding answer choices, enabling us to associate written reasoning with particular error types. Teaching experts (i.e. faculty and instructors) then manually wrote feedback tailored to the content of these essays and error types.

Such nuanced feedback annotation is naturally a challenging task for LLMs. For example, students who selected the correct answer might still write imperfect strategy essays by not explicitly mentioning the force direction. Conversely, a semantically similar essay written by another student might lead to the direction error. Human experts are more competent in capturing these subtle differences between strategy essays and answer choices. Figure 4 illustrates an example of the annotation process, where experts identify missing concepts regardless of whether the student’s final answer was correct.

Once the feedback annotation is completed for a small representative set (50 examples), they serve as the few-shot examples for in-context learning, enabling the real-time classification of error types

Method	Accuracy	Macro F1
Fine-tuned Classifier	50.00 $\pm$ 0.001	28.39 $\pm$ 0.001
Zero-shot LLM	21.00 $\pm$ 0.004	12.74 $\pm$ 0.006
Zero-shot LLM w/ Secondary	37.00 $\pm$ 0.003	28.34 $\pm$ 0.005
Few-shot LLM	42.42 $\pm$ 0.001	33.50 $\pm$ 0.002
Few-shot LLM w/ Secondary	<b>60.61</b> $\pm$ 0.002	<b>54.24</b> $\pm$ 0.003

Table 1: Classification performance of different models in predicting error types (i.e. correct, direction, position, position-direction) given strategy essay. LLMs perform better when they are prompted with few-shot examples and Chain-of-Thought reasoning while considering secondary label.

using student’s strategy essays. Recognizing that strategy essays are often non-trivial to categorize into potential error types, we move beyond single-label prediction; we prompt the LLMs to identify both the primary and secondary most likely labels. The model then produces feedback via Chain-of-Thought reasoning (Wei et al., 2022), conditioned on this expanded label space to ensure more robust instructional support. Detailed explanations regarding prompts and JiT intervention framework are provided in Appendix C.

## 5 Experiments and Results

### 5.1 Error Type Predictions

In this experiment, we measure the accuracy of error type prediction across different models. The task is formulated as a multi-class classification where the model gets a strategy essay as input and predicts one of four classes: correct, or one of the three error types—direction, position, and position-direction. Table 1 describes the accuracy and macro F1 score of the models.

As a baseline, we first implement a sentence classifier using a pre-trained BERT-based checkpoint using the quiz data from Fall 2024 semester. Since the number of data instances is small (1.1K), the trained model shows zero accuracy on labels with insufficient amount of training instances, results in a very low macro F1 score.

As the number of training data instances is limited, we use high-performance LLMs to classify the error types without fine-tuning. We try prompting the LLMs with zero-shot and few-shot settings (3 examples per label). Although adding few-shot examples improve the performance, the accuracy and macro F1 scores are not strong enough. After we prompt the model to consider the secondary most likely labels, the performance increases signif-

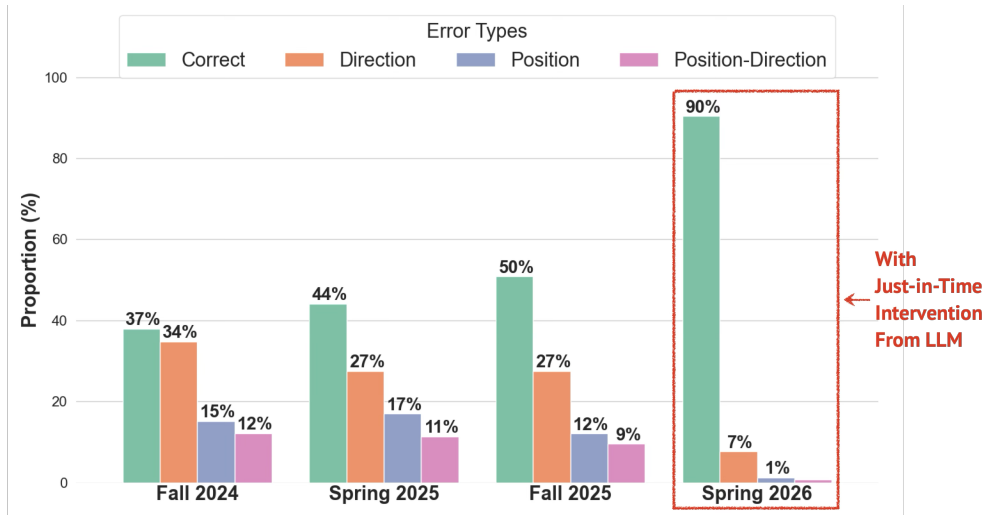


Figure 5: Student’s performance in the same quiz question over four semesters. When the students solved the problem with the LLMs providing JiT adaptive intervention, the overall portion of correct students boosted significantly.

icantly, implying that the LLM can generate more robust feedback with expanded label space. Detailed explanations regarding prediction models are described in Appendix B.

## 5.2 Quiz with JiT Interventions

We administered JiT interventions during a specific quiz in the Spring 2026 semester. The problem and its corresponding multiple-choice options are illustrated in Figure 3. Students accessed the quiz through their standard Learning Management System, with the LLM feedback interface integrated directly into the problem description. A comprehensive view of the user interface and the interaction flow is provided in Figure 13 (see Appendix C). Upon engaging with the system, students were informed that the LLM would not provide direct solutions or answers. Instead, they were instructed to use the framework to get advice on specific concepts requiring further attention, based on the content of their strategy essays.

## 5.3 Student Performance Improvements

To measure the effectiveness of LLM-generated feedback, we compare student performance on the same quiz question over four consecutive semesters. As shown in Figure 5, more than half of the class in previous semesters made errors. In contrast, the error rate drops to less than 10% with the help of JiT LLM intervention.

Notably, the feedback remains highly effective despite sub-optimal performance of LLMs in predicting error types. We attribute this to domain-

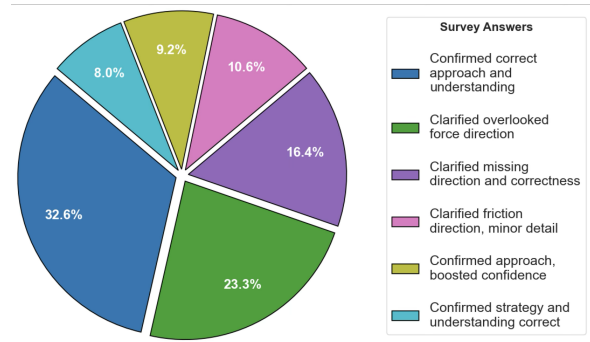


Figure 6: Survey results regarding why LLM’s feedback was helpful. Aside from confirmation, students indicate that LLM was useful in clarifying overlooked / missing aspects in the essays

knowledge grounding, which incorporated expert-generated feedback that addresses potentially missing or incorrect concepts regardless of the predicted error class.

After students solve the quiz with LLM feedback, we ask whether the feedback was helpful and why. Approximately 77.59% of the students reported the feedback was helpful; their qualitative responses are clustered using HDBSCAN, and shown in Figure 6.

There are two major themes emerged from the clusters. Around half of the students valued the feedback for concept verification and confirmation, while the other half stated that the LLMs clarified overlooked or missing items in their strategy essays. This illustrates the adaptiveness of JiT LLM feedback; for students with a correct understanding,

Metric	Value
Total instances	1,042
Conversational instances	209 (20.05%)
Mean # conv. turns (std)	2.77 (1.52)
Min/Max # conv. turns	2 / 14
Conv. turn dist. skewness	3.57
<b>% of Model Prediction as Correct</b>	
Turn 1 (Initial essay)	42.79%
Last turn (Final essay)	71.64%

Table 2: Statistics of conversational instances. The skewness of the number of conversation turn distributions is computed by Fisher-Pearson coefficient of skewness where the value greater than 2 is normally considered extremely skewed.

the LLM provides essential validation, whereas for those prone to errors, it guides them rethink or pay more attention to specific concepts.

#### 5.4 Learning through Conversations

One of the emergent behaviors of students in interacting with feedback-generating LLM is that they have conversations with the LLMs to improve their understanding. This observation is impressive considering that students are not directed or incentivized to iteratively chat with the LLMs. Students are notified as solving the quiz with LLMs will give them a bonus credit regardless of the correctness of their answers. Nevertheless, there are around 20% of the students who voluntarily spent more time interacting with the LLM so that they could improve their learning experience. Basic statistics of the conversational instances are reported in Table 2, and more detailed statistics are reported in Appendix D.

Figure 7 shows an example of conversation between a student and LLM. As the student receives LLM’s feedback that tackles missing or incorrect points in the strategy essay, they add these components to their strategy essay. By continuing conversations, students incrementally add missing points and gain more comprehensive understandings. Table 2 illustrates such instances: while only 43% of students’ initial essays were predicted as correct, the portion boosted to 72% after having conversations.

We visualize learning trajectories by analyzing the sequence of conversation turns with the LLM. This involves tracking the intermediate strategy essays and their corresponding predicted error types throughout the interaction. Figure 8 illustrates how students modify their strategy essays and how the

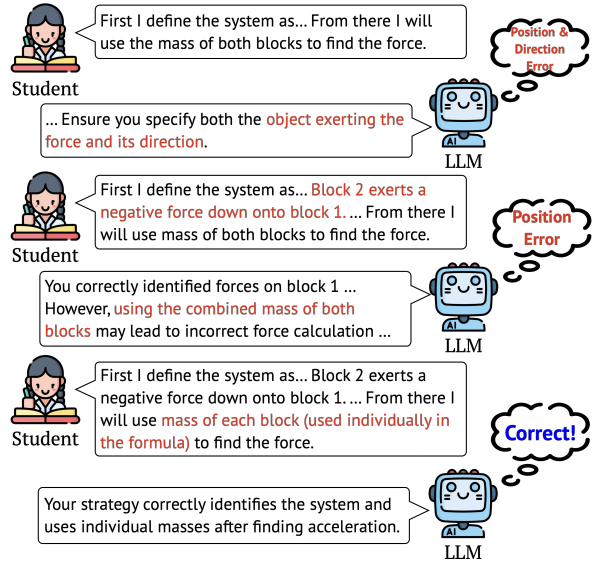


Figure 7: **Example Conversation** of a student modifying their strategy essays with the help of LLM. LLM generates feedback based on its error type prediction, and informs different aspects as the student fixes errors one by one.

predicted error types change while they have conversations with LLMs. The figure represents a subset where students’ strategy essays are initially classified as having a direction error.

As conversation proceeds, students keep modifying their essays by applying LLM’s advice, and there exists multiple learning trajectories based on how they improve their essays. In the end, approximately 80% of the students corrected their essays and the LLM classified them as ‘correct’. Other students, even though their final version of essays are not classified as ‘correct’, they solved the quiz correctly. We assume that this is the case where students skip the final validation step after having their perspectives corrected. Overall, 91.43% of the students (32 out of 35) solved the quiz correctly by iteratively getting feedback from LLM and developing their understandings.

We conducted a further analysis to validate the hypothesis that students who engage more actively with the LLM feedback improve their strategy essays better. We defined “activeness” through two metrics: reflective latency (the time in seconds between conversation turns) and degree of revision (the change in word count between consecutive essay versions). Specifically, we posit that spending more time refining an essay or making more substantial text changes indicates a more active adaptation of the feedback.

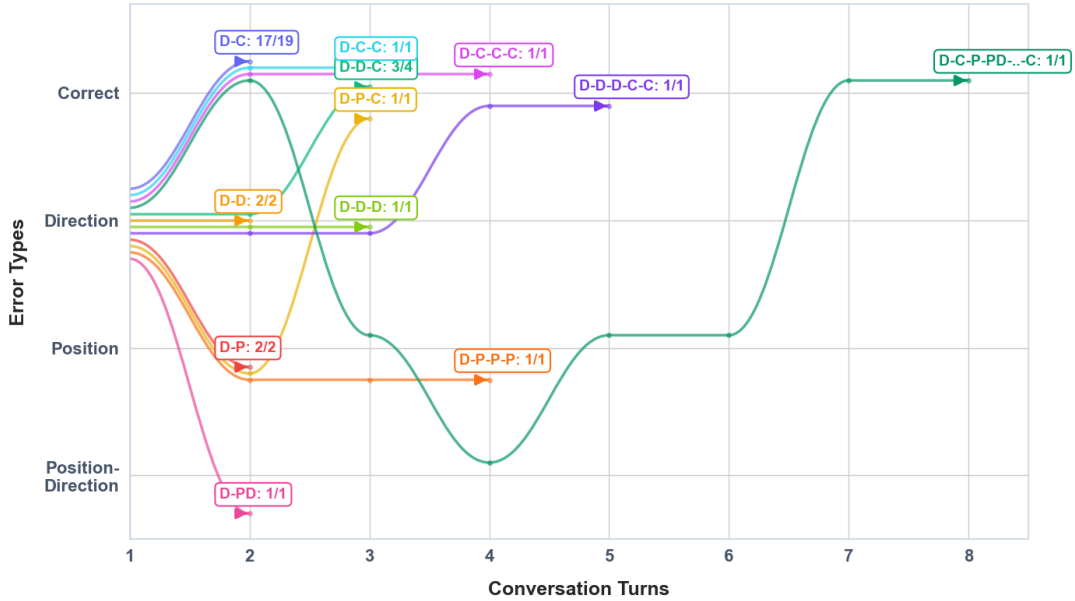


Figure 8: **Learning Trajectories** for students starting with Direction error. Each trace represents a sequence of essay error types as predicted by the feedback-LLM at each conversation turn. Terminal markers are labeled with the specific Path Sequence and the Path Outcome. For instance, ‘D-C: 17/19’ indicates that out of 19 students whose essays the LLM classified as ‘Direction error’ and then ‘Correct’, 17 got correct answers in the end.

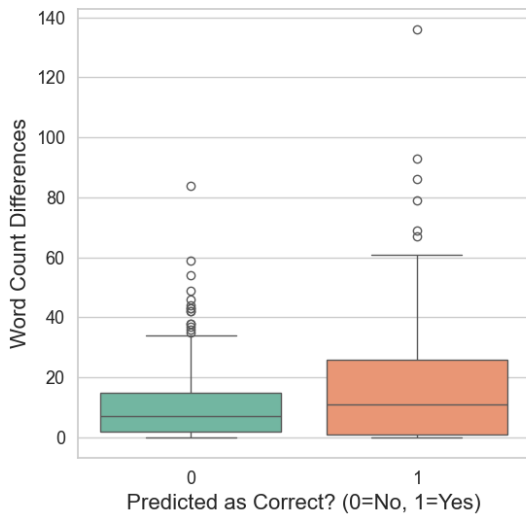


Figure 9: **Correlation between word count differences and essay correctness.** We compute the correlation between the word count differences in strategy essays between each conversation utterances and the correctness of essay. When the students make more changes to their essays, they are more likely getting their essays correct.

To evaluate these hypotheses, we compute the Pearson correlation coefficients. The results indicate a negligible relationship between time spent and essay correctness ( $r = 0.06$ ,  $p = 0.27$ ). In contrast, the correlation between word count dif-

ference and essay correctness is statistically significant ( $r = 0.18$ ,  $p < 0.001$ ). Figure 9 illustrates how word count differences are distributed differently for the essays that are classified as correct or not. These findings suggest that simply spending more time on the interface is insufficient; rather, the degree of revision is the primary factor of success. This implies that while the LLM provides the necessary feedback, the likelihood of achieving a correct understanding depends heavily on the student’s willingness to actively reformulate their reasoning.

## 6 Conclusion

In this paper, we explore how LLMs can be utilized efficiently in higher education through Just-in-Time (JiT) adaptive interventions. By analyzing strategy essays from students, our framework identifies potential errors and provides adaptive feedback grounded in domain knowledge. Experimental results show that our framework significantly improves the overall student performance by providing validation and guiding re-evaluation of incorrect concepts. Qualitative analysis further demonstrates that students engage in conversations with LLMs to improve their understanding, highlighting the potential for LLMs to serve as sophisticated pedagogical tools.

## Limitations

One of the limitations of this work is the sub-optimal accuracy of the error type classification. Predicting error types solely from strategy essays is a non-trivial task even for humans. Students often omit important details, whether due to oversight or lack of conceptual understanding. LLM feedback falls into a catastrophic failing mode when the LLM incorrectly classifies an essay as ‘correct’ and does not provide necessary interventions as a result of that. While we mitigate this risk by considering second most-likely label, the reliability and robustness of the framework can be improved with better classification. In future iterations, we could improve it by using relevant metadata (e.g. student’s level of knowledge, tone and vocabulary used in the essay, etc.), or by enforcing a more structured essay format to ensure all necessary concepts are articulated.

Another limitation is the small number of JiT interventions, as the framework was applied to only a single quiz during the semester. Although we observe significant performance improvements with the LLM feedback, a broader deployment across multiple topics and quiz types is necessary to validate these findings. Expanding the scope would allow for a more granular analysis of which specific subject matters or problem types benefit most from LLM-based interventions, providing deeper insights in how LLM feedback should be designed.

Our student preference survey shows that while the LLM is highly effective for novice-level feedback, it struggles to generate feedback that is more suitable to students with advanced knowledge; students with over-the-average performance still preferred LLM feedback that is targeted to novice-level over the advanced-level. This opens a new direction for designing LLM feedback in future studies to provide minimal interventions considering the knowledge level of students.

From a technical point of view, approximately 4% of the students experienced server failures due to the high volume of synchronous API calls, mainly due to the size of the course being very large ( $N > 1,000$ ). To make the framework more scalable, future implementations need to address rate-limiting and asynchronous processing. Furthermore, the current LLM lacks a memory function that help efficient conversations; the current version processes updated strategy essays in isolation without the context from previous versions. In

future work, we aim to incorporate student-specific memory and conversation history that reduces the number of turns required for a student to reach a correct understanding.

## Ethics Consideration

This research was declared ‘Exempt’ under IRB protocol number IRB-2023-391. The participant pool consisted of undergraduate students enrolled in a large-scale introductory Physics course. All personally identifiable information, including student names and university IDs, was removed from the data before being processed by the LLMs. Student essays used for qualitative analysis in this paper have been further anonymized, truncated, and paraphrased.

Our framework contains risks when generating feedback to students, primarily because of hallucinations and incorrect classification of error types. To mitigate these risks, the LLM was grounded with expert-annotated domain knowledge as well as its secondary predictions. Furthermore, we prioritized equity of treatment and algorithmic fairness throughout the study. The LLM was prompted using a standardized, invariant template for all participants, ensuring that the feedback was controlled solely by the student’s essay rather than any external variables such as student’s personal information or prior performance levels. By maintaining a uniform prompting strategy, we ensured that no specific group of students was prioritized or discriminated against, providing consistent pedagogical support to the entire students.

We disclose that LLMs are used for checking grammatical errors and exploring paraphrases. All core research contributions including motivation, experiment designs, and results analyses were developed by the authors.

## Acknowledgments

This research is supported in part by U.S. National Science Foundation grant 2300645. Opinions expressed are of the authors and not of the Foundation.

## References

Patricia Albacete, Pamela Jordan, Sandra Katz, Irene-Angelica Chounta, and Bruce M McLaren. 2019. The impact of student model updates on contingent scaffolding in a natural-language tutoring system. In

- International conference on artificial intelligence in education*, pages 37–47. Springer.
- Daman Arora, Himanshu Singh, and 1 others. 2023. Have llms advanced enough? a challenging problem solving benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7527–7543.
- Cameron Brooks, Annemaree Carroll, Robyn M Gillies, and John Hattie. 2019. A matrix of feedback for learning. *Australian Journal of Teacher Education (Online)*, 44(4):14–32.
- Annette Burgess, Christie van Diggele, Chris Roberts, and Craig Mellis. 2020. Feedback in the clinical setting. *BMC medical education*, 20(Suppl 2):460.
- Micheline TH Chi, Paul J Feltovich, and Robert Glaser. 1981. Categorization and representation of physics problems by experts and novices. *Cognitive science*, 5(2):121–152.
- Wei Dai, Jionghao Lin, Hua Jin, Tongguang Li, Yi-Shan Tsai, Dragan Gašević, and Guanliang Chen. 2023. Can large language models provide feedback to students? a case study on chatgpt. In *2023 IEEE international conference on advanced learning technologies (ICALT)*, pages 323–325. IEEE.
- Jennifer L Docktor, Natalie E Strand, José P Mestre, and Brian H Ross. 2010. A conceptual approach to physics problem solving. In *AIP Conference Proceedings*, volume 1289, pages 137–140. American Institute of Physics.
- Robert J Dufresne, William J Gerace, Pamela Thibodeau Hardiman, and Jose P Mestre. 1992. Constraining novices to perform expertlike problem analyses: Effects on schema acquisition. *The Journal of the Learning Sciences*, 2(3):307–331.
- H Foster. 2024. The impact of formative assessment on student learning outcomes: A meta-analytical review. *Academy of Educational Leadership Journal*, 28(S1):1–3.
- Concha Furnborough and Mike Truman. 2009. Adult beginner distance language learner perceptions and use of assignment feedback. *Distance Education*, 30(3):399–418.
- Syed Furqan Abbas Hashmi and N. Sanjay Rebello. 2025. Analyzing undergraduate problem-solving in physics through interaction with an ai chatbot. In *Physics Education Research Conference 2025*, PER Conference, pages 184–189, Washington, DC.
- John Hattie and Shirley Clarke. 2018. *Visible learning: feedback*. Routledge.
- John Hattie and Helen Timperley. 2007. The power of feedback. *Review of educational research*, 77(1):81–112.
- Michael Henderson, Tracii Ryan, David Boud, Phillip Dawson, Michael Phillips, Elizabeth Molloy, and Paige Mahoney. 2021. The usefulness of feedback. *Active Learning in Higher Education*, 22(3):229–243.
- Qinjin Jia, Jialin Cui, Haoze Du, Parvez Rashid, Ruijie Xi, Ruochi Li, and Edward Gehringer. 2024. Llm-generated feedback in real classes and beyond: Perspectives from students and instructors. In *Proceedings of the 17th international conference on educational data mining*, pages 862–867.
- Slava Kalyuga. 2007. Expertise reversal effect and its implications for learner-tailored instruction. *Educational psychology review*, 19(4):509–539.
- Greg Kestin, Kelly Miller, Anna Klales, Timothy Milbourne, and Gregorio Ponti. 2025. Ai tutoring outperforms in-class active learning: An rct introducing a novel research-based design in an authentic educational setting. *Scientific Reports*, 15(1):17458.
- Avraham N Kluger and Angelo DeNisi. 1996. The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological bulletin*, 119(2):254.
- Sunjun Kweon, Sooyohn Nam, Hyunseung Lim, Hwajung Hong, and Edward Choi. 2025. A large-scale real-world evaluation of an llm-based virtual teaching assistant. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 850–864.
- William J Leonard, Robert J Dufresne, and Jose P Mestre. 1996. Using qualitative problem-solving strategies to highlight the role of conceptual knowledge in solving problems. *American Journal of Physics*, 64(12):1495–1503.
- Jose P Mestre, Robert J Dufresne, William J Gerace, Pamela T Hardiman, and Jerold S Touger. 1993. Promoting skilled problem-solving behavior among beginning physics students. *Journal of research in science teaching*, 30(3):303–317.
- John Orlando. 2016. A comparison of text, voice, and screencasting feedback to online students. *American Journal of Distance Education*, 30(3):156–166.
- Madeline Ortiz-Rodríguez, Ricky W Telg, Tracy Irani, T Grady Roberts, and Emily Rhoades. 2005. College students’ perceptions of quality in distance education: The importance of communication. *Quarterly Review of Distance Education*, 6(2):97–105.
- Xinyu Pang, Ruixin Hong, Zhanke Zhou, Fangrui Lv, Xinwei Yang, Zhilong Liang, Bo Han, and Changshui Zhang. 2025. Physics reasoner: Knowledge-augmented reasoning for solving physics problems with large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11274–11289.

- Tung Phung, Victor-Alexandru Pădurean, Anjali Singh, Christopher Brooks, José Cambronero, Sumit Gulwani, Adish Singla, and Gustavo Soares. 2024. Automating human tutor-style programming feedback: Leveraging gpt-4 tutor model for hint generation and gpt-3.5 student model for hint validation. In *Proceedings of the 14th learning analytics and knowledge conference*, pages 12–23.
- Edd Pitt and Lin Norton. 2017. ‘now that’s the feedback i want!’ students’ reactions to feedback on graded work and what they do with it. *Assessment & Evaluation in Higher Education*, 42(4):499–516.
- Margaret Price, Karen Handley, Jill Millar, and Berry O’donovan. 2010. Feedback: all that effort, but what is the effect? *Assessment & Evaluation in Higher Education*, 35(3):277–289.
- Carina M Rebello and Yuri Piedrahita Uruena. 2019. Using a hybrid of argumentation and problem solving prompts to facilitate undergraduates’ problem solving performance and confidence. In *The 13th Conference of the European Science Education Research Association (ESERA)*.
- Tracii Ryan, Michael Henderson, and Michael Phillips. 2016. Written feedback doesn’t make sense’: Enhancing assessment feedback using technologies. In *International conference of the Australian Association for Research in Education*, pages 1–11.
- Chandralekha Singh and David Rosengrant. 2003. Multiple-choice test of energy and momentum concepts. *Am. J. Phys.*, 71(6):607–617.
- Karan Taneja, Pratyusha Maiti, Sandeep Kakar, Pranav Gururuprasad, Sanjeev Rao, and Ashok K Goel. 2024. Jill watson: A virtual teaching assistant powered by chatgpt. In *International Conference on Artificial Intelligence in Education*, pages 324–337. Springer.
- Jonathan Tuminaro and Edward F Redish. 2007. Elements of a cognitive model of physics problem solving: Epistemic games. *Physical Review Special Topics—Physics Education Research*, 3(2):020101.
- Erin Walker, Nikol Rummel, Sean Walker, and Kenneth R Koedinger. 2012. Noticing relevant feedback improves learning in an intelligent tutoring system for peer tutoring. In *International Conference on Intelligent Tutoring Systems*, pages 222–232. Springer.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Naomi E Winstone, Robert A Nash, James Rowntree, and Michael Parker. 2017. ‘it’d be useful, but i wouldn’t use it’: barriers to university students’ feedback seeking and recipience. *Studies in Higher Education*, 42(11):2026–2041.
- Benedikt Wisniewski, Klaus Zierer, and John Hattie. 2020. The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in psychology*, 10:487662.
- Thomas Wolsey. 2008. Efficacy of instructor feedback on written work in an online program. In *International Journal on E-learning*, volume 7, pages 311–329. Association for the Advancement of Computing in Education (AACE).
- Zheyuan Zhang, Daniel Zhang-Li, Jifan Yu, Linlu Gong, Jinchang Zhou, Zhanxin Hao, Jianxiao Jiang, Jie Cao, Huiqin Liu, Zhiyuan Liu, and 1 others. 2025. Simulating classroom education with llm-empowered agents. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10364–10379.

## A Feedback Preference Survey

We survey student’s LLM feedback preference in the Fall 2025 semester for two quizzes. Students are instructed that they will get extra credits for participating in the survey, and the number of participants for each iteration was 1,074 and 1,160. The two options for feedback is generated with the prompt described in Figure 11; it first evaluates the student’s strategy essays using expert-annotated rubrics, then generate two different versions of feedback using gpt-5.1, one for novice and the other for advanced.

Once the feedback generation is completed, we hosted a separate survey webpage from the student’s standard Learning Management System as shown in Figure 10. The two options are randomly swapped to minimize any potential biases. Students are instructed to read the quiz problem and essay they authored to remind themselves of the context, and choose one of the feedback texts they prefer with a set of pre-defined reasons (students can choose multiple reasons). For both surveys, the most selected reason was “Helps me better understand the concept”, followed by “Better explains the errors in my strategy” and “Aligns better with my level of knowledge”.

## B Error Type Prediction Model Details

For fine-tuned classifiers, we randomly split the data into 80-10-10 ratio for train-validation-test sets. We use DeBERTa-v3-small model with learning rate of  $2e-5$  and batch size of 16. The model is fine-tuned and tested on a single NVIDIA A100 GPU with 80GB memory, takes around 10 minutes to fine-tune 5 epochs.

**Quiz 11 Survey. Which Feedback Do You Prefer?**  
Review the feedback options for your strategy essay and choose an option which you find most helpful.

**Quiz Question**  
In the Atwood's Machine on the inclined plane as shown, the rope does not slip over the pulley with bearings that have negligible friction. The masses of the two blocks shown are known. The inclined plane makes a known angle with the horizontal. The coefficient of friction between the block  $m_1$  and the inclined plane is known. The pulley is a disk of known mass and known radius. What is the linear acceleration,  $a$ , of each block?

**Your Essay**  
To find the acceleration of each block, I must first set up the Momentum Principle for each block, which states that the net force acting on the block is equal to its mass times its acceleration. However, since that gives me only two equations with more than two unknowns, I must also use the Angular Momentum Principle for the pulley, which states that the net torque on the pulley is equal to its moment of inertia times its angular acceleration. I can write the angular acceleration in terms of the linear acceleration of each block since the rope is not slipping. Then I will plug in known variables, and use substitution to solve the system of equations for the acceleration of each block.

**Feedback Options**

**Feedback A**  
You've identified the correct governing principles and the no-slip kinematic relation, but your plan is missing some critical structural details—most notably the need to treat the tensions on either side of the pulley as distinct and to explicitly include friction and weight components in your force balances. Before plugging in variables, try to write three fully specified equations: two linear momentum equations with properly resolved components (including friction for  $m_1$  and weight components along the incline) and one rotational equation with torque from the two different tensions and the correct disk moment of inertia. As you refine your strategy, ask yourself: where exactly does the torque on the pulley come from if the tensions were equal, and how does the frictional interaction on the incline qualitatively affect the sign and magnitude of the acceleration you expect?

**Feedback B**  
You've made a very strong start by correctly naming the two big ideas: using Newton's Second Law (momentum principle) for the blocks and rotational Newton's Second Law (angular momentum principle) for the pulley. One important next step is to be more explicit about the forces: write separate force equations for each block that include gravity (weight), friction on the incline, and the tension pulling each block, and also remember that the tension on each side of the pulley can be different so that there is a net torque on the pulley. It's great that you know to connect angular acceleration to linear acceleration with the no-slip condition; if you now carefully list all forces and write three equations (one for each block and one for the pulley), you'll have everything you need to solve for the acceleration with confidence.

**1. Choose better feedback**

A > B

**2. Why do you prefer this feedback?**  
Please select all that apply. At least one option is required.

- Helps me better understand the concept
- Better explains the errors in my strategy
- Aligns better with my level of knowledge
- Other

Please specify

Submit

For LLM-based inference, we used the deepseek-r1:70B checkpoint, hosted on-premise at the authors' institution, to make sure the student's essay data is not shared outside.

All fine-tuning and inferences are performed with 2 trials and the average score was reported with the confidence interval in Table 1.

### C JiT Feedback Implementation

When the LLM is deployed for generating JiT feedback, we host a web app using Flask API, which is later embedded to the Learning Management System that students use for the course. The web app is a simple text box where students put their strategy essays and clicks a button to get the feedback from the LLM. In the back-end, we use pre-defined few-shot examples with Chain-of-Thought reasoning so that the model considers primary and secondary class of the strategy essay. The example of an actual prompt and the quiz page students have accessed is shown in Figure 12 and Figure 13.

### D Conversation Analysis

From \ To	Correct	Position	Direction	Pos-Dir
<b>Correct</b>	<b>0.6797</b>	0.1328	0.1250	0.0625
<b>Position</b>	0.3491	<b>0.3585</b>	0.1415	0.1509
<b>Direction</b>	<b>0.5278</b>	0.1250	<b>0.2917</b>	0.0556
<b>Pos-Dir</b>	<b>0.3750</b>	0.2969	0.0781	<b>0.2500</b>

Table 3: Transition probabilities between error types across conversation turns

In this section, we report additional analyses on conversational instances. Table 3 shows the transition probabilities between conversation turns. When the student's strategy essay at the  $i$ -th conversation turn is predicted as correct, it is still going to be classified as the same label with the 67.97% chances, but there exists slightly more than 30% chances the essay is classified as one of the error types. While majority of the strategy essays with direction and position-direction errors are likely developed into correct strategy, essays with position error is more likely going to stay at the same error type. This implies that LLM's ability to guide students avoid errors differ based on the types.

Figure 10: Survey webpage. Students answer two questions; choosing better feedback and the reason why.

### Prompt Sample: Generating Post-hoc Feedback for Preference Survey

You are an expert College Teaching Assistant and Pedagogical Specialist. Your goal is to provide feedback on a student's "Strategy Essay" regarding a quiz problem. First, analyze the student's written strategy to infer their Knowledge Level based on the complexity and accuracy of their writing. Then, regardless of their actual level, generate two distinct versions of feedback (one adapted for a "Novice" and one for an "Advanced" student) to provide options for instruction. Generate each feedback in 3 sentences.

#### Input Data Provided:

1. The Quiz Problem: The specific question the student attempted to solve.
2. Student's Strategy Essay: The student's written explanation of how they approached the problem.
3. Expert's Strategy Essay: A sample strategy essay written by an expert, that can serve as a rubric for evaluating student's essay

#### Instructions for Adaptation:

- If Knowledge Level is "NOVICE":
  - Tone: Highly encouraging, supportive, and patient. Use the "Sandwich Method" (Praise -> Gentle Correction -> Praise).
  - Vocabulary: Avoid jargon where possible. If technical terms are used, briefly define them or use analogies.
  - Focus: Focus on identifying misconceptions and solidifying the fundamental concept.
  - Scaffolding: Break down the next steps into small, manageable pieces. Guide them toward the correct starting point.
  - Goal: To build confidence and ensure they understand the basic mechanism of the problem.
- If Knowledge Level is "ADVANCED":
  - Tone: Professional, peer-to-peer, challenging, and concise.
  - Vocabulary: Use precise, high-level academic/technical terminology appropriate for the field.
  - Focus: Focus on optimization, edge cases, and connecting this problem to broader course concepts. Point out subtle logic flaws or inefficiencies.
  - Scaffolding: Do not hand-hold. Ask Socratic questions that force the student to re-evaluate their own logic.
  - Goal: To refine their reasoning, improve efficiency, and deepen metacognition.

#### Feedback Structure (Output Format) Please output the response in the following json format:

```
{ "Essay_Evaluation": "1-2 sentences summarizing their approach based on the expert sample",  
  "Inferred_Level": "Either Novice or Advanced",  
  "Feedback": {  
    "Novice": "The core feedback paragraph. Apply the Instructions for Adaptation logic for NOVICE here strictly.",  
    "Advanced": "The core feedback paragraph. Apply the Instructions for Adaptation logic for ADVANCED here strictly.",  
  } }
```

[INPUT DATA]

**The Quiz Problem:** quiz\_description\_input

**Student's Strategy Essay:** student\_strategy\_input

**Expert's Strategy Essay:** expert\_strategy\_with\_rubrics

Figure 11: The full system prompt used for generating post-hoc feedback

### Prompt Sample: Generating Just-in-Time Feedback for Quiz Problem Solving

You are an expert Physics Education Researcher and Cognitive Scientist. Your goal is to analyze student essays describing their strategy for solving a problem. You must determine if the student's proposed strategy will lead to the Correct Answer or a specific type of Misconception.

**TASK DESCRIPTION:** I will provide you with a physics problem statement and a "Strategy Essay" written by a student.

1. Analyze the student's essay. Look for specific keywords or logical steps.
2. Identify if they are applying correct principles/concepts or falling into misconceptions.
3. Output your final reasoning and label as strictly valid JSON.

#### Label Categories & Definitions

- 'correct': The reasoning and final answer are physically sound.
- 'direction': The student makes an error where their answer is in the exact opposite direction of the correct vector OR The student does not specifically state the direction they want to apply.
- 'position': The error arises from calculating forces or properties on a different object than the one asked OR The student does not specifically state the particular object they want to calculate for.
- 'position-direction': Both a 'position' error and a 'direction' error occur simultaneously in the student's response OR The student does not specifically state the direction nor the object they want to consider.

#### Few-Shot Examples

[Examples of student's sample strategy essay, its error type, and feedback generated by human experts]

**TEST CASE** {quiz\_problem}

{STUDENT\_STRATEGY\_ESSAY}

**INSTRUCTIONS:** Analyze the "Student Essay" above and provide the JSON output.

#### OUTPUT FORMAT:

{ "classification": Choose one from [correct, direction, position, position-direction],

"confidence": How confident the classification is (scale of 1 to 5, 5 being most confident),

"secondary\_classification": What would be the second most likely label?

"feedback": Feedback message (in 50 words) to students based on their essay. If their essay is classified to make a certain type of mistake, the feedback should highlight that aspect. If it is classified as correct, the feedback can be general and notify some common mistakes that can be made. If the secondary classification is different from classification, take both classes into consideration.

}

#### Constraint

- Output ONLY the json.
- Do NOT provide reasoning, explanations, or introductory text.

Figure 12: The full system prompt used for generating JiT feedback

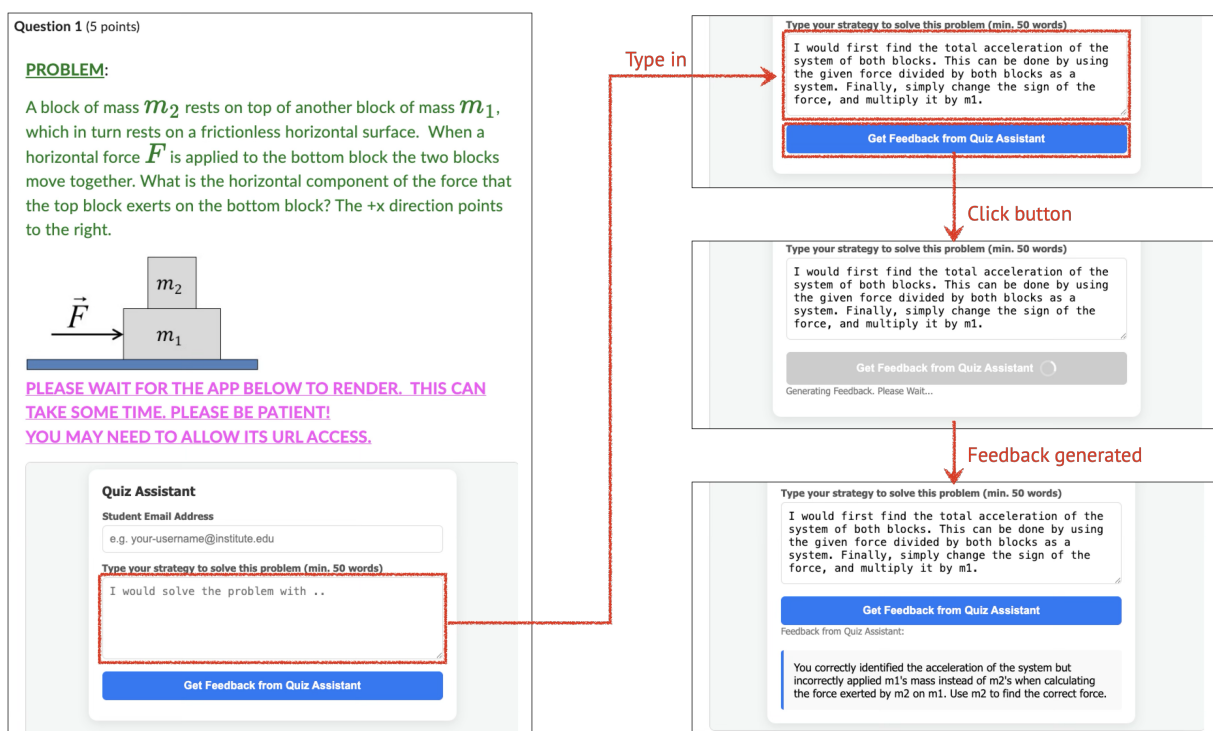


Figure 13: Overall flow of JiT feedback. The feedback snippet is placed below the quiz question. Students type in their strategy essays, click a button, and gets a feedback after a few seconds.