International Society of the Learning Sciences

Capturing Collaborative Competency with GPT-4o and ENA

Yoonjae Lee, Department of Intelligence and Information, Seoul National University, yunjae.lee@snu.ac.kr
Christine Kwon, Human-Computer Interaction Institute,Carnegie Mellon University, ckwon2@andrew.cmu.edu
Sarah Seoh, Department of Intelligence and Information, Seoul National University, sarahseoh@snu.ac.kr
Gahgene Gweon, Department of Intelligence and Information, Seoul National University, ggweon@snu.ac.kr
John Stamper, Human-Computer Interaction Institute, Carnegie Mellon University, jstamper@cs.cmu.edu
Carolyn Rosé, Language Technologies Institute, Carnegie Mellon University, cprose@cs.cmu.edu

Abstract: Collaboration is a critical learning skill employed across many educational domains. Since moments of collaborative competencies frequently occur in student discourse, there has been a growing body of work in automatic collaborative process analysis. With recent advancements in Large Language Models (LLMs), it became feasible to automate the collaborative process analysis through simple prompting methods. In this study, we develop CoComTag, an LLM-powered approach using GPT-4o that captures students' collaborative competency using four prompting techniques. Our findings demonstrate that CoComTag shows substantial agreement with humans (kappa of 0.67). Such a result shows that LLM yields comparable performance with prior studies using only prompting strategies without additional fine-tuning. Next, using qualitative error analysis and epistemic network analysis (ENA), we examine how CoComTag differs from humans in detecting collaborative competency. Our analysis shows that ENA provides visual cues on where collaborative competencies co-occur, revealing the differences between LLM and human annotation patterns.

Introduction and background

As individual learning outcomes are insufficient for evaluating the quality of collaboration (Kent & Cukurova, 2020), significant efforts have been made to understand the process of student collaboration through discourse (Zhang et al., 2022; Samadi et al., 2024; Snyder et al., 2024). For example, Zhang et al. (2022) examined different aspects of collaborative problem-solving processes between high and low academic performance groups. The data collected from the collaboration processes provides opportunities to extract variables that help explain the effect of experimental manipulations on collaborative learning outcomes.

Since 2005 (Dönmez et al., 2005), there has been a growing body of work in automatic collaborative process analysis (Fiacco et al., 2018; Flor & Andrews_Todd, 2022). Accordingly, the field of CSCL has developed a wide range of coding schemes and rating schemes (Borge & Rosé, 2021). Though substantial progress has been made in the CSCL field on these automated approaches, issues with respect to accuracy, generalizability to new domains, and validity remain. In this paper, we explore how Large Language Models (LLMs) might contribute to this ongoing line of research. With earlier LLMs, such as BERT and RoBERTa, there have already been some strides towards domain generality as well as generality across learner populations for collaborative processes (Fiacco et al., 2018; Fiacco et al., 2021). Some studies also focused on improving the accuracy by training earlier LLMs (Fiacco et al., 2021; Ganesh et al., 2024; Samadi et al.,2024; Bosch et al., 2024). Now with even more powerful LLMs, such as recent OpenAI models, we expect the capabilities to be greater. Thus, there is a need to explore recent LLMs' capability in automatic collaborative process analysis, particularly given their promising performance through prompting without the need for additional training.

While some studies show the potential and promise of LLMs to annotate moments of collaboration, performance variability still remains a challenge for current LLMs (Wang et al., 2024), leaving human annotations more highly valued. Hence, it is important to understand the differences in reasoning between human-driven and LLM-driven labels of collaborative competencies, given the complex nature of collaboration. In addition to exploring the performance of recent LLMs, we introduce an analytical approach that uses an epistemic network analysis (ENA) to compare underlying patterns of labeled collaborative competency between human and LLM-generated annotations. Traditionally, ENA has been used to quantitatively and qualitatively analyze complex CPS interactions between human roles within groups, revealing interactivity, codependence, and temporal dynamics in collaboration (Zhang et al., 2022). For instance, prior work applied an ENA-based approach to analyze online student discussions (Ferreira et al., 2022) and socio-spatial collaborative learning in classroom settings (Yan et al., 2023). Using the nuanced ability of ENA, we analyze significant contrasts between human-driven and LLM-driven annotations. Thus, in this paper, we aim to answer these research questions: *1) Can an LLM detect collaborative competency through prompting and 2) Can epistemic network analysis provide insight on how LLM annotations differ from human annotations?*

To answer these research questions, we developed CoComTag, an LLM-based collaborative competency annotation system. CoComTag employs four prompting techniques: Persona of a teacher, Predict addressee, Chain-of-thought, and Example cases. Using CoComTag, we examined the differences between the human and LLM annotations of collaborative competency in conversations among graduate students working together on various data visualization exercises. We conduct our investigation through three analytical approaches: performance analysis, qualitative error analysis, and epistemic network analysis. Based on performance analysis, we show that CoComTag exhibits the potential to capture collaborative competency (kappa value of 0.67). However, both qualitative error analysis and ENA show that CoComTag still needs more improvement in fully understanding the context and nuanced intentions of moments of collaborative competencies. Our study provides a comprehensive analysis of an LLM's ability to capture collaborative competency and an in-depth investigation analyzing differences between human-driven and LLM-driven labels in the context of collaborative learning.

CoComTag

In this study, we design CoComTag, an LLM-based annotation system that captures collaborative competency in student discourse by using various prompting techniques. To identify moments of collaborative competency, we employed a collaborative competency rubric based on the generalized competency model proposed by Sun et al. (2020). The generalized competency model captures collaborative competency by explicitly integrating cognitive and social aspects of CPS skills. We chose this model as the foundation for our annotation scheme because it provides specific indicators applicable to text-based conversations and has been validated across various populations and contexts. To measure the performance of LLM at different levels, we added an overall competency level to the generalized competency model. As shown in Table 1, the overall competency consists of three main facets: Constructing shared knowledge, Negotiation/Coordination, and Maintaining team function. Each facet comprises two sub-facets that include multiple indicators.

Table 1

The Collaborative Competency Rubric Based on the Generalized Competency Model Proposed by Sun et al. (2020)

Overall	Facet	Sub-facet	Indicators
Collaborative competency	(f1) Constructing shared knowledge	(sf1) Shares understanding of problems and solutions	Proposes specific solutions Talks about givens and constraints of a specific task Builds on others' ideas to improve solutions
		(sf2) Establishes common ground	Confirms understanding by asking questions/paraphrasing Interrupts or talks over others as intrusion (R)
	(f2) Negotiation/ Coordination	(sf3) Responds to others' questions/solutions	Respond when spoken to by others Makes fun of, criticizes, or is rude to others (R) Provide reasons to support/refute a potential solution Makes an attempt after discussion
		(sf4) Monitors execution	Talk about results Brings up giving up a challenge (R)
	(f3) Maintaining team function	(sf5) Fulfills individual roles on the team	Visibly not focused on tasks and assigned roles (R) Initiates off-topic conversation (R) Joins off-topic conversation (R)
		(sf6) Takes initiatives to advance collaboration processes	Asks if others have suggestions Asks to take action before anyone asks for help Compliments or encourages others

Note: "R" next to an indicator means that it is reverse coded.

To build CoComTag, we selected GPT-4o as our backbone model because it is one of OpenAI's most recent large language models. With its strong performance across diverse general tasks, GPT-4o allows us to experiment with various prompting techniques without the need for model fine-tuning. CoComTag processes user utterances along with five preceding utterances as context, then performs multi-class classification by predicting one of seven labels: six sub-facets and not-competent. To ensure the consistency of the results, we set the temperature parameter of the GPT-4o model to 0. To capture the collaborative competency, we designed a base prompt along with four prompting techniques and iteratively refined the prompts to eliminate awkward phrasing and better specify different situations.

The base prompt consists of four components that contain foundational information essential for labeling the data. "Dialogue context" provides the five preceding utterances of the given utterance along with contextual

information about the collaborative learning situation as shown in Table 2. "Categories" offers detailed explanations of each category. "Output format" is a placeholder-based structure that guides the LLM's responses. "Instruction" contains the main instruction of the task which in our case is to annotate collaborative competency. To enhance collaborative competency detection, we implemented four prompting techniques using the LLM as follows:

- *Persona of a teacher:* Studies suggest that assigning an appropriate persona can improve the performance of an LLM on a given task (Kong et al., 2024). In education, adding the persona of a teacher improved the performance of an LLM in evaluating collaborative learning (Ganesh et al., 2024). Thus, we assigned the LLM with the persona of a teacher who is assessing collaborative competency.
- *Predict addressee:* Our data consists of multi-party conversation, which in nature has a non-linear structure, and implicit addressees (Ganesh et al., 2023). Therefore, recognizing the addressee is important in understanding multi-party conversations. Based on these characteristics of multi-party conversation, we prompt LLM to predict the addressee of each utterance before generating the label.
- *Chain-of-thought:* Chain-of-thought is a method that is widely used when prompting an LLM. This helps the model to break down the task into step-by-step processes. He et al. (2024) proposed a two-step method called "explain-then-generate" that applies chain-of-thought to annotation tasks. Similarly, we adopted a two-step method, where the model first generates the intention of the speaker and then generates a label based on that intention.
- *Example cases:* Few-shot learning is a form of in-context learning where the LLM is provided with examples related to the task in the prompt (Brown et al., 2020). Due to the diverse patterns of collaboration among different groups, instead of providing a chunk of conversation as an example, we generalized example cases. Two levels of examples, micro and macro, were added for each sub-facet.

Table 2

Description and Examples of the Base Prompt and Four Prompting Techniques

Prompt type	Description	Example
base	[Dialogue context] The context of the collaborative learning situation and the five preceding utterances	A small group of students collaborate to solve 3 data visualization tasks. …
	[Categories] Explanation of each category following the human annotation guideline	(sf1): these are the cases where a student proposes or improve specific solution… (sf2): these are the cases where …
	[Output format] The output format LLM should follow	Generate the output in the following format: {"utt": "<Speaker's utterance>", … "category": "<Category>" }
	[Instruction] The main instruction of the task which is to annotate collaborative competency	Student 1 says … What category does Student 1's intention falls into?
Persona of a teacher	Assigning LLM with a persona of a teacher	You are a teacher who is assessing the students' collaborative competency. To do so, …
Predict addressee	Prompting LLM to predict the addressee before generating the label	For each utterance, generate who the addressee is.
Chain-of-thought	Prompting LLM to predict the intention then generate the label	Generate the intention of each utterance. Then, find the category that best fits the intention.
Example cases	Micro-level: Examples of each sub-facet. These examples are added in the [Categories] section of the base prompt	(sf6) … Some examples a student can say are "great", "sweet", "nice", "cool", or "awesome".
	Macro-level: Examples of how utterances correspond to a specific sub-facet in a generalized situation	The speaker is asking questions: If the speaker is asking for suggestions, label as (sf6). If the speaker is asking a question about …

Our comprehensive annotation system, CoComTag, combines the base prompt with all four prompting techniques. For experimental purposes, we create five system variants denoted as CoComTag_*[prompt_type]*. The variants combine the base prompt with a single prompting technique. For example, CoComTag_persona combines the base prompt with the 'Persona of a teacher' prompt.

Method

In this study, we examine how collaborative competency captured by CoComTag differs from those captured by humans. We used conversation data collected from a collaborative data visualization task. The dataset contains chat logs from a graduate foundational data science course conducted in a private R1 university in the United States. In this course, students were divided into groups of one to five members with an average of 3.82 members to solve data visualization problems collaboratively. Each problem consisted of three tasks. At the beginning of each task, a conversational agent randomly assigns each student to one of the following roles: driver, navigator, or researcher. The students access a web page that consists of a shared JupyterLab and a chat room. Students solve the common task by communicating with other students through the chat room, assisted by the conversational agent. The conversational agent gives students warnings when tasks are nearing time limits and automatically reassigns roles throughout the exercise. A total of 43 group data were collected over three semesters. Since we focused on collaboration among students, we excluded files where there was only one student throughout the whole exercise. As a result, a total of 39 files of group data and 3,242 utterances were used for the annotation.

To obtain moments of collaborative competencies captured by humans, two researchers carried out the annotation process at the indicator level. The unit of annotation was each utterance, where the utterance is defined as concatenated chat logs separated by the change of speakers. For each utterance, five previous utterances were given as context. Utterances were labeled with multiple indicators if two or more indicators were mapped to different spans of the utterance. Before the annotation, two researchers created an initial annotation guideline for each of the indicators. The researchers labeled 20% of the randomly selected data and iteratively updated the guideline. As a result, the inter-rater reliability between the two researchers reached a kappa value of 0.79 on the indicator level. The remaining data was divided into half and annotated by each of the two researchers on indicator level. Finally, the labeled indicators were grouped at the sub-facet level. In total, 51.9% of the total utterances were labeled as collaborative competency. The percentage for each sub-facet was sf1: 33.4%, sf2: 14.2%, sf3: 24.5%, sf4: 14.8%, sf5: 0.7%, sf6: 12.4%.

To identify the collaborative competencies captured by LLMs, we used CoComTag to annotate the data at the sub-facet level. As in the human annotation process, the unit of annotation was each utterance, and five previous utterances were given as context. CoComTag predicted one of the seven categories, six sub-facets and not-competent, that best fits for each utterance. The predicted labels were post-processed. Namely, for the facet level, we combine sub-facets into corresponding facets. For example, if an utterance was either labeled as sf1 or sf2, the label for the facet level would be f1. For the overall level, all 6 sub-facets were combined into a single label, ‘competent’.

Analytical approach

We conducted three types of analyses to examine the CoComTag's ability to detect collaborative competency. First, we conduct a performance analysis to analyze the performance of the CoComTag focusing on the effect of four prompting techniques. Second, we conduct a qualitative error analysis to analyze error cases that LLMs are prone to incorrectly detect when capturing collaborative competency. Finally, we conduct an epistemic network analysis to compare underlying patterns of collaborative competencies captured by humans and CoComTag.

For the performance analysis, we first explore the performance of the CoComTag on three levels: overall, facet, and sub-facet level. Second, we examine the effect of the four prompting techniques on sub-facet level by comparing the performance of CoComTag_base with four CoComTag variants: persona, addressee, cot, example. To measure the performance of CoComTag, we compare the human-annotated result and CoComTag-annotated result using three metrics: accuracy, macro-average f1-score, and Cohen's kappa. Due to the inconsistent nature of LLM, we repeated the experiment 3 times. Here, 90.9% of the results were identical across all three attempts, 8.8% had two identical results, and only 10 cases produced three different results. We report the best scores with the standard deviation and the average score of the predicted results. The average score is obtained using the majority vote method. In the 10 cases that produced three different results, we randomly selected the label.

For the qualitative error analysis, we analyzed the error cases of the LLM predictions to explore where they differ from human annotations. To this end, three researchers reviewed the CoComTag-annotated result of 10 randomly selected files. The CoComTag-annotated result contained the addressee, the speaker's intention, and the category. The researchers examined each annotated utterance where the LLM and humans had different labels based on the annotation guideline. The researchers then combined the frequent errors to categorize the cases.

For the epistemic network analysis (ENA), we compare the networks generated from human-annotated data with those generated from CoComTag-annotated data. ENA is a network-based method that reflects the frequency of temporal connections between the codes. To gain a more profound understanding of how the aspects of collaborative competencies captured by the LLM differ from human annotators, we compare the centroid of each network and qualitatively analyze the visualized networks using rENA package in R (Shaffer & Ruis, 2017). To observe if there is a significant statistical difference between the centroids of human and LLM networks, we

employ the Wilcoxon Signed-Rank test as the data did not pass the Shapiro-Wilk normality test; For the x-axis, human-annotated data (W=0.80, p=3.81e-13) and LLM-annotated data (W=0.95, p=8.56e-05) did not pass the Shapiro Wilk test. Also, for the y-axis, human-annotated data (W=0.79, p=3.37e-13) and CoComTag-annotated data (W=0.90, p=1.35e-08) did not pass the Shapiro Wilk test.

Result and discussion

Performance analysis

Table 3 shows the performance of the CoComTag on three levels: overall, facet, and sub-facet level competency. For the overall competency, the CoComTag showed an agreement with humans with a kappa value of 0.67, which is considered a substantial agreement (Landis & Koch, 1977). Specifically, the best-case kappa value was 0.67 (*sd=0.13*), the best-case accuracy was 0.84 (*sd=0.06*) and the best-case f1-score was 0.83 (*sd=0.07*). For facet level, the best-case Kappa value was 0.66 (*sd=0.11*), the best-case accuracy was 0.79 (*sd=0.07*) and the best-case F1-score was 0.51 (*sd=0.1*). For the sub-facet level, the best-case Kappa value was 0.63 (*sd=0.11*), the best-case accuracy was 0.76 (*sd=0.07*) and the best-case F1-score was 0.42 (*sd=0.1*).

Table 3

The Performance of CoComTag Compared to Human Annotations on 3 Levels: Overall, Facet, Sub-Facet Level

	Acc (avg / best)	F1-score (avg / best)	Kappa (avg / best)	Human-Human kappa
Overall level	0.83 / 0.84	0.82 / 0.83	0.65 / 0.67	0.84
Facet level	0.78 / 0.79	0.50 / 0.51	0.64 / 0.66	0.85
Sub-facet level	0.75 / 0.76	0.41 / 0.42	0.62 / 0.63	0.85

Table 4 shows the performance of CoComTag and its five variants on sub-facet level. The four CoComTag variants, persona/addressee/cot/example, showed higher performance compared to CoComTag_base. Among these four variants, CoComTag_example showed the highest performance improvement on all metrics compared to other variants. In previous literature, GPT-4 showed a higher performance with few-shot learning compared to zero-shot learning (Ganesh et al., 2024). This might be related to how well LLMs can generate results with in-context learning, particularly with more explicit examples.

Table 4

The Performance of CoComTag and its Five Variants on Sub-Facet Level

	Acc (avg / best)	F1-score (avg / best)	Kappa (avg / best)
CoComTag_base	0.65 / 0.66	0.36 / 0.36	0.53 / 0.53
CoComTag_persona	0.66 / 0.67	0.38 / 0.38	0.54 / 0.55
CoComTag_addressee	0.66 / 0.67	0.38 / 0.38	0.54 / 0.55
CoComTag_cot	0.67 / 0.68	0.37 / 0.37	0.54 / 0.56
CoComTag_example	0.72 / 0.73	0.41 / 0.41	0.60 / 0.61
CoComTag	**0.75 / 0.76**	**0.41 / 0.42**	**0.62 / 0.63**

On the other hand, CoComTag_persona, CoComTag_addressee, and CoComTag_cot showed less improvement in accuracy and kappa value compared to CoComTag_example. For CoComTag_persona, a possible explanation for this limited improvement is that socio-demographic persona may have unintentionally influenced the LLM's performance (Gupta et al., 2024). Gupta et al. (2024) suggested that persona-assigned agents may be more error-prone in complex tasks due to biases inherent in LLMs. The persona we implemented may not have been well-suited for complex task of detecting collaborative competency. This suggests a need to explore more suitable personas for capturing students' collaborative competency. In addition, the LLM's insufficient understanding of multi-party conversations is one potential reason why CoComTag_addressee and CoComTag_cot did not show improved accuracy. Multi-party conversations have complex structure characterized by non-linearity and implicit addressees (Ganesh et al., 2023). Simply prompting the LLM to predict the addressee or intention may not have been enough to help CoComTag grasp the complexity of multi-party conversations.

Qualitative error analysis

We conducted a deeper analysis of the LLM's prediction results to assess how CoComTag differed from humans when detecting collaborative competency. We report three cases where CoComTag differed in capturing collaborative competency compared to human annotators: Qual_EA1) Challenges in capturing the nuanced

intention of moments of collaborative competency, Qual_EA2) Different prioritization of intention between human and LLM annotators, and Qual_EA3) Lack of understanding of the context.

Qual_EA1) Challenges in capturing the nuanced intention of moments of collaborative competency: CoComTag annotations differed from human annotations in that CoComTag struggled to capture the nuanced intentions of moments of collaborative competency. Some utterances may initially seem misleading as their meaning may differ from how they are expressed. In our analysis, we noticed that CoComTag fails to capture the true intentions of moments of collaborative competency due to the superficial forms. For instance, there were several misaligned labels between human and CoComTag on utterances phrased as questions. CoComTag showed a tendency to consider utterances phrased as questions to be sf2 (Establishes common ground), which includes asking questions to confirm understanding. For example, student 2 says, "What would be the aggregation_function?". In this case, humans labeled sf6 (Takes initiative to advance collaboration process). However, the CoComTag prioritized the superficial format of question and predicted sf2. One possible way to address this misalignment is to adopt a more structured prompt. In our study, we prompted the model to generate the intention of the speaker. A previous study conducted by Dutt et al. (2024) showed that using three types of rationale -Intention, Assumptions, and Implicit Information- demonstrates a positive impact on detecting implicit social meanings. Adopting such methods for analyzing collaborative learning may further improve the LLM's ability to capture nuanced intentions of moments of collaborative competency.

Qual_EA2) Different prioritization of intention between Human and LLM annotators: CoComTag and humans prioritized different intentions. In some cases, a single span of a student's utterance played multiple roles. According to the annotation guideline, if the same span of utterance had multiple roles, the label corresponding to the main intention was applied. However, there were instances where humans and the CoComTag prioritized different aspects. For instance, student 3 said "So how should we proceed for this?" and student 1 responded, "I think the plot type will be histogram. and the column will be genre column". Here, student 1's utterance has two intentions: proposing a specific solution and responding to student 3. In this scenario, CoComTag prioritized the intention of proposing a specific solution and chose sf1 (Shares understanding of problems and solutions), while humans prioritized the intention of responding to student 3 and chose sf3 (Responds to others' questions/solutions). This may be alleviated by adopting a new annotation method where multiple competencies can be mapped to the same span of utterance. There were also instances where an utterance contained both statements that map to collaborative competencies and those that do not. In such cases, CoComTag occasionally prioritized statements that do not map to collaborative competencies whereas humans prioritized the statements that map to competencies. For example, for the utterance "Hi! I am researcher this round", CoComTag predicted the intention of the "Hi" as a greeting, and the rest of the utterance as "stating their role for the current round". Here, "Greeting" is an intention that does not map to any competencies, while "stating their role for the current round" corresponds to sf1, which considers moments on "Talks about givens and constraints of a specific task". However, CoComTag predicted this utterance as not competent. CoComTag might have been distracted by other parts of the utterance and incorrectly predicted the label. To address this error, we can consider breaking down the unit of analysis into smaller units such as sentences or propositional units (Csanadi et al., 2018).

Qual_EA3) Lack of understanding of the context: CoComTag demonstrated lower contextual understanding compared to humans. Given the non-linear conversation structure of our dataset, utterances may refer to earlier statements in the conversation rather than immediately preceding ones. Such complexity can hinder an LLM from fully understanding the context of each utterance. We observed that CoComTag often shows a lack of understanding of the context in two cases: 1) inaccurate prediction of the addressee, 2) fail to grasp the previous context. In the first case, CoComTag inaccurately predicted the addressee when the labeled utterances refer to earlier statements. This can lead to incorrect classifications of utterances or failure in capturing sf3 (Responds to others' questions/solutions). For example, when the agent aims a question to the students, the student's responses are directed to the conversational agent. According to human guidelines and the prompt that is given to CoComTag, the utterances directed to the agent should not be labeled as competent. However, CoComTag occasionally misidentified the addressee to be other students or the entire group, resulting in labeling the associated utterance as a competency, especially as sf1 (Shares understanding of problems and solutions) or sf4 (Monitors execution). In the second case, CoComTag failed to grasp the previous context and was unable to capture sf3 correctly. Based on human guidelines, the utterance of a student should be labeled as sf3 when a student responds to others' questions or agrees to a proposed solution. However, CoComTag occasionally labels such utterances as not competent. Conversely, CoComTag sometimes misidentifies not-competent utterances, which respond to others that are not a question or solution, as sf3. For example, in one case, student 3 says "Don't worry about roles right now none of us have them", followed by student 5 stating, "Oh okay". Here, student 5 is replying to student 3's utterance which is neither a question nor a solution, but CoComTag incorrectly predicts

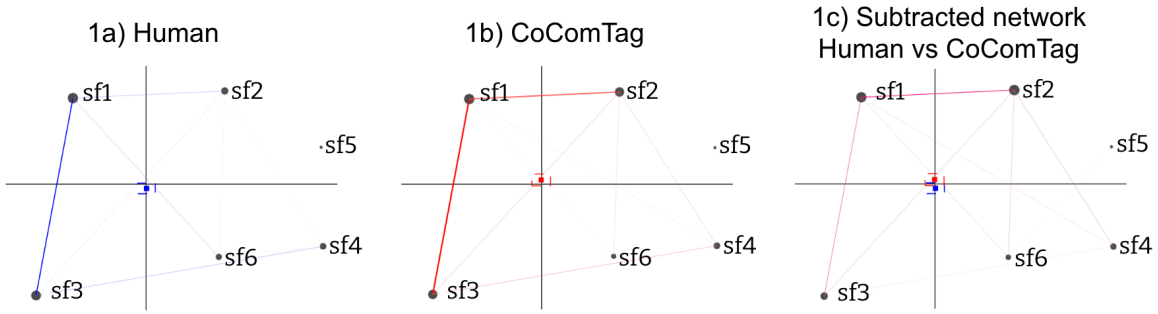
the utterance of student 5 as sf3. To address this issue, one potential approach is to finetune the LLM or employ modeling techniques that can enhance the model's understanding of complicated multi-party conversations.

Epistemic network analysis

The statistical difference between the centroids of the two different networks visualized from the human-annotated data and the CoComTag-annotated data was as follows. According to the Wilcoxon Signed-Rank test, there was no significant difference in the centroids' x-axis values (p-value = 0.102), but a significant difference in their y-axis values (p-value = 0.005). This indicates that although CoComTag shows moderate agreement to human labels based on the kappa value, CoComTag may still face challenges in capturing the co-occurrence of the collaborative competencies. These results are also shown in the epistemic networks of human and CoComTag annotated data in Figure 1. The network of human-annotated data, Figure 1a, shows the strongest connection of co-occurrence between sf1 (Shares understanding of problems and solutions) and sf3 (Responds to others' questions/solutions). The network of CoComTag-annotated data, Figure 1b, shows the strongest connection of co-occurrence between sf1 and sf3, followed by sf1 and sf2 (Establishes common ground). The subtracted network in Figure 1c, shows the difference between the epistemic network of human-annotated data and CoComTag-annotated data. As shown in the edges visualized in Figure 1c, there were two main differences between human and CoComTag annotated results: ENA_1) Co-occurrence of sf1 and sf2, ENA_2) Co-occurrence of sf1 and sf3.

Figure 1

Epistemic Networks of Human- and CoComTag-Annotated Data and the Difference Between Two Networks



Note: Red squares represent CoComTag-annotated results, and blue squares represent human-annotated results.

ENA_1) Co-occurrence of sf1 and sf2: The main difference between the human and CoComTag annotated results was the co-occurrence of sf1 and sf2. One possible explanation for the difference is that CoComTag prioritizes question format utterances as sf2. When active communication occurred between students, the exchanges of question-formatted utterances became more frequent. However, in such cases, CoComTag tends to predict sf2 (Establishes common ground) for utterances framed as questions. For example, after an active discussion, a student says, “let’s try it?”. Humans classified this utterance as sf3, which captures moments of making an attempt, while CoComTag classified it as sf2, which captures moments of seeking confirmation by asking questions. Such prioritization of question format by LLM may also partially explain why in Figure 1b, sf2 shows more frequent connections with other sub-facets such as sf3, sf4, and sf6, compared to Figure 1a.

ENA_2) Co-occurrence of sf1 and sf3: CoComTag and human annotators also showed differences in capturing the temporal co-occurrence of sf1 (Shares understanding of problems and solutions) and sf3 (Responds to others' questions/solutions). Sf1 and sf3 co-occurred when a proposed solution was followed by others' agreement, reasoning, or further attempts. In capturing these moments, there were two major cases where humans and CoComTag showed differences. The first case was different prioritization among multiple intentions. For example, student 1 proposes an answer as “i think the agg function should be sum” and student 2 corrects the answer by responding “sum' not sum”. In this case, humans labeled student 2’s utterance as sf1, considering it as proposing a solution, while CoComTag labeled it as sf3, interpreting it as a response to another student’s solution. The second case was when CoComTag predicted the addressee as a student instead of the conversational agent. Labeling utterances directed to the conversational agent, rather than a student, should be labeled as being not-competent. For example, when the agent gives students a question and asks them to explain their reasoning at the end of each exercise, the answers to such questions are directed to the conversational agent and thus labeled as not-competent. However, CoComTag occasionally incorrectly predicted the addressee to be a student, rather than the conversational agent, and labeled the utterances as competent rather than as not-competent.

Insights on LLM performance and ENA analysis

Our study demonstrated that 1) CoComTag, which is based on LLM, shows comparable performance with prior studies with prompting only and without fine-tuning, and 2) ENA provides visual cues on where collaborative competency co-occur, revealing how the LLM and human annotations differ. First, our findings demonstrated comparable or greater LLM coding performance relative to previous studies that used fine-tuning approaches. A study by Samadi et al. (2024) used fine-tuned models to capture collaborative competency using a different CPS framework from our study, which consists of 8 categories. They showed an accuracy of 0.73. Our result on sub-facet level which consists of 6 categories shows an accuracy of 0.75 which is comparable to the prior study. Another study by Fiacco et al. (2021) fine-tuned models to capture three constructs of collaborative process and showed an average kappa of 0.66. Our result shows a comparable performance of a kappa value of 0.64 on facet level. Ganesh et al. (2024) used GPT-4 to automatically detect collaborative competency using the generalized competency model. They showed an average F1-score of 0.36 on facet level. Our result shows improvement in performance by showing an average F1-score of 0.51 on facet level. Our results suggest that with appropriate prompting techniques, LLM can achieve comparable performance to a fine-tuned language model. Former methods of fine-tuning required additional domain-specific dataset to train the automatic collaborative process analysis models. However, LLMs demonstrate a possibility of domain generalizability by achieving comparable performance without the necessity of training a domain-specific model. One explanation can be that recent LLMs potentially possess inherent knowledge across diverse domains because they are trained on extensive datasets.

Second, we show that ENA provides visual cues on the co-occurring moments of collaborative competency to reveal how the LLM and human annotations differ. Using ENA can be an efficient approach in understanding why the differences occur between the two annotations because it shows which collaborative competency could be examined in more detail. Interestingly, the findings from the ENA, which can be regarded as a top-down approach, correspond to the findings from the qualitative error analysis, which is a bottom-up approach. Specifically, ENA_1 finding showed that the LLM tends to prioritize the question-format utterances as sf2 (Establishes common ground), which aligns with the observations made in the Qual_EA1. Also, ENA_2 finding identified that differences in temporal co-occurrence between sf1 (Shares understanding of problems and solutions) and sf3 (Responds to others' questions/solutions) predicted by humans and the LLM stem from conflicting priorities among multiple intentions and incorrect addressee predictions. These findings align with Qual_EA2 and Qual_EA3 respectively. ENA is known to provide more insight into socio-cognitive learning activities by allowing researchers to capture temporal co-occurrences between collaborative competencies, visualize those co-occurrences, and conduct statistical comparisons between different groups of learners (Csanadi et al., 2018). Thus, compared to qualitative error analysis, researchers can visually perceive the main differences between the collaborative competency captured by humans and the LLM annotator. Also, because ENA considers the co-temporal moments of collaborative competency, the method can provide guidance on aspects of conversations that need to be examined in more depth.

Conclusion

In this paper, we address the potential of LLMs to capture collaborative competencies with a thorough prompt design to observe if LLMs are capable of understanding or measuring the collaboration process. We explore the potential of LLMs as annotators by designing a system called CoComTag that implements four types of prompting techniques. Our results indicate that while some differences exist, CoComTag shows substantial agreement with human annotators. This study has several limitations. The results of our study are limited to the collaborative context of graduate students in the CS field. Therefore, there is a need to apply these methods to a wider range of populations and contexts. Additionally, we analyzed the effect of four prompting techniques, and there is a need to explore additional prompting methods as there may be more prompting techniques that can help LLMs capture collaborative moments more accurately.

Despite such limitations, this study provides valuable insights through qualitative error analysis and epistemic network analysis, identifying the differences between LLM and human annotators. Specifically, the LLM annotation system, CoComTag, struggled with understanding multi-party conversations and had difficulties capturing the nuanced collaborative intentions. Although LLMs need further improvement before functioning as standalone tools, our research demonstrates their potential to assist teachers in detecting meaningful student collaboration. Our research is particularly relevant for programming-oriented courses, where collaboration is challenging to detect due to the structured and close-ended nature of many tasks, yet meaningful interactions still occur through technical contexts. Future research might explore multimodal data to support teachers in identifying student collaboration in classroom settings, incorporating facial expressions or audio features in addition to the linguistic features examined in this study. Through examining multiple facets of collaboration, we aim to develop a more comprehensive understanding of the critical skill of collaboration in education.

References

Borge, M., & Rosé, C. (2021). Quantitative approaches to language in CSCL. In *International Handbook of Computer-Supported Collaborative Learning* (pp. 585–604). Springer International Publishing.

Bosch, N., Williams-Dobosz, D., & Perry, M. (2024). Measuring Help-seeking in Online Course Discussion Forums with Privacy-preserving Large Language Models. In Proc. CSCL 2024, pp. 189-192.

Csanadi, A., et al. (2018). When coding-and-counting is not enough: using epistemic network analysis (ENA) to analyze verbal data in CSCL research. International Journal of CSCL, 13(4), 419–438.

Dönmez, P., Rosé, C., Stegmann, K., Weinberger, A., & Fischer, F. (2005). *Supporting CSCL with automatic corpus analysis technology*. 125–134.

Dutt, R., Wu, Z., Shi, K., Sheth, D., Gupta, P., & Rose, C. P. (2024). Leveraging machine-generated rationales to facilitate social meaning detection in conversations. ACL, (Volume 1: Long Papers), 6901–6929.

Ferreira, M. A. D., Ferreira Mello, R., Kovanovic, V., Nascimento, A., Lins, R., & Gasevic, D. (2022, March). NASC: Network analytics to uncover socio-cognitive discourse of student roles. In *LAK22: 12th International Learning Analytics and Knowledge Conference* (pp. 415-425).

Fiacco, J., & Rosé, C. (2018). Towards domain general detection of transactive knowledge building behavior. Proceedings of the Fifth Annual ACM Conference on Learning at Scale, 1–11.

Flor, M., & Andrews-Todd, J. (2022). Towards automatic annotation of collaborative problem-solving skills in technology-enhanced environments. Journal of Computer Assisted Learning, 38(5), 1434–1447.

Ganesh, A., Chandler, C., D'Mello, S., Palmer, M., & Kann, K. (2024). Prompting as panacea? A case study of in-context learning performance for qualitative coding of classroom dialog. In Proceedings of the 17th International Conference on Educational Data Mining, 835–843.

Ganesh, A., Palmer, M., & Kann, K. (2023). A Survey of Challenges and Methods in the Computational Modeling of Multi-Party Dialog. Proceedings of the 5th Workshop on NLP for Conversational AI (pp. 140–154).

Gupta, S., Shrivastava, V., Deshpande, A., Kalyan, A., Clark, P., Sabharwal, A., & Khot, T. (2024). Bias runs deep: Implicit reasoning biases in persona-assigned LLMs. The Twelfth International Conference on Learning Representations.

He, X., et al. (2024). AnnoLLM: Making large language models to be better crowdsourced annotators. In Proc. 2024 North American Chapter of the Association for Computational Linguistics, p165–190.

Kent, C., & Cukurova, M. (2020). Investigating Collaboration as a Process with Theory-driven Learning Analytics. Journal of Learning Analytics, 7(1), 59–71.

Kong, A., Zhao, S., Chen, H., Li, Q., Qin, Y., Sun, R., Zhou, X., Wang, E., & Dong, X. (2024). Better zero-shot reasoning with role-play prompting. In Proc. 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Vol 1. pp. 4099–4113.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), 159–174.

Samadi, M. A., Jaquay, S., Lin, Y., Tajik, E., Park, S., & Nixon, N. (2024). Minds and machines unite: Deciphering social and cognitive dynamics in collaborative problem solving with AI. Proceedings of the 14th Learning Analytics and Knowledge Conference, 885–891.

Shaffer, D. W., & Ruis, A. R. (2017). Epistemic network analysis: A worked example of theory-based learning analytics. In Handbook of Learning Analytics (pp. 175–187).

Snyder, C., Hutchins, N. M., Cohn, C., Fonteles, J. H., & Biswas, G. (2024). Analyzing students collaborative problem-solving behaviors in synergistic STEM+C learning. *LAK24*, 540–550.

Sun, C., Shute, V. J., Stewart, A., Yonehiro, J., Duran, N., & D'Mello, S. (2020). Towards a generalized competency model of collaborative problem solving. Computers & Education, 143, 103672.

Wang, X., Kim, H., Rahman, S., Mitra, K., & Miao, Z. (2024). Human-LLM collaborative annotation through effective verification of LLM labels. Proceedings of the CHI Conference on Human Factors in Computing Systems, 1–21.

Yan, L., Martinez-Maldonado, R., Zhao, L., Li, X., & Gasevic, D. (2023, March). SeNA: Modelling socio-spatial analytics on homophily by integrating social and epistemic network analysis. *LAK23:* pp. 22-32.

Zhang, S., Gao, Q., Sun, M., Cai, Z., Li, H., Tang, Y., & Liu, Q. (2022). Understanding student teachers' collaborative problem solving: Insights from an epistemic network analysis (ENA). *Computers & Education*, *183*, 104485.

Acknowledgements

This work was supported by the SNU-Global Excellence Research Center establishment project and IITP grant funded by the MSIT [NO.2021-0-01343, Artificial Intelligence Graduate School Program (Seoul National University)]