International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Artikel Review Brit J Educational Tech - 2024 - Cohn - A Multimodal Approach to Support Teacher Researcher and AI Collaboration in STEM C

Mr. Muhammad Sintur[1], Prof. Dr. Masra Latjompoh[2], Prof. Dr. Mursalin[3], Ashwin T. S[4], Clayton Cohn[5], Caitlin Snyder[6], Joyce Horn Fonteles[7], Gautam Biswas[8]

[1]Headmaster, Education Authorities, SMAN 6 Sigi
[2,3]Lecturer, College, Gorontalo State University

Abstract

Abstract: Recent advances in generative artificial intelligence (AI) and multimodal learning analytics (MMLA) have allowed for new and creative ways of leveraging AI to support K12 students' collaborative learning in STEM+C domains. To date, there is little evidence of AI methods supporting students' collaboration in complex, open-ended environments. AI systems are known to underperform humans in (1) interpreting students' emotions in learning contexts, (2) grasping the nuances of social interactions and (3) understanding domain-specific information that was not well-represented in the training data. As such, combined human and AI (ie, hybrid) approaches are needed to overcome the current limitations of AI systems. In this paper, we take a first step towards investigating how a human-AI collaboration between teachers and researchers using an AI-generated multimodal timeline can guide and support teachers' feedback while addressing students' STEM+C difficulties as they work collaboratively to build computational models and solve problems. In doing so, we present a framework characterizing the human component of our human-AI partnership as a collaboration between teachers and researchers. To evaluate our approach, we present our timeline to a high school teacher and discuss the key insights gleaned from our discussions. Our case study analysis reveals the effectiveness of an iterative approach to using human-AI collaboration to address students' STEM+C challenges: the teacher can use the AI-generated timeline to guide formative feedback for students, and the researchers can leverage the teacher's feedback to help improve the multimodal timeline. Additionally, we characterize our findings with respect to two events of interest to the teacher: (1) when the students cross a difficulty threshold, and (2) the point of intervention, that is, when the teacher (or system) should intervene to provide effective feedback. It is important to note that the teacher explained that there should be a lag between (1) and (2) to give students a chance to resolve their own difficulties. Typically, such a lag is not implemented in computer-based learning environments that provide feedback.

INTRODUCTION

Recent studies show that STEM+C (Science, Technology, Engineering, Math and Computing) open-ended

learning environments enhance conceptual understanding and practice (eg, Hutchins, Biswas, Maróti, et al., 2020). These environments promote the synergistic learning of science and computing concepts by engaging students in contextualized, problem-based learning (Astuti et al., 2021; Grover & Pea, 2013). The next-generation science standards (NGSS) advocate for the integration of science and engineering, and emphasize computa- tional thinking (CT) as a critical skill (National Research Council, 2012; NGSS, Next genera- tion science standards: For states, by states, 2013). However, integrating multiple knowledge domains introduces complexities, which may cause students to encounter difficulties span- ning multiple domains (Basu et al., 2016; Hutchins, Biswas, Maróti, et al., 2020).

Building computational models of scientific processes is effective in fostering integrated STEM+C learning (Sengupta et al., 2013). In the C2STEM environment, students collabo- ratively translate scientific concepts into computational models using block-based program- ming, guided by domain-specific modelling languages (DSML) that support the interplay between science and computing (Hutchins, Biswas, Maróti, et al., 2020).

In previous work, researchers have studied how students synergistically learn kinematics and computational thinking while working on a *Truck Task* in C2STEM. Pairs of students model a truck's accelerated motion, speeding up from rest, cruising at a speed limit and decelerating to stop at a stop sign. Students apply kinematic laws that connect physics concepts like position, velocity and acceleration with computing concepts such as variables, loops and conditional statements to build a computational model that simulates the truck's movement. C2STEM facilitates this process by providing a domain-specific, block structured environment that allows students to construct computational models, and offers execution, visualization and debugging tools to help students build and refine their models (Hutchins, Biswas, Maróti, et al., 2020; Hutchins, Biswas, Zhang, et al., 2020). An example solution for the Truck Task is provided in Figure 1.

Practitioner notes

What is already known about this topic

- Collaborative, open-ended learning environments enhance students' STEM+C conceptual understanding and practice, but they introduce additional complexities when students learn concepts spanning multiple domains.
- Recent advances in generative AI and MMLA allow for integrating multiple da- tastreams to derive holistic views of students' states, which can support more informed feedback mechanisms to address students' difficulties in complex STEM+C environments.
- Hybrid human-AI approaches can help address collaborating students' STEM+C difficulties by combining the domain knowledge, emotional intelligence and social awareness of human experts with the general knowledge and efficiency of AI.

What this paper adds

- We extend a previous human-AI collaboration framework using a hybrid intelligence approach to characterize the human component of the partnership as a researcher-teacher partnership and present our approach as a teacher- researcher-AI collaboration.
- We adapt an AI-generated multimodal timeline to actualize our human-AI collabo- ration by pairing the timeline with videos of students encountering difficulties, en- gaging in active discussions with a high school teacher while watching the videos to discern the timeline's utility in the classroom.
- From our discussions with the teacher, we define two types of *inflection points* to address students' STEM+C difficulties—the *difficulty threshold* and the *interven- tion point*—and discuss how the

feedback latency interval separating them can inform educator interventions.

- We discuss two ways in which our teacher-researcher-AI collaboration can help teachers support students encountering STEM+C difficulties: (1) teachers using the multimodal timeline to guide feedback for students, and (2) researchers using teachers' input to iteratively refine the multimodal timeline.

Implications for practice and/or policy

- Our case study suggests that timeline gaps (ie, disengaged behaviour identified by off-screen students, pauses in discourse and lulls in environment actions) are particularly important for identifying inflection points and formulating formative feedback.
- Human-AI collaboration exists on a dynamic spectrum and requires varying de- grees of human control and AI automation depending on the context of the learn- ing task and students' work in the environment.
- Our analysis of this human-AI collaboration using a multimodal timeline can be extended in the future to support students and teachers in additional ways, for example, designing pedagogical agents that interact directly with students, devel- oping intervention and reflection tools for teachers, helping teachers craft daily lesson plans and aiding teachers and administrators in designing curricula.

While working on the Truck Task, students face challenges such as translating kinematic equations into computational structures and calculating the lookahead distance for deceler- ation. As an example, not using the correct conditional forms when modelling the lookahead

FIGURE 1 C2STEM *Truck Task* (Snyder et al., 2024).

distance to slow down and stop the truck can lead to the truck moving backwards, and students struggle to associate this movement with negative velocity values. Additionally, stu- dents often misconstrue the differences between updating variables incrementally (‘change by’ block) and setting variable values (‘set’ block). These issues underscore the complexity of STEM+C learning and the need to understand student thought processes and problem- solving approaches, and help them overcome their difficulties. To gain insights into student behaviours and outcomes, we collect and analyse multimodal data including video (via Open Broadcaster Software; OBS), screen recordings, audio (captured with lapel microphones and transcribed using Otter.ai; https://otter.ai/) and C2STEM system logs to understand stu- dents'

collaborative problem-solving behaviours. This comprehensive data collection en- ables nuanced analysis of student interactions and learning experiences.

Multimodal learning analytics (MMLA) literature underscores the value of integrating mul- tiple data types to enhance the understanding of student learning processes, often reveal- ing subtleties not apparent in unimodal data (Liu et al., 2019; Olsen et al., 2020; Vrzakova et al., 2020). Despite its advantages, multimodal data analysis presents challenges, such as the labor-intensive tasks of collecting, cleaning, labelling, preprocessing and analysing (Kubsch et al., 2022; Liu et al., 2018, 2019). Recent advances in generative AI offer some relief by streamlining the labelling and analysis of multimodal data, and providing deeper insights into student behaviours. However, AI models often struggle with domain-specific tasks that fall outside of the AI model's training (Cohn, 2020). Additionally, they underper- form humans at interpreting emotions in context and understanding the nuances of social in- teractions (Järvelä et al., 2023). Therefore, we advocate for a *hybrid approach* that combines human expertise in domain knowledge, emotional intelligence and social context with the efficiency of AI for multimodal analysis, ensuring a comprehensive understanding of student learning in rich STEM+C environments.

Our research is driven by the goal of enhancing human capabilities, not replacing them (Akata et al., 2020). We draw inspiration from Järvelä et al. (2023), who demonstrate the ben- efits of human-AI collaboration in analysing socially shared regulation in learning (SSRL). By extending their approach to include *researcher-teacher partnerships*, we highlight the need for expertise of diverse stakeholders (Coburn & Penuel, 2016; Thompson et al., 2017) and aim to deepen the understanding of students' collaborative behaviours and SSRL processes (Holmlund et al., 2018). In our human-AI collaboration model, we integrate the perspectives of teachers and researchers to create AI tools that provide actionable insights into class- room studies. This collaborative loop involves teachers identifying pedagogical needs that inform the development of AI tools by researchers, which are then refined based on teach- ers' classroom experiences.

This paper seeks to address the following exploratory research question: *How can hu- man-AI collaboration using an AI-generated multimodal timeline assist teachers in identifying student challenges in STEM+C learning and crafting supportive feedback*? We adopt a case study approach and extend a hybrid intelligence framework (Akata et al., 2020) to character- ize human-AI collaboration with two sets of human stakeholders—teachers and research- ers. To facilitate this collaboration, we use an *AI-generated multimodal timeline* inspired by previous work (Fonteles et al., 2024) to identify student challenges during the C2STEM Truck Task. This timeline amalgamates diverse data types: students' emotional responses, synergy scores, social interaction metrics, prosodic audio cues, verbatim conversation tran- scripts, prior physics and computing knowledge, log-segmented and -contextualized sum- maries from a large language model (Snyder et al., 2024), synchronized video and screen recordings, and detailed logs of student actions within the C2STEM environment. Unlike prior work, our timeline considers students' collaborative dynamics (via social interaction metrics, discourse summaries and speaker diarization) and includes the use of synergy scores and prior domain knowledge. The timeline presented in this work is the direct result of our previous co-design efforts (Cohn, Snyder, et al.,2024) with a high school physics and computer science (CS) teacher (the Teacher).

We document our interactions with the Teacher using video and audio recordings as we review the timeline, analysing these recordings to extract key insights. The Teacher, who de- signed and taught the C2STEM kinematics curriculum for a high school physics class, has 20 years of teaching experience

across all levels of high school (and some middle school) in both low- and high-performing schools, and holds a Bachelor of Science in Electrical Engineering. In his own words, his pedagogical efforts centre on developing students' tech- nical skills and fostering their interest in STEM with the goal of ensuring that a broader, more diverse range of students leave high school prepared for and interested in pursuing engi- neering fields. To support this goal, he began collaborating with several Vanderbilt University laboratories 10 years ago to develop integrated CS curricula, addressing the significant gap in CS education in Tennessee.

To answer our research question, we use the AI-generated timeline to help the Teacher better understand students' STEM+C difficulties as they work on the C2STEM Truck Task and analyse the video and audio recordings to study the researcher, teacher and AI inter- actions by memoing key findings (Hatch, 2002). Our goal is to support the teacher in gener- ating actionable insights by identifying students' difficulties and helping students overcome them. In the following sections, we discuss our hybrid intelligence approach to AI–human collaboration, present the results of our exploratory analysis and discuss directions for fu- ture research.

HUMAN-AI COLLABORATION IN EDUCATION

Understanding student learning is a multifaceted endeavour that encompasses metacogni- tive, cognitive and social dimensions (Baker, 2015; Snyder et al., 2019; Zimmerman, 2002). To gain a comprehensive understanding of students' learning processes, teachers and re- searchers must draw on a broad array of data sources. Teachers in classrooms, for exam- ple, interpret students' verbal and non-verbal cues and emotional states during interactions, and are conscious of their students' backgrounds and preferences. Research has shown that multimodal data, including students' speech, actions and video, can provide a more comprehensive view of learning than individual data modalities (Blikstein & Worsley, 2016; Emerson et al., 2020; Nasir et al., 2021). However, the complexity of analysing such data necessitates methods, often AI-based, to structure and coordinate the data collection pro- cess using multiple sensors, and requires developing tools that analyse the data and provide information that can enhance student learning. Given the recent advances in AI, a natural question arises about the balance between AI-supported automation, and human oversight and decision-making (Cukurova, 2024).

Human-AI collaboration in education can adopt varying degrees of human control, in contrast to AI automation (illustrated in Figure 2). Intelligent tutoring systems (ITSs) typically involve high AI automation with limited human control and provide personalized learning in a structured format (Graesser et al., 2012). Conversely, open-ended learning environments (OELEs) offer greater human control by varying the levels of exploration and discovery that students are exposed to (Hannafin et al., 2014). AI is often used to generate feedback on students' problem-solving processes (Basu et al., 2017; Munshi et al., 2023). Traditional classroom supports like dashboards represent a lower level of AI automation, where AI analyses and presents data, leaving interpretation and decision-making to teachers and students (Hutchins & Biswas, 2024; Molenaar & Knoop-van Campen, 2019).

This spectrum also applies to analysis techniques (also illustrated in Figure 2). Fully au- tomated AI analysis, such as engagement classifiers for student videos, minimizes human involvement (Sümer et al., 2023). In contrast, techniques like large language model (LLM) prompt engineering can require more human input to refine and assess AI outputs (Cohn, Hutchins, et al., 2024). Human-led analysis methods, such as Interaction Analysis (Hall & Stevens, 2015) or hand-coding discourse, can be augmented by AI to reduce data pro- cessing time without replacing human judgement. AI is often leveraged to support,

but not replace, human-led analysis methods. As an example, AI algorithms can automatically transcribe student speech to decrease analysis time, but the transcripts are then hand-coded using qualitative analysis methods (Bokhove & Downey, 2018).

The effectiveness of learning supports and analysis techniques hinges on a synergistic relationship between humans and AI. Hybrid intelligence approaches advocate for AI to aug- ment human intellect, supporting goals unattainable by either alone (Akata et al., 2020). In education, human-AI collaboration can enhance understanding and support effective learn- ing. Our work demonstrates this collaboration in action, showing how it can improve under- standing students' difficulties and inform the development of future learning supports.

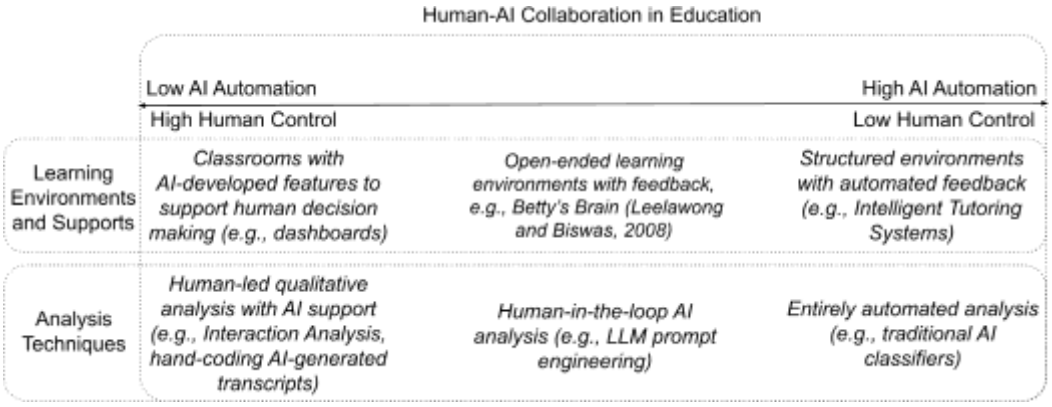

FIGURE 2 Human-AI collaboration in education spectrum.

MULTIMODAL TIMELINE: HUMAN-AI COLLABORATION IN PRACTICE

To comprehend student learning in STEM+C, we leverage a multimodal visual timeline representation originally developed in Fonteles et al. (2024). The timeline integrates multiple data sources—webcam footage, screen recordings, audio of group discussions and C2STEM system logs—and aligns derived information from the data temporally, offering a detailed view of student interactions with the system and with each other over time. This facilitates seamless navigation and visualization of events that combine multiple modalities. Figure 3 depicts the collaborative efforts of teachers, researchers and AI in creating and utilizing the timeline. For this project, we adapted the timeline by Fonteles et al. (2024) based on previous findings by Cohn, Snyder, et al. (2024) that suggest visual timelines may aid teachers in providing instructional supports to students, and also based on our own expertise

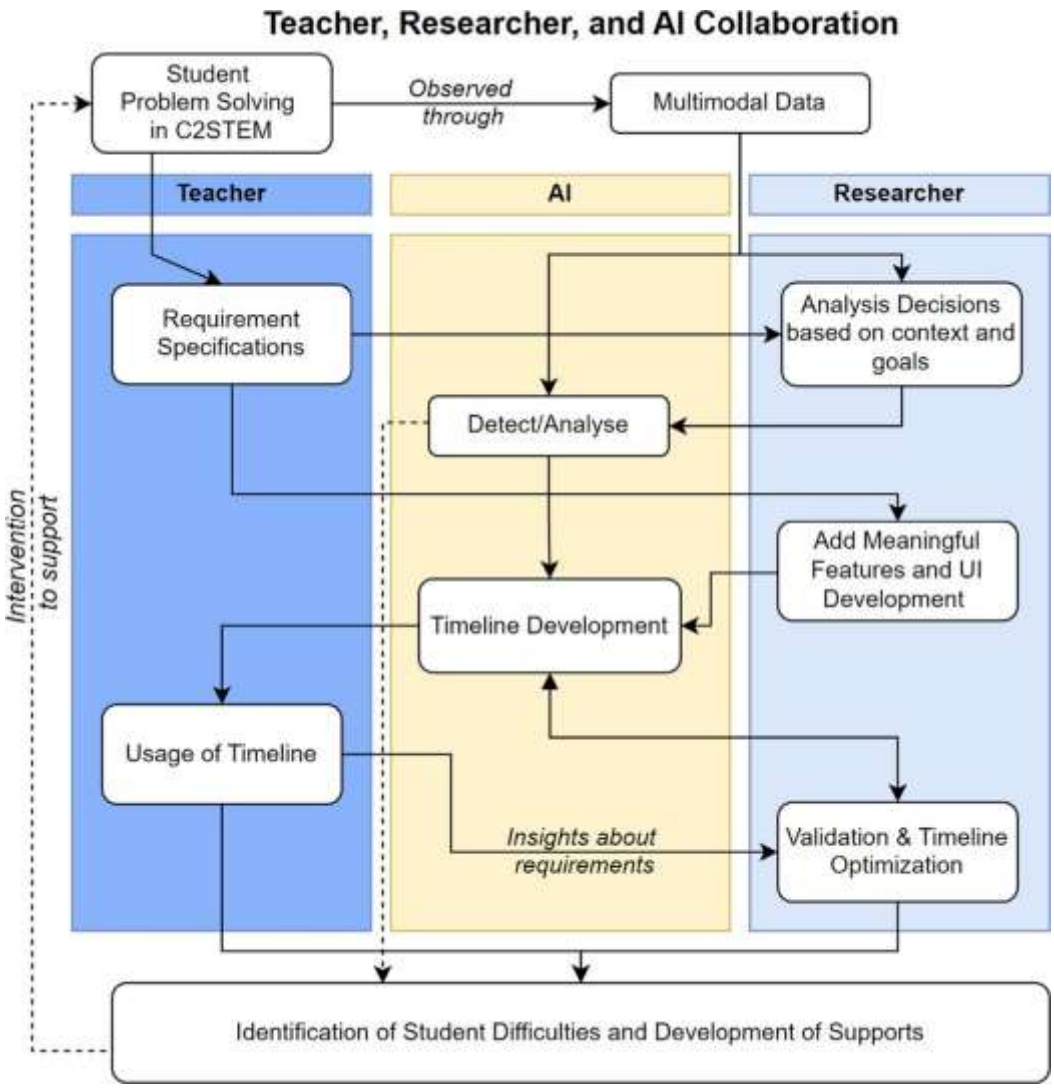

FIGURE 3 Teacher, researcher and AI collaboration framework.

with multimodal data and AI analysis. The timeline serves as a tool for both researchers and teachers to meaningfully interpret the analysed data.

As outlined in fourth section, the Teacher employs the timeline to address student chal- lenges and provides insights to refine it for subsequent applications. The timeline provides a holistic perspective on students' engagement with the task and helps pinpoint pivotal mo- ments in students' model building, social interactions and affective states. By analysing tem- poral patterns, we can detect key instances of students' struggles and successes. Teachers can use this information to extract actionable insights and develop additional instructional strategies and formative feedback to support their students (Hutchins & Biswas, 2024).

Figure 4, adapted from Järvelä et al. (2023), illustrates the multimodal analysis meth- ods within the timeline and how human-AI collaboration can be applied to enhance our un- derstanding of student interactions, with teacher feedback shaping the analysis approach. Students' problem-solving activities are observed across multiple datastreams, specifically logged actions, audio, video and student assessments. Each datastream can be processed by either AI, a human or a combination of both. Within our framework, the 'Human' role can be fulfilled by a teacher, researcher or the two working together. In this case study, the researcher was responsible for the initial data processing (eg, *Validation and*

Correction of Affect and Transcripts), while the Teacher was responsible for *Analysing and Sense Making* and *Identification of Student Difficulties and Development of Support*. Logged data were used to construct a hierarchical task-oriented process of student actions, which we used to segment our videos (Snyder et al., 2024). This segmentation, along with audio transcription and diarization, was then utilized for segment-specific LLM summarization. Video processing

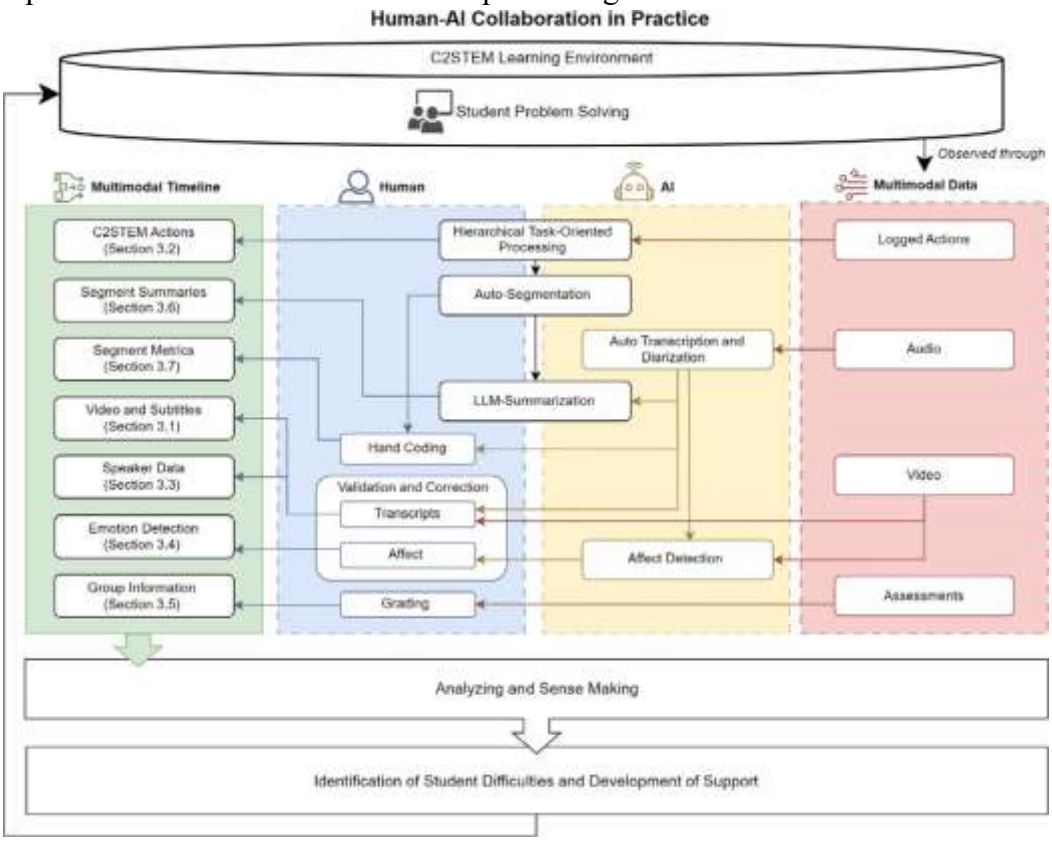

FIGURE 4 Human-AI collaboration in practice; a multimodal analysis approach.

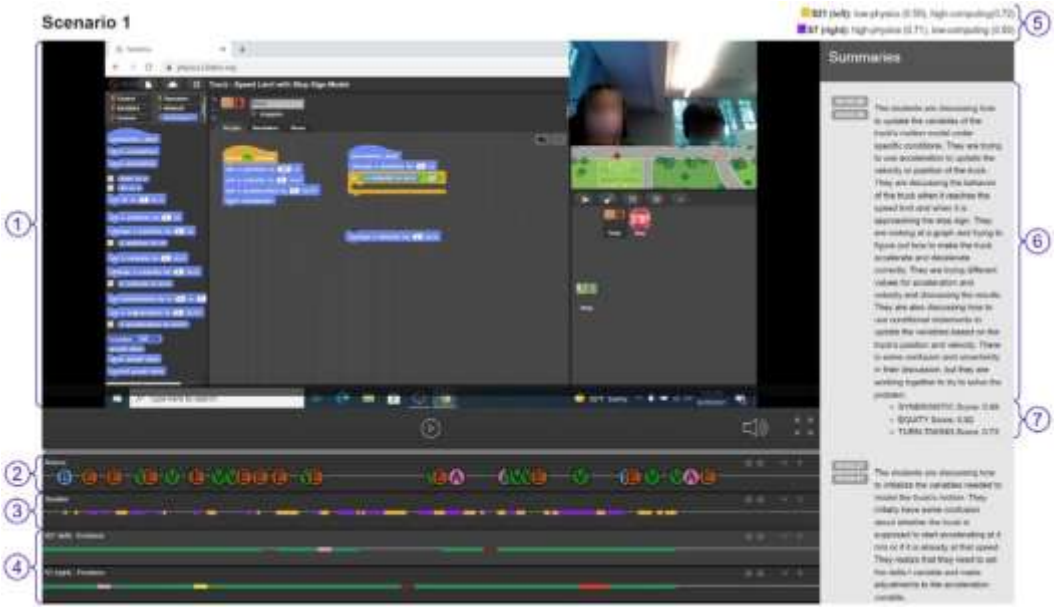

FIGURE 5 AI-generated multimodal visual timeline adapted from Fonteles et al. (2024).

was employed to detect affective states. Both AI-generated transcripts and affective states were subsequently validated and corrected by a human. These validated modalities were then integrated into a

multimodal timeline, which served as a visualization tool for the teacher to identify students' difficulties and inform his support decisions.

Figure 5 shows the timeline with the human-validated modalities from Figure 4. The in- tegrated temporal visual representation provides an augmented understanding of student interactions paired with video watching. The subsections detail the timeline's components and corresponding collaborative analysis methodologies.

Video and subtitles (1 in Figure 5)

The video data consist of screen recordings and webcam footage, captured via OBS and laptop webcams, with subtitles from transcribed student speech captured via lapel micro- phones. Lapel microphones were used to collect audio data instead of laptop webcams to reduce background noise, and we used Otter.ai to automatically transcribe and diarize the students' conversations. Transcripts were manually corrected by researchers, blending human oversight with AI automation.

C2STEM actions (2 in Figure 5)

The timeline provides contextual insights through a 'track' that displays visual representa- tions of students' high-level C2STEM actions, categorized using a hierarchical task-oriented structure adapted from (Emara et al., 2021). This structure classifies student actions into five interpretable categories:

- *Build*: Students add components to their model.
- *Adjust*: Students move, edit or remove blocks from their model.
- *Draft*: Students move or edit blocks not connected to their executable model (akin to com- menting code).
- *Execute*: Students run their executable model.
- *Visualize*: Students use data tools or inspect variable values.

While developing the hierarchical task-oriented structure involves human expertise, the processing of C2STEM actions within these categories is fully automated.

Speaker data (3 in Figure 5)

Speaker data are derived from audio data diarized via Otter.ai (with human correction) and shown on the timeline with coloured segments indicating individual students as active speakers. This facilitates the analysis of conversational dynamics, including conversational pauses. Prosodic audio analysis also contributes insights into conversational dynamics and speech patterns, particularly when studied in combination with the transcript subtitles that provide semantic information about these interactions. Just as in section 'Video and sub- titles (1 in Figure 5)', this process blends AI automation with human oversight, combining AI-generated diarization with human validation.

Emotion detection (4 in Figure 5)

The timeline uses colour-coded segments to depict students' detected emotions. Since these study data were collected during the COVID pandemic, students wore masks. In addition, students sometimes moved away from the webcam's view. To make our multimodal emotion recognition more robust, we combined video and audio data. For video frames, we initially used multi-task cascaded convolutional networks (MTCNNs) for face detection, followed by face re-identification to distinguish individuals (Gupta et al., 2018; Zhang et al., 2016). We then applied HSEmotion for predicting valence and arousal

(Savchenko, 2023), com- plemented by Lawpanom et al.'s (2024) model for masked facial expression detection. We employed Wagner et al.'s (2023) method to analyse speech valence and arousal. Non-verbal audio cues, such as pitch and intensity, contributed to assessing emotional valence and arousal, while sentiment analysis of transcribed text provided additional context.

We fused the valence and arousal data from speech and video using decision-level fu- sion, aligning the speech data with the corresponding video frames. The values of valence and arousal ranged from −1 to +1 and were automatically categorized into four emotional quadrants based on Russell's circumplex model (Russell, 1980). When both valence and arousal values are positive, they fall under the first quadrant. If arousal is positive and va- lence is negative, it is considered the second quadrant. If both are negative, it is the third quadrant, and if valence is positive and arousal is negative, it falls under the fourth quad- rant. This classification is widely used in the literature, and we followed the same approach (Mollahosseini et al., 2017; Posner et al., 2005; Toisoul et al., 2021).

In the absence of speech, facial expressions were summarized at five-second intervals (Fonteles et al., 2024). Apart from quadrant-based classification, we also recorded the exact valence and arousal values. The literature indicates that these values can be used to de- rive learning-centred emotions. We used these values to classify emotions into Engaged Concentration, Boredom, Confusion, Frustration and Delight (Akpanoko et al., 2024; Akpanoko & Biswas, 2024; Fonteles et al., 2024). We then integrated an LLM (GPT-4; OpenAI et al., 2024) to assign learning-centred emotions to each utterance, which were fused with the video and speech data to generate the final emotion predictions. For the textual component, we provided the transcribed text (obtained via Otter.ai, as mentioned in sections 'Video and subtitles (1 in Figure 5)' and 'Speaker data (3 in Figure 5)') to the LLM via the OpenAI API. Our prompts contained LLM task instructions, study content details (eg, the specific subject matter discussed), C2STEM task context details, and the sentence for learning-centred emotion classification.

Although we had the learning-centred emotions for the entire dataset, we did not have ground truth data to verify the accuracy of the model. To address this, we validated our emo- tion recognition process by having two human annotators independently review a randomly selected >20% of the instances for each emotion. The inter-rater reliability among the two annotators, measured by Cohen's Kappa, was 0.94 overall, and a perfect 1 for confusion and frustration, indicating a high level of agreement and validating our affective state rec- ognition. Our approach balances human control with AI automation, leveraging AI to derive valence, affect and speech transcriptions, and using human annotators (ie, researchers) to validate our emotion recognition process and develop the LLM prompts.

Group information (5 in Figure 5)

Students' prior knowledge influences their social skills and domain-specific collaborative problem-solving (CPS) behaviours (Yang et al., 2015). To better understand their interac- tions, we incorporate students' prior physics and computing knowledge into the timeline. This knowledge was classified as high or low based on pretest scores relative to the median. Although grading was done manually by our research team (indicating high human control), advances in LLM-based grading suggest there is potential for increased AI automation in the future (Cohn, Hutchins, et al., 2024).

Segment summary information (6 in Figure 5)

We segmented students' multimodal data temporally using a context-specific method based on the model component the students were working on (Snyder et al., 2024), rather than an arbitrary segmentation method devoid of educational context (Knight et al., 2017). An LLM then summarized these segments,

providing an overview of the video content and capturing the evolution of discussions and task interactions. This LLM-based summarization, which required human oversight, represents a midpoint on the spectrum between human control and AI automation.

Segment metrics (7 in Figure 5)

We included three segment-level metrics to capture the nature of student dialogue and collaborative dynamics. The research team manually coded utterances for physics and computing content, calculating a *synergy score* to quantify the interweaving of science and computing content (values near 1 indicate high synergy; values near 0 suggest domain focus).

We do not currently use AI to calculate synergy scores, but recent work suggests LLMs may be leveraged towards this end (Cohn, Snyder, et al., 2024). We also measured social collaboration measures like *turn-taking* and *equity* (Rummel et al., 2009) using diarized ut- terances to assess students' contribution balance.

CASE STUDY METHOD

We adopted a case study approach to demonstrate how the multimodal timeline further facilitates human-AI collaboration. In our study, we used the timeline as the basis for researcher-teacher-AI interactions to (1) understand students' model building and problem- solving work in the C2STEM environment, (2) identify specific points where students are having difficulties and (3) make decisions on when to intervene and how to help students overcome their difficulties. To design and implement the multimodal timeline, we leveraged previous findings and teacher insights from Cohn, Snyder, et al. (2024). We selected two scenarios for two groups of high school students working on the C2STEM Truck Task in a classroom study (see first section). The research team then worked with the Teacher in two 90-minute sessions to review the two video segments and analyse the data presented in the timeline. The sessions were recorded and then reviewed by the research team to extract key findings that we discuss in fifth section.

Scenario details

Scenario 1 involved students S21 (with low physics and high computing prior knowledge) and S7 (with high physics and low computing prior knowledge). The students worked to- gether and contributed equally to the discussions. In this scenario, the group was working on modelling the truck's cruising motion. They assessed their model utilizing the data tools (ie, the graph). After having difficulty interpreting the graph, they recognized that the truck's acceleration never switched to 0 but instead stayed at the initial acceleration value. This was because they did not set the acceleration value to 0 while updating the velocity variable. One student suggested the error was due to not initializing the 'delta_t' variable. Though the group had the physics knowledge to recognize that acceleration impacts velocity and must be set to 0 in order for the truck to cruise at a constant velocity, they had difficulty in debug- ging the update velocity construct.

Scenario 2 featured students S2 (with low prior knowledge in both domains) and S8 (with high prior knowledge in both domains). Throughout the scenario, S8 took the lead in discussing the model building process but involved S2 by asking questions and verbally expressing her thoughts. In the specific segment, we analysed, the students were model- ling the motion of the truck slowing down to stop at a stop sign. They began by editing their conditional statement that governed when the truck should start to

decelerate (ie, via the lookahead distance). They encountered difficulties using the kinematic equations to com- pute the lookahead distance and instead used a trial-and-error approach to determine the value. The errors in their model resulted in the truck moving backwards. Confused as to why this happened, the group changed a different conditional statement that aimed to stop the simulation once the truck reached the stop sign. Overall, the group used an ineffective de- bugging strategy to update the lookahead distance. They also lacked the physics knowledge to recognize that the truck began to slow down too early, causing the velocity value to reach 0 before the stop sign, after which the negative acceleration led to a negative velocity that caused the truck to move backwards.

Teacher-researcher-timeline interactions

In the first meeting, we introduced the Teacher to our multimodal timeline, discussing its components and some of the possible inferences that we could derive from it. The Teacher first observed each scenario without the timeline and then interacted with the timeline to identify student difficulties and consider possible interventions. The discussion was gener- ally open-ended, but we provided some guidance by asking the Teacher the following ques- tions: (1) how do you identify students' difficulties? (2) how do you decide whether or not to intervene? and (3) if you intervene, what guides your interventions?

In the second meeting, we went through the two scenarios again but focused on the feed- back the Teacher would give the students. We extended the video viewing past the original stopping points to give the Teacher a more complete picture of how the students worked further to address their difficulties. In both scenarios, the students summoned a researcher who provided help. We inquired about the Teacher's views on the researcher's feedback and his own feedback strategies. We also wanted to know what guided the Teacher's feedback strategies. At the end of each interview, we sought the Teacher's impressions of the timeline and his suggestions for improving its utility in identifying and supporting student difficulties in C2STEM.

RESULTS

Pursuant to the Figure 3 framework, our findings reveal two primary ways in which our teacher-researcher-AI collaboration using the multimodal timeline can help identify students' challenges and craft supportive feedback: (1) the Teacher can use the multimodal timeline to derive a more holistic view of students' problem-solving processes from which to better guide his feedback, and (2) the researchers can leverage the Teacher's feedback towards iteratively refining the multimodal timeline to make it more useful to the Teacher. We discuss both of these in the subsections that follow.

Teacher: Using AI timeline to support student feedback

During discussions about the AI-generated multimodal timeline, the Teacher identi- fied key moments for its use to support student learning. We refer to these moments as *inflection points* (Munshi et al., 2023) and highlight two pivotal ones: (1) a *difficulty threshold* when students encounter a challenge, and (2) an *intervention point* when a teacher decides to provide feedback. AI-based pedagogical agents often merge these inflection points, using strategy-based triggers for intervention (eg, Basu et al., 2017). Contrary to this approach, the Teacher emphasized the importance of allowing students to navigate through difficulties. He discussed the idea of *productive failure* and students' need to *develop debugging skills*—key elements in open-ended learning environments (Kapur, 2008; Land,

2000).

An important consideration is the gap between the difficulty threshold and interven- tion point, that is, the *feedback latency interval* that allows teachers to observe student behaviour and craft tailored feedback based on their observations of students' domain understanding and problem-solving approaches. In our case, this involved the teacher reviewing students' timelines after the fact, but in the future we envision teachers sitting at their classroom desks, monitoring students' work on aggregated timelines as they work on their problem-solving tasks in the C2STEM environment. Importantly, leveraging the feedback latency interval to more effectively support students has broader implications for teachers' instructional practices, which we discuss in sixth section. We explore the dynamics of the difficulty threshold and intervention point in the paragraphs that follow, along with the specific modalities the Teacher found most useful for informing his inflec- tion point determinations and decisions.

Scenario 1

Using our multimodal timeline, we can leverage the collaboration between the teacher, researchers and AI to identify the point at which students begin to experience difficulties in C2STEM. With both Scenarios (see section 'Scenario details'), the Teacher was asked to identify what he believed to be the difficulty threshold, and he highlighted specific com- ponents (ie, modalities) of the timeline that he felt best helped him identify these inflection points:

…the words on its own, the timeline, looks like one of the good indicators here is that they're actually talking more…but with very few actions…

I mean, watching the screen, I could also see that they were just puzzling over what was going on there without doing anything…

…so the lack of action might be a good indicator…

In these examples from Scenario 1, the Teacher recognized the transcript, speech (audio), students' video and actions in the C2STEM environment as influencing his determination of the difficulty threshold. He also recognized frustration and boredom (disengaged behaviour) as being particularly useful for making his determination (eg, 'I'd like to know if there's some way to identify…that frustration point, somebody like disengaging from the computer…'). Disengagement included situations where the students looked off screen (identified by gaps in affect detection; '…he's frustrated, and she's off screen here…'). Students repeating ac- tion sequences without making adjustments to the model (eg, repetitive 'visualize + execute' sequences) could also indicate a lack of progress, he remarked ('…it was just running the same thing over and over…').

In Scenario 1, students encountered the difficulty threshold when they demonstrated a misunderstanding about the relationship between position, velocity and acceleration while translating these relations into computational form. The students used C2STEM's graph tool to assess the truck's velocity and acceleration but were unable to figure out how to make the truck slow down. In this instance, the teacher relied heavily on the students' conversations to identify the difficulty threshold: '…they are misaligning in their speech (the acceleration, the velocity)…'. Leveraging the feedback latency interval, the Teacher remarked that the stu- dents' difficulties 'are occurring in lulls of their interest or their competence' (determined by their speech). He then inquired about whether the students had figured out how to make the truck cruise at a constant velocity. At this point, he realized the students '…are not changing velocity, and they're not actually using the velocity to mark the positions…', which was illus- trated via the students' computational model on the screen.

The Teacher emphasized timeline gaps as being particularly informative. He mentioned that these gaps are important for detecting difficulty thresholds and could similarly be used to inform his specific intervention decisions. His continued focus on these gaps from both this and previous work (Cohn, Snyder, et al., 2024) is notable:

…I think going just away from the screen…just the gap is enough of an indicator already if that is in fact away from the screen…I still think just in general, being away from the screen means something else is going on. That means enough to me right now…

…I'd like to know if there's some way to identify that kind of, that lull…

…that was the lull where I should be intervening because now they're, they are just stuck…

Given the 'lulls' (ie, timeline gaps), and the students' failure to connect position and ve- locity computationally, the Teacher ultimately decided he would intervene. It is worth noting that the Teacher supplemented the use of this timeline with his own teaching experience, stating, '…listening to them, I feel instinctually called to intervene'. He used the timeline to *augment* (not supplant or suppress) this experiential instinct by considering facets of the stu- dents' behaviour and understanding to help guide his intervention strategy. As discussed, the Teacher referenced several reasons, derived from several modalities, for wanting to intervene (eg, timeline gaps and conceptual misunderstandings in their speech). However, he also used the timeline's screen recording to recognize that the students were exhibiting an effective debugging strategy by using the graph tool ('…they've been using the graph effectively…'). He weighed all of this information before ultimately deciding the students in Scenario 1 required an intervention, at which point he chose his intervention point by scroll- ing to a particular point in the video, saying, '…it's in this range right here that I feel like I need to intervene…'.

Scenario 2

In Scenario 2, the students first tested different lookahead distance values using a trial- and-error approach. They were unable to debug the truck moving backwards systemati- cally, that is, by isolating variables to pinpoint the source of the truck's erroneous behaviour. Interestingly, the Teacher remarked that students' 'delight' on the timeline could play a role in identifying the students' difficulty threshold, as they exhibited positive valence (likely 'sur- prise') when their model performed in a manner other than intended ('…we hit the same delight and then kind of, dropping curve, into inattention, which actually makes me think that delight is the surprise…you found something unexpected…'). The Teacher also noted boredom and off-screen behaviour as signs of disengagement ('…the [student] on the right's affect is boredom, while the student on the left is off screen…'). Just as in Scenario 1, the Teacher used the timeline to help formulate his understanding of the difficulty students were experiencing ('I got the sense that that was as much because she was irritated, it wasn't working, and was just trying something'). Unlike Scenario 1, the Teacher's focus on student affect in Scenario 2 suggests a reliance on emotion over discourse for gauging difficulty thresholds (although the Teacher did refer to students' emotions in both Scenarios).

The Teacher opted not to intervene in Scenario 2. He stated he would allow students to continue problem-solving, as he believed their issue lay in their computational model development (derived from the screen recording) rather than their physics understanding (derived from discourse audio; 'I don't think that necessarily means anything was wrong with their physics, I think it does mean that they hadn't fleshed out the computational model in the code for themselves'.). Despite encountering the difficulty, the students continued to problem-solve and investigate the source of their misunderstanding. The Teacher felt the students should be afforded the opportunity to continue debugging their code as long as

they continued to interact with the environment meaningfully (ie, by performing new actions in the environment and not random or repeated ones; 'I don't think I'd want to intervene here unless something about that never went back to a new action'.).

610 |		COHN ET AL.
TABLE 1	Timeline modalities considered by the Teacher for identifying inflection	points in both Scenarios.
	Difficulty threshold	**Intervention point**
Scenario 1	***Student Emotions***	***Discourse Audio***
	Student Video	***Environment Screen***
	Discourse Transcript	***Recording***
	Discourse Audio	
	Environment Actions	
Scenario 2	***Student Emotions***	***Discourse Audio***
	Student Video	***Environment Screen***
		Recording
		Environment Actions

Note: Modalities used in more than one context are colour-coded accordingly. Items in **bold** highlight similarities across Scenarios for that particular inflection point.

Use of data modalities and timeline features to support student problem-solving

The Teacher considered multiple modalities and timeline features to identify both sets of in- flection points (ie, difficulty thresholds and intervention points), and these modality sets dif- fered across Scenarios. Table 1 presents the primary modalities considered by the Teacher in various contexts, based on our findings above.

Table 1 illustrates the diverse array of modalities the Teacher considered while identifying inflection points across Scenarios. The shift in modalities considered by the Teacher in dif- ferent contexts highlights the dynamic interpretability of the timeline. In each Scenario, and for each inflection point, the Teacher was able to focus on the modalities he deemed most informative. For instance, the Teacher relied on Student Emotions and Student Video for identifying difficulty thresholds in both Scenarios. Similarly, he focused on Discourse Audio and Environment Screen Recordings for determining intervention points. This suggests that the Teacher prioritized students' affective states, and their visual appearances and interac- tions, more heavily while determining if they were experiencing difficulties. Conversely, the content of the students' discourse and their computational models (shown via the screen recording) played a larger role in his intervention decisions.

Just as in previous work (Cohn, Snyder, et al., 2024), the Teacher emphasized the impor- tance of being able to *visualize* student data, which was the original impetus for our timeline ('…it's just a good indicator; it's a nice visual way to see when things went wrong…'). This is further evidenced by his reliance on the Student Video, Student Emotions and Environment Screen Recording modalities in **bold** in Table 1. Additionally, the Teacher suggested creat- ing graphical representations of student knowledge states: '…a matrix or even a web idea, and you're just trying to figure out how far in any direction they can take the knowledge that would be like this holistic idea. Like this is what their

knowledge is…'. This suggests that the Teacher values visual tools not only for tracking students' progress and understanding their difficulties, but also their potential to offer comprehensive, multidimensional views of students' domain knowledge.

Crucially, the Teacher expressed interest in being aware of difficulty thresholds (even when interventions were not required), especially if they recur across groups, to inform po- tential future interventions: '…I wouldn't want to intervene here, but would want to know that this had happened, and how much it happened across groups'. In a real-time class- room setting, we envision the Teacher monitoring students' timelines from his desk. When students encounter a difficulty threshold, that specific timeline could be highlighted on the screen to alert the Teacher to the specific student(s) encountering the difficulty. At that point, the Teacher could use the timeline to observe students and decide whether and how to intervene. This approach would also allow for aggregated metrics and reporting, as we could use AI to analyse information from *every* timeline to alert the teacher to difficulties and misunderstandings that permeate the classroom.

In both Scenarios, the Teacher used insights from the AI-generated multimodal timeline to determine intervention timing and type, differentiating between classroom-level interventions that address the entire class and group-level interventions tailored to specific groups or indi- viduals. Classroom-level interventions were deemed important by the Teacher, as a single group's struggles often reflect widespread misconceptions. For example, when students in Scenario 1 misunderstood the velocity-acceleration relationship, the Teacher preferred to offer corrective, formative feedback to *all* students instead of just those encountering the difficulty:

…the baseline stuff's not happening, and so that feels like a thing that the whole class also needs to have reminded of…so from a formative standpoint, I feel like I need to go back and rehash the simulation steps and how the acceleration affecting velocity expecting position…

…I think I would suspect…that there's plenty of other people that need similar fixes…

At the group level, the Teacher considered intervening in Scenario 1 when students strug- gled with velocity and acceleration concepts, as previously discussed. In Scenario 2, he chose not to intervene, allowing students to problem-solve independently (though they even- tually sought help on their own). The Teacher discussed effective feedback strategies for intervention points, such as prompting students to articulate their difficulties and goals ('…I think I'd still want them to have to articulate first, just to get an idea of what's going on…'). He suggested recalling past successes to connect current challenges to previous knowledge: '…I think calling back to previous knowledge to try to get them to connect something that had been done successfully and try to remember why it was successful is the right move…'. He also recommended crafting reflective questions based on observed difficulties to guide students before intervening, deciding whether feedback should address domain knowledge or problem-solving strategies ('What's the right question to ask in that context to get them thinking about what they're doing'?).

Researcher: Using teacher feedback to improve AI timeline

Just as we analysed how the Teacher used the AI-generated timeline to support students, we studied how researchers can leverage the Teacher's feedback to identify additional useful features for the timeline, as well as refine existing ones, with the goal of improv- ing the timeline's utility and functionality to more effectively support students. The teacher frequently highlighted how he found AI-generated features, actualized through the timeline modalities, useful to support student learning. In particular, he

focused on the intuitiveness of the timeline's visual components and AI's ability to provide a holistic view of student problem-solving:

…it's very intuitive just to follow along…

…I think it's a crucial point, and it's also the inflection point marked again, by AI…so…I think that's a good sign that the AI is picking up on things changing…

…the segmentation was nice to compare what the AI was doing…

In addition to the Teacher using the timeline to identify difficulty thresholds and inform intervention points, he also alluded to using the timeline to plan subsequent class instruction (orchestration). In instances where classroom-level interventions are required, but immediate interventions may be disruptive to students' problem-solving, the Teacher highlighted how he could address students' difficulties post hoc by solidifying their understanding of domain knowledge and encouraging the use of more effective problem-solving strategies:

…I think that given those misunderstandings, and the code, that would be some- thing I would watch for would be my formative feedback to want to go back and address on the next lesson with a whole class… I think I would want to do that with everyone just to make sure that we're all fol- lowing the same steps that would essentially be something I would do before we tackle this again the next day…

Overall, our findings demonstrate that the Teacher found the timeline useful for support- ing students through classroom orchestration. This goes beyond identifying their inflection points and crafting real-time feedback and suggests that the teacher-researcher-AI collab- oration may similarly be worth investigating in future work for curating lesson plans or even designing and refining curricula to better meet student needs.

Our discussions with the Teacher also revealed several insights into how we can improve the timeline. While the Teacher remarked that he found the LLM summaries useful as a reference, he revealed that he did not rely on them for identifying inflection points or guiding interventions. Instead, he found other timeline components (such as the emotion tracks, synergy and social metrics and transcripts) more useful:

…I'm not opposed to summaries per se, it's just I don't find myself reading them often…

…even just the timestamps and scores, but surely turn-taking and synergy scores would probably be more than enough…

…it's nice to see the scores at the bottom…

…like all the text, I tend to go to the visual bars at the bottom first…

The Teacher's underutilization of LLM summaries during this case study is in stark con- trast to previous work, where he relied on students' discourse segment summaries to char- acterize students' synergistic learning (Cohn, Snyder, et al., 2024). This suggests that, in addition to the Teachers' modality preferences differing across scenarios and inflection points (as shown in Table 1), they can similarly differ across tasks depending on the specific needs of the teacher. For this case study, the Teacher remarked that the LLM summaries would be more useful for real-time feedback if they instead characterized students' difficul- ties (ie, what exactly the students were struggling with), as opposed to summarizing their discourse and actions.

The Teacher also expressed interest in knowing which student was more active in an as- signed task, stating it would be helpful to know if one student was doing most of the talking or controlling the laptop more (‘…who's running the mouse? Is there a way to tell that’?). Knowing this, he could encourage the less involved student by asking him or her reflective questions while allowing the more involved student

to continue leading and imparting infor- mation to the less involved partner. In addition, he recommended the *transcripts* be clickable and scrollable, and that they should highlight the specific physics and computing concepts students discuss: '…I do like to use the transcript to skip…I can scan for the words I'm looking for and if they were already highlighted in some fashion…have a filter to jump to the regions of the video I want to pay attention to…'.

Overall, these findings demonstrate how the human-AI partnership with a multimodal timeline can produce effective teacher-researcher-AI collaboration and drive the iterative refinement of technology-based multimodal analytics to support teachers in addressing stu- dents' STEM+C difficulties. From the Teacher's quotes, we can conclude that he found multiple timeline features useful. The Teacher also suggested improvements. His use of

the timeline for identifying inflection points and guiding feedback, along with his timeline critiques, have implications beyond the specific use case we describe in this work (illustrated in Figure 4). We discuss these implications in sixth section.

DISCUSSION AND CONCLUSIONS

Our teacher-researcher-AI collaboration offers insights into leveraging a multimodal time- line for identifying student difficulties and formulating feedback in STEM+C learning. The Teacher's distinction between difficulty thresholds and intervention points was very reveal- ing, as was his using the feedback latency interval to inform decisions on whether and how to intervene. In another example of a human-AI partnership, intelligent tutoring systems are often designed to trigger instructional support *immediately* upon the detection of a prede- termined event (Azevedo et al., 2022; Munshi et al., 2023; Sottilare et al., 2014), treating the difficulty threshold and intervention point as a single inflection point. However, little research has investigated whether refining ITS triggering mechanisms to make use of the lag be- tween these two events would positively affect student behaviours and outcomes.

This deliberate lag also has implications beyond AI-driven educational development and may speak more broadly to teachers' instructional practices. In their work characterizing the *micro-zone of proximal development* (ie, the optimal moment for intervention), Shvarts and Abrahamson (2019) discuss a tutor who chooses not to intervene immediately despite the student's actions deviating from the tutor's perceived ideal actions. By waiting, the tutor allowed the student to solve the problem on her own and was able to select a more optimal intervention point to encourage student reflection and introduce a new conceptualization of the problem to advance the student's understanding. This suggests that defining the diffi- culty threshold and intervention point as two distinct inflection points may have benefits be- yond human-AI educational partnerships, extending to instructional support practices more broadly and calling for further research exploration.

Overall, the Teacher found the AI-generated timeline to be useful in understanding student problem-solving behaviours and their difficulties, and providing more informed feedback, within the context of our framework. As discussed in third section (illustrated in Figure 4), this paper relied more heavily on researchers to address the initial 'Human' tasks (tinted blue in Figure 4; eg, *Transcript Validation and Correction*) that drove our timeline's creation and im- plementation, while the researchers primarily relied on the Teacher to address the *Analysing and Sense Making* and *Identification of Student Difficulties and Development of Support* components. Importantly, we envision our Figure 3 framework extending beyond the specific use case illustrated in Figure 4 to include continued co-design where teachers and research- ers work in tandem to perform tasks like *Auto-Segmentation*, *LLM-Summarization* and *Hand*

Coding, allowing teachers greater agency in the design process (Hutchins & Biswas, 2024; Sarmiento & Wise, 2022). Just as our human-AI collaboration spectrum in Figure 2 illustrates the balance between AI automation and human control, a similar spectrum could be used to highlight the balance between teacher and researcher contributions, and highlight how vari- ous contexts may demand different inputs from both sets of stakeholders.

The Teacher's enthusiasm for using visual representations as mechanisms for under- standing learners' problem-solving behaviours is also informative, supporting findings from previous work (Cohn, Snyder, et al., 2024) that inspired our timeline's creation. Given his reliance on the timeline's visual components, one promising avenue for further research is exploring the use of learner models (Dillenbourg & Self, 1992) and knowledge graphs (Abu-Rasheed et al., 2024) to model both classroom- and group-level representations of students' understanding visually. Echeverria et al. (2019) extend this idea of visualizing stu- dents' domain understanding to also include collaboration. Their approach, *collaborative translucence*, presents key features of group activity visually, which they ground theoreti- cally (in the physical, epistemic, social and affective dimensions of group activity) and con- textually (using domain-specific concepts in the domain of nursing training). We see a similar application to students collaborating in C2STEM, based on the Teacher's insights.

Another point that the Teacher focused on both here and in previous work was using gaps in problem-solving (like pauses or disengagement) to support students. His emphasiz- ing this point across multiple interviews and tasks underscores the significance he places on these gaps for understanding students' problem-solving behaviours, contextualizing the difficulties they encounter and guiding intervention decisions and strategies. Cossavella and Cevasco (2021) previously studied the importance of 'filled' pauses (ie, using filler words such as 'uh' and 'um') in the construction of coherent representations of spoken lan- guage discourse. Historically, these pauses were largely considered performance errors (Chomsky, 1965); however, Cossavella and Cevasco (2021) point out that filled pauses are often used to focus listeners' attention on upcoming speech, which is not necessarily indica- tive of a lack of knowledge or understanding. This speaks to the importance of distinguishing productive pauses (ie, those encouraging further engagement) from those signalling disen- gagement or a lack of conceptual understanding.

The Teacher's emphasis on student disengagement and its identification across multiple modalities motivates the need for collecting and using multimodal data for analysing student learning in open-ended learning environments. His interest in effectively navigating through videos (see Section 5) highlights an opportunity for future work to enhance the AI-generated timeline using LLM-based event identification, which we are actively pursuing in our current research. By allowing teachers to query the LLM for instances of student disengagement, these moments can be automatically marked on the timeline for further exploration to help differentiate productive pauses from conceptual misunderstandings. Leveraging AI-driven insights from teachers to refine educational technologies extends beyond our timeline and can similarly be used to embed adaptive feedback mechanisms in learning environments, automatically responding to students' needs and reducing the instructor's workload in the classroom. This automated scaffolding allows teachers to engage more deeply in analysing and enhancing their pedagogical approaches. As a result, they can focus their attention on struggling students who require direct human interventions, ensuring that all students progress through the curriculum effectively. While many ITSs employ adaptive scaffolding to varying degrees (Anwar et al., 2022; Azevedo et al., 2022; Koike et al., 2021), iteratively refining scaffolding using teacher input remains largely unexplored.

Limitations and future work

Our case study interviews with the Teacher mark an exploratory first step in examining the teacher-researcher-AI partnership and are not intended to present generalizable findings. Instead, we aim to shed light on the notable gap in educational research that characterizes human-AI collaboration in terms of AI and two key stakeholders by demonstrating that in- corporating all three perspectives allows (1) teachers to leverage AI tools to provide more informed feedback to students, and (2) researchers to leverage teacher input for improving AI tools for teachers. We also acknowledge that the Teacher's pedagogical approach highly influences his timeline usage and intervention strategy (eg, leveraging productive failure). This raises questions with regard to the specificity–scalability trade-off in *human-centred design* (HCD), which involves balancing the degree of user specification and customizability against the broader need for scalable solutions (Lyon et al., 2020). Teachers' needs vary, and their sense-making relies heavily on pedagogical standpoint. In the future, we will ex- plore the generalizability and scalability of this partnership with other teachers whose peda- gogical standpoints may differ from that of our Teacher.

Additionally, our case study does not involve a teacher using this tool while working in a real-world classroom. In the future, we envision our timeline dashboard being available to help teachers monitor students' problem-solving behaviours during class instruction and immediately after, alerting teachers to students encountering difficulties and allowing teach- ers to make more informed decisions on whether and how to intervene. Going forward, we will (1) leverage the Teacher's insights from this case study and others to refine our timeline through continued co-design with both teachers and students; (2) investigate the ways in which our timeline may be deployed live in a real-life classroom and (3) develop other AI-driven approaches to identify difficulty thresholds and provide explicit, actionable guidance for educator use in the classroom (as indicated by the dotted lines in Figure 3). We will also extend our analysis supporting the multimodal timeline to inform teacher reflection and de- velop intervention tools that support student problem-solving, whether via a timeline dash- board for teachers, or pedagogical agents that interact directly with students and provide adaptive feedback informed by teacher input.

Finally, the way in which humans interact with technology is not unilateral. As discussed, how teachers interact with technology is largely a product of pedagogical standpoint (eg, the specific timeline modalities our Teacher focused on), and students are likely to adapt to what they perceive as intelligent systems' routines, abilities and weaknesses. While rec- ommending future research directions for spoken language interaction with robots, Marge et al. (2022) state, ‘people invariably form mental models of the artifacts they interact with’. As technology evolves, so do humans' behaviours and expectations, which in turn inspires further technological innovation (eg, mechanisms to reduce ChatGPT hallucinations, tox-icity and misuse). Both parties influence and adjust to each other, adapting and changing based on the other's actions and needs. This *co-regulation* creates a complex and dynamic interaction, emphasizing the need to investigate how introducing a multimodal timeline (and other technologies) in the classroom affects both teacher and learner behaviour.

Conclusion

This case study characterizes human-AI collaboration in terms of a partnership between teachers, researchers and AI. Using our multimodal timeline, the teacher-researcher-AI col- laboration offers a comprehensive view of students' experiences while working on tasks, enabling researchers and teachers

to identify inflection points and derive actionable insights. Our findings encourage further exploration into the potential for human-AI collaboration in education. By examining temporal patterns and trends, it is possible to identify critical mo- ments of difficulty, success and collaboration. Connections between data modalities offer a deeper understanding of the interplay between affective states, social dynamics and learning outcomes. Emotion labelling enables nuanced analysis of students' emotional states and their impact on learning, highlighting opportunities for emotional regulation and support. We believe that this work will inspire future multimodal research exploring the richness and ben- efits of human-AI collaboration in terms of teachers, researchers and other stakeholders.

FUNDING INFORMATION

This work is supported under National Science Foundation awards IIS-2327708 and DRL- 2112635. Any opinions, findings and conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

1. Abu-Rasheed, H., Abdulsalam, M. H., Weber, C., & Fathi, M. (2024). Supporting student decisions on learning recommendations: An LLM-based Chatbot with knowledge graph contextualization for conversational ex- plainability and mentoring. *Joint proceedings of LAK 2024 workshops, co-located with 14th International Conference on Learning Analytics and Knowledge (LAK 2024)*. https://ceur-ws.org/Vol-3667/GenAILA- paper8.pdf
2. Akata, Z., Balliet, D., De Rijke, M., Dignum, F., Dignum, V., Eiben, G., Fokkens, A., Grossi, D., Hindriks, K., Hoos, H., Hung, H., Jonker, C., Monz, C., Neerincx, M., Oliehoek, F., Prakken, H., Schlobach, S., Van Der Gaag, L., Van Harmelen, F., … Welling, M. (2020). A research agenda for hybrid intelligence: Augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. *Computer*, *53*(8), 18–28. https://doi.org/10.1109/MC.2020.2996587
3. Akpanoko, C. E., Ashwin, T. S., Cordell, G., & Biswas, G. (2024). Investigating the relations between students' affective states and the coherence in their activities in open-ended learning environments. In *Proceedings of the 17th International Conference on Educational Data Mining* (pp. 511–517). International Educational Data Mining Society. https://doi.org/10.5281/zenodo.12729872
4. Akpanoko, C. E., & Biswas, G. (2024). The interplay of affective states and cognitive processes in an open-ended learning environment: A case study. In *Proceedings of the 18th International Conference of the Learning Sciences-ICLS 2024* (pp. 873–880). International Society of the Learning Sciences, Buffalo, NY, USA.
5. Anwar, A., Haq, I. U., Mian, I. A., Shah, F., Alroobaea, R., Hussain, S., Ullah, S. S., & Umar, F. (2022). Applying real-time dynamic scaffolding techniques during tutoring sessions using intelligent tutoring systems. *Mobile Information Systems*, *2022*(1), 6006467.
6. Astuti, N. H., Rusilowati, A., & Subali, B. (2021). STEM-based learning analysis to improve students' problem solving abilities in science subject: A literature review. *Journal of Innovative Science Education*, *10*(1), 1. https://doi.org/10.15294/jise.v9i2.38505
7. Azevedo, R., Bouchet, F., Duffy, M., Harley, J., Taub, M., Trevors, G., Cloude, E., Dever, D., Wiedbusch, M., Wortha, F., & Cerezo, R. (2022). Lessons learned and future directions of metatutor: Leveraging multichan- nel data to scaffold self-regulated learning with an intelligent tutoring system.

Frontiers in Psychology, *13*, 813632.

8. Baker, M. J. (2015). Collaboration in collaborative learning. *Interaction Studies*, *16*(3), 451–473. https://doi.org/ 10.1075/is.16.3.05bak
9. Basu, S., Biswas, G., & Kinnebrew, J. S. (2017). Learner modeling for adaptive scaffolding in a computational thinking-based science learning environment. *User Modeling and User-Adapted Interaction*, *27*(1), 5–53. https://doi.org/10.1007/s11257-017-9187-0
10. Basu, S., Biswas, G., Sengupta, P., Dickes, A., Kinnebrew, J. S., & Clark, D. (2016). Identifying middle school stu- dents' challenges in computational thinking-based science learning. *Research and Practice in Technology Enhanced Learning*, *11*(1), 13. https://doi.org/10.1186/s41039-016-0036-2
11. Blikstein, P., & Worsley, M. (2016). Multimodal learning analytics and education data mining: Using computational technologies to measure complex learning tasks. *Journal of Learning Analytics*, *3*(2), 2. https://doi.org/10. 18608/jla.2016.32.11
12. Bokhove, C., & Downey, C. (2018). Automated generation of 'good enough' transcripts as a first step to transcrip- tion of audio-recorded data. *Methodological Innovations*, *11*(2), 2059799118790743. https://doi.org/10.1177/ 2059799118790743
13. Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press.
14. Coburn, C. E., & Penuel, W. R. (2016). Research–practice partnerships in education: Outcomes, dynamics, and open questions. *Educational Researcher*, *45*(1), 48–54. https://doi.org/10.3102/0013189X16631750
15. Cohn, C. (2020). BERT efficacy on scientific and medical datasets: A systematic literature review. *College of Computing and Digital Media Dissertations*, *24*, 50–56. https://via.library.depaul.edu/cdm_etd/24
16. Cohn, C., Hutchins, N., Le, T., & Biswas, G. (2024). A chain-of-thought prompting approach with LLMs for evaluat- ing students' formative assessment responses in science. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(21), 21. https://doi.org/10.1609/aaai.v38i21.30364
17. Cohn, C., Snyder, C., Montenegro, J., & Biswas, G. (2024). Towards a human-in-the-loop LLM approach to col- laborative discourse analysis. In A. M. Olney, I. A. Chounta, Z. Liu, O. C. Santos, & I. I. Bittencourt (Eds.), *Artificial intelligence in education. Posters and late breaking results, workshops and tutorials, industry and innovation tracks, practitioners, doctoral consortium and blue sky. AIED 2024. Communications in Computer and Information Science* (Vol. *2151*). Springer. https://doi.org/10.1007/978-3-031-64312-5_2
18. Cossavella, F., & Cevasco, J. (2021). The importance of studying the role of filled pauses in the construction of a coherent representation of spontaneous spoken discourse. *Journal of Cognitive Psychology*, *33*(2), 172–186. https://doi.org/10.1080/20445911.2021.1893325
19. Cukurova, M. (2024). The interplay of learning, analytics and artificial intelligence in education: A vision for hybrid intelligence. *British Journal of Educational Technology*, 1–20. https://doi.org/10.1111/bjet.13514
20. Dillenbourg, P., & Self, J. (1992). A framework for learner modelling. *Interactive Learning Environments*, *2*(2), 111–137. https://doi.org/10.1080/1049482920020202
21. Echeverria, V., Martinez-Maldonado, R., & Buckingham Shum, S. (2019). Towards collaboration translucence: Giving meaning to multimodal group data. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–16). Association for Computing Machinery (ACM).
22. Emara, M., Hutchins, N. M., Grover, S., Snyder, C., & Biswas, G. (2021). Examining student

regulation of collabo- rative, computational, problem-solving processes in open-ended learning environments. *Journal of Learning Analytics*, *8*(1), 49–74.

23. Emerson, A., Cloude, E. B., Azevedo, R., & Lester, J. (2020). Multimodal learning analytics for game-based learn- ing. *British Journal of Educational Technology*, *51*(5), 1505–1526. https://doi.org/10.1111/bjet.12992
24. Fonteles, J., Davalos, E., Ashwin, T. S., Zhang, Y., Zhou, M., Ayalon, E., Lane, A., Steinberg, S., Anton, G., Danish, J., Enyedy, N., & Biswas, G. (2024). A first step in using machine learning methods to enhance inter- action analysis for embodied learning environments. In A. M. Olney, I. A. Chounta, Z. Liu, O. C. Santos, & I. Bittencourt (Eds.), *Artificial intelligence in education. AIED 2024. Lecture notes in computer science* (Vol.
25. *14830*). Springer. https://doi.org/10.1007/978-3-031-64299-9_1
26. Graesser, A. C., Conley, M. W., & Olney, A. (2012). Intelligent tutoring systems. In *APA educational psychology handbook, Vol 3: Application to learning and teaching* (pp. 451–473). American Psychological Association. https://doi.org/10.1037/13275-018
27. Grover, S., & Pea, R. (2013). Computational thinking in K–12: A review of the state of the field. *Educational Researcher*, *42*(1), 38–43. https://doi.org/10.3102/0013189X12463051
28. Gupta, S. K., Ashwin, T. S., & Reddy Guddeti, R. M. (2018). CVUCAMS: Computer vision based unobtrusive class- room attendance management system. In *2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT)* (pp. 101–102). Institute of Electrical and Electronics Engineers (IEEE).https://doi.org/ 10.1109/ICALT.2018.00131
29. Hall, R., & Stevens, R. (2015). Interaction analysis approaches to knowledge in use. In *Knowledge and interaction*.
30. Routledge.
31. Hannafin, M. J., Hill, J. R., Land, S. M., & Lee, E. (2014). Student-centered, open learning environments: Research, theory, and practice. In J. M. Spector, M. D. Merrill, J. Elen, & M. J. Bishop (Eds.), *Handbook of research on educational communications and technology* (pp. 641–651). Springer. https://doi.org/10.1007/978-1-4614- 3185-5_51
32. Hatch, J. A. (2002). *Doing qualitative research in education settings*. State University of New York Press. https:// muse.jhu.edu/pub/163/monograph/book/4583
33. Holmlund, T. D., Lesseig, K., & Slavit, D. (2018). Making sense of "STEM education" in K-12 contexts. *International Journal of STEM Education*, *5*(1), 32. https://doi.org/10.1186/s40594-018-0127-2
34. Hutchins, N. M., & Biswas, G. (2024). Co-designing teacher support technology for problem-based learning in middle school science. *British Journal of Educational Technology*, *55*(3), 802–822. https://doi.org/10.1111/ bjet.13363
35. Hutchins, N. M., Biswas, G., Maróti, M., Lédeczi, Á., Grover, S., Wolf, R., Blair, K. P., Chin, D., Conlin, L., Basu, S., & McElhaney, K. (2020). C2STEM: A system for synergistic learning of physics and computational thinking. *Journal of Science Education and Technology*, *29*(1), 83–100. https://doi.org/10.1007/s10956- 019-09804-9
36. Hutchins, N. M., Biswas, G., Zhang, N., Snyder, C., Lédeczi, Á., & Maróti, M. (2020). Domain-specific modeling languages in computer-based learning environments: A systematic approach to support science learning through computational modeling. *International Journal of Artificial Intelligence in Education*, *30*, 537–580.

37. Järvelä, S., Nguyen, A., & Hadwin, A. (2023). Human and artificial intelligence collaboration for socially shared regulation in learning. *British Journal of Educational Technology*, *54*(5), 1057–1076. https://doi.org/10.1111/ bjet.13325
38. Kapur, M. (2008). Productive failure. *Cognition and Instruction*, *26*(3), 379–424. https://doi.org/10.1080/07370
39. 000802212669
40. Knight, S., Wise, A. F., & Chen, B. (2017). Time for change: Why learning analytics needs temporal analysis.
41. *Journal of Learning Analytics*, *4*(3), 3. https://doi.org/10.18608/jla.2017.43.2
42. Koike, K., Fujishima, Y., Tomoto, T., Horiguchi, T., & Hirashima, T. (2021). Learner model for adaptive scaffolding in intelligent tutoring systems for organizing programming knowledge. In S. Yamamoto & H. Mori (Eds.), (Vol. *12766*). Human interface and the management of information. Information-rich and intelligent environments. HCII 2021. Lecture notes in computer science, Springer. https://doi.org/10.1007/978-3-030-78361-7_6
43. Kubsch, M., Caballero, D., & Uribe, P. (2022). Once more with feeling: Emotions in multimodal learning analytics. In M. Giannakos, D. Spikol, D. Di Mitri, K. Sharma, X. Ochoa, & R. Hammad (Eds.), *The multimodal learn- ing analytics handbook* (pp. 261–285). Springer International Publishing. https://doi.org/10.1007/978-3-031- 08076-0_11
44. Land, S. M. (2000). Cognitive requirements for learning with open-ended learning environments. *Educational Technology Research and Development*, *48*(3), 61–78. https://doi.org/10.1007/BF02319858
45. Lawpanom, R., Songpan, W., & Kaewyotha, J. (2024). Advancing facial expression recognition in online learning education using a homogeneous ensemble convolutional neural network approach. *Applied Sciences*, *14*(3),
46. 3. https://doi.org/10.3390/app14031156
47. Leelawong, K., & Biswas, G. (2008). Designing learning by teaching agents: The Betty's brain system. *International Journal of Artificial Intelligence in Education*, *18*(3), 181–208.
48. Liu, R., Stamper, J., Davenport, J., Crossley, S., McNamara, D., Nzinga, K., & Sherin, B. (2019). Learning linkages: Integrating data streams of multiple modalities and timescales. *Journal of Computer Assisted Learning*, *35*(1), 99–109. https://doi.org/10.1111/jcal.12315
49. Liu, R., Stamper, J. C., & Davenport, J. (2018). A novel method for the in-depth multimodal analysis of student learning trajectories in intelligent tutoring systems. *Journal of Learning Analytics*, *5*(1), 1. https://doi.org/10. 18608/jla.2018.51.4
50. Lyon, A. R., Brewer, S. K., & Areán, P. A. (2020, November). Leveraging human-centered design to implement modern psychological science: Return on an early investment. *The American Psychologist*, *75*(8), 1067– 1079. https://doi.org/10.1037/amp0000652
51. Marge, M., Espy-Wilson, C., Ward, N. G., Alwan, A., Artzi, Y., Bansal, M., Blankenship, G., Chai, J., Daumé, H., III, Dey, D., Harper, M., Howard, T., Kennington, C., Kruijff-Korbayová, I., Manocha, D., Matuszek, C., Mead, R., Mooney, R., Moore, R. K., … Yu, Z. (2022). Spoken language interaction with robots: Recommendations for future research. *Computer Speech & Language*, *71*, 101255.
52. Molenaar, I., & Knoop-van Campen, C. A. N. (2019). How teachers make dashboard information actionable. *IEEE Transactions on Learning Technologies*, *12*(3), 347–355.

https://doi.org/10.1109/TLT.2018.2851585
53. Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2017). Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, *10*(1), 18–31.
54. Munshi, A., Biswas, G., Baker, R., Ocumpaugh, J., Hutt, S., & Paquette, L. (2023). Analysing adaptive scaf- folds that help students develop self-regulated learning behaviours. *Journal of Computer Assisted Learning*, *39*(2), 351–368. https://doi.org/10.1111/jcal.12761
55. Nasir, J., Kothiyal, A., Bruno, B., & Dillenbourg, P. (2021). Many are the ways to learn identifying multi-modal behavioral profiles of collaborative learning in constructivist activities. *International Journal of Computer- Supported Collaborative Learning*, *16*(4), 485–523. https://doi.org/10.1007/s11412-021-09358-2
56. National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. National Academies Press. https://doi.org/10.17226/13165
57. *Next generation science standards: For states, by states*. (2013). National Academies Press. https://doi.org/10.
58. 17226/18290
59. Olsen, J. K., Sharma, K., Rummel, N., & Aleven, V. (2020). Temporal analysis of multimodal data to predict col- laborative learning outcomes. *British Journal of Educational Technology*, *51*(5), 1527–1547. https://doi.org/ 10.1111/bjet.12982
60. OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., … Zoph, B. (2024). GPT-4 Technical Report (*arXiv:2303.08774*). *arXiv*. https://doi.org/10.48550/ arXiv.2303.08774
61. Posner, J., Russell, J. A., & Peterson, B. S. (2005). The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology*, *17*(3), 715–734.
62. Rummel, N., Spada, H., & Hauser, S. (2009). Learning to collaborate while being scripted or by observing a model. *International Journal of Computer-Supported Collaborative Learning*, *4*(1), 69–92. https://doi.org/10.1007/ s11412-008-9054-4
63. Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161– 1178. https://doi.org/10.1037/h0077714
64. Sarmiento, J. P., & Wise, A. F. (2022). Participatory and co-design of learning analytics: An initial review of the literature. In *LAK22: 12th International Learning Analytics and Knowledge Conference* (pp. 535–541). Proceedings of Machine Learning Research (PMLR).
65. Savchenko, A. (2023). Facial expression recognition with adaptive frame rate based on multiple testing correc- tion. In *Proceedings of the 40th International Conference on Machine Learning* (pp. 30119–30129). https:// proceedings.mlr.press/v202/savchenko23a.html
66. Sengupta, P., Kinnebrew, J. S., Basu, S., Biswas, G., & Clark, D. (2013). Integrating computational thinking with K-12 science education using agent-based computation: A theoretical framework. *Education and Information Technologies*, *18*(2), 351–380. https://doi.org/10.1007/s10639-012-9240-x
67. Shvarts, A., & Abrahamson, D. (2019). Dual-eye-tracking Vygotsky: A microgenetic account of a teaching/learn- ing collaboration in an embodied-interaction technological tutorial for mathematics.

Learning, Culture and Social Interaction, *22*, 100316.

68. Snyder, C., Hutchins, N., Biswas, G., Emara, M., Grover, S., & Conlin, L. (2019). Analyzing students' synergis- tic learning processes in physics and CT by collaborative discourse analysis. In K. Lund, G. P. Niccolai,
69. E. Lavoué, C. Hmelo-Silver, G. Gweon, & M. Baker (Eds.), *A wide lens: Combining embodied, enactive, extended, and embedded learning in collaborative settings, 13th International Conference on Computer Supported Collaborative Learning (CSCL) 2019* (Vol. *1*, pp. 360–367). International Society of the Learning Sciences.
70. Snyder, C., Hutchins, N. M., Cohn, C., Fonteles, J. H., & Biswas, G. (2024). Analyzing students collaborative problem-solving behaviors in synergistic STEM+C learning. In *Proceedings of the 14th Learning Analytics and Knowledge Conference* (pp. 540–550). Association for Computing Machinery (ACM). https://doi.org/10. 1145/3636555.3636912
71. Sottilare, R. A., Graesser, A., Hu, X., & Goldberg, B. (Eds.). (2014). *Design recommendations for intelligent tutor- ing systems: Volume 2—Instructional management*. US Army Research Laboratory.
72. Sümer, Ö., Goldberg, P., D'Mello, S., Gerjets, P., Trautwein, U., & Kasneci, E. (2023). Multimodal engagement analysis from facial videos in the classroom. *IEEE Transactions on Affective Computing*, *14*(2), 1012–1027. https://doi.org/10.1109/TAFFC.2021.3127692
73. Thompson, K. D., Martinez, M. I., Clinton, C., & Díaz, G. (2017). Considering interest and action: Analyzing types of questions explored by researcher-practitioner partnerships. *Educational Researcher*, *46*(8), 464–473. https://doi.org/10.3102/0013189X17733965
74. Toisoul, A., Kossaifi, J., Bulat, A., Tzimiropoulos, G., & Pantic, M. (2021). Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nature Machine Intelligence*, *3*(1), 42–50.
75. Vrzakova, H., Amon, M. J., Stewart, A., Duran, N. D., & D'Mello, S. K. (2020). Focused or stuck together: Multimodal patterns reveal triads' performance in collaborative problem solving. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge* (pp. 295–304). Association for Computing Machinery (ACM). https://doi.org/10.1145/3375462.3375467
76. Wagner, J., Triantafyllopoulos, A., Wierstorf, H., Schmitt, M., Burkhardt, F., Eyben, F., & Schuller, B. W. (2023). Dawn of the transformer era in speech emotion recognition: Closing the valence gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(9), 10745–10759. https://doi.org/10.1109/TPAMI.2023.3263585
77. Yang, W.-T., Lin, Y.-R., She, H.-C., & Huang, K.-Y. (2015). The effects of prior-knowledge and online learning approaches on students' inquiry and argumentation abilities. *International Journal of Science Education*, *37*(10), 1564–1589. https://doi.org/10.1080/09500693.2015.1045957
78. Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded con- volutional networks. *IEEE Signal Processing Letters*, *23*(10), 1499–1503. https://doi.org/10.1109/LSP.2016. 2603342
79. Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory Into Practice*, *41*(2), 64–70. https://doi.org/10.1207/s15430421tip4102_2