Measuring Cognitive Engagement in Collaborative Discourse with an Extended ICAP Framework: Comparing Human Annotation, In-Context Learning, and Reflective LLM Agents

Lan Anh Do[1*] **(lan_anh.do@tufts.edu) Hanling Jiang**[2*] **(hanling.jiang@tufts.edu)**
Shuchin Aeron[2] **(shuchin.aeron@tufts.edu) Ayanna K. Thomas**[1] **(ayanna.thomas@tufts.edu)**
[1]Department of Psychology, Tufts University
[2]Department of Electrical and Computer Engineering, Tufts University
[*]These authors contributed equally as co–first authors.

Abstract

Collaboration supports learning and problem-solving, but its effectiveness depends on cognitive engagement during discourse. This study applies an extended 7-point ICAP framework based on the Interactive, Constructive, Active, and Passive modes to characterize variation in cognitive engagement during collaborative dialogue. Engagement was coded by trained human annotators and compared with large language model (LLM)–based labeling approaches, including in-context learning (ICL), zero-shot prompting, and self-reflective agents. Interrater reliability among human annotators was robust across framework refinement stages ($\kappa = 0.906$–0.998), higher than the moderate agreement observed for ICL-based annotation ($\kappa = 0.541$–0.609). The human-refined framework improved agreement among human annotators ($\Delta\kappa = 0.10$), but produced only modest gains for ICL-based LLMs ($\Delta\kappa < 0.04$). Agent-refined frameworks improved cross-model agreement but remained below the human-refined framework. These findings highlight the promise of agent-based approaches and the importance of continued interaction between theory-guided human annotation and LLM-based methods in future work.

Keywords: Collaborative Learning, Interrater Reliability, Large Language Models, In Context Learning, LLM Agents

Introduction

Collaboration is widely recognized as a powerful context for learning and problem solving (Nokes-Malach, Zepeda, Richey, & Gadgil, 2019). It has the potential to enhance not only group performance on the task at hand but also subsequent individual problem solving (Laughlin, Carey, & Kerr, 2008; Lippold, Schulz-Hardt, & Schultze, 2021). However, simply working in groups is not always sufficient to produce meaningful gains (Do & Thomas, 2025). Productive collaboration depends on the extent to which learners engage in high-quality dialogue that supports the generation, refinement, and integration of ideas (Barron, 2003; Dillenbourg, 1999).

In the present study, we measured cognitive engagement in collaborative discourse using complementary approaches based on trained human annotation and large language models (LLMs). Human annotators can identify theoretically meaningful distinctions in discourse through contextual interpretation and structured coding frameworks (Chi, 1997; Krippendorff, 2018), whereas LLM-based methods offer strong potential for scalable analysis across large datasets (Gilardi, Alizadeh, & Kubli, 2023; Tan et al., 2024). We evaluated the reliability of LLM-based approaches for characterizing cognitive engagement and supporting refinement of the engagement framework relative to trained human annotation.

Cognitive Engagement in Collaborative Discourse

A prominent theoretical framework for characterizing cognitive engagement is the ICAP model, which distinguishes Passive, Active, Constructive, and Interactive modes of engagement based on whether learners contribute information beyond what is already available and whether they coordinate their reasoning with others (Chi, 2009; Chi & Wylie, 2014). Passive engagement involves attending without overt contribution, whereas active engagement involves participation that supports the discussion without introducing new understanding such as expressing agreement or repeating what was already stated. Constructive engagement involves extending understanding by generating explanations, summaries, interpretations, or questions. Interactive engagement involves sustained exchanges in which learners build on one another's reasoning. These modes form an ordered hierarchy of engagement, with interactive engagement associated with the most robust learning outcomes, followed by constructive, active, and passive engagement (Menekse, Stump, Krause, & Chi, 2013).

In collaborative problem-solving settings, important variation can arise within the same mode of participation, as learners' contributions differ in their depth and quality of engagement. For example, within the passive category, remaining silent reflects a different level of involvement than providing brief acknowledgments (e.g., "yes" or "uh-huh"), which signal attention and help maintain coordination among group members (Schegloff, 1982). Questions also vary in how much they advance understanding: simple clarification or fact-checking questions reflect active engagement, whereas higher-level "how" or "why" questions extend the discussion and signal movement toward constructive engagement (Graesser & Person, 1994; King, 1992). Similarly, contributions can range from partially developed ideas to well-supported explanations articulated with a clear rationale, with more elaborated explanations associated with greater learning (Lachner, Russ, Hübner, Sibley, & Scheiter, 2025). Interactive engagement likewise varies in depth, depending on whether learners briefly respond to others' ideas or engage in sustained exchanges that iteratively refine interpretations. Short sequences of constructive turn-taking reflect emerging coordinated reasoning, whereas extended back-and-forth exchanges reflect more sustained co-construction of shared understanding (Stahl, Koschmann, & Suthers, 2006).

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

To represent these gradations while preserving the theoretical ordering of engagement proposed in the ICAP framework, engagement in the present study was coded using a 7-point scale anchored to the four engagement modes. Introducing intermediate levels between adjacent categories allowed us to capture differences in participation when instances did not align cleanly with a single mode. This approach enables more fine-grained analyses of how engagement unfolds over time and provides a foundation for examining how subtle differences in participation may contribute to variation in learning outcomes during collaborative problem solving.

Large Language Models for Text Analysis

LLMs have increasingly been applied as scalable tools in psychological research, particularly for text-based labeling tasks in which models assign psychological categories to qualitative responses or discourse segments (Bermejo, Gago, Gálvez, & Harari, 2025; Lu, Zhang, & Wang, 2025). For example, prior work has shown that a fine-tuned GPT-3 model can outperform traditional manual coding in capturing orchestration actions in collaborative learning scenarios, highlighting the potential of LLM-based coding in complex psychological and educational contexts (Amarasinghe, Marques, Ortiz-Beltrán, & Hernández-Leo, 2023).

One capability of LLMs that has recently received particular attention is in-context learning (ICL) (Brown et al., 2020). This approach allows models to perform unseen tasks without retraining or updating weights by conditioning on a number of examples provided in the prompt. In some cases, ICL can achieve strong performance that approaches or even matches the performance of fine-tuning (Yin et al., 2024). Despite its effectiveness, it also has important limitations. Previous research has shown that ICL is highly sensitive to prompt design, as the selection, ordering, and quality of examples may lead to various outcomes (Liu et al., 2021). Furthermore, standard ICL lacks explicit, systematic, and persistent mechanisms to support complex, multi-step annotation pipelines that require ongoing iteration, feedback, and reflection.

In contrast to the one-shot prompting paradigm of standard ICL, LLM-based agents support more interactive workflows that enable decision-making and multi-step reasoning (Xi et al., 2023). These agents can maintain state across turns, perform iterative reasoning, and take actions, which may make them better suited for complex labeling tasks and more complicated systems. Accordingly, recent work has explored the use of LLM agents to support annotation processes. For example, *MEGAnno+* proposes a collaborative annotation framework in which humans and LLMs jointly generate reliable labels (Kim, Mitra, Chen, Rahman, & Zhang, 2024). Similarly, recent work introduces a multi-agent system that coordinates task assignment, data labeling, and quality management within the annotation pipeline (Qin et al., 2025). Motivated by these developments, the present study compares human annotation, ICL-based labeling, and LLM agent-based approaches for coding cognitive engagement in collaborative discourse.

The Present Study

The present study used an extended 7-point scale based on the ICAP framework to capture different levels of cognitive engagement during collaborative problem solving. Prior work has shown that extensions of the ICAP framework can help capture finer-grained variation in learners' cognitive engagement in collaborative learning contexts (Vosniadou et al., 2023). Building on this line of work, trained coders in the present study annotated learners' behaviors and iteratively refined category boundaries and coding criteria across two versions.

To compare human annotation with automated methods, we evaluated LLM-based labeling under both ICL and agent-based settings. LLM-based approaches can achieve relatively high accuracy using ICL and fine-tuning in cognitive research involving text analysis (Yin et al., 2024; Amarasinghe et al., 2023). However, ICL relies on single-pass prompting, which may limit agreement with human annotation. Work on agentic LLM systems has explored interactive workflows to improve annotation efficiency, cost control, and coordination (Kim et al., 2024; Qin et al., 2025). Yet these approaches typically do not model iterative reflection and refinement of the coding framework in response to ambiguous cases, as human annotators do. To address this gap, we propose an agentic LLM system that assigns labels while identifying ambiguities and iteratively refining the coding framework. We conducted a comparative evaluation of three methods—human annotation, ICL-based labeling, and agent-based LLM annotation—to examine differences in annotation reliability and framework refinement. Our code is available at `https://github.com/HanlingJ/human-icl-llm-agent`.

Method

Task and Dataset

We measured cognitive engagement during collaborative problem solving by analyzing discourse from three-person groups solving a logical inference task in which ten letters (A–J) were randomly assigned numerical values (0–9) without replacement. Participants inferred these hidden assignments through iterative reasoning across multiple trials (Do & Thomas, 2025). On each trial, groups proposed an arithmetic expression using letters and $+$ or $-$ signs (e.g., $A+B$), received the result in letter form (e.g., CE), and then guessed the value of one letter with immediate feedback. This structure enabled trial-level coding of individual cognitive engagement as groups progressed toward solving the full mapping.

Participants were undergraduate students at the authors' institution. All participants provided informed consent, and their discussions were video-recorded and transcribed. The final dataset included 42 group conversations (averaging 11 minutes and six trials per group).

Human Annotation of Cognitive Engagement

Two trained annotators classified cognitive engagement in the collaborative discourse. We used an extended 7-point scale

based on the ICAP framework to capture intermediate levels of engagement between the four original modes (Chi & Wylie, 2014). Annotators labeled participants' cognitive engagement on each trial of the task, including both the expression and the subsequent guess.

The 7-point scale ranged from low to high engagement. Level 1 (Passive) reflected listening without overt contribution. Level 2 captured minimal responses while primarily listening (e.g., brief acknowledgments such as "uh-huh"). Level 3 (Active) included observable participation such as expressing agreement or repeating a peer's statement. Level 4 (Initiating Constructive) represented emerging constructive engagement, such as asking higher-order "how" or "why" questions or introducing new input without elaborated reasoning. Level 5 (Constructive) involved proposing a new expression or guess with reasoning. The highest levels reflected sustained interactive engagement. Level 6 consisted of two-turn exchanges between constructive participants, and Level 7 consisted of three or more back-and-forth turns in which group members jointly developed or defended ideas. A turn was defined as all utterances produced by one speaker until another speaker began speaking.

Annotators independently coded group conversations and met regularly to resolve disagreements through discussion. This annotation process unfolded in three stages, reflecting the gradual refinement of the 7-point scale (Appendix D).

Stage 1 (Video data 1–10) included 306 labeling and reasoning samples. During this stage, the first author worked with the two annotators to establish and refine the rules and criteria for the 7-point Likert scale adapted from the ICAP framework. This version is referred to as **Criteria Version 1**. The first author served as a mediator, helping the annotators resolve disagreements and reach consensus.

Stage 2 (Video data 11–31) included 660 labeling and reasoning samples. The two annotators worked independently without mediator involvement, applying the criteria developed during Stage 1. They met after every two videos to discuss their annotations and resolve disagreements on their own, without additional guidance.

Stage 3 (Video data 32–42) included 315 labeling and reasoning samples. Following Stage 2, the first author and two coders collaboratively developed **Criteria Version 2**, with more detailed coding criteria. All videos in Stage 3 were coded using this revised framework. Annotators were asked to reference specific codebook criteria to justify their classifications of each behavior.

LLM Model Selection

Two models were used in this study: GPT-4o (OpenAI et al., 2024) and GPT-5.2 (Singh et al., 2025). Given the complexity of the task, we selected relatively advanced models, as prior work suggests they demonstrate stronger reasoning and produce more structured outputs.

In-context Learning

ICL was used to adapt the LLM to the cognitive engagement annotation task using a few examples provided in the prompt, without retraining or parameter updates. This approach allowed us to compare machine-based annotation with human annotators and examine whether different versions of the engagement criteria produced similar improvements in agreement across methods.

Data Splitting. For ICL, our trained human annotators segmented the video transcripts into trials by identifying the start and end of each trial. These trial-level segments were then used as inputs to the LLMs.

Algorithm 1: ICL for Annotation

Input: Criteria C; labeled examples $\{(x_i, y_i, r_i)\}$; data x.

Output: Predicted label $\hat{y}$, reasoning $\hat{r}$.

1. Build prompt.

$P \leftarrow I \,\|\, C$ `// instructions and criteria`

foreach *example* (x_i, y_i, r_i) **do**

  $P \leftarrow P \,\|\, (x_i, y_i, r_i)$ `// add ICL examples`

$P \leftarrow P \,\|\, x$ `// add the data sample`

2. Ask the LLM.

$\hat{y}, \hat{r} \leftarrow LLM f_\theta(P)$ `// LLM generates the answer`

Prompt and Examples. Our prompt followed this format: task description, output rules, the 7-point scale, labeled examples and test samples, and a final emphasis on the output rules (Appendix A). The ICL examples were drawn from the human coders' consensus codebook, in which all disagreements were resolved and final labels were jointly agreed upon by both annotators. These annotations were used as the ground truth for constructing ICL examples and evaluating model performance. Examples were randomly sampled from a separate transcript that was strictly disjoint from the test transcript, ensuring no overlap between the example and test data.

Self-Reflective Agent

Inspired by the iterative refinement process of the two human coders, we designed an LLM-based agent that performed both annotation and framework refinement. After labeling each data sample, the agent performed reflection to identify ambiguities in the current framework and made decisions about whether to modify, add, delete, hold, or merge classes before proceeding to subsequent samples (Appendix C).

The agent first labeled each sample, then identified weaknesses in the current framework and generated reflections and revision suggestions (Appendix B). At each iteration, the agent retained short-term memory of the three most recently labeled samples and the three most recent versions of the criteria, which served as context for reflection and labeling the next sample. This iterative memory enabled progressive self-refinement beyond a single-prompt labeling setup.

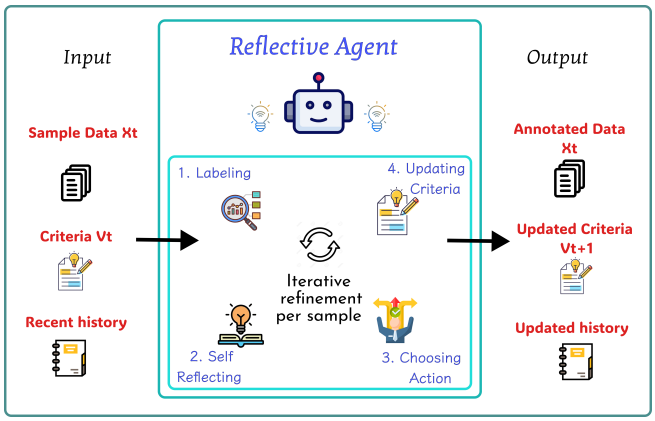

Figure 1: Reflective LLM Agent System With Memory Workflow.

Algorithm 2: LLM-based Agent with Reflection and Criteria Update

Input: Initial criteria $C^{(0)}$; data $\{x_t\}_{t=1}^{T}$.
Output: Annotations $\{(x_t, \hat{y}_t, \hat{r}_t)\}$; Updated criteria $C^{(final)}$.

for $t \leftarrow 1$ **to** T **do**

1. Label current sample
$P_{label} \leftarrow instructions, criteria$ C, $history$ H_{buf} with most recent K samples, and sample x_t
$(\hat{y}_t, \hat{r}_t) \leftarrow LLM f_\theta(P_{label})$

2. Reflect on criteria
Build P_{ref} from C, H_{buf}, and $(x_t, \hat{y}_t, \hat{r}_t)$
$\Delta \leftarrow LLM f_\theta(P_{ref})$ `// choose action`

3. Update criteria
$C \leftarrow \text{APPLYCHANGES}(C, \Delta)$

4. Update history
Append $(x_t, \hat{y}_t, \hat{r}_t)$ to H_{buf}

return All annotations, H history with all versions of the criteria and reflection, the final criteria $C^{(final)} \leftarrow C$.

Results

Interrater Reliability

We used three agreement coefficients: quadratic weighted kappa (QWK) (Fleiss & Cohen, 1973), Krippendorff's α (Hayes & Krippendorff, 2007), and Randolph's κ (Randolph, 2005). QWK captures chance-corrected agreement on the extended 7-point scale. Krippendorff's α is a standard reliability index in content analysis. Randolph's κ is a free-marginal reliability coefficient that is less sensitive to class imbalance than fixed-marginal kappa measures. Together, these metrics provide converging evidence for the reliability analyses.

We compared the original ICAP framework with the extended 7-point scale using zero-shot learning (Table 1). Interrater reliability between the two models increased substantially when using the 7-point scale (ΔQWK = 0.261, ΔRandolph's κ = 0.427, ΔKrippendorff's α = 0.352), suggesting that the extended framework supports more consistent annotation. Additional results with the 7-point scale across human annotators, ICL, and LLM agents are reported in Table 2.

Table 1: Interrater Reliability Between GPT-5.2 and GPT-4o With Zero-Shot Labeling Across Two ICAP Versions.

Metric	7-point Scale	Original ICAP
QWK	0.874	0.613
Randolph's κ	0.609	0.182
Krippendorff's α	0.839	0.487

Human–Human and Agent–Agent Reliability. Interrater agreement between the two human annotators (H^1H^2) was high in Stage 1 (QWK = 0.906; Krippendorff's α = 0.873) and increased further as the framework was refined, reaching near-perfect agreement by Stage 3 (QWK = 0.998; α = 0.999). Overall, reliability between the two human coders was very high (QWK = 0.974; Randolph's κ = 0.894).

Agreement between the two LLM agents (A^1A^2) was substantial but lower than human–human agreement. During the refinement process, the agents reached QWK = 0.841. After applying their separately refined frameworks under zero-shot prompting, agreement ranged from QWK = 0.655 to 0.790, suggesting variability in annotation consistency across foundation models. More advanced models may support more stable and reliable framework refinement.

Human–Machine Reliability. For machine-based annotation using GPT-5.2 with ICL prompting, human–machine QWK values ranged from 0.593 to 0.655 across stages and criteria versions, with Krippendorff's α around 0.575–0.641. The human-refined framework produced only small changes in reliability from Criteria Version 1 to Version 2 (overall ΔQWK $\approx$ 0.01–0.04), in contrast to the larger gains observed for human–human agreement. A similar pattern held for GPT-4o, with overall QWK remaining around 0.585–0.631. Differences between ICL and zero-shot prompting were also modest and sometimes favored zero-shot, indicating that few-shot prompting with human examples did not consistently improve human–machine reliability on this task.

Machine–Machine Reliability. Machine–machine reliability was higher than human–machine reliability but remained lower than human–human agreement. Under the ICL setting, QWK was 0.882 for Criteria Version 1 and 0.851 for Version 2, with corresponding Krippendorff's α values of 0.847 and 0.810, respectively. Zero-shot prompting using human-refined framework produced slightly higher agreement (QWK = 0.894 and 0.874), suggesting that in-context examples did not improve annotation consistency once a clear framework was provided. Overall, agreement remained stable across prompting conditions, with only modest reliability

Table 2: Interrater Reliability Across Human Coders, ICL and LLM Agents throughout Criteria Refinement Process

Agreement between	ICL setting	Model	Condition	QWK	Randolph's-κ	Krippendorff's-α
H^1H^2	N/A		Stage 1	**0.906**	**0.705**	**0.873**
			Stage 2	0.986	0.937	0.981
			Stage 3	**0.998**	**0.986**	**0.999**
			Overall	0.974	0.894	0.966
A^1A^2	N/A		Refinement Stage 1	0.841	0.600	0.823
$H^\star M$	ICL	GPT-5.2	Cri-V1 Stage 1,2	**0.655**	0.354	**0.641**
			Cri-V2 Stage 3	**0.593**	0.351	**0.575**
			Cri-V1 Overall	0.631	**0.327**	0.622
			Cri-V2 Overall	0.635	**0.369**	0.622
	Zero-shot	GPT-5.2	Cri-V1 Stage 1,2	**0.649**	0.315	**0.624**
			Cri-V2 Stage 3	**0.518**	**0.255**	**0.580**
			Cri-V1 Overall	0.629	0.300	0.613
			Cri-V2 Overall	0.608	**0.353**	0.594
	ICL	GPT-4o	Cri-V1 Stage 1,2	**0.631**	0.340	**0.616**
			Cri-V2 Stage 3	**0.585**	**0.417**	0.589
			Cri-V1 Overall	0.607	**0.323**	0.594
			Cri-V2 Overall	0.589	0.367	**0.569**
	Zero-shot	GPT-4o	Cri-V1 Stage 1,2	**0.640**	0.323	**0.626**
			Cri-V2 Stage 3	**0.523**	**0.332**	**0.509**
			Cri-V1 Overall	0.624	**0.315**	0.617
			Cri-V2 Overall	0.576	0.341	0.559
M^1M^2	ICL		Cri-V1 Overall	0.882	**0.541**	0.847
			Cri-V2 Overall	**0.851**	0.594	**0.810**
	Zero-shot		Cri-V1 Overall	**0.894**	0.547	**0.853**
			Cri-V2 Overall	0.874	**0.609**	0.839
	Zero-shot		Cri-A^1 overall	**0.655**	**0.406**	**0.636**
			Cri-A^2 overall	0.790	0.588	0.762

Note: H^1 and H^2 denote the human coders. A^1 and A^2 denote the GPT-4o and GPT-5.2–based agents. "Cri-V1" and "Cri-V2" refer to the human-refined criteria versions, whereas "Cri-A^1" and "Cri-A^2" refer to the criteria refined by the GPT-4o and GPT-5.2 agents. $H^\star M$ indicates agreement between the human consensus annotations and an LLM, whereas M^1M^2 indicates GPT-4o vs. GPT-5.2 agreement. "Stage" refers to phases of human annotation, and "Overall" aggregates results across all videos and stages. Purple represents the highest agreement values under each condition, and red represents the lowest.

gains from the human-refined framework.

Human Annotation Disagreements and Framework Refinement

Human disagreements occurred most frequently between Levels 4 and 5 (24% of all disagreement pairs), followed by Levels 5 and 6 (17%) and Levels 3 and 4 (16%), although this pattern may partly reflect the higher frequency of Level 5 behaviors. The confusion matrices further showed bidirectional confusion across levels (Figure 2). When one annotator coded Level 4, the other coded Level 5 in 11–23% of cases. A similar pattern was observed between Levels 5 and 6, with 13–22% of Level 6 ratings coded as Level 5 by the other annotator. Together, these findings indicate greater ambiguity in the mid-level categories than at the endpoints of the scale.

Consistent with this pattern of disagreement, human annotators made the most revisions to the mid-level categories, adding more detailed criteria to distinguish among Levels 3–5 (Appendix D). Across the agent refinement process, Levels 3–5 likewise underwent more modifications than the other levels (Figure 3), mirroring the refinement trajectory observed for human coders. Although agents could take multiple types of refinement actions, they only modified existing levels and did not add, merge, or remove any categories, preserving the seven-level structure throughout refinement.

Discussion

The present study highlights both the promise of LLM-based agent annotation and important differences in how engagement coding frameworks operate for human versus LLM-

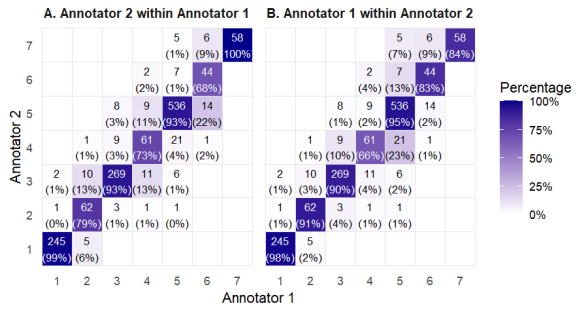

Figure 2: Confusion Matrix Between Human Annotators. Note: Panel A shows how Annotator 2 classified items within each Annotator 1 level, and Panel B shows how Annotator 1 classified items within each Annotator 2 level.

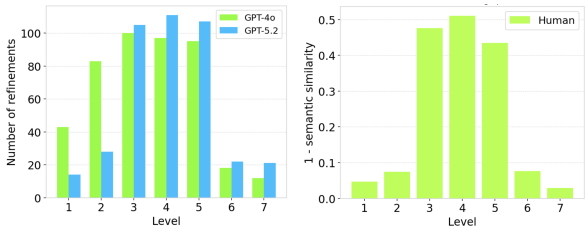

Figure 3: Agent Automated Refinement and Human Criteria Updates. Note: The left plot shows refinement counts per engagement level for the GPT-5.2 and GPT-4o agents, while the right plot shows $(1 - \text{CosSim})$ between the Criteria Version 1 and the Criteria Version 2 for each engagement level.

based annotators. Human coders showed consistently high interrater reliability across the three stages of annotation, indicating that continued training and clarification of category definitions improved agreement. LLM-based annotation also showed moderate to relatively high interrater reliability in machine–machine and agent–agent agreement. However, agreement between human and LLM-based annotations remained more moderate, indicating that consistency within LLM-based annotation does not necessarily translate into closer alignment with human coding. Furthermore, applying the same human-refined criteria produced only small and sometimes inconsistent changes in reliability for LLM annotators across annotation stages, suggesting that refinements optimized for human interpretation do not directly translate into comparable improvements for models through modification of the task instruction in the prompt alone.

Our results also indicate that ICL does not necessarily enhance annotation consistency for LLMs in this setting. Across criteria versions and model configurations, ICL and zero-shot prompting produced similar levels of agreement, and in several cases zero-shot prompting slightly outperformed ICL. This pattern suggests that few-shot examples derived from human annotations may not provide a stronger signal than clear task instructions alone for complex, theory-based engagement coding tasks.

When comparing human-developed criteria with agent-refined frameworks, we found that both supported substantial agreement between LLMs, though with greater variability under agent refinement. Cross-model agreement remained relatively high under human-developed criteria (QWK $\approx$ 0.85–0.89), but ranged more widely under agent-refined criteria (QWK $\approx$ 0.66–0.79). Notably, both human annotators and agents concentrated revisions on the mid-range engagement levels, particularly Levels 3–5 (Figure 3). This pattern suggests greater ambiguity in the mid-level categories, whereas the lower and upper levels remained relatively stable. These findings suggest that refining mid-level category boundaries may improve extended ICAP frameworks in future work.

Limitations and Future Directions

The first limitation concerns differences in the information available to human versus LLM-based annotators. Human coders had access to both video recordings and transcripts, whereas LLMs were restricted to text-only inputs. Although human annotators were instructed to focus on explicit verbal behaviors, they were naturally exposed to cues such as tone, timing, and nonverbal behavior while watching the videos. The restriction of LLMs to text-only inputs was due to the high computational cost of processing long video data with large vision–language models and may have reduced the sensitivity of LLM-based annotation to multimodal aspects of engagement. A second limitation concerns differences in how framework refinement unfolded for human annotators and LLM agents. Human refinement occurred across three stages, whereas LLM agents refined the framework at each iteration. Additionally, human annotators frequently discussed disagreements during coding, whereas agent-based refinement relied solely on each agent's own analysis. The high interrater reliability observed for human coders may partly reflect these opportunities for discussion. Future work should adopt more comparable refinement procedures and include independent annotators outside the consensus process to better evaluate their relative contributions.

Future work could also extend this research in two directions to further enhance scalability and annotation support. First, it is important to examine whether LLMs can be fine-tuned using the human-refined framework and annotated data, to assess whether models can acquire implicit knowledge not fully captured by the explicit criteria. Second, future research could explore LLM agent systems from two perspectives: hybrid systems in which human coders and LLM agents collaborate, and multi-agent systems in which multiple agents coordinate without human involvement. Both approaches may help refine the coding framework, identify challenging cases, and support annotation and quality control. A key question is whether LLMs can make reliable autonomous decisions and align with human judgments, which has important implications for redefining the roles of LLMs and human annotators in text analysis. Ultimately, these approaches may improve reliability while reducing the workload of human annotators.

Acknowledgments

This work was supported by the National Science Foundation (NSF) under Grant No. 2428640, *GCR: Towards a Convergent Understanding of the Dynamics of Uncertainty in Individuals and Groups with a Focus on STEM Education*. In this work, Lan Anh Do contributed to conceptualization, development of the human-refined framework, human-annotation training and supervision, data collection, data curation, formal analysis, and writing (original draft, review, and editing). Hanling Jiang contributed to the LLM-based methodology, software development and validation, experimental implementation and evaluation, data curation, formal analysis, and writing (original draft, review, and editing). Ayanna Thomas and Shuchin Aeron contributed to conceptualization, supervision, and writing (review and editing). AI tools were used for code development and to assist with proofreading.

References

Amarasinghe, I., Marques, F., Ortiz-Beltrán, A., & Hernández-Leo, D. (2023). Generative pre-trained transformers for coding text data? an analysis with classroom orchestration data. In O. Viberg, I. Jivet, P. J. Muñoz-Merino, M. Perifanou, & T. Papathoma (Eds.), *Responsive and sustainable educational futures* (Vol. 14200, pp. 32–43). Cham: Springer. doi: 10.1007/978-3-031-42682-7_3

Barron, B. (2003). When smart groups fail. *The journal of the learning sciences*, *12*(3), 307–359.

Bermejo, V. J., Gago, A., Gálvez, R. H., & Harari, N. (2025). Llms outperform outsourced human coders on complex textual analysis. *Scientific Reports*, *15*, 40122. doi: 10.1038/s41598-025-23798-y

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... others (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877–1901.

Chi, M. T. (1997). Quantifying qualitative analyses of verbal data: A practical guide. *The journal of the learning sciences*, *6*(3), 271–315.

Chi, M. T. (2009). Active-constructive-interactive: A conceptual framework for differentiating learning activities. *Topics in cognitive science*, *1*(1), 73–105.

Chi, M. T., & Wylie, R. (2014). The icap framework: Linking cognitive engagement to active learning outcomes. *Educational psychologist*, *49*(4), 219–243.

Dillenbourg, P. (1999). *Collaborative learning: Cognitive and computational approaches. advances in learning and instruction series.* ERIC.

Do, L. A., & Thomas, A. K. (2025). In-person collaboration, but not virtual, enhances problem-solving efficiency and knowledge transfer from groups to individuals. *Mind, Brain, and Education*, *19*(3), 130–139.

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, *33*(3), 613–619.

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, *120*(30), e2305016120.

Graesser, A. C., & Person, N. K. (1994). Question asking during tutoring. *American educational research journal*, *31*(1), 104–137.

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, *1*(1), 77–89.

Kim, H., Mitra, K., Chen, R. L., Rahman, S., & Zhang, D. (2024). *Meganno+: A human-llm collaborative annotation system.* Retrieved from https://arxiv.org/abs/2402.18050

King, A. (1992). Facilitating elaborative learning through guided student-generated questioning. *Educational psychologist*, *27*(1), 111–126.

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. Sage publications.

Lachner, A., Russ, H., Hübner, N., Sibley, L., & Scheiter, K. (2025). When does learning by non-interactive teaching work? a large-scale analysis of learner characteristics in a classroom setting. *Educational Psychology Review*, *37*(3), 88.

Laughlin, P. R., Carey, H. R., & Kerr, N. L. (2008). Group-to-individual problem-solving transfer. *Group Processes & Intergroup Relations*, *11*(3), 319–330.

Lippold, M., Schulz-Hardt, S., & Schultze, T. (2021). Gi transfer in multicue judgment tasks: Discussion improves group members' knowledge about target relations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *47*(3), 532.

Liu, J., Shen, D., Zhang, Y., Dolan, B., Carin, L., & Chen, W. (2021). *What makes good in-context examples for gpt-3?* Retrieved from https://arxiv.org/abs/2101.06804

Lu, Y.-L., Zhang, C., & Wang, W. (2025). *Systematic bias in large language models: Discrepant response patterns in binary vs. continuous judgment tasks.* Retrieved from https://arxiv.org/abs/2504.19445

Menekse, M., Stump, G. S., Krause, S., & Chi, M. T. (2013). Differentiated overt learning activities for effective instruction in engineering classrooms. *Journal of Engineering Education*, *102*(3), 346–374.

Nokes-Malach, T. J., Zepeda, C. D., Richey, J. E., & Gadgil, S. (2019). Collaborative learning: The benefits and costs.

OpenAI, :, Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., ... Malkov, Y. (2024). *Gpt-4o system card.* Retrieved from https://arxiv.org/abs/2410.21276

Qin, M., Zhu, R., Xia, M., Chenchenkai, Zhu, Z., Lin, M., ... Wang, H. (2025, November). CrowdAgent: Multi-agent managed multi-source annotation system. In I. Habernal, P. Schulam, & J. Tiedemann (Eds.), *Proceedings of the 2025 conference on empirical methods in natural language processing: System demonstrations* (pp. 925–942). Suzhou, China: Asso-

ciation for Computational Linguistics. Retrieved from https://aclanthology.org/2025.emnlp-demos.72/ doi: 10.18653/v1/2025.emnlp-demos.72

Randolph, J. J. (2005). Free-marginal multirater kappa (multirater k [free]): An alternative to fleiss' fixed-marginal multirater kappa. *Online submission*.

Schegloff, E. A. (1982). Discourse as an interactional achievement: Some uses of 'uh huh'and other things that come between sentences. *Analyzing discourse: Text and talk*, *71*(93).

Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., ... Wang, Z. (2025). *Openai gpt-5 system card.* Retrieved from https://arxiv.org/abs/2601.03267

Stahl, G., Koschmann, T. D., & Suthers, D. D. (2006). *Computer-supported collaborative learning*. na.

Tan, Z., Li, D., Wang, S., Beigi, A., Jiang, B., Bhattacharjee, A., ... Liu, H. (2024). Large language models for data annotation and synthesis: A survey. In *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 930–957).

Vosniadou, S., Lawson, M. J., Bodner, E., Stephenson, H., Jeffries, D., & Darmawan, I. G. N. (2023). Using an extended icap-based coding guide as a framework for the analysis of classroom observations. *Teaching and Teacher Education*, *128*, 104133.

Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., ... Gui, T. (2023). *The rise and potential of large language model based agents: A survey.* Retrieved from https://arxiv.org/abs/2309.07864

Yin, Q., He, X., Leong, C. T., Wang, F., Yan, Y., Shen, X., & Zhang, Q. (2024, November). Deeper insights without updates: The power of in-context learning over fine-tuning. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Findings of the association for computational linguistics: Emnlp 2024* (pp. 4138–4151). Miami, Florida, USA: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2024.findings-emnlp.239/ doi: 10.18653/v1/2024.findings-emnlp.239

Appendix A: In Context Learning

Prompt

In Context Learning Prompt = ICL task description + output rules + the 7-point framework + labeled examples and test samples + final emphasis on the output rules

Example:

```
You will be given a classification task
involving student engagement in a group
discussion while playing the letters-to-number
game.  Your goal is to analyze each conversation
and, for each subject, assign a 7-point Likert
scale score and provide a brief reasoning based
on their statements.  You will first see a few
examples.  Then, you will be given a new task to
label.

Please follow these rules:  ·For each
**Subject**, assign a Likert scale score (1-7)
and provide a concise reasoning for your choice.
·Follow the structure in the example.  ·Only
provide labels and reasoning for trial(s) shown
in TASK. **Do not** make up or add any trial
that are not included.

Insert examples and test samples here

Please follow these rules:  · For each
**Subject**, assign a Likert scale score (1-7)
and provide a concise reasoning for your choice.
· Follow the structure in the example.  · Only
provide labels and reasoning for trial(s) shown
in TASK. **Do not** make up or add any trial
that are not included.  For each subject:  1.
Assign a Likert score refer to your criteria.
2.  Provide short reasoning.  3.  Output your
annotation follow the format:
Trial n expression
subject_1 (Subject ID): label = x, reasoning =
one line
subject_2 (Subject ID): label = x, reasoning =
one line
subject_3 (Subject ID): label = x, reasoning =
one line

Trial n guess
subject_1 (Subject ID): label = x, reasoning =
one line
subject_2 (Subject ID): label = x, reasoning =
one line
subject_3 (Subject ID): label = x, reasoning =
one line
Again, only provide labels and reasoning for
trial(s) shown in TASK. **Do not** make up or
add any trial that are not included.

"""
```

Appendix B: Example of Agent Reflection

Reflection

The current codebook leaves gaps that caused interpretive decisions:

(a) proposals without justification are not explicitly placed (we treated them as 4)

(b) weak heuristics blur 4 vs 5

(c) 6–7 rely on vague 'turn-taking' language that can mislabel parallel contributions as interactive

(d) the 1–3 boundary between silence, backchannel, and simple task talk is underspecified. The proposed updates supply clear decision rules and examples aligned with the observed

utterances (e.g., 'random guess' as 4; 'mid-range' without why as 4; constraint-based reasoning as 5) and make interaction levels reliably detectable by requiring explicit builds across turns.

Appendix C: Actions Assigned to the Agent for Framework Refinement

- Modify: refine an existing scale definition
- Add: introduce a new scale level
- Merge: merge multiple scales into a target scale
- Delete: remove a scale entirely
- Rename: reindex a scale (e.g., shifting numbers) or rename keys
- Hold: no changes.

Note: If no action was detected, *hold* was automatically selected. Although the agent was allowed to take different actions, in our experiments it consistently chose only to modify one or more of the seven levels and never proposed adding, merging, or deleting any level.

Appendix D: Criteria Versions

Criteria Version 1

Level 1 (Passive):
The student does not say anything.

Level 2 (Between Passive and Active):
Minimal responses like 'uh-huh'.

Level 3 (Active):
Agrees, repeats others, asks simple clarifications.

Level 4 (Between Active and Constructive):
Adds reasoning or asks 'why'/'how'.

Level 5 (Constructive):
Suggests expressions or guesses with reasoning, explains situation.

Level 6 (Between Constructive and Interactive):
Two turn-taking between two constructive participants.

Level 7 (Interactive):
Three or more turn-taking between two constructive participants.

Criteria Version 2

Level 1 (Passive):
Not saying anything.

Level 2 (Between Passive and Active):
Has a minimal response, such as "uh-huh."

Level 3 (Active):
Agreed and/or repeated what others said.
Asked simple questions for basic clarification (e.g., "What did you just say?" or whether they could divide the letters).
Simple correction (e.g., when all members agreed that C was 6, but the host mistakenly entered 7, and someone corrected her).
Expressions and guesses based on others' suggestions and kept track of the group's progress by taking notes, without contributing new information beyond what others already knew.

Level 4 (Between Active and Constructive):
Agreed and added their own reasoning.
Asked advanced questions that led to constructive information, such as "how" and "why" types.
Provided some new input (e.g., an alternative expression; or says the expression should go up to J) but did not elaborate the reasoning.
Attempted to provide new input, but failed to complete their reasoning.
Disagreed without elaborating the reasoning.

Level 5 (Constructive):
Suggested a new expression or guess with reasoning (or implicitly conveys reasoning that is understood by members). This includes suggestions such as making a random guess.
Answered questions.
Explained or summarized the situation to a member.
Identified possible values or ruled out impossible values for a specific letter or number.
Disagreed with reasoning.

Level 6 (Between Constructive and Interactive):
Two turn-taking between two constructive participants.

Level 7 (Interactive):
Three or more turn-taking between two constructive participants.