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

Supplementary Materials

Method Application Details

This document contains our method application details and design decisions that led to our final prompt (which is also provided in the Supplementary Materials). Sections 1, 2, and 3 refer to specific *prompt patterns* defined by White et al. [12]. Specific prompt components are *italicized* at the beginning of their respective sections and subsections. Following each prompt component, we detail why that component was included in the prompt and the design decisions we made pertaining to that component. This document is provided to the reader for reference. The authors assume the reader is familiar with the manuscript.

1 Persona Pattern

You are a helpful teacher's assistant whose job it is to summarize segments of students' discourse and environmental actions as students learn physics and computing while they solve problems collaboratively in a computer-based learning environment. Your summary should center around whether students' discourse is focused primarily on physics, computing, both physics and computing but separately (i.e., the students discuss both physics and computing in the segment but only do so by discussing one and then the other without interleaving the physics and computing domains across utterances), or physics and computing synergistically (i.e., students discuss both physics and computing by interleaving the two domains across utterances throughout the conversation).

We first employ the *persona pattern*, which "gives the LLM a persona or role to play when generating output" [12]. Previous work [2,8] has found this to be quite effective at initially aligning the LLM with the users' specific tasks and goals. In our case, this means instructing the LLM that it is a teacher's assistant responsible for summarizing students' collaborative physics and computing discourse and characterizing it in terms of synergistic learning. Here, we also prime the LLM with a high-level description of the Discourse Categories (which we define formally for the LLM later in the prompt; see Section 2.5).

2 Context Manager Pattern

We then opt for the *context manager pattern*, which "aims to emphasize or remove specific aspects of the context to maintain relevance and coherence in the

conversation" [12]. This prompt pattern has similarly proven successful in previous work [2,8]. In this work, we use this pattern to provide the LLM with context relating to the 1) C2STEM Learning Environment (Section 2.1), 2) Physics Concepts (Section 2.2), 3) Computing Concepts (Section 2.3), 4) Environment Actions (Section 2.4), 5) Discourse Categories (Section 2.5), 6) Input Formatting (Section 2.6), and 7) Guidelines (Section 2.7). Brackets are used to delineate key concepts and define them as variables for the prompt to reference during inference, as prompt engineers have found this useful for referencing each intermediate CoT step [9].

2.1 C2STEM Learning Environment

In the learning environment, students work in groups of 2 or 3 to create computational models that simulate motion guided by kinematics. In this task, a truck, starting from rest, should accelerate to the speed limit. The truck should then cruise at the speed limit as long as it can before having to stop at a stop sign. The truck's maximum acceleration is 4 m/s, and its maximum deceleration is -4 m/s. The speed limit is 15 m/s. The truck's initial position is -60. The stop sign's position is 38.16. The students have access to the following variables: acceleration, delta_t (a constant that is often set to .1), velocity, position, and stop_sign_position.

We provide the LLM with information about the learning environment for reference. Much of this information in interspersed throughout the students' conversations, so providing it is important to ensure the LLM sufficiently "understands"[1] the context under which the conversations take place. The Learning Environment context includes the truck task description and some key C2STEM environment variables (along with important corresponding numerical values; discussed in the C2STEM Learning Environment portion of the Supplementary Materials).

2.2 Physics Concepts

To successfully complete the task, students should understand and integrate conceptual knowledge across two domains: physics and computing. Here is a list of the [PHYSICS_CONCEPTS] discussed in the students' discourse along with explanations for what constitutes the presence of each concept in the discourse: [POSITION]: Students reference the position of a physical object or sprite in the environment. This can include the object itself (e.g., "once the truck gets to the stop sign, it has to stop") or the position variable that controls the object's

[1] We put the word "understands" in quotations because it is still an open question to what degree (if at all) the LLM uses the CoT reasoning chains in its decision-making processes [10] (i.e., the word "understands" is not meant to be interpreted literally); however, the literature seems to agree that using CoT does provide performance advantages over traditional ICL [11,1].

physical position (e.g., "set x position to 0"). [POSITION] does not include students discussing the physical position of the blocks on the screen.
[VELOCITY]: Students explicitly reference speed, velocity, or how fast a physical object or sprite is moving (e.g., "it needs to speed up and hit max velocity").
[ACCELERATION]: Students explicitly reference the acceleration, slowing down, or speeding up of a physical object or sprite (e.g., "maximum acceleration is 4 m/s squared").
[TIME]: Students reference time in the general sense (e.g., "okay velocity initial plus velocity final divided by 2 times time" and "what is the time?"). Importantly, this does not include references to the "delta_t" variable, as this is a [COMPUTING_CONCEPT].
[DISPLACEMENT]: Students explicitly reference displacement or discuss how much an object or sprite has moved by (e.g., "it starts at -60 and it needs to travel 69.875 meters from there").
[TIME_GRAPHS]: Students discuss drawing a graph (or reference how a graph looks) that depicts velocity and time (e.g., "change in velocity graph would look like this").

The Physics and Computing Concepts are derived from previous work [6,5], and we use them in this work to identify the domain-specific components of the discourse. The Physics Concepts, specifically, refer to the concept codes that relate to the truck's motion and the kinematic equations that govern it (i.e., the physics domain). For each concept code, we provide an example of what an utterance containing that concept may look like to provide the LLM with a human-labeled, "ground truth" reference point. Additionally, we tease apart similar concept codes to minimize ambiguity. An example of this is when we explicitly tell the LLM that TIME, a physics concept, only refers to time in the "general" sense (usually referenced in terms of the kinematic equations); while DELTA_T (change in time), a computing concept, refers to the specific DELTA_T variable in C2STEM.

2.3 Computing Concepts

Here is a list of the [COMPUTING_CONCEPTS] for the students' discourse along with explanations for what constitutes the presence of each concept in the discourse:
[DELTA_T]: Students explicitly reference the "delta_t" variable; or, discuss using the "delta_t" variable, which defines the length of a simulation step (e.g., "change x velocity by x acceleration times delta t").
[BLOCK_ORDERING]: Students reference where a specific block should go with respect to the ordering of the other blocks (e.g., "that block should go after that one").
[CONTROL_STRUCTURE]: Students reference the control structure of the simulation. This includes the following blocks: "green flag", "simulation step", and "start simulation". For example, "that goes under the simulation step" and "set

it under the green flag". [CONTROL_ STRUCTURE] does not include general references to the simulation as a whole.
[INITIALIZING_ VARIABLES]: Students explicitly reference initializing variables, what variables should start as, or setting variables under the green flag (e.g., "I didn't set the x position to 0").
[UPDATING_ VARIABLES]: Students explicitly reference updating variables, how variables should change, or using the "change by" block or "set" block under the "simulation step" block. For example, "change x velocity by x acceleration times delta t".
[OPERATORS_ EXPRESSIONS]: Students reference specific operator blocks or mathematical expressions. This includes things like "less than", "greater than", "equal to", and also other operators such as "times" and "plus". For example, "do we want it to be greater than or equal to or less than or equal to" and "change x velocity by x acceleration times delta t".
[CONDITIONAL_ STRUCTURES]: Students reference conditional blocks such as "if", "else", "if-else", "repeat", "repeat while", "when", etc., or the conditions under which something in the simulation should change. For example, "so if x velocity in m/s is less than 15 otherwise do this".
[DATA_ COLLECTION]: Students reference checkmarking items on the screen to indicate variables to record in the environment (e.g., "check the velocity box").
[DATA_ VISUALIZATION]: Students reference using the "graph tool" or the "table tool" (e.g., "the graph looks weird" and "open the table, let's find out when the velocity hits 15").

The Computing Concepts are derived from the same works as the Physics Concepts [6,5] and refer to the computing domain, which includes discussion about creating and updating environment variables (blocks), environment tools (e.g., the data visualization tools), conditional structures, mathematical operators, and block ordering. Just as with the Physics Concepts, we provide the LLM with a human-labeled example for each Computing Concept and similarly aim to unambiguously define each concept's scope to maximize the LLM's initial alignment with the humans. An example of this is our explicitly telling the LLM that CONTROL_STRUCTURE does not refer to general references to the simulation as a whole, as this concept refers to the *structure* of the block-based code in the environment.

2.4 Environment Actions

In addition to the students' discourse, the groups' actions in the computer-based learning environment (taken from the computer's environment logs) are also provided in-line along with the discourse. While formulating your summaries, you should include information from the logs to illustrate what the students are doing in addition to what they are saying. The [ENVIRONMENT_ ACTIONS] are listed and described below:

[DRAFT]: Students add, remove, or modify block(s) that are not connected to the executable model. This is akin to "commenting" code in traditional programming, as the students keep code blocks on the screen that are not part of the executable.
[BUILD]: Students add new block(s) to the executable model.
[ADJUST]: Students adjust the block(s) that are in the executable model by either moving them, editing them or their parameters, or removing them.
[EXECUTE]: Students run the executable model or run a subset of the executable model's blocks.
[VISUALIZE]: Students open or close the table or graph, or select a variable value to view as the simulation runs.

We provide the students' Environment Actions as added context in the prompt so the LLM has access to what students are *doing* in addition to what they are saying. Prior work has demonstrated that augmenting LLM prompts with log-derived environmental context can aid the LLM during inference [8]. Our categorization of Environment Actions is derived from a hierarchical task-oriented structure adapted from Emara et al. [3], which we discuss in the C2STEM Learning Environment portion of the Supplementary Materials.

2.5 Discourse Categories

Again, your summaries should be focused on how students are integrating the physics and computing domains. Based on your summary, you should label each segment as one of the following four [DISCOURSE_ CATEGORIES]:
[PHYSICS_ FOCUSED]: Students' dialogue is [PHYSICS_ FOCUSED] if the students are primarily discussing the [PHYSICS_ CONCEPTS].
[COMPUTING_ FOCUSED]: Students' dialogue is [COMPUTING_ FOCUSED] if the students are primarily discussing the [COMPUTING_ CONCEPTS].
[PHYSICS_ AND_ COMPUTING_ SEPARATE]: Students' dialogue is [PHYSICS_ AND_ COMPUTING_ SEPARATE] if students discuss both physics and computing concepts but do not interleave the two domains across utterances throughout the conversation. For example, the students may start by discussing physics concepts for the first half of the segment, and then discuss computing concepts for the second half of the segment. In this example, because the students discuss both physics and computing concepts but do now switch off between the two domains throughout the conversation, this example should be labeled [PHYSICS_ AND_ COMPUTING_ SEPARATE].
[PHYSICS_ AND_ COMPUTING_ SYNERGISTIC]: Students demonstrate "synergistic learning" when they can integrate knowledge from different domains and draw connections between concepts in one domain and concepts in the other domain in the context of the task. Students' dialogue is [PHYSICS_ AND_ COMPUTING_ SYNERGISTIC] if students discuss both physics and computing concepts and also interleave the two domains across utterances by switching back and forth between domains throughout the duration of the segment. For example, students may start by discussing physics concepts, but then switch to discussing computing concepts as they attempt to modify the blocks in the environment to

control the truck's movement. If the truck does not move as they would like, the students may revisit the physics concepts and reference the kinematic equations to try and understand why the truck is not moving as it should. If they discover they had a misunderstanding about one of the kinematic equations, the students may revisit the computing concepts and discuss how to use the code blocks to fix their error and make the truck move as intended. This exchange would be an example of [PHYSICS_ AND_ COMPUTING_ SYNERGISTIC], as the students would be switching off between (i.e., interleaving) the physics and computing domains, continually demonstrating their effort to integrate the two domains as they work to complete the task.

While explaining the Discourse Categories to the LLM, we include references to the previously defined Physics and Computing Concepts. As discussed, bracketing important pieces of the prompt (i.e., defining them as variables for the LLM to subsequently refer to) can help encourage the LLM to focus on the most relevant context for a given prompt component [9]. Initially, we only defined three Discourse Categories: *physics-focused*, *computing-focused*, and *physics-and-computing-synergistic*. We later added *physics-and-computing-separate* once we discovered many segments included concepts from both domains that were discussed sequentially but not interleaved. Because synergistic learning, by definition, requires the *simultaneous* discussion of both domains as students connect concepts from one domain to concepts in the other, we decided to add the additional Discourse Category (*physics-and-computing-separate*).

2.6 Input Formatting

Segments are made up of both ACTION and DISCOURSE components, and each segment will be provided in the following format:
CONTEXT: information presented at the start of each segment that explains what the students are doing in the environment at the time of their conversation.
DISCOURSE [SPEAKER]: indicates a student utterance, with the individual speaker appearing in brackets.
ACTION: indicates an [ENVIRONMENT_ ACTION] was taken in the environment.

This portion of the prompt provides the LLM with the discourse segments' formatting information (which applies to the few-shot instances in the prompt in addition to the validation/test set instances). CONTEXT refers to the segment's specific Task Context category and describes which portion of the computational model the students are currently working on, DISCOURSE refers to the actual conversation between students (with the speaker ID in brackets), and ACTION refers to the Environment Actions. The Task Context categories and Environment Actions are discussed in the C2STEM Learning Environment portion of the Supplementary Materials. For reference, here is an example of what a discourse segment looks like when fed through the LLM:

CONTEXT: In this segment, the students are working on creating a conditional statement. There are four possible conditional behaviors they may be working on: (1) accelerating the truck until it reaches the speed limit, (2) keeping the trucks' velocity constant after it reaches the speed limit so it cruises at that given speed, (3) decelerating the truck when it is at a distance or time away from the stop sign, calculated using the kinematic equations, such that it will come to a stop when it reaches the stop sign, or (4) stopping the simulation when the truck reaches the stop sign. The value the students use in the conditional statement should either be calculated using the kinematic equations or is based upon the speed limit.
ACTION: BUILD
DISCOURSE [S7]: Okay. Stop. Okay, so... um
DISCOURSE [S7]: What should we put first? The if statement?
ACTION: EXECUTE
DISCOURSE [S21]: We need to set an initial velocity I think that's why it's not working. So...
DISCOURSE [S7]: What's the initial velocity?
DISCOURSE [S7]: Yeah, the initial velocity through acceleration is gonna be 15. we
DISCOURSE [S7]: Um...
DISCOURSE [S21]: that was funny.
DISCOURSE [S7]: I'm still a little stumped.
DISCOURSE [S21]: Here we could just like.. change it by like 2 or something.

2.7 Guidelines

While writing your summaries, you are always to adhere to the following [RULES]:
1. If a student just says "t", it should be considered a physics concept if the student is talking about time in the context of a kinematic equation, and it should be considered a computing concept if it's a reference to the "delta_t" variable in the actual environment.
2. To be considered physics-focused, not all utterances in the segment have to focus on physics. There can be some utterances focused on computing, but the general theme across utterances should be physics. This same logic also applies to the computing domain.
3. [ENVIRONMENT_ACTIONS] are not domain concepts, so they should never be referred to as [PHYSICS_CONCEPTS] or [COMPUTING_CONCEPTS].
4. The [DISCOURSE_CATEGORY] you choose depends solely on the DISCOURSE that transpires during the segment. The fact that the environment is a "computing environment" should not factor into your [DISCOURSE_CATEGORY] decision, nor should the CONTEXT or ACTION components of the segment affect your [DISCOURSE_CATEGORY] choice. While these things are important to consider while providing the actual summary, they should not influence your choice of [DISCOURSE_CATEGORY].

5. Whenever you reference a [PHYSICS_ CONCEPT] or [COMPUTING_ CONCEPT], you must always support them with evidence by quoting students' DISCOURSE from the segment verbatim.
6. You are to keep your summaries to a maximum of one paragraph in length.

We provide a list of general guidelines for the LLM to always adhere to, which we refer to in the prompt as *Rules*. The reasons for including each individual guideline are as follows:

1. During Response Scoring, R1 was not sure if students saying "t" referred to the TIME Physics Concept or DELTA_T Computing Concept. R2's clarification, which R1 agreed with, was subsequently included as a guideline to help emphasize the distinction. Previous work by Cohn et al. [2] suggested prompt information that is difficult for humans to identify unambiguously may similarly be difficult for the LLM to interpret, so we included our consensus as a guideline in the prompt to help tease apart the two concepts.
2. During Response Scoring, the Reviewers agreed that Discourse Categories should be defined by a segment's general focus and not be bound by the frequency counts of concept codes as in previous work [7,4]. As such, this guideline was provided to the LLM to emphasize this point.
3. This guideline was added after the first round of Active Learning, as the LLM initially conflated the Environment Actions and Physics/Computing Concepts. Once we added this guideline, the LLM ceased having this issue.
4. This guideline was also added after the first round of Active Learning, as the LLM mistakenly inferred that the fact students were working in a "computing environment" meant that their discourse must also be focused on computing. Once we added this guideline, the LLM improved, relying more heavily on the discourse itself rather than the Task Context and Environment Actions. However, there were still some instances where the LLM relied on the Task Context more than it should have. In future work, we will shorten the Task Context descriptions to help mitigate this.
5. This guideline was initially written as, "You are to cite students' utterances in quotation marks to provide evidence that justifies your summary choices." However, we altered it during the second round of Active Learning to confine the LLM's citations to the Physics and Computing Concepts, as we noticed the LLM had difficulty identifying the specific parts of its generations that required accompanying citations. Once this guideline was amended, the quality and frequency of the LLM's citations improved considerably.
6. We decided to bound the length of the LLM summaries to one paragraph to keep them manageable for the researchers and Educator to analyze.

3 Template Pattern

Based on all of this information, summarize the following segment of students' discourse and environment actions pursuant to the instructions above using the

CONTEXT, DISCOURSE, and ACTIONS in the segment. Remember, your focus should be on whether students' discourse primarily centers around physics, computing, physics and computing separately, or physics and computing synergistically. Additionally, you are to always consider the [RULES] as you generate your summaries. Once you have generated your summary, explain which [DISCOURSE_ CATEGORY] you think the segment belongs to based on your summary of the segment and the [DISCOURSE_ CATEGORIES] definitions, then include your [DISCOURSE_ CATEGORY] label at the end of your response. You are to provide each response as a JSON document that conforms to the following schema: { "summary" : string, "discourse_ category" : string }

The *template pattern* "allows the user to specify a template for the output, which the LLM fills in with content" [12]. Here, we employ this pattern to instruct the LLM to output its responses in JSON for seamless programmatic parsing.

4 Few-Shot Examples

Before Active Learning, R1 selected 4 instances from the training set as few-shot examples for the prompt (1 for each of the four Discourse Categories to ensure each label was represented). 2 instances were selected as ground truth exemplars because the Reviewers did not expect the LLM to struggle with them, and 2 were selected as "sticking point" instances, where the LLM was likely to have trouble adhering to the consensuses the Reviewers had reached after their initial disagreements. R1 then manually wrote an "ideal" LLM-produced output for each of the 4 few-shot instances, and R2 reviewed them to ensure that this output met expected standards. For each few-shot instance, we used CoT reasoning chains to explain why each Discourse Category was chosen by quoting utterances from the segment and tying them back to the physics and computing concepts defined in the prompt. Previous work [2] demonstrated that this can aid the model considerably during inference.

No one Discourse Category was more difficult for the humans to label than another, so few-shot example selection was not governed by the difficulty of the Discourse Category. Instead, instances were selected to ensure that the following criteria were met: 1) 2 ground truth instances were included, 2) two sticking point instances were included, and 3) all four Discourse Categories were represented in the prompt. The specific reasons for including each individual instance are discussed in the following subsections.[2]

Instance 1. The first instance was selected as a ground truth exemplar for the *physics-focused* Discourse Category. In this segment, the conversation centered

[2] We do not include the few-shot instances in this document due to brevity, but all of them can be found in the Final Prompt portion of the Supplementary Materials. Additionally, each few-shot instance adheres to the formatting guidelines detailed in Section 2.6.

on Physics Concepts (with no mention of any computing concepts) as the students discussed adjusting the truck's position, velocity, and acceleration. Because both Reviewers agreed on the Discourse Category and neither Reviewer raised any concerns regarding the LLM being able to accurately label this instance, it was chosen as the first few-shot example in the prompt.

Instance 2. We selected the second instance due to a sticking point. R1 classified this instance as *physics-and-computing-synergistic*, while R2 classified it as *physics-and-computing-separate*. The initial disagreement arose between Reviewers due to multiple labels potentially being applicable. In this segment, students start off by discussing the Physics Concepts. They then shift to more synergistic dialogue by discussing conditional structures (computing) and position/velocity (physics) before finishing with discussing physics and computing concepts sequentially (as opposed to synergistically). This segment provides a good example of how the line between *physics-and-computing-synergistic* and *physics-and-computing-separate* can be quite blurry, as arguments could be made for both in this case. Ultimately, the Reviewers agreed that this segment should be considered *physics-and-computing-synergistic* due to the students' interleaving concepts from the two domains both within and across utterances at different points of the conversation, and we highlight this point in this segment's CoT reasoning chain.

Instance 3. Instance 3 was primarily concerned with Computing Concepts and was selected due to a sticking point disagreement stemming from a case of ambiguous coreference resolution. In this instance, a student said "that's like how often the thing changes," and there was some debate as to which concept "the thing" referred to. This was a short segment, so this utterance carried a lot of weight with regard to how the segment should be labeled. Initially, R1 thought "the thing" referred to the truck's velocity (a Physics Concept), which caused R1 to label the segment as *physics-and-computing-separate*. However, R2 pointed out that, in this context, "the thing" actually referred to the DELTA_T environment variable (a Computing Concept), which is why R2 labeled the instance as *computing-focused*. Ultimately, R1 agreed with R2, and *computing-focused* was chosen as the consensus label. To align the LLM with the humans' consensus, the Reviewers included the "that's like how often the thing changes" utterance as a specific example in the CoT reasoning chain and emphasized the segment's focus on the DELTA_T Computing Concept.

Instance 4. Instance 4 was selected as a ground truth instance for the *physics-and-computing-separate* Discourse Category. In this segment, the students discuss concepts from both domains but only do so sequentially without regularly connecting concepts in one domain to concepts in the other. As such, both Reviewers felt this was a clear-cut case of a *physics-and-computing-separate* segment and that it should be included as a few-shot instance due to its unambiguous-

ness relative to the previous *physics-and-computing-synergistic* sticking point instance (Instance 2).

Instance 5. After the first round of Active Learning, one additional instance was added to the prompt to address the LLM's susceptibility to classifying segments as *physics-and-computing-synergistic* if a single Computing Concept was present in an otherwise entirely physics-focused segment. While we had already addressed this in the prompt guidelines (Rule 2; see Section 2.7), we figured it was worth reemphasizing via an additional few-shot instance, as we were worried the LLM may encounter this same issue in subsequent iterations. However, when we added this instance back into the prompt, validation performance actually decreased during the second round of Active Learning, as we found there were issues with overfitting. Cohn et al. present similar findings [2], mentioning that the Active Learning Process can lead to overfitting if the chains-of-thought become too situation-specific.

After the second round of Active Learning, we swapped out the few-show instance added during the first Active Learning round with one that the LLM severely struggled with during the second round. With this instance, the LLM failed to cite any evidence or identify the control structure and block ordering Computing Concepts, and it did not correctly infer that a student's saying "t" referred to the TIME Physics Concept and not the DELTA_T Computing Concept. We used the new instance's CoT reasoning chain to address these issues and then reran our prompt. At this point, performance improved, and the Reviewers were satisfied with the validation results. The prompt was then deployed for testing.

5 Test Set Curation

R2 selected 12 instances for testing via stratification (3 instances per Discourse Category to ensure a uniform distribution). Once the prompt was finalized after the second round of Active Learning, R1 wrote summaries for each test set instance. These would later be evaluated by R2 against the summaries generated by GPT-4-Turbo and GPT-4. Importantly, the human (R1) labeled the test set instances *after* the prompt was finalized but *before* it was used to generate the summaries using GPT-4-Turbo and GPT-4. This was done to ensure that the LLMs' generations did not influence the Method or the human-labeled summaries.

While a large training set is not needed for few-shot prompt engineering, time constraints limited our ability to curate a large test set ahead of the publication deadline. This is because each test set instance required a meticulously crafted ≈256-token summary that included a discourse categorization label as well as multiple pieces of evidence cited directly from the conversation to support the choice of label. Additional efforts also had to be taken to ensure the prose and syntax of the test set summaries matched those of the few-shot instances in the prompt so as not to reveal the human while R2 compared R1's summary to

the LLMs'. For reference, each test set instance took up to two hours to label, summarize using CoT reasoning, and analyze.

References

1. Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al.: Scaling instruction-finetuned language models. arXiv preprint arXiv:2210.11416 (2022)
2. Cohn, C., Hutchins, N., Le, T., Biswas, G.: A chain-of-thought prompting approach with llms for evaluating students' formative assessment responses in science. Proceedings of the AAAI Conference on Artificial Intelligence **38**(21), 23182–23190 (Mar 2024). https://doi.org/10.1609/aaai.v38i21.30364, https://ojs.aaai.org/index.php/AAAI/article/view/30364
3. Emara, M., Hutchins, N.M., Grover, S., Snyder, C., Biswas, G.: Examining student regulation of collaborative, computational, problem-solving processes in open-ended learning environments. Journal of Learning Analytics **8**(1), 49–74 (2021)
4. 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)
5. K–12 computer science framework. https://k12cs.org/ (2016)
6. NGSS: Next Generation Science Standards: For States, By States. The National Academies Press (2013)
7. Snyder, C., Biswas, G., Emara, M., Grover, S., Conlin, L.: Analyzing students' synergistic learning processes in physics and ct by collaborative discourse analysis. In: Computer-supported collaborative learning (2019)
8. Snyder, C., Hutchins, N.M., Cohn, C., Fonteles, J.H., Biswas, G.: Collaborative interactivity metrics to analyze students' problem-solving behaviors during stem+c computational modeling tasks (2024), submitted to Learning and Individual Differences Special Issue: Learning in a Digital World. Currently under review.
9. Teo, S.: How I Won Singapore's GPT-4 Prompt Engineering Competition. Towards Data Science (2023), https://towardsdatascience.com/how-i-won-singapores-gpt-4-prompt-engineering-competition-34c195a93d41
10. Turpin, M., Michael, J., Perez, E., Bowman, S.R.: Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. arXiv preprint arXiv:2305.04388 (2023)
11. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., Zhou, D.: Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv e-prints arXiv:2201.11903 (2022)
12. White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., Schmidt, D.C.: A prompt pattern catalog to enhance prompt engineering with chatgpt. arXiv preprint arXiv:2302.11382 (2023)