논문 주제 고민

하고 싶은 방향

- 학습자(아동+청소년) 대상 AI 접목(에듀테크)
- CSCL = Computer-Supported Collaborative Learning
- 음성 또는 자연어처리 ,HCI 상관 X

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

2025.12.14
HyorinJung

Index

Introduction

- Collaboration is a key learning skill in many educational settings
- Individual outcomes alone cannot capture collaboration quality
- CSCL research analyzes collaborative discourse to understand learning processes
- Automated collaboration analysis has progressed, but accuracy and generalizability remain challenges

Introduction

- Earlier models (e.g., BERT, RoBERTa) required fine-tuning with domain-specific data
- Recent LLMs (e.g., GPT-4o) show strong general reasoning ability
- Limited research on LLMs using prompting only for collaboration analysis

👉Need to examine how LLM annotations compare to human annotations

Research Questions

- Can an LLM detect collaborative competency using prompting only?
- How do LLM annotations differ from human annotations?
- Can ENA reveal structural differences between LLM and human coding?

Background

- CoComTag: LLM-based collaborative competency annotation system
- Backbone model: GPT-4o
- No fine-tuning; prompting-only approach
- Labels each utterance using surrounding conversational context

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.

Background

Prompting Techniques

- Teacher persona to frame evaluation
- Addressee prediction for multi-party dialogue
- Chain-of-thought to infer speaker intention
- Example-based prompting (most effective)

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 …

Dataset and Method

Task Setup

- Groups of 1–5 students (avg. 3.82)
- Shared JupyterLab + chat environment
- Roles dynamically assigned by agent: Driver, Navigator, Researcher

Dataset

- 3 semesters, 39 group sessions
- 3,242 utterances used for analysis
- Single-student sessions excluded

Annotation & Analysis

Human Annotation

- 2 researchers, indicator-level labeling
- Unit: Utterance (+ 5 previous utterances as context)
  Multiple indicators allowed per utterance
- Inter-rater reliability: Cohen's $\kappa = 0.79$
- 51.9% of utterances labeled as collaborative competency

LLM Annotation (CoComTag)

- Sub-facet level prediction
- 7 classes: 6 sub-facets + not-competent
- Post-processed to facet and overall levels

Annotation & Analysis

Analyses Conducted

1. **Performance Analysis**
    - Compare prompting strategies
    - Metrics: Accuracy, Macro-F1, Cohen's κ

2. **Qualitative Error Analysis**
    - Identify where LLM diverges from humans

3. **Epistemic Network Analysis (ENA)**
    - Compare competency patterns between humans and LLM

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

- Overall: CoComTag reliably detects whether collaborative competence is present
- Facet: Performs well in identifying broad collaboration categories
- Sub-facet: Most challenging level due to subtle intentions and specific indicators

👉 CoComTag performs strongly at detecting overall collaboration but struggles with fine-grained and nuanced sub-facet distinctions.

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**

👉 Incorporating **example-based** and combined prompting strategies significantly improves CoComTag’s ability to classify collaborative competence, reaching human-comparable agreement levels.

Qualitative Error Analysis

Qual_EA1: Difficulty Capturing Nuanced Intentions

- CoComTag often relies on surface linguistic forms rather than underlying intent.
- Utterances phrased as questions were frequently misclassified.

Example

- *“What would be the aggregation function?”*
  - Human: Takes initiative to advance collaboration (sf6)
    CoComTag: Establishes common ground (sf2)

Interpretation

- CoComTag focused on the question form, missing the initiative-taking intention.

Qualitative Error Analysis

Qual_EA2: Different Prioritization of Intentions

- A single utterance may serve multiple collaborative roles.
- Humans and CoComTag often prioritize different intentions.

Example

- *“I think the plot type will be histogram.”*
    - Human: Responds to others’ questions (sf3)
    - CoComTag: Shares problem understanding/solution (sf1)

Additional Issue

- CoComTag sometimes focuses on non-competency elements (e.g., greetings) and overlooks competency-related content.

Implication

- Errors may be reduced by allowing multiple labels per utterance or using smaller units of analysis.

Qualitative Error Analysis

Qual_EA3: Limited Contextual Understanding

- CoComTag struggles with non-linear, multi-party conversations.
- Two main issues were observed:
  - Misidentification of the addressee
  - Failure to connect to earlier conversational context

Example

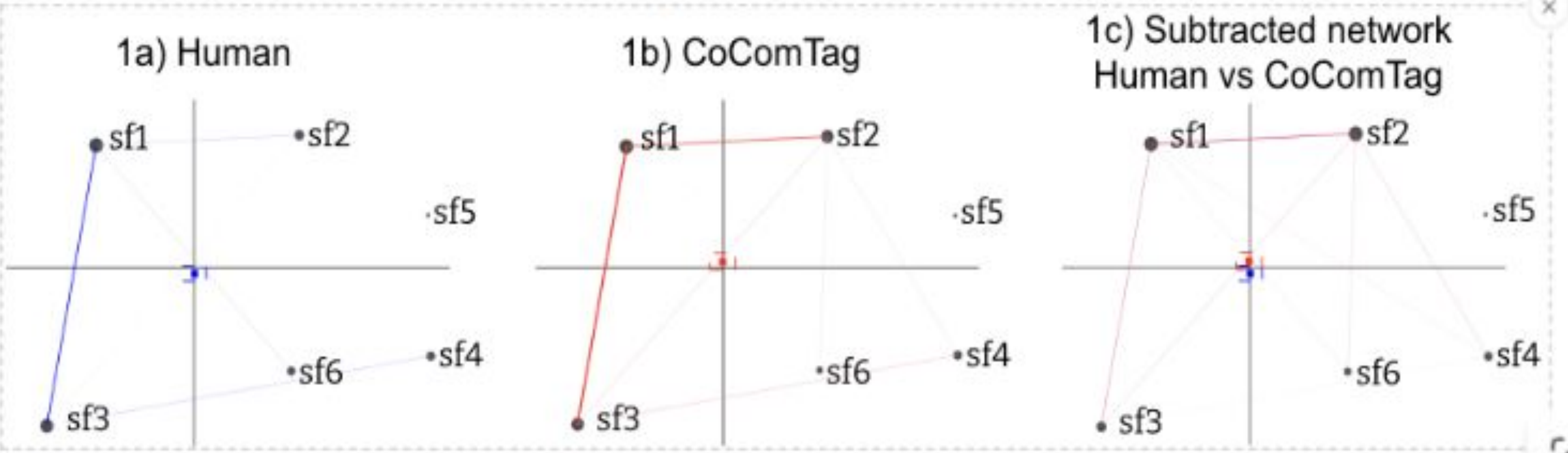
- *“Oh, okay.”*
  - Human: Not a collaborative competency
  - CoComTag: Responds to others (sf3)

Interpretation

- CoComTag incorrectly assumed the utterance was responding to a question or solution.

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.

Sub-facet Definitions of Collaborative Competency

sf1 – Shares understanding of problems and solutions
Shares understanding of the problem or proposes solutions
(e.g., suggesting concrete solutions, building on others' ideas)

sf2 – Establishes common ground
Builds shared understanding among team members
(e.g., asking clarification questions, unintended interruptions (R))

sf3 – Responds to others' questions/solutions
Responds to others' questions or proposed solutions
(e.g., answering questions, supporting or challenging ideas, trying solutions after discussion)

sf4 – Monitors execution
Monitors and reflects on task progress or outcomes
(e.g., discussing results, mentioning giving up on a challenge (R))

sf5 – Fulfills individual roles on the team
Performs or manages individual roles within the team
(e.g., off-task behavior (R), initiating or joining off-topic conversations (R))

sf6 – Takes initiative to advance collaboration processes
Takes initiative to move collaboration forward
(e.g., inviting others' ideas, prompting action before help requests, giving praise or encouragement)

ENA_1: Co-occurrence of sf1 and sf2

- CoComTag overestimates the co-occurrence of sf1 and sf2 compared to humans.
- This aligns with *Qual_EA1*: CoComTag tends to interpret question-form utterances as sf2 (common ground), even when humans see them as initiative-taking behaviors.

ENA_2: Co-occurrence of sf1 and sf3

- Differences also appear in the temporal co-occurrence of sf1 and sf3.
- These discrepancies relate to:
    - Different prioritization of multiple intentions (Qual_EA2)
    - Limited understanding of context and addressee (Qual_EA3)

👉 ENA visualizations show that CoComTag captures overall collaboration patterns similar to humans, **but reveals subtle yet important differences in how relationships among sub-facets**—especially sf2, sf1, and sf3—are interpreted.

Conclusion & Implications

- LLMs show strong potential as collaboration analysis assistants
- Useful for supporting teachers and researchers
- Not yet suitable as fully autonomous annotators
- Future work: broader contexts and multimodal data

Appendix

- Cohen’s Kappa (κ): 두 명 이상 평가자 (또는 인간-AI 시스템) 간 합치도 (agreement) 측정 지표
- 단순 일치율이 아닌 **우연에 의한 일치 (chance agreement)**를 보정

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$