Towards A Human-in-the-Loop LLM Approach to Collaborative Discourse Analysis

Supplementary Materials

Teacher Feedback

This document contains a more in depth look at our conversation with the Educator and is provided to the reader for reference. The authors assume the reader is familiar with the manuscript.

As a first step toward adopting a human-in-the-loop approach to support classroom teachers in generating actionable insight into their instructional activities, we interviewed a knowledgeable high school physics and computer science teacher who has a STEM+C background (the Educator) and discussed four of the GPT-4-Turbo-generated summaries. This Educator helped develop the C2STEM curriculum and taught it in a semester-long 11th-grade Physics Honors class [2]. During the discussion, we described the Method and shared segments of students' problem-solving actions, transcribed discourse, and corresponding LLM-generated summaries (the Educator was never exposed to any summaries' human labels). Additionally, for a more complete picture, we also provided him with measures related to the social processes of collaborative problem-solving for each segment. This included two measures: (1) *equity*, i.e., the balance in the amount that each student contributes to a conversation segment; and (2) *turn-taking*, i.e., how much students' conversations are interleaved (indicating they listen and respond to each other). We defer to previous work [3] for how these metrics are calculated. Because this work is an exploratory first step, we only conducted a single interview with one teacher for this paper. Going forward, we will conduct additional, more thorough interviews with teachers to gain a more comprehensive understanding of how best to employ human-in-the-loop LLM approaches to support classroom teachers and their students.

The summaries evaluated by the Educator included one from each of the four Discourse Categories. We asked the Educator to (1) highlight information he believed was salient to students' STEM+C problem-solving; (2) identify missing information from the summary; (3) determine if he agreed with the characterization of the summary as one of the four Discourse Categories; and (4) discuss ways in which the summaries could be useful to him in the classroom and what improvements he would like to see. During the discussion, R2 memoed key insights [1] from the educator, which we later used to help answer RQ2. Table 1 provides an overview of the Educators' comments on two of the summaries, one in which the Educator agreed with the LLM (Example 1) and one in which the Educator disagreed with the LLM (Example 2). We highlight these two contrasting examples to discuss the benefits of the LLM summaries and opportunities for improvement before describing exploratory designs for leveraging these summaries to provide actionable insights for teacher feedback.

Meta Summary	**Educator Comments**
Example Segment 1	
In this segment, the students were working on creating the conditional structure to model the motion of the truck slowing down to a stop at the stop sign as they spend time talking about which kinematic equations to use. The LLM categorized this segment as *physics-and-computing-separate*, as the concepts are discussed in sequence as separate elements rather than intertwined in the discussion.	– At first glance, the Educator thought the discussion was synergistic, but the LLM summary helped him refine his own definition of what synergy means: conceptual understanding that is moving forward in both domains. – The LLM highlighted a discussion of an "if" statement that the Educator missed on his first review.
Example Segment 2	
In this segment, the students were working on creating a conditional structure to model the motion of the truck. They identified that they needed to set an initial velocity and wanted the velocity to increase to 15 m/s. However, they often paused and said "um" or "I'm still a little stumped" while trying to figure out how to translate this into computational from. The LLM categorized the segment as *physics-focused*.	– The Educator disagreed with the LLM and categorized this segment as *computing-focused*. – The LLM correctly identified students' verbalization of physics concepts, but the segment's focus is on the students' lack of understanding of how to represent this in the computational model (i.e., *computing-focused*). – The LLM summary does not mention utterances such as "um" or "I'm still a little stumped" but instead focuses on the concepts the students *do* verbalize.

Table 1: Contrasting examples of Educator comments on LLM summaries.

Overall, the Educator agreed with the LLM-generated summary content and characterization in three out of the four segments and identified several benefits of these summaries to support his teaching. In *Example Segment 1*, the students worked on creating a conditional structure to model the motion of the truck slowing down to a stop at the stop sign. They first talked about using the relevant kinematic equations. The LLM categorized this segment as *physics-and-computing-separate*, as the concepts were discussed separately one after another rather than being intertwined in the discussion. After reviewing this segment, the Educator pointed out that the LLM summary fully described the students' problem-solving and even helped him refine his own internal definition of synergistic dialogue. In his initial review, the Educator thought the students' discussion integrated both physics and computing concepts synergistically. However, upon reevaluation and considering the LLM's summarization, he concurred that the students were not effectively interleaving their knowledge of the two domains, which would have improved their solution development. The Educator also acknowledged that the LLM highlighted a part of the students' discourse about conditional statements that he had initially overlooked, aiding him in better understanding the students' conceptual discussion. Similarly, the Educator agreed with the LLM-generated summaries and characterizations of *Example Segments 3* and *4*.

However, the Educator identified issues with the LLM-generated summary in *Example Segment 2*, where the students were working on creating a conditional structure to model the speedup motion of the truck. They identified that they needed to set an initial velocity and that the velocity should increase to 15 m/s. However, they often paused and said "um" or "I'm still a little stumped" while trying to figure out how to translate this into the appropriate computational structure. In this case, the Educator disagreed with the LLM's characterization of the segment as *physics-focused* and instead labeled it as *computing-focused*. He argued that the students were trying to figure out how to translate their physics understanding of updating the velocity value into computational form. The Educator also noted that the LLM's summary omitted utterances indicating confusion (e.g., "Um..."), and therefore, overly emphasized the successfully verbalized physics concepts without recognizing the students' primary focus on grappling with computing concepts. Additionally, the Educator suggested including information about students' speaking duration with the specific components in the summary, pointing out that the group's frequent pauses while discussing computational blocks hinted at a lack of computing understanding. Instances where the Educator identified crucial elements missed by the LLM in its summaries were seen as opportunities for Active Learning to enhance the LLM's performance in future work.

The Educator had several suggestions and ideas for how these summaries could be used to generate actionable insights to support students' STEM+C learning. First, the Educator suggested the LLM be used to generate a graphical representation (i.e., a timeline) to capture students' conceptual understanding of targeted physics and computing concepts. Such a timeline would provide the instructor with a succinct overview of students' progress and identify specific segments to view in more detail to then discuss with the students. Second, the Educator suggested that social metrics would help highlight when a teacher may need to intervene to get quieter students to be more engaged in group discussions. He suggested that the conceptual summary, in combination with the social metrics, would provide insight into which conceptual component of the task to engage the student in, especially in cases where the students' difficulties persisted. Last, the Educator highlighted the LLM's ability to identify off-task dialogue but emphasized a need to know how long students were off-task, as he would not intervene if the students were taking short breaks.

References

1. Hatch, J.A.: Doing qualitative research in education settings. SUNY Press (2002)
2. Hutchins, N.M., Biswas, G., Maróti, M., Lédeczi, Á., Grover, S., et al.: C2stem: A system for synergistic learning of physics and computational thinking. Journal of Science Education and Technology **29**, 83–100 (2020)
3. Snyder, C., Hutchins, N.M., Cohn, C., Fonteles, J.H., Biswas, G.: Analyzing students collaborative problem-solving behaviors in synergistic stem+c learning. In: Proceedings of the 14th Learning Analytics and Knowledge Conference (2024)