

Measuring Collaborative Reasoning with LLMs

Dawei Xie, Tochukwu Eze, Marcelo Worsley

Northwestern University

{daweixie, td.eze}@u.northwestern.edu

marcelo.worsley@northwestern.edu

Abstract

We present Collaborative Reasoning State (CRS), a framework representing multi-party reasoning as contributions, argumentative relations, and derived participation structure. Evaluating five closed-weight and open-weight LLMs on 152 student discussions, we find that reasoning is classifiable but not reliably segmentable, with identification being the bottleneck for both models and human annotators. Prompting improves labeling far more than identification. Downstream, participation shares are robust to detection error, whereas fine-grained structure remains difficult to recover. These findings have implications for the design of LLM-based measurement and learning analytics for collaborative reasoning.

1 Introduction

Consider five consecutive turns from a student team deciding how to structure their web application:

OLIVIA: *And then also, how do you want to structure it? Do you want to have it like when you click on each ...*

EMMA: *Sidebar tab. It actually goes to a different URL or if it just ... it just renders a new thing in the main area.*

OLIVIA: *I think we should just keep everything on the same page. It's definitely a lot easier. We don't have to work with like routing or anything.*

SOPHIA: *I think it could just split into different pages, right? Like, everyone can work on a different page.*

EMMA: *So we can have a side panel and when you click on it, it imports like an external page into the working window ... we can write them in separate files, but it won't go to a different URL.*

The five turns contain several distinct reasoning moves: an *issue* raised and co-constructed across two speakers, a *proposal* that addresses it, a supporting *claim*, another *proposal* challenging the previous one, and a final *proposal* that synthesizes them. A system that supports or assesses collaborative discussion needs to represent these moves: *Was the single-page proposal challenged?* (Yes; Sophia

counter-proposed splitting into pages for parallel work.) *What is on the table now?* (A merged proposal for their design choice: one page but with sections loaded from separate component files.) *Who has built on whose ideas?* (Emma synthesized Olivia's and Sophia's positions.) Figure 1 renders the CRS for this excerpt.

In collaborative learning, group reasoning is a central target of assessment and feedback (Weinberger and Fischer, 2006; Michaels et al., 2008). As LLMs are increasingly used to summarize educational dialogues (Jain et al., 2023), generate formative feedback (Dai et al., 2024), and support learning analytics (Misiejuk et al., 2025), a question becomes pressing: **how reliably can collaborative reasoning be measured, and can LLMs serve as measurement instruments?** Existing formalisms capture related but different structures: RST (Mann and Thompson, 1988) and SDRT (Asher and Lascarides, 2003) model coherence relations, argumentation mining (Stab and Gurevych, 2017) models argumentative components primarily in monologic text, and common ground tracking (Khebour et al., 2024) monitors shared beliefs, evidence, and questions under discussion.

We introduce **Collaborative Reasoning State (CRS)**, a discourse representation that tracks, as a discussion unfolds, typed contributions (CLAIM, ISSUE, PROPOSAL), reasoning relations between them (SUPPORTS, CHALLENGES, REFINES, SYNTHESIZES, ADDRESSES), and the derived participation structure.

To evaluate whether current language models can recover CRS structure, we build a benchmark from 152 natural multi-party student discussions from a human-computer interaction (HCI) class, annotate them with CRS to form a reference standard, and evaluate five LLMs spanning closed-weight and open-weight models across three prompting conditions. We find that *labeling* the reasoning moves a model has already found is tractable, while

identifying which contributions and relations are there at all is substantially harder. The best model reaches only 0.54 Contribution F1 and 0.25 Relation F1. The labeling–identification gap persists across prompting conditions and model tiers, and also propagates into derived structure. Thread partitions, built from the weakly detected relations, are recovered poorly (same-thread pairwise F1 0.25–0.31), whereas participation (who contributed, and in what proportion) survives detection error and is recovered well.

The paper contributes (1) the CRS framework as an operational representation of collaborative reasoning; (2) an evaluation over 152 classroom discussions, five LLMs, and three prompting conditions, with quantified inter-annotator reliability; and (3) findings that the unreliable part of measuring collaborative reasoning is *identifying* the units (segmentation) rather than *labeling* them (classification), which clarifies where current LLMs can and cannot be trusted for multi-party discourse understanding and informs the design of AI-integrated assessment and learning analytics systems.

2 Related Work

Discourse processing offers several representations of discourse structure. RST (Mann and Thompson, 1988) models coherence in text through rhetorical relations, while SDRT (Asher and Lascarides, 2003) extends discourse interpretation to dialogue. Argumentation mining work, such as Stab and Gurevych (2017), models argumentative components and support or attack relations, primarily in monologic text. Dialogue state tracking (Williams et al., 2016) represents task-oriented user goals, while common ground tracking (Khebour et al., 2024) tracks shared beliefs and questions rather than explicit reasoning relations among contributions. Multi-party corpora such as AMI and Deli-Data (Carletta et al., 2005; Karadzhov et al., 2023) annotate dialogue acts, summaries, or deliberation cues rather than typed reasoning graphs. In learning sciences, argumentative knowledge construction and group cognition (Weinberger and Fischer, 2006; Stahl, 2006) motivate CRS but have not been operationalized as a speaker-attributed contribution–relation framework for LLM-based analysis.

Recent work examines related tasks. Tran et al. (2025) use few-shot prompting to jointly segment and label within-turn argument components, reporting difficulty finding exact boundaries while com-

ponent order is often preserved. CRS instead models typed relations among sub-turn contributions across speakers and derives thread and participation structure. Ravi et al. (2026) study conversational-thread identification in synchronous small-group talk and show that thread information improves LLM coding. CRS derives threads from typed reasoning relations and evaluates their recoverability.

3 The CRS Framework

The CRS represents the evolving state of a group’s collaborative reasoning. At any point during a discussion, the CRS captures what contributions are on the table, how they connect into lines of reasoning, and who has contributed what to each line.

3.1 Theoretical Grounding

CRS draws primarily on **argumentative knowledge construction** (Weinberger and Fischer, 2006), where we take four dimensions of productive collaborative discourse: an *epistemic* dimension (task-relevant knowledge construction), an *argument* dimension (how arguments are constructed, warranted, and countered), a dimension of *social modes of co-construction* (the degree to which participants operate on each other’s reasoning), and a *participation* dimension (who contributes how much). We additionally draw on **grounding in communication** (Clark and Brennan, 1991), which describes how interlocutors establish mutual understanding through an iterative process of providing evidence of uptake. A contribution is *grounded* when sufficient evidence of understanding exists. CRS extends this from dyadic to multi-party settings, with cross-speaker reasoning relations serving as explicit evidence of uptake.

3.2 Core Ontology

The CRS has three components: **contributions**, **relations**, and **participation structure**. Contributions and relations are identified directly from the transcript, whereas participation structure is derived from them as a pair of sub-objects, including threads (groups of related contributions) and speaker profiles (per-thread authorship and role). We describe each in turn.

Contributions. A contribution is the atomic unit of the CRS, representing a single propositional move in the group’s reasoning. Contributions are smaller than turns (one turn may contain several) but larger than clauses (a contribution may span

multiple clauses expressing a unified proposition). Each contribution carries a *type* from three categories that correspond to distinct speech-act functions in collaborative reasoning discourse:

- CLAIM: asserting that something is the case.
- ISSUE: raising a problem, concern, or open point to be resolved.
- PROPOSAL: suggesting an action, decision, or direction.

The three are made deliberately coarse in that finer distinctions (e.g., separating evidence from claim) could be encoded in the relation structure rather than the type, and pilot annotation showed that finer-grained type sets reduced inter-annotator agreement without adding analytic value.

Relations. A relation is a directed edge from a source contribution to a target one, representing the reasoning move the source makes with respect to the target. Five relation types capture the core argumentative moves in collaborative discussion:

- SUPPORTS: provides a reason, grounds, or evidence for the target.
- CHALLENGES: raises an objection or counter-consideration.
- REFINES: narrows, qualifies, or improves the target without changing stance.
- SYNTHESIZES: combines two or more contributions into a new formulation (requires ≥ 2 targets).
- ADDRESSES: responds to an ISSUE without itself explicitly supporting or challenging.

A contribution can participate in multiple relations to different targets; it can also have zero incoming relations (e.g., a question opening a new line). Relations are attributed to the speaker who produced the source contribution.

Participation structure. The third component has two parts that are jointly derivable from the relation graph. A **thread** groups contributions addressing a common line of reasoning, so that contributions with no incident relation form singleton threads; threads surface which lines of reasoning the discussion has pursued and which have stalled as isolated points. A **speaker profile** records, for each (speaker, thread) pair, which threads the speaker initiated or contributed to and what types

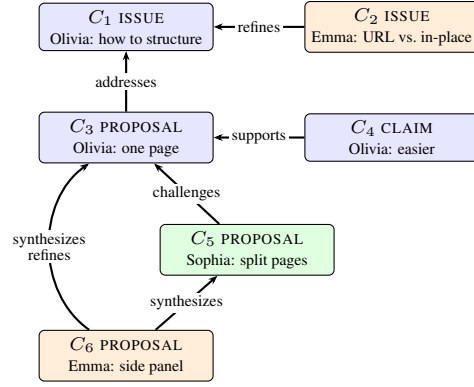


Figure 1: **A CRS example.** The six contributions form a single thread (one connected component), with per-thread participation distribution {Olivia:3, Emma:2, Sophia:1}.

of contributions they produced there. Together, they support computing participation equity (e.g., the distribution of contributions across speakers) and characterizing each speaker’s role across the discussion.

4 Study Design

4.1 Corpus

We study 152 naturally occurring multi-party student discussions collected over 8 weeks from a university HCI class (34.2 turns and 3.2 participants per discussion on average; range 5–90 turns). The discussions are mostly task-oriented where students reason together about design decisions. The corpus is pseudonymized and collected under an IRB protocol. The 152 discussions are a cohort drawn after excluding transcripts with intrinsic data quality flaws.

4.2 Annotation and Reliability

Two annotators with backgrounds in discourse analysis and collaborative learning independently annotated every discussion with full CRS structure. Since the two annotators segment at different granularities, we measure alignment-based agreement where contributions are matched by SBERT cosine ≥ 0.7 (Reimers and Gurevych, 2019), constrained to the same speaker and turn. Disagreements on contribution segmentation, types, relation endpoints, and relation types were then resolved in adjudication sessions. The reconciled annotations constitute the gold standard used throughout.

Annotators segmented at the level of propositional moves. A turn was split only into units that could each stand alone (e.g., a PROPOSAL and its

supporting reason, linked by SUPPORTS); repetition, emphasis, discourse management, and inseparable facets of a single idea were not split, and contributions were never merged across speakers.

Crucially, two reliability patterns are surfaced: **Conditional labeling agreement is substantial:** on contributions identified by both annotators, type agreement is $\kappa=0.77$ and relation-type agreement is $\kappa=0.71$, both in the “substantial” band (Lanidis and Koch, 1977; Artstein and Poesio, 2008). **Identification agreement is weak:** Contribution F1 is only 0.44. The annotators also differ sharply in granularity, marking 30.6 versus 55.3 contributions per discussion ($1.8\times$). This indicates a unit-of-analysis problem: annotators agree on what a reasoning move *is* but not on how finely to cut the transcript into moves. Boundary comparison confirms this: of one annotator’s spans, 37% exactly match the other’s and 57% are nested within or contain one, while only 3% cross boundaries, indicating that disagreement is refinement or coarsening rather than placement—the unitizing problem that γ (Mathet et al., 2015) was designed to measure. The reconciled gold standard is coarser than either annotator: 22.8 contributions per discussion (3,467 total, 3,260 of them with a reference type; 50% CLAIM, 32% PROPOSAL, 18% ISSUE) and 851 relations (670 with a reference type).

4.3 LLM Evaluation Setup

We evaluate five LLMs spanning closed-weight and open sourced models: **GPT-5**, **Claude Opus-4.1**, **Gemini-2.5-Pro**, **Llama-4-Maverick**, and **Qwen3-235B**. We evaluate how well current LLMs perform the detection step of CRS tracking, which requires recovering contributions and relations from raw multi-party transcripts, under three prompting conditions of increasing richness: `transcript_only` (output format and label names only), `schema_only` (full ontology definitions and constraints), and `schema_few_shot` (ontology definitions with four worked examples).

All runs requested schema-constrained output. For GPT-5, Opus-4.1, and Gemini-2.5-Pro the API enforces the schema during decoding. For Llama-4-Maverick and Qwen3-235B, served via OpenRouter, the API cannot enforce a schema, so the schema was supplied in the prompt and the free-form output was parsed afterwards. Section 5.3 shows that this difference does not affect the results. More generally, the five models differ in size, architecture, serving route, and decoding regime

(Appendix A), and thus the results should be interpreted as differences between the tested configurations rather than model families or tiers.

Our experiments answer four questions. (Q1) Can LLMs correctly *identify* which spans of the transcript are contributions, and which pairs stand in a reasoning relation? (Q2) When they do identify a contribution or relation, can they *label* its type correctly? (Q3) Does richer prompting improve identification, labeling, both, or neither? (Q4) How faithfully does the derived state (thread structure, participation profiles) propagate from the detected contributions and relations?

Task. Given a speaker-attributed transcript, produce the CRS contributions (each with speaker, turn, type, and verbatim span) and the CRS relations (each with source, target, and type).

Metrics. Let G and P be the gold and predicted contributions for a discussion, and $A \subseteq P \times G$ the greedy one-to-one alignment formed by matching pairs in descending SBERT cosine of their summary strings, restricted to pairs that share speaker and turn and whose similarity is at least $\tau=0.7$. Let M_P, M_G be the matched sides. Let G_R denote the gold relations whose endpoints were both identified, P_R the predicted relations whose endpoints were both aligned (mapped through the alignment), and $A_R \subseteq P_R$ those whose mapped (source, target) pair coincides with a gold relation.

- **Contribution F1:** $2PR/(P + R)$ with precision $|M_P|/|P|$ and recall $|M_G|/|G|$;
- **Type accuracy:** $|\{(p, g) \in A : \text{type}(p) = \text{type}(g)\}|/|A|$ over aligned pairs whose gold unit carries a reference type;
- **Relation F1:** F1 on A_R ignoring relation type, with precision $|A_R|/|P_R|$ and recall $|A_R|/|G_R|$;
- **Relation-type accuracy:** the fraction of A_R whose type matches.

All metrics are macro-averaged over discussions. Type accuracy and relation-type accuracy are computed only where the aligned set is non-empty, and model-level scores average the three prompting conditions equally. We report means over the 152 discussions with 2,000-sample bootstrap 95% CIs (resampling discussions); the model ranking is stable across $\tau \in \{0.6, 0.7, 0.75, 0.8\}$ (Appendix C).

5 Results

Figure 2 illustrates per-model performance, which surfaces two patterns: **(A) Labeling is tractable.** Closed-weight models type aligned contributions at ≈ 0.92 and relations at ≈ 0.92 . **(B) Identification is hard:** the best Contribution F1 is 0.54 and the best Relation F1 only 0.25. The asymmetry holds within every tier (closed-weight: 0.92 vs. 0.47–0.54; open-weight: 0.68–0.71 vs. 0.33–0.39), a gap (type accuracy – Contribution F1) of **+0.32 to +0.44 for every model**. Holding one annotator’s raw annotation fixed, we score the other annotator and each model against it. Against the coarser annotator (A1), Gemini (0.468) and GPT-5 (0.462) agree with A1 no worse than the second annotator does (0.444; paired 95% CIs on the difference include zero), while the other three models fall short. Against the finer annotator (A2), every model falls short (best: GPT-5, 0.390).

5.1 The Labeling–Identification Asymmetry

Type accuracy is conditional on each model’s own alignment (coverage 32–59% of the gold standard, mildly enriched for the easiest type, PROPOSAL). To remove that conditioning, we gave each model the gold contributions (span, speaker, turn; no type) and asked for types only. On all 3,260 reference units with a reference type, accuracy is 0.897 (GPT-5), 0.880 (Gemini), 0.845 (Llama), and 0.825 (Qwen), against a majority-class baseline of 0.50 (Figure 2, green markers), so the gap measured on identical units is +0.34 to +0.52. Closed-weight models still label more accurately (paired +0.04 to +0.07, all CIs exclude zero), but the open models’ apparent labeling deficit is segmentation, not labeling (≈ 0.22 on self-identified units, ≈ 0.05 on identical ones). Their disadvantage lies mainly in finding units (Contribution F1 gap 0.15).

Relation detection inherits most of its error from contribution identification. Given the gold contributions and types, relation recall rises from 0.07–0.35 to 0.54–0.86, while precision against the gold standard stays 0.20–0.30 (Table 1). That precision understates the models: 37% of their apparent false positives were edges at least one annotator marked, and counting those as correct raises precision to 0.42–0.66. Models thus disagree with the gold standard about *which* edges, as the annotators do with each other, so we treat recall as the primary relation metric. Typing a co-identified relation is reliable (0.71–0.96 by type, Table 7).

Prompting does not close the identification gap. Adding the schema (Table 2) raises type accuracy by +0.17 on average (0.72→0.89), driven by the open models acquiring the taxonomy (Qwen 0.46→0.86, Llama 0.42→0.84; closed-weight models are near ceiling, Appendix B), whereas identification improves by only +0.04 (paired 95% CI [+0.03, +0.05]). Worked examples slightly reduce both (−0.02 each, CIs exclude zero). Schema documentation conveys the taxonomy but not where to segment.

5.2 Derived State: What Survives Detection Error

We apply the deterministic derivation rules identically to the gold and predicted graphs, where threads are connected components of the relation graph and participation is the distribution of contributions across speakers (Table 3). Two derived measures behave very differently. **Participation is recovered well.** We report the Jensen–Shannon divergence $JSD(p, g) = \frac{1}{2} KL(p \| m) + \frac{1}{2} KL(g \| m)$, with $m = (p + g)/2$, between the predicted and gold distributions of contributions over speakers ($JSD \in [0, 1]$, lower is better). JSD is 0.017–0.037 for every model, against a turn-share baseline (each speaker’s share of turns; no model at all) of 0.035, a uniform baseline of 0.073, and an annotator-to-annotator scale of 0.008 (annotator-to-gold 0.012–0.017). Who contributed and in what proportion, survives even imperfect detection, because misses are roughly proportional across speakers. **Thread structure is recovered poorly:** same-thread pairwise F1 is 0.25–0.31 (null = 0), and ARI is uninformative here because 81% of gold threads are singletons and an all-singleton prediction already scores 0.25–0.47 (Table 3). Aggregate, participation-based measures are thus far more *automatable* than fine-grained argument-structure measures.

5.3 Robustness and Error Analysis

Baseline. A one-contribution-per-turn baseline reaches Contribution F1 0.16 and type accuracy ≤ 0.42 (Appendix D); the best LLMs exceed it by +0.38 on identification and +0.50 on labeling. **By type** (Table 4), PROPOSAL is easiest to identify (recall 0.57) and CLAIM/ISSUE hardest (~ 0.40); ADDRESSES is the hardest relation to detect. **Failure modes** differ: GPT-5 and Gemini over-segment, Llama over-produces with low precision, and Opus gets about the right count but segments differently.

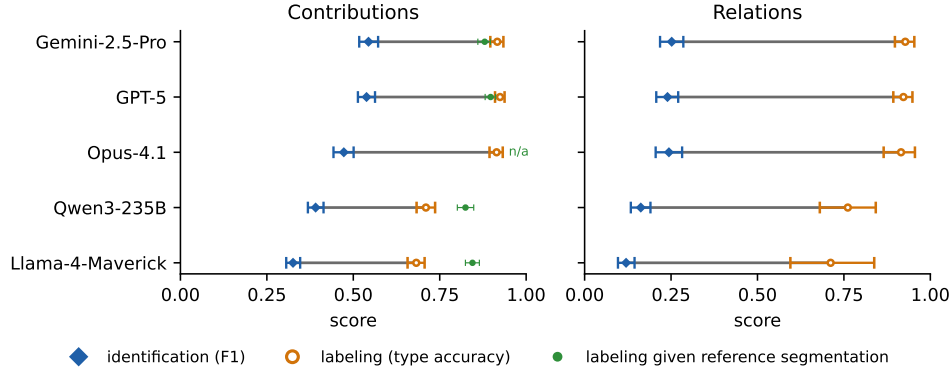


Figure 2: **Detection performance.** Each model (rows, sorted by Contribution F1) appears as a diamond for *identification* and an orange circle for *labeling*, joined by a line marking the gap. **Left (contributions):** identification is Contribution F1, labeling is type accuracy. Green dots: type accuracy when the model is given the reference segmentation (Opus-4.1 was retired before that experiment). **Right (relations):** identification is relation-presence F1, labeling is relation type accuracy. Means over the 152 discussions (averaged over the three prompting conditions) with 95% bootstrap CIs (2,000 resamples of discussions).

Model	Given reference segmentation			Run-to-run SD	$\Delta F1$ [95% CI]
	Type acc.	Rel. recall	Rel. precision		
Gemini-2.5-Pro	.880	.864	.296 / .662	.010	−.020 [−.040, −.000]
GPT-5	.897	.800	.251 / .548	.008	+.002 [−.015, +.018]
Qwen3-235B	.825	.573	.201 / .437	.025	+.000 [−.020, +.022]
Llama-4-Maverick	.845	.538	.201 / .421	.009	+.004 [−.017, +.025]

Table 1: **Decomposition and robustness checks.** *Given reference segmentation:* type accuracy given the gold segmentation (the 3,260 contributions); relation recall and precision given the gold contributions and their types (strict / crediting edges marked by either annotator). *Run-to-run SD:* SD of the three-condition-average Contribution F1 over three generations on a length-stratified 40-discussion subsample. $\Delta F1$: paired difference in Contribution F1 between schema-as-prompt free decoding and provider-enforced decoding, schema_only, 146–152 discussions.

Condition	Identif. (Contrib. F1)	Label. (Type)
transcript_only	0.43	0.72
schema_only	0.48	0.89
schema_few_shot	0.45	0.87

Table 2: Prompting affects labeling far more than identification (avg. over models). Schema documentation lifts type accuracy by +0.17 and identification by +0.04 (paired 95% CIs exclude zero for both).

Robustness. Absolute scores depend on the alignment threshold (best Contribution F1 $0.67 \rightarrow 0.36$ from $\tau=0.6$ to 0.8), but the ranking is identical at every $\tau \in \{0.6, 0.7, 0.75, 0.8\}$ and the labeling–identification gap stays positive for every model, widening as matching tightens (+0.16 to +0.63; Appendix C). The ranking is also unchanged across gold-standard variants.

Stability. Re-running all three conditions twice more on a length-stratified 40-discussion subsample (four models; Opus-4.1 was retired), the three-condition Contribution F1 varies by SD 0.008–

0.010 (Qwen 0.025; Table 1), inside the reported CIs, with an identical ranking in every run. Open models’ unit density also varies within a model across identical re-runs (± 2.5 contributions per discussion for Qwen vs. ± 0.1 for GPT-5).

Decoding regime. Constrained decoding can degrade reasoning (Tam et al., 2024). Using the same discussions, we compared provider-enforced schema decoding with schema prompting under free decoding. Paired Contribution-F1 differences are within ± 0.02 for all four models (Table 1); three CIs include zero and Gemini’s borderline -0.020 favors enforcement. The open models’ schema gain appears in both regimes, and the labeling–identification asymmetry is unlikely to be driven by decoding constraints.

6 Discussion and Conclusion

A reliability result with validity consequences. The consistent finding across two human annotators and five models is that, in this corpus, collaborative

Model	Thread pairwise F1 $\uparrow$	ARI / null	Participation JSD $\downarrow$
Gemini-2.5-Pro	.282	.304 / .275	.017
GPT-5	.278	.286 / .253	.018
Opus-4.1	.311	.361 / .347	.037
Qwen3-235B	.252	.330 / .399	.029
Llama-4-Maverick	.249	.403 / .466	.024
Turn-share baseline (no model)	—	—	.035
Uniform baseline	—	—	.073
Annotator A1 / A2 vs. gold	—	—	.012 / .017

Table 3: Derived-state agreement with the gold standard (avg. over conditions). Thread agreement is computed on aligned units; ARI is shown only to document its degeneracy against the all-singleton null. Participation JSD is reported against no-model baselines and the annotator scale.

Model	Identification recall			Labeling acc.		
	CLAIM	ISSUE	PROPOSAL	CLAIM	ISSUE	PROPOSAL
Gemini-2.5-Pro	.50	.47	.68	.92	.96	.92
GPT-5	.53	.52	.70	.90	.95	.94
Opus-4.1	.38	.36	.58	.93	.95	.91
Qwen3-235B	.33	.34	.51	.61	.66	.88
Llama-4-Maverick	.27	.31	.40	.64	.57	.81

Table 4: Per-type identification recall vs. labeling accuracy. PROPOSAL is consistently easiest to *find*; labeling is high for closed-weight models across all types and a weakness for open models on CLAIM/ISSUE.

reasoning is reliably *classifiable* but not reliably *segmentable*. Human annotators agree on what a contribution is, yet mark it at $1.8\times$ different rates, and models inherit the same difficulty. For measurement, this relocates the validity problem. The threat is not that systems mislabel reasoning moves, but that they disagree on which moves exist.

Implications for AI-integrated assessment.

First, *structural plausibility is not structural accuracy*. A model can produce internally well-typed CRS such that every span classified correctly while missing much of the actual structure, so agreement with a single LLM is not a sufficient quality signal, necessitating a human-validated layer. Second, *address identification as a separate engineering problem*. The prompting strategies tested here improved labeling but yielded only modest gains in identification. We therefore suggest practitioners targeting identification also explore complementary approaches such as supervised segmentation or human-in-the-loop workflows. Third, *choose measures by what survives detection error*. Participation-share analytics are recoverable with the stronger off-the-shelf models, whereas argument-structure analytics (threads, relation graphs) are not yet reliable enough to stand alone. An assessment system can responsibly report “who participated and in what proportion”,

while “how the reasoning was structured” remains low-confidence and should be flagged as such.

CRS as a useful yet partial target. CRS captures reasoning structures in multi-party collaborative discussions that existing discourse frameworks may overlook, and our study shows it is partially recoverable with current models. However, the identification ceiling, shared by human practitioners and LLMs, suggests that future progress may benefit from models adapted or calibrated to a target segmentation granularity, or from measurement practices that report granularity-robust quantities.

7 Limitations

The corpus is a single educational genre (8 weeks of discussions in a HCI class) and, under IRB, cannot be fully released, which limits external replication to the framework; we mitigate by reporting robustness across thresholds and reference-standard variants. Throughout the study, we evaluate batch detection on full transcripts. We note that online, turn-by-turn tracking would be a more challenging setting, and that real-time tracking affordances are important for educational applications to inform effective intervention. Criterion validity—linking CRS structure to an external indicator of collaboration quality—and the downstream utility evaluation of CRS-based analytics are left to future work.

Ethics Statement

The corpus consists of pseudonymized student discussions used under an IRB protocol permitting analysis and reporting. We report only aggregate metrics and no full transcript content, and we do not redistribute the data. Data collection was conducted with informed consent.

One methodological risk to flag is that the CRS framework is a partial representation of group dynamics. Applications that treat it as a full description of a group's dynamics risk flattening affective, relational, and non-propositional aspects of participation. We position CRS as one analytic layer, to be used with human oversight and with measures chosen for their demonstrated reliability (e.g., participation over fine-grained structure).

Acknowledgments

We thank the anonymous reviewers for their constructive feedback, which improved the analyses and solidified the framing of this paper. We also thank Bookyung Shin, Chenghong Lin, Rhea Kaw, and Prakeerth Prasad for their help with data collection, as well as the students who participated in the discussions analyzed in this study.

References

- Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational linguistics*, 34(4):555–596.
- Nicholas Asher and Alex Lascarides. 2003. *Logics of conversation*. Cambridge University Press.
- Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. 2005. The ami meeting corpus: A pre-announcement. In *International workshop on machine learning for multimodal interaction*, pages 28–39. Springer.
- Herbert H Clark and Susan E Brennan. 1991. Grounding in communication. *Perspectives on socially shared cognition*, pages 127–149.
- Wei Dai, Yi-Shan Tsai, Jionghao Lin, Ahmad Aldino, Hua Jin, Tongguang Li, Dragan Gašević, and Guanliang Chen. 2024. Assessing the proficiency of large language models in automatic feedback generation: An evaluation study. *Computers and Education: Artificial Intelligence*, 7:100299.
- Raghav Jain, Tulika Saha, Jhagrut Lalwani, and Sriparna Saha. 2023. Can you summarize my learnings? towards perspective-based educational dialogue summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3158–3173.
- Georgi Karadzhov, Tom Stafford, and Andreas Vlachos. 2023. Delidata: A dataset for deliberation in multi-party problem solving. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–25.
- Ibrahim Khalil Khebour, Kenneth Lai, Mariah Bradford, Yifan Zhu, Richard A Brutti, Christopher Tam, Jingxuan Tu, Benjamin A Ibarra, Nathaniel Blanchard, Nikhil Krishnaswamy, and James Pustejovsky. 2024. Common ground tracking in multimodal dialogue. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3587–3602.
- J Richard Landis and Gary G Koch. 1977. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, pages 363–374.
- William Mann and Sandra Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.
- Yann Mathet, Antoine Widlöcher, and Jean-Philippe Métyvier. 2015. The unified and holistic method gamma (γ) for inter-annotator agreement measure and alignment. *Computational Linguistics*, 41(3):437–479.
- Sarah Michaels, Catherine O'Connor, and Lauren B Resnick. 2008. Deliberative discourse idealized and realized: Accountable talk in the classroom and in civic life. *Studies in philosophy and education*, 27(4):283–297.
- Kamila Misiejuk, Sonsoles López-Pernas, Rogers Kalisa, and Mohammed Saqr. 2025. Mapping the landscape of generative artificial intelligence in learning analytics: A systematic literature review. *Journal of Learning Analytics*, 12(1):12–31.
- Prerna Ravi, Dong Won Lee, Beatriz Flámia, Jasmine David, Brandon Hanks, Cynthia Breazeal, Emma Anderson, and Grace Lin. 2026. Leveraging large language models to identify conversation threads for collaborative learning analysis. *Journal of Educational Data Mining*, 18(1):387–434.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3982–3992.
- Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Gerry Stahl. 2006. *Group cognition: Computer support for building collaborative knowledge (acting with technology)*. The MIT Press.

Zhi Rui Tam, Cheng-Kuang Wu, Yi-Lin Tsai, Chieh-Yen Lin, Hung-yi Lee, and Yun-Nung Chen. 2024. Let me speak freely? a study on the impact of format restrictions on large language model performance. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1218–1236.

Nhat Tran, Diane Litman, and Amanda Godley. 2025. Using large language models to analyze students’ collaborative argumentation in classroom discussions. In *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Full Papers*, pages 111–125.

Armin Weinberger and Frank Fischer. 2006. A framework to analyze argumentative knowledge construction in computer-supported collaborative learning. *Computers & education*, 46(1):71–95.

Jason D Williams, Antoine Raux, and Matthew Henderson. 2016. The dialog state tracking challenge series: A review. *Dialogue & Discourse*, 7(3):4–33.

A Model configurations

Table 5 lists the tested configuration of each model; GPT-5’s API accepts only its default temperature, so it ran at 1.0 while the other models ran at 0.1.

B Full per-condition results

Table 6 reports all four metrics for every model and prompting condition (macro over 152 discussions), with the denominators underlying relation-type accuracy; Table 7 breaks that accuracy down by relation type.

C Robustness: alignment threshold

Table 8 gives Contribution F1 and the labeling–identification gap (type accuracy – Contribution F1) at SBERT thresholds $\tau \in \{.6, .7, .75, .8\}$. Absolute F1 falls as τ rises (stricter matching), but the model ranking is identical at every threshold, and the gap is positive for every model at every τ , widening as matching tightens.

D Per-turn baseline

This baseline emits one contribution per speaker-turn (turn text as the span), scored against the same reference. Contribution F1 is 0.164. Type accuracy depends on the fixed type assigned: always-claim 0.42, always-proposal 0.40, always-issue 0.18, random-weighted 0.37.

Model	Identifier	Temp.	Max out	Reasoning	Schema
GPT-5	gpt-5-2025-08-07	1.0*	32,000	default effort (on)	enforced
Claude Opus-4.1	claude-opus-4-1-20250805	0.1	32,000	extended thinking off	enforced
Gemini-2.5-Pro	gemini-2.5-pro	0.1	32,000	thinking on (dynamic)	enforced
Llama-4-Maverick	Llama-4-Maverick-17B-128E-Instruct	0.1	32,000	none	prompt-guided
Qwen3-235B	Qwen3-235B-A22B-Instruct-2507	0.1	32,000	none (instruct variant)	prompt-guided

Table 5: Tested configurations. All calls requested temperature 0.1, except GPT-5, which supports only its default temperature (1.0). No seed or top- p was set; reasoning effort was left at each provider’s default.

Model		Contrib. F1	Type acc.	Rel. F1	Rel. type acc.	n	d
Gemini-2.5-Pro	transcript_only	.556	.919	.253	.916	220	88
	schema_only	.536	.914	.236	.911	214	85
	schema_few_shot	.540	.917	.266	.955	218	84
GPT-5	transcript_only	.523	.915	.239	.906	217	91
	schema_only	.558	.922	.225	.926	205	83
	schema_few_shot	.534	.936	.255	.935	196	86
Opus-4.1	transcript_only	.462	.906	.220	.895	126	65
	schema_only	.498	.930	.283	.916	167	78
	schema_few_shot	.458	.908	.229	.935	107	56
Qwen3-235B	transcript_only	.340	.458	.066	.579	24	19
	schema_only	.427	.855	.222	.823	106	58
	schema_few_shot	.407	.817	.200	.882	71	50
Llama-4-Maverick	transcript_only	.292	.417	.031	.500	13	10
	schema_only	.355	.841	.164	.888	65	43
	schema_few_shot	.330	.789	.165	.747	45	31

Table 6: Per-condition detection results (macro over 152 discussions). Relation-type accuracy is computed over n co-identified typed relations in d evaluable discussions; cells with small n carry correspondingly wide uncertainty. Type accuracy jumps from transcript_only to schema_only for open models (the schema effect on labeling); identification moves little across conditions.

ADDRESSES	SUPPORTS	CHALLENGES	REFINES	SYNTHESIZES
0.96 (497)	0.96 (583)	0.87 (162)	0.85 (696)	0.71 (56)

Table 7: Relation-type accuracy by gold type on co-identified relations, pooled over models and conditions (n in parentheses). SYNTHESIZES is the smallest sample and the only relation requiring ≥ 2 targets; the dominant confusion is REFINES→SUPPORTS (77 cases).

Model	Contribution F1				Gap			
	.6	.7	.75	.8	.6	.7	.75	.8
Gemini-2.5-Pro	.667	.544	.460	.362	+.25	+.37	+.46	+.56
GPT-5	.677	.538	.442	.332	+.24	+.39	+.49	+.60
Opus-4.1	.597	.473	.390	.296	+.31	+.44	+.52	+.63
Qwen3-235B	.533	.391	.297	.205	+.16	+.32	+.43	+.53
Llama-4-Maverick	.453	.326	.246	.173	+.22	+.36	+.43	+.51

Table 8: Contribution F1 and labeling–identification gap vs. alignment threshold τ . The ranking is stable and the gap positive at every τ .