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

Supplementary Materials

LLM Strengths, Weaknesses, and Trade-offs

This document is an extended version of the findings from our analysis of the LLM's strengths, weaknesses, and trade-offs while applying the Method to summarize and characterize students' synergistic learning in their collaborative discourse. It is provided to the reader for reference. The authors assume the reader is familiar with the manuscript.

Strengths

GPT-4-Turbo consistently generated high-quality summaries that effectively captured students' collaborative learning. Similar to prior research [2], the LLM consistently followed prompt instructions, cited relevant discourse pieces similarly to humans (often citing the exact same discourse pieces as R1), and adhered to the CoT reasoning chains outlined in the few-shot examples. Notably, the LLM seamlessly integrated the additional context, including Discourse Categories, environment variables, physics and computing concepts, and XYZ environment activities, into its summaries. For instance, both R1 and GPT-4-Turbo (as well as GPT-4) pointed to the same piece of student discourse, "If x velocity in m/s is... greater than.. 15," as evidence of physics-focused discussion. Such instances were frequent, showcasing the LLM's adept handling of nuanced context. This also led to GPT-4-Turbo providing longer, more detailed summaries than both the human and GPT-4.

GPT-4-Turbo demonstrated several strengths, including its ability to selectively extract relevant information from segments for summarization. It exhibited accurate coreference resolution, correctly identifying entities like physics and computing concepts when ambiguous pronouns such as "it" or "that" were used. Notably, the LLM excelled in adhering to the CoT reasoning chains in the few-shot examples, effectively recognizing ambiguous segments and discerning when multiple Discourse Categories might be applicable. The LLM also showcased 1) adept zero-shot identification of concepts defined in the prompt but not used in the reasoning chains, and 2) a considerable reduction in hallucinations. Neither Reviewer noticed a single *context inconsistency* [3] hallucination (i.e., the LLM contradicts the context defined in the prompt) in our test set. GPT-4-Turbo also exhibited fewer *instruction inconsistency* hallucinations (where the LLM deviates from the user's instructions) [3] compared to GPT-4. However, our sample size was too small to generalize this finding, so we will investigate this on a larger scale in future work.

Another strength is the LLM's ability to pinpoint specific items the human may have missed. The Educator remarked that he initially believed a segment to be *physics-and-computing-synergistic*. After reading the LLM's summary, he realized (as the LLM pointed out) that the students were merely discussing the domains sequentially and not interleaving and forming connections between the cross-domain concepts. The Educator agreed with the LLM that the correct Discourse Category was *physics-and-computing-separate*. In another example, one of R1's summaries was mislabeled due to human error. In this instance, both LLMs agreed about the human error and pointed out R1's misclassification. In fact, in 8/12 instances, both GPT-4-Turbo and GPT-4 agreed in their labeling of the segment, and in only one of those instances did R2 disagree with the LLMs' consensus while scoring. This suggests using *self-consistency*, which is a prompt engineering approach that extends chain-of-thought prompting by sampling different LLM reasoning chains and employing a majority voting scheme to arrive at the final result [4], may be beneficial.

Weaknesses

Despite GPT-4-Turbo's capabilities, there are notable areas for improvement. Some of these issues mirror those found in previous research [2], but to a lesser extent with GPT-4-Turbo in this study. Examples include the occasional non-adherence to the prompt rules (instruction inconsistency) and relying on specific keywords and phrases, as seen in the repetitive use of "In this segment..." at the beginning of each summary, despite this phrase being used infrequently in the few-shot examples. This repetition stems from the heavy use of the phrase in the *Task Context* (See the *Method Application Details* portion of the Supplementary Materials) at the segment's outset, a trade-off discussed shortly.

The autoregressive nature of LLMs also introduces challenges related to reliance on keywords and phrases. As the LLM considers every token generated previously in subsequent iterations, any hallucinations (or misinterpretations of the prompt or its own generation) can propagate forward and compromise the overall integrity of its response. An instance of this occurred when the LLM fixated on two physics concepts during summarization when the segment was almost entirely focused on the computing domain. The LLM's initial focus on the physics concepts caused it to label the segment as *physics-and-computing-synergistic*, even though both Reviewers and GPT-4 all considered the segment to be *computing-focused*. This also happens when the LLM *fails* to recognize the presence of specific concepts in the discourse. For example, the LLM did not refer to several computing concepts in one segment. Initially, this did not appear to be an issue since the LLM was not instructed to highlight *every* physics and computing concept in the discourse due to the large number of concepts interspersed throughout each segment. However, the LLM's oversight in not highlighting the computing concepts led it to inaccurately label the segment as *physics-focused*. This contradicted the consensus of both Reviewers and GPT-4, who agreed the segment was *physics-and-computing-synergistic*.

Both Reviewers and the Educator have identified areas for improvement in the LLM's performance. One important aspect is the LLM's limited integration of actions from XYZ environment in its summaries, often addressing them superficially without connecting them to the broader discourse context. However, we can correct this behavior by developing more integrated reasoning chains for the few-shot instances in future work. The Educator also suggested the LLM consider segments' temporality. He proposed incorporating "pause" identification and duration from prosodic audio via timestamps, which may help teachers identify students' difficulties. Moreover, highlighting instances where students expressed uncertainty, like saying "Um..." or "I'm still a little stumped," could serve as valuable alerts for teachers and help them identify students' difficulties.

Trade-offs

GPT-4-Turbo exhibits behaviors that, while desirable, can compromise other favorable characteristics. First, its impressive attention to detail (an improvement over GPT-4) can inadvertently make shorter segments appear longer, potentially giving a false impression of in-depth discussions by the students. Second, its effective use of the Task Context enables accurate inferences in ambiguous situations. However, the LLM occasionally relies excessively on this information, placing more emphasis on context than the discourse itself (particularly in shorter segments with fewer utterances), which can lead to incorrect discourse categorization. In future work, we plan to shorten the Task Context descriptions to maintain the LLM's focus on the actual discourse.

Interestingly, the LLM demonstrated desirable behaviors that it was not taught in the prompt (either via the task instructions or few-shot CoT reasoning chains). In one example, GPT-4-Turbo alerted us to students being off task when the prompt did not instruct the model to do so ("However, the conversation quickly diverges into personal anecdotes unrelated to the task at hand.") The Educator liked this idea, emphasizing its value in providing actionable feedback for teachers. However, a critical question arises about the extent to which we grant the LLM freedom to generate outputs that deviate from user-labeled few-shot examples. While allowing some latitude enables the LLM to leverage its extensive training and general language knowledge towards generating creative and meaningful responses, it may also lead to undesired behavior, as observed in previous work [1] when the LLM included domain concepts in summaries that *should have been* in the discourse but were not actually present.

References

1. Anonymous, et al.: Anonymous title. Anonymous Publication pp. 0–999 (2024)
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 (2024), `https://arxiv.org/abs/2403.14565`, 14th Symposium on Educational Advances in Artificial Intelligence (EAAI). (in press)

3. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., et al.: A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. arXiv preprint arXiv:2311.05232 (2023)
4. Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., Zhou, D.: Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171 (2022)