

How Undergraduates Interact with a Generative AI Peer: Eight Participation Roles and the Conditions for Effective Collaboration in University Mathematics

Matias Rojas

matias.rojas@uga.edu

University of Georgia

Nicolás Guerra

Diego Portales University

Jonathan Frez

Diego Portales University

Brayan Díaz

Utah State University

Xiaoming Zhai

University of Georgia

Research Article

Keywords: Human–GenAI interaction, Generative AI in higher education, Collaborative problem solving, Participation typology, LLM peer, Automated assessment, Formative feedback, PISA 2015 CPS framework

Posted Date: September 1st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9851111/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

1 How Undergraduates Interact with a
2 Generative AI Peer: Eight Participation Roles
3 and the Conditions for Effective Collaboration
4 in University Mathematics

5 Author Name^{1*}

6 ¹[Affiliation withheld for double-blind review]

7 ^{*}[Corresponding author contact withheld for double-blind review]

8 *Submitted to the International Journal of Educational Technology in Higher Education*
9 *(IJETHE)*

10 ISSN: 2365-9440 | DOI prefix: 10.1186/s41239

11 Open Access — Creative Commons Attribution 4.0 International (CC BY 4.0)

Abstract

Generative AI is rapidly entering higher-education classrooms, yet remarkably little is known about how individual students actually interact with an AI peer during open-ended collaborative problem solving. This study examines the full range of human–GenAI interaction patterns that emerge when undergraduates collaborate with a large language model (LLM)-based peer to solve calculus optimization problems, and identifies the interaction conditions under which that collaboration is educationally productive.

Thirty students used a web-based platform that pairs each participant with an LLM peer through an information-splitting (jigsaw) task structure, making collaboration structurally necessary rather than optional. Full conversation and interaction logs were captured and coded using a multi-agent automated scoring pipeline grounded in the PISA 2015 Collaborative Problem Solving (CPS) matrix. A qualitative case analysis of eight purposively selected conversations then identified eight distinct participation roles: the *Co-Regulator*, the *Recalibrating Collaborator*, the *Low-Effort Guesser*, the *Delegating Director*, the *Misaligned Follower*, the *Low-Engagement Answer-Seeker*, the *Efficient Knowledge Seeker*, and the *Early Disengager*. These roles reveal that human–GenAI collaboration is far from monolithic: students who sustain a shared cognitive state with the AI peer, externalize their reasoning, and negotiate intermediate steps achieve substantially better outcomes ($OR \approx 10.83$) than those who delegate, disengage, or interact at a transactional level.

Quantitative trace-based clustering confirmed two interaction profiles: a higher-interaction cluster where 71% of students solved both tasks, and a lower-interaction cluster where only 19% did. The automated scorer showed promising validity evidence at broad process and competence levels (human–LLM $\kappa = .60$; human–human $\kappa = .80$). Students perceived the AI feedback as useful (86.7%) and the peer as easy to use (80.0%).

These findings provide a theoretically grounded typology of student–GenAI interaction in higher education and show that effective AI-peer collaboration requires deliberate design of both the task structure and the feedback loop. Implications for profile-sensitive AI-peer design and formative feedback are discussed.

Keywords: Human–GenAI interaction; Generative AI in higher education; Collaborative problem solving; Participation typology; LLM peer; Automated assessment; Formative feedback; PISA 2015 CPS framework

1 Introduction

Generative AI tools are entering higher-education classrooms faster than researchers can characterize how students actually use them. Yet the central question for learning technology is not whether students interact with AI, but *how* — and whether those interaction patterns constitute educationally productive collaboration. When an LLM functions as a peer rather than a tutor or a search engine, a new and largely unexplored set of questions emerges: Do students co-regulate with the AI, externalize their reasoning, and maintain a shared cognitive state? Or do they delegate, disengage, or engage in a transactional exchange that bypasses the collaborative process entirely? These questions are practically urgent for instructors deploying AI-peer tools in STEM courses, where the structure of collaboration — not just the presence of AI — determines whether students learn (Law et al., 2025; OECD, 2017). They also carry theoretical stakes: human–GenAI collaboration does not map neatly onto existing frameworks for human–human or human–scripted-agent interaction, because LLMs can sustain open-ended, contextually responsive dialogue in ways that previous systems could not (Oppenheimer et al., 2025; Yue et al., 2024). Understanding the range of interaction patterns that emerge, the mechanisms that distinguish productive from unproductive ones, and what task and feedback designs promote the former is therefore a central research priority for educational technology in higher education.

This paper addresses that priority by studying human–GenAI interaction in the context of collaborative calculus problem solving, an ecologically realistic and academically consequential activity. Collaborative Problem Solving (CPS) is increasingly recognized as a core 21st-century competency, spanning cognitive dimensions (planning, executing, monitoring, and adapting strategies) and social-regulatory dimensions (perspective-taking, participation, coordination, and co-regulation) (Chai et al., 2024; Hesse et al., 2015; Sun et al., 2020). In mathematics, these dimensions are particularly consequential: problem-solving requires not only executing procedures but also articulating reasoning, interpreting representations, and negotiating strategies with a partner (Felmer, 2023). At the same time, measuring how CPS actually unfolds during open-ended dialogue has remained methodologically difficult. Existing large-scale approaches trade authenticity for scalability by constraining interaction to scripted agent dialogue (as in PISA 2015), which may suppress key mechanisms of collaboration such as negotiation, sustained argumentation, and emergent coordination (Graesser et al., 2020; Herborn et al., 2020). Human–human settings preserve richer dynamics but require time-intensive hand-coding that is prohibitively costly at scale (Hao et al., 2025). LLM-based peers create a new option: they can sustain open-ended, naturalistic dialogue while also producing machine-readable interaction logs that support automated, process-based measurement (Forsyth et al., 2024; Hou et al., 2025). This combination makes it possible, for the first time, to study how students actually interact with an AI peer at a level of behavioral specificity

that goes beyond “used the chatbot” or “solved the task.”

The rapid development of large language models (LLMs) offers a promising approach to addressing this gap. Unlike earlier scripted agent systems, LLMs can generate coherent, contextually appropriate responses that sustain multi-turn dialogue, allowing interactions that more closely resemble collaborative reasoning and communication (Forsyth et al., 2024; Oppenheimer et al., 2025; Yue et al., 2024). This capability enables a new class of “human–AI collaboration” environments in which the AI does not act solely as a tutor or evaluator but as a peer collaborator that participates in joint problem-solving (Oppenheimer et al., 2025). Recent work in authentic collaborative settings further argues that generative AI should be understood not merely as a tool, but as part of hybrid human–AI knowledge-building systems whose effectiveness depends on communication, role structure, and the orchestration of joint activity (Law et al., 2025). Importantly, LLMs also create an opportunity on the measurement side: the same generative and interpretive capabilities that support natural dialogue can be leveraged to automate the coding of communication data, reducing dependence on manual annotation and opening the door to scalable, log-based CPS assessment in open-ended settings (Hao et al., 2025). In parallel, recent work has explored the use of automated planning and intelligent support to provide real-time scaffolding for regulatory skills in collaborative contexts (Author et al., 2022), suggesting that technology can both elicit and support CPS behaviors.

However, leveraging LLMs for CPS assessment raises key design and measurement questions. If an AI agent is both a participant in the collaboration and part of the assessment pipeline, the system must be designed to ensure that interactions are sufficiently authentic to elicit meaningful CPS behaviors, while remaining structured enough to support interpretation and measurement. One established approach to eliciting collaboration is to design tasks that enforce positive interdependence, for example, through information-splitting (jigsaw) structures in which partners must share complementary resources to succeed (Andrews-Todd & Forsyth, 2020; Author et al., 2021). When implemented in a peer dialogue interface, information-splitting can promote behaviors aligned with CPS indicators — such as resource sharing, role coordination, and building shared understanding — by making collaboration functionally necessary rather than optional.

In this paper, we present the development and evaluation of a web-based platform that enables undergraduate students to collaborate with an LLM peer to solve two calculus-based optimization problems. The LLM peer is prompt-engineered to play a role that requires interdependence and sustained dialogue in order to reach a correct solution. The platform is grounded in the PISA 2015 CPS framework, targeting a matrix of social and cognitive competencies (OECD, 2017). Beyond facilitating collaboration, the system captures detailed interaction logs and dialogue and provides automated CPS feedback. To enable scalable measurement in open-ended settings, we propose a PISA-grounded, message-level CPS scoring pipeline in which a three-agent mixture-of-experts classifies

each student turn along social and cognitive dimensions before synthesizing a final code and quality rating. These message-level codes are then aggregated into task-level and student-level CPS indices that support both research analysis and formative feedback.

To address these gaps, this study examines how an LLM-based peer environment can be designed and analyzed to elicit and assess collaborative problem solving in undergraduate mathematics. More specifically, we focus on a setting that combines information-splitting, open-ended peer dialogue, and process-based measurement in order to examine not only whether students solve the tasks, but also how collaboration unfolds throughout the interaction. The study therefore distinguishes among task-level CPS differences, student-level interaction profiles, and qualitative participation patterns, while also providing initial validation evidence for the automated scoring pipeline used to support these analyses.

Accordingly, we investigate the following research questions:

- **RQ1 (Perceived feasibility):** How do students perceive the usability and usefulness of collaborating with an LLM peer and receiving automated CPS feedback?
- **RQ2 (Task-level CPS differences and interaction profiles):** What task-level CPS differences and student-level interaction profiles emerge from the platform traces, and how are they related to achievement?
- **RQ3 (Qualitative participation patterns):** How do representative students enact collaboration with the LLM peer, and how do qualitative participation patterns help interpret the quantitative interaction profiles?

2 Review of the Literature

This review covers (i) CPS frameworks and the authenticity–scalability trade-off in assessment, (ii) trace- and process-based CPS measurement, (iii) LLMs as peers and the quality of human–GenAI interaction, and (iv) LLMs as automated coders of CPS evidence.

2.1 CPS frameworks, assessment, and the authenticity–scalability trade-off

CPS is defined as the capacity to engage effectively in a process in which two or more agents share understanding and effort to reach a solution (OECD, 2017). The PISA 2015 CPS framework operationalizes this as a skills matrix crossing three collaborative competencies (establish and maintain shared understanding (1); take appropriate action (2); establish and maintain team organization (3)) with four cognitive processes (explore and understand (A); represent and formulate (B); plan and execute (C); monitor and reflect (D)), yielding 12 cells that can be inferred from interaction evidence such as information sharing, coordination moves, and monitoring behavior (Graesser et al., 2018; OECD,

2017). Complementary frameworks (e.g., ATC21S) converge on the dual cognitive–social-regulatory structure of CPS (Chai et al., 2024; Griffin & Care, 2014; Hesse et al., 2015), motivating the use of dialogue traces as the primary evidence source.

A central measurement challenge is the authenticity–scalability trade-off. PISA 2015 addressed scalability through Human–Agent (H–A) interactions with simulated partners, but constrained agent designs (menu-based chat, scripted dialogue) suppress negotiation and emergent coordination, limiting ecological validity (Graesser et al., 2020; Herborn et al., 2020; Rosen, 2015). Human–Human (H–H) settings preserve richer discourse but require intensive coding and introduce partner-variability noise (Scoular & Care, 2020). Importantly, Author et al. (2021) showed that neither format is inherently superior: valid CPS scores depend on aligning task design, evidence opportunities, and coding logic — a principle that motivates the information-splitting and automated scoring approach used here.

2.2 Trace-based CPS measurement and LLMs as collaborative peers

Process-oriented measurement leverages digital traces (dialogue, clickstreams, action sequences) rather than final outcomes to infer CPS indicators (Chai et al., 2024; Mislevy et al., 2003). This is conceptually sound but practically difficult: mapping open-ended dialogue to CPS codes has historically required labor-intensive annotation (Andrews-Todd & Forsyth, 2020; Flor et al., 2016; Hao et al., 2017), an ongoing bottleneck that has forced assessments to constrain interaction formats at the expense of authenticity (Hao et al., 2025).

LLMs offer a way out of this bind on both sides. As *peers*, they can sustain contextually coherent, multi-turn dialogue that enables open-ended collaboration — acting as a partner rather than a tutor (Oppenheimer et al., 2025; Yue et al., 2024). Forsyth et al. (2024) showed that prompt-engineered LLM conversations are more naturalistic than scripted assessments, and Hou et al. (2025) demonstrated a multi-agent assessment architecture that jointly manages dialogue and evidence collection. As *coders*, the same generative capabilities can be turned to automated classification of interaction data with limited task-specific training, enabling an “LLM-as-a-judge” paradigm for CPS (Hao et al., 2025; Li et al., 2024). Design risks remain — overly direct assistance reduces productive struggle (Sætra, 2022), and hallucinations and scoring inconsistencies require validation (Chu et al., 2025; Emirtekin, 2025) — but the dual role of LLMs as both elicitors and coders of CPS evidence creates a new class of measurement-enabled learning environments that this paper exploits.

2.3 Human–GenAI interaction quality: patterns, risks, and productive conditions

The growing use of generative AI in higher education has prompted research into how students actually engage with AI-based tools, and what interaction patterns lead to productive versus unproductive outcomes. A consistent finding across contexts is that exposure to AI assistance does not uniformly improve learning or collaboration; rather, effects are highly contingent on how students position the AI and how much cognitive work they externalize during the interaction (Oppenheimer et al., 2025; Yue et al., 2024). Studies examining AI use in academic tasks have identified meaningful variability in interaction depth: some students engage in genuine back-and-forth dialogue that externalizes reasoning and invites contingent responses, while others interact transactionally — issuing requests and accepting outputs without engaging critically with the AI’s contributions (Chu et al., 2025; Sætra, 2022).

A central concern in this literature is *over-reliance* and *automation bias*: the tendency for students to accept AI-generated outputs uncritically, reducing the cognitive effort they invest in problem solving (Emirtekin, 2025). Over-reliance appears especially prevalent in tasks where the AI can plausibly produce a correct-looking answer without requiring the student to understand the underlying logic — a risk that is acute in STEM domains, where procedural outputs may look correct but obscure conceptual gaps (Chu et al., 2025). Complementary work shows that the degree of productive struggle preserved during AI-assisted tasks matters for learning: when AI tools are designed to absorb the full procedural burden, students may succeed on the immediate task while retaining shallow conceptual understanding (Sætra, 2022).

These concerns motivate deliberate task and interaction design. Information-splitting structures, in which students and AI hold complementary task-relevant information, are one approach to making delegation structurally costly: if the AI lacks the student’s information, the student cannot simply outsource reasoning without first contributing it (Andrews-Todd & Forsyth, 2020; Author et al., 2021). Similarly, Law et al. (2025) found in an interdisciplinary hackathon context that generative AI contributed most productively when integrated into shared communicative routines rather than used as a stand-alone answer engine. The implication for AI-peer design is that structural features of the interaction — role definitions, information asymmetry, and scaffolding style — shape whether students co-regulate with the AI or merely consume its outputs.

A related challenge concerns *the quality of interaction evidence* available to instructors. Current AI-use studies typically rely on outcome measures (task scores, survey ratings) or coarse behavioral indicators (session duration, number of queries). These are informative but insufficient for understanding the mechanism: whether a student succeeded *with* the AI through genuine co-construction, or *around* it through strategic minimal engagement

(Baker et al., 2004; Muldner et al., 2011). Process-based measurement that codes the type, content, and sequence of student contributions during AI-assisted collaboration can reveal these distinctions and provide the basis for profile-sensitive instructional support. The present study contributes to this research gap by providing a behaviorally rich typology of human–GenAI interaction roles grounded in a validated CPS framework.

2.4 LLM-enabled automated coding and assessment of CPS traces

Beyond serving as collaborators, LLMs also have potential as automated coders of interaction traces — an especially consequential possibility for CPS assessment, where open-ended dialogue is a primary source of evidence. Prior trace-based approaches often relied on supervised models trained on extensive labeled data or on feature-engineered representations, which can miss semantic nuance in social-regulatory language (Flor et al., 2016; Hao et al., 2017). Recent work argues that LLMs can classify complex behaviors from text with limited task-specific training, enabling an “LLM-as-a-judge” paradigm in which models evaluate interaction segments according to explicit criteria (Li et al., 2024). A key claim in this line of work is that, given a well-defined coding framework, LLMs can code CPS-related communication with performance comparable to that of human raters, thereby reducing the scalability barrier that has historically constrained assessment design (Hao et al., 2025).

However, LLM-driven scoring raises new questions about validity and reliability. Although LLMs can be efficient, they may exhibit biases, sensitivity to prompting, and inconsistency across runs, which motivates careful validation and transparent reporting practices (Emirtekin, 2025). In addition, when LLMs are embedded as both participants (peers) and assessors, researchers must consider how system design shapes the evidence being scored and how scoring rules relate to the intended construct (Mislevy et al., 2003). These issues are particularly important for CPS, where the construct involves both cognitive and social dimensions and where interaction patterns can be influenced by agent behavior.

A recent methodological direction is to decompose complex coding decisions into multiple smaller decisions rather than relying on a single monolithic rater. In CPS contexts, this decomposition is theoretically sensible because the PISA matrix itself crosses social competencies with cognitive processes. A mixture-of-experts approach can therefore ask one model to judge the collaborative dimension, another to judge the cognitive-process dimension, and a third to synthesize them into a final CPS code. This design may improve transparency, reduce the decision space for each agent, and make disagreement or ambiguity more diagnosable than in a single-pass classifier.

Finally, trace-based assessment is increasingly positioned not only for summative scoring but also for formative scaffolding. When a system can infer collaboration states from traces, it can potentially intervene to support regulation and coordination (Author et al.,

2022, 2023). This motivates “closed-loop” measurement-enabled learning environments in which the collaborative agent elicits evidence, the scoring mechanism interprets it, and the system provides targeted feedback to learners. The present study builds on this trajectory by combining (i) an interaction design that enforces interdependence and elicits CPS-relevant evidence (Andrews-Todd & Forsyth, 2020; Author et al., 2021) with (ii) a multi-agent automated scoring approach grounded in the PISA CPS framework (OECD, 2017), and (iii) a mixed-methods feasibility study that examines scorer validation, task-level CPS differences, interaction profiles, and qualitative participation patterns.

3 System and Task Design

This section describes the platform as a measurement-enabled learning system: it (i) orchestrates an LLM-mediated peer collaboration experience, (ii) captures fine-grained interaction traces, and (iii) produces automated CPS scoring outputs and feedback aligned to the PISA 2015 CPS framework. We focus on the elements essential to interpreting the method and the study results, rather than implementation-level engineering details.

3.1 Design goals

The system was designed to address three requirements that are typically difficult to satisfy simultaneously in CPS assessment:

1. **Authentic collaboration:** enable open-ended dialogue that supports explanation, negotiation, and coordination.
2. **Standardized partner behavior:** keep the “peer” experience sufficiently consistent across learners so that observed differences reflect learners’ interaction patterns rather than uncontrolled partner effects.
3. **Trace-based measurement:** capture interaction evidence (dialogue + actions) at sufficient granularity to support automated, indicator-aligned CPS scoring and feedback.

To meet these requirements, the platform combines an LLM-based peer collaborator with an automated CPS analysis pipeline that operates on the conversation and interaction logs after (and, optionally, during) task completion.

3.2 System overview

At a high level, the platform is organized as a multi-agent measurement pipeline that connects the student’s problem-solving experience to automated CPS evidence and feedback. The core idea is a *separation of roles*: an LLM-based peer participates in the collaboration, while distinct agents/services orchestrate, log, analyze, and report interaction evidence.

Two agents shape the collaborative interaction: the **Responding Agent** participates as a Calculus I peer, alternating responsibility with the student to propose deductions or calculations while avoiding complete single-turn solutions; and the **Provocateur Agent**, which issues brief, empathetic prompts (e.g., “What does your part of the data tell you?”) to elicit information sharing, shared understanding, and monitoring moves without providing mathematical content. A background safety filter screens incoming messages, and a Verifying Agent monitors evolving consensus on the mathematical solution state. All dialogue and key workspace actions are time-stamped and logged for downstream CPS scoring. Figure 3 summarizes the end-to-end flow.

3.3 Data captured for measurement

The platform captures two complementary categories of traces:

1. **Dialogue traces:** the full sequence of student and LLM peer messages, preserving conversational context (e.g., requests for clarification, coordination moves, agreement/disagreement, monitoring statements).
2. **Interaction/action traces:** time-stamped events that describe learners’ engagement with the workspace (e.g., task navigation, reading/consulting resources, submissions, and interactions with embedded task artifacts such as dynamic graphs or GeoGebra applets when instrumented). These traces support process-based analyses (e.g., interaction profiles) and provide context for interpreting dialogue (e.g., whether a student consulted a resource before proposing a plan).

Together, these traces enable two downstream measurement functions used in this paper: (i) profile discovery from log features (e.g., activity levels, messaging patterns, time-on-task), and (ii) automated CPS scoring aligned to PISA indicators using the dialogue as the primary evidence source.

3.4 Task design: information-splitting to enforce interdependence

The platform includes two calculus-based optimization tasks designed to elicit CPS behaviors through information splitting (jigsaw), a structure that distributes complementary resources among collaborators, requiring mutual exchange and coordination for success.

Across both tasks, the design principle is the same:

- Each student and the LLM peer has access to different pieces of task-relevant information (e.g., partial constraints, contextual facts, intermediate results, or hints).
- Neither party can reliably solve the problem on its own.
- Progress requires CPS-relevant behaviors such as sharing information, establishing shared understanding, coordinating a plan, and monitoring/repairing misunderstandings.

Task 1: Multi-zone optimization In the first task, students solve an optimization problem that involves multiple regions or conditions (e.g., distinct “zones”) that must be coordinated to identify a global optimum. Information splitting is implemented by providing the student with constraints and representations, while the LLM peer has access to complementary information needed to formulate and verify the full optimization setup. The task encourages:

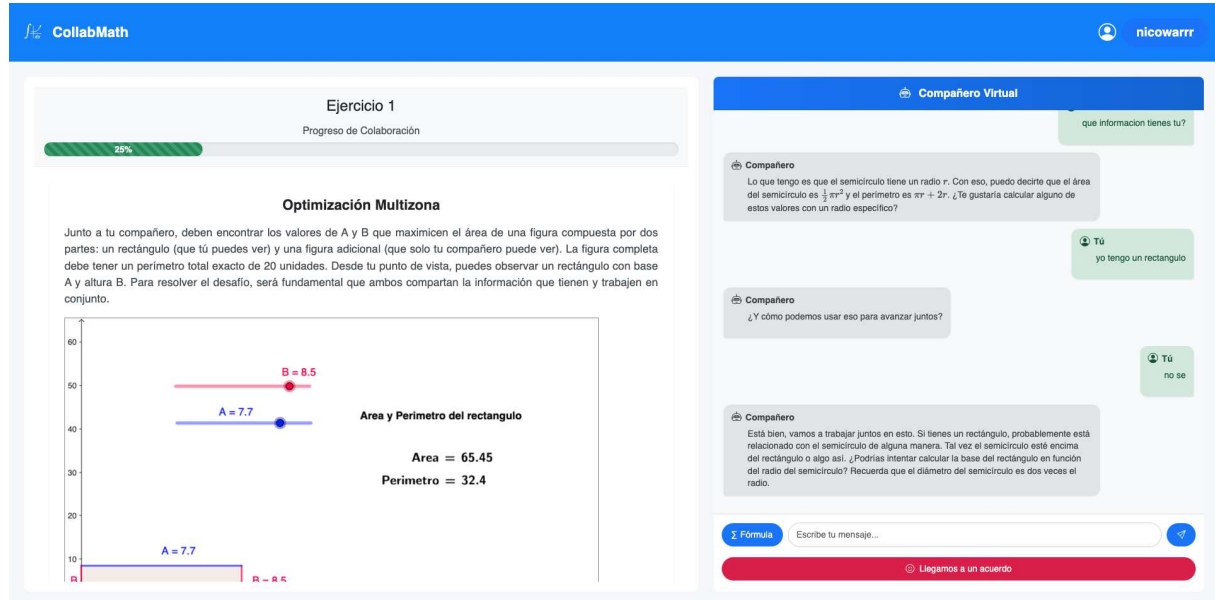


Figure 1: Multi-agent architecture and data flow between the student workspace, the LLM-based peer, and the automated CPS analyzer.

- **Representing and formulating** jointly (agreeing on variables/constraints)
- **Planning and executing** collaboratively (assigning subtasks, checking derivatives/conditions)
- **Monitoring and reflecting** (verifying boundary cases and correcting errors)

Task 2: Optimization with an unknown function In the second task, students work with a scenario in which a key function is not explicitly provided upfront, requiring inference, interpretation of representations, or coordination across partial information. As in Task 1, the student and the peer hold complementary information resources. The task is designed to elicit:

- **Exploring and understanding** (clarifying what is known vs unknown)
- **Discovering perspectives and resources** (requesting missing information)
- **Maintaining team organization** (tracking assumptions, aligning next steps)

By using two tasks with the same underlying interaction logic (interdependence + open-ended dialogue) but different mathematical structures, the platform aims to elicit

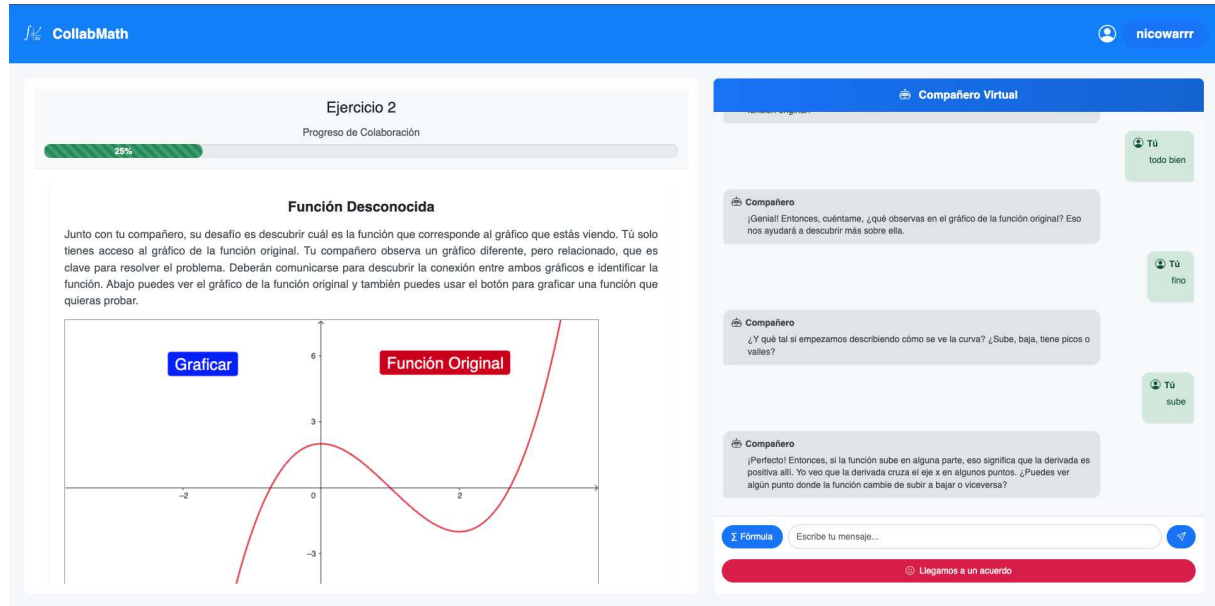


Figure 2: Multi-agent architecture and data flow between the student workspace, the LLM-based peer, and the automated CPS analyzer.

324 CPS evidence across multiple problem contexts while keeping measurement aligned to a
 325 common indicator framework.

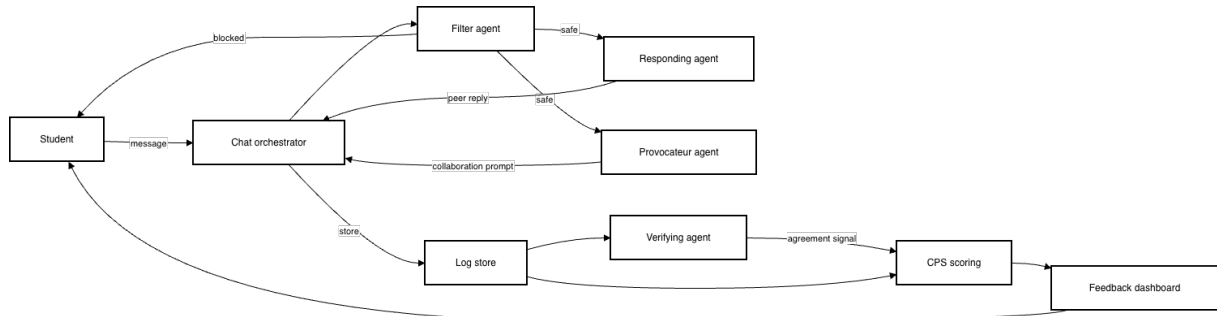


Figure 3: Functional overview of the multi-agent peer and measurement pipeline used in this study (orchestration, safety gating, peer interaction, trace capture, verification, CPS scoring, and feedback reporting).

326 4 Automated CPS Scoring Method

327 The pipeline converts open-ended student–LLM dialogue into PISA-grounded CPS evi-
 328 dence. The PISA 2015 matrix (four cognitive processes $\times$ three collaborative competen-
 329 cies) is operationalized as a closed set of 12 message-level codes (A1–D3), where the letter
 330 denotes the cognitive process and the number the competence. The full codebook with
 331 examples and tie-breaking rules is in Appendix A.

4.1 Three-agent scoring pipeline and aggregation

Three LLM-based agents operate sequentially on each student message: (1) a **Social expert** assigns the dominant collaborative competence; (2) a **Cognitive expert** assigns the dominant cognitive process; (3) a **Judge/synthesizer** fuses the two into a final code (A1–D3) and assigns an ordinal quality score (1 = weak evidence, 2 = clear evidence, 3 = strong evidence). All agents receive the full cumulative prior dialogue so that short or ambiguous turns (e.g., *yes*, *ok*, formula fragments) can be resolved in context.

The pipeline produces three nested output levels: **message-level** (code, quality, agent rationales for auditability), **task-level** (counts, shares, and integrated code indices per task), and **student-level** (global and subdimensional CPS summaries across both tasks). For each code c and task t , an integrated index combines frequency share and normalized quality with equal weighting (0.5/0.5); execution-heavy codes receive 0.4/0.6 weighting to account for opportunity dependency. The primary global CPS index is the mean of the 12 code-level integrated indices. Full index formulas are provided in the online supplementary materials.

5 Method

This study reports a pilot, mixed-methods, within-subjects evaluation of an LLM-mediated peer collaboration environment for two Calculus I optimization tasks.

5.1 Participants and setting

Participants were undergraduate students at a private Chilean university enrolled in or having completed Calculus I, recruited via institutional LMS and email. Of 59 registrants, 30 completed both tasks and the post-activity questionnaire and were included in the analyses.

5.2 Procedure

The study protocol was ethically approved; informed consent was obtained before registration. In a single online session, participants completed two calculus optimization tasks (displayed problem statement, open-ended chat collaboration, final submission), then a post-activity questionnaire. Information splitting ensured that the student and the LLM peer held complementary data, making dialogue necessary. All dialogue and workspace actions were time-stamped and logged; automated CPS scoring was applied post-session as described in Section 4.

5.3 Measures

All questionnaire items used a 5-point Likert agreement scale. Perceptual measures included: perception of AI-generated feedback (7 items), perception of the virtual agent (14 items), collaboration self-efficacy (9 items), and attitudes toward collaboration (12 items). All four scales showed adequate internal consistency and unidimensional structure in this pilot sample ($\alpha > .95$ for all scales; EFA details available upon request). Task success was coded 0/1 per task and as a total success sum (0–2). Interaction trace features (total actions, messages, average message length, time-on-task) were aggregated across both tasks. Automated CPS metrics were derived as described in Section 4. A gold-standard subset of 76 student messages from four conversations was independently coded by two human raters for scorer validation.

Task success. For each task, success was coded as 0/1 based on whether the participant reached the correct solution. We additionally computed a total success indicator reflecting whether the participant solved both tasks correctly, as well as a total success sum (0–2).

Interaction trace features. From platform logs, we derived behavioral indicators per task and in aggregate, including: (a) total number of recorded workspace actions, (b) number of student chat messages, (c) average student message length, and (d) time-on-task.

Automated CPS metrics. Using the scoring pipeline described in Section 4, we computed message-level CPS codes and quality scores for student turns, task-level counts/shares and integrated CPS indices, and student-level CPS summaries across the two tasks. In the quantitative analyses we use (i) a primary global CPS index, (ii) process-level and competence-level shares, and (iii) selected auxiliary indicators such as overall normalized quality and code-distribution entropy.

Human-coded benchmark subset. To evaluate the scorer, we used a gold-standard subset of four conversations comprising 76 student messages. Two human raters independently coded these messages with the same message-level codebook and resolved disagreements through consensus. This subset was used to compute human–human and human–LLM agreement.

5.4 Method of analysis

Quantitative analyses used Python (pandas, scipy, scikit-learn, statsmodels), $\alpha = .05$ two-sided, and nonparametric tests given the modest sample size. The analysis proceeded in four stages. **Stage 1** (Section 5.5) established scorer validation against the human-coded benchmark. **Stage 2** compared Task 1 and Task 2 using Wilcoxon signed-rank and McNemar tests. **Stage 3** identified interaction profiles by k -means clustering on z-standardized trace features (actions, messages, message length, time-on-task), evaluated profile quality

with silhouette, Calinski–Harabasz, and Davies–Bouldin indices, then related profiles to achievement using Fisher’s exact test and nested logistic models (self-efficacy baseline; plus dialogue volume; plus global CPS; plus CPS subdimensions). Linear models examined cluster membership as a predictor of the primary CPS index and total achievement sum. **Stage 4** (qualitative) conducted thematic memoing on eight purposively sampled conversations, selected to preserve variation in achievement and CPS composition (Creswell & Plano Clark, 2017), focusing on regulatory quality, scaffolding uptake, and reasoning externalization.

5.5 Scorer validation

We examined whether the multi-agent scorer produced CPS labels that aligned reasonably with human judgment before using it as the primary analytic instrument for the study’s substantive questions. Two expert raters independently coded a stratified benchmark subset of 76 student messages drawn from four conversations (spanning both tasks, both interaction clusters, and multiple achievement outcomes). Table 1 summarizes agreement at three levels of granularity.

Table 1: Agreement for the gold-standard benchmark subset at three levels of granularity.

Comparison level	n	Accuracy	Cohen’s κ	Macro-F1
Human–human (12 codes)	76	0.829	0.801	0.789
Human–human (process level)	76	0.921	0.880	0.872
Human–human (competence level)	76	0.895	0.841	0.897
Human–LLM (12 codes)	76	0.658	0.599	0.491
Human–LLM (process level)	76	0.737	0.606	0.710
Human–LLM (competence level)	76	0.855	0.781	0.858

Note. The table reports agreement for the full 12-code CPS labeling scheme and for two broader summaries: the cognitive-process level (explore and understand (A)–monitor and reflect (D)) and the collaborative-competence level (establish and maintain shared understanding (1)–establish and maintain team organization (3)).

Human–human agreement was substantial to near-perfect across all three levels, confirming the benchmark itself as a reliable gold standard. The automated scorer performed best at broader construct levels: agreement was moderate at the full 12-code level ($\kappa = .60$) and substantially stronger at the collaborative-competence level ($\kappa = .78$), indicating that the pipeline more reliably captures *which kind* of collaborative contribution is being made than every fine-grained code boundary.

Figure 4 illustrates that most disagreement occurred between adjacent competence categories. We therefore use the scorer in the remainder of the paper as a process-sensitive analytic instrument at the process and competence levels, while treating the present evidence as initial rather than definitive.

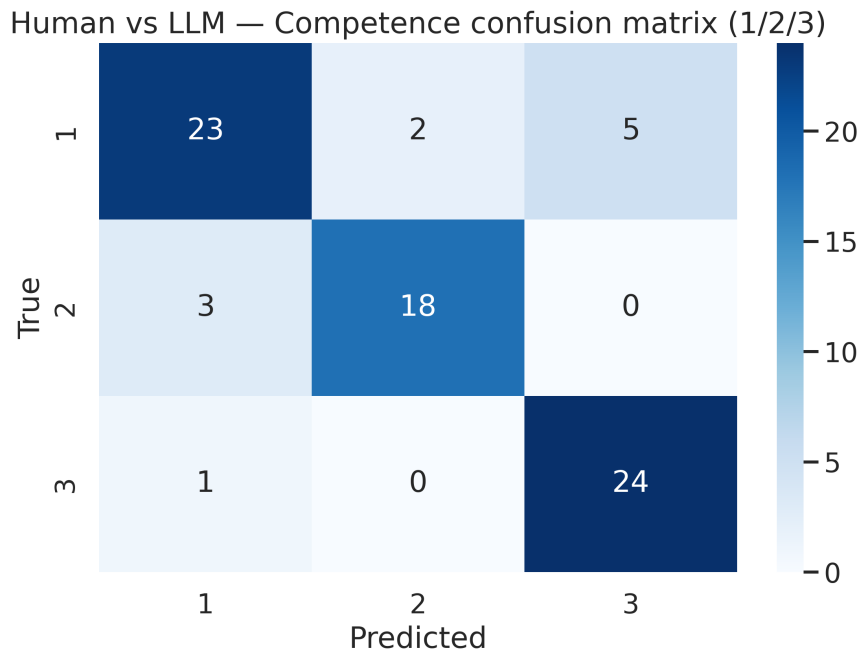


Figure 4: Confusion matrix for the collaborative-competence level in the human–LLM validation subset.

6 Results

Results address the three substantive research questions in sequence. Unless otherwise noted, analyses use the final sample of participants who completed both tasks and the post-activity measures ($n = 30$). Total achievement is coded as 1 when a participant solved *both* tasks correctly and 0 otherwise.

6.1 RQ1: Perceived feasibility

Participants reported positive perceptions of both the virtual peer and the feedback. In the post-activity questionnaire, 86.7% of participants agreed that the automated feedback was useful, and 80.0% agreed that the virtual peer was easy to use. Table 2 summarizes descriptive statistics for the main perceptual measures by total-achievement group. Overall ratings were positive across groups, with only modest mean differences. Students did not merely “tolerate” the system: mean ratings for both the agent and the feedback were above 4 on a 5-point scale, which indicates generally favorable reactions even in this early pilot. The scale checks reported in Section 5.3 also suggest that these questionnaire scores behaved coherently enough to be treated as interpretable summaries of perceived feedback quality, agent quality, collaboration self-efficacy, and collaboration attitudes.

Table 3 contextualizes these perceptions with interaction traces and the new primary CPS index. Participants who solved both tasks showed substantially more engagement on average (more actions and messages, longer messages, and more time on task) and also higher primary CPS scores. Students who solved both tasks did not just finish with more

Table 2: Descriptive statistics for perceptual measures by total achievement (1 = solved both tasks correctly).

Measure	Achievement = 0 ($n = 17$)		Achievement = 1 ($n = 13$)	
	M	SD	M	SD
Calculation self-efficacy	3.235	0.752	3.000	1.000
Feedback perception	4.068	1.083	4.033	1.009
Agent perception	4.029	1.045	4.205	1.023
Collaboration self-efficacy	4.151	0.985	4.085	1.042
Attitude toward collaboration	3.876	0.898	4.214	1.005

Note. Perceptual variables are 5-point post-activity ratings. Higher values indicate more positive perceptions, stronger self-efficacy, or more favorable attitudes.

correct answers; they also tended to stay longer, write more, and make their reasoning more visible in the interaction logs. In this sense, the descriptive results already point toward a process difference, not only an outcome difference.

Table 3: Descriptive statistics for trace-based interaction measures and the primary CPS index by total achievement.

Measure	Achievement = 0 ($n = 17$)		Achievement = 1 ($n = 13$)	
	M	SD	M	SD
Total actions (both tasks)	47.294	40.444	73.538	37.629
Total messages (both tasks)	15.882	11.554	24.846	13.843
Avg. message length (both tasks)	22.828	13.131	36.592	22.070
Total time-on-task (min; both tasks)	17.739	15.454	26.621	15.570
Primary automated CPS index	17.224	5.754	21.339	8.061

Note. Trace variables aggregate student behavior across the two tasks. The primary automated CPS index is the student-level integrated CPS summary derived from the multi-agent coding pipeline.

6.2 Task-level CPS patterns

The two tasks elicited distinct task-level CPS profiles. Task 2 yielded higher success than Task 1 (McNemar $p = .035$; Wilcoxon $p = .022$), while the two tasks differed sharply in interaction and CPS composition. The two tasks did not behave like interchangeable items from the same pool: they pulled for different kinds of collaboration and differed in difficulty. Table 4 reports selected within-participant contrasts.

Task 1 elicited more messages, actions, and execution-oriented moves associated with plan and execute (C) and take appropriate action (2), whereas Task 2 elicited more exploratory behavior associated with explore and understand (A) and more shared-understanding behavior associated with establish and maintain shared understanding (1), and it yielded higher success. Task 1 looked more like a heavy “working-through-the-math” task, whereas Task 2 elicited fewer turns but proportionally more exploration and shared-understanding moves and was easier to complete.

Table 4: Selected within-participant differences between Task 1 and Task 2 (Wilcoxon signed-rank tests). Positive mean differences indicate higher values in Task 2; negative values indicate higher values in Task 1.

Measure	Median E1	Median E2	Mean diff (E2–E1)	<i>p</i>
Actions	47.50	1.00	-56.47	< .001
Messages	12.00	2.00	-7.70	< .001
plan and execute (C) share	51.19	0.00	-27.94	< .001
explore and understand (A) share	21.76	50.00	28.64	< .001
Primary task-level CPS index	15.63	9.72	-4.16	.001
take appropriate action (2) share	15.00	0.00	-14.07	.010
establish and maintain shared understanding (1) share	28.99	50.00	15.40	.037
Success	1.00	1.00	0.30	.022

Note. Task-level CPS variables summarize the distribution of message-level codes within each task. Shares are percentages of student turns coded in each process or competence category.

6.3 RQ2: Interaction profiles and achievement

This section reports two related but distinct quantitative results. First, we examine *task-level CPS differences*, that is, how the two exercises elicited different distributions of coded collaborative and cognitive behavior. Second, we identify *student-level interaction profiles* from the aggregated platform traces and examine how those profiles relate to achievement.

To identify interpretable student-level interaction profiles, we clustered participants using four standardized interaction-trace features aggregated across the two tasks: total actions, total messages, average message length, and total time on task. The two-cluster solution was clearly the most stable and interpretable (average silhouette = 0.536; Calinski–Harabasz = 46.10; Davies–Bouldin = 0.696), and we therefore retained it as the primary quantitative profile solution.

Table 5 shows the resulting profile means. Cluster 1 reflected a *lower-interaction* profile; Cluster 2 reflected a *higher-interaction* profile characterized by substantially more sustained dialogue, longer messages, longer time on task, and higher primary CPS scores.

Figure 5 provides a two-dimensional projection of this retained solution. Although the figure is only a visualization of the higher-dimensional clustering space, it helps show that the retained profiles are not arbitrary partitions of nearly identical students; rather, they correspond to a clear lower-interaction/higher-interaction contrast across the standardized trace variables.

The association between interaction profile and total achievement was statistically significant ($\chi^2(1) = 6.43$, $p = .011$; Fisher’s exact $p = .009$). Students in the higher-interaction profile were substantially more likely to solve both tasks (10/14) than students in the lower-interaction profile (3/16), yielding an odds ratio of approximately 10.83 (95% CI: [1.7, 108.3]). Students in the higher-interaction profile solved both tasks at a rate of

Table 5: Interaction profile means for the retained two-cluster solution.

Measure	Low interaction ($n = 16$)	High interaction ($n = 14$)
Total actions	25.63	96.43
Total messages	10.19	30.71
Avg. message length	17.69	41.48
Total time-on-task (min)	9.69	35.19
Primary automated CPS index	14.30	24.39
take appropriate action (2) share	10.12	25.85
establish and maintain team organization (3) share	53.57	38.72
explore and understand (A) share	43.27	24.11
monitor and reflect (D) share	3.39	16.15
Total achievement (0/1)	0.19	0.71
Total achievement sum (0–2)	1.06	1.71

Note. Interaction profiles were derived from standardized trace variables (actions, messages, average message length, and time-on-task). CPS rows report profile-level averages of the student-level summaries.

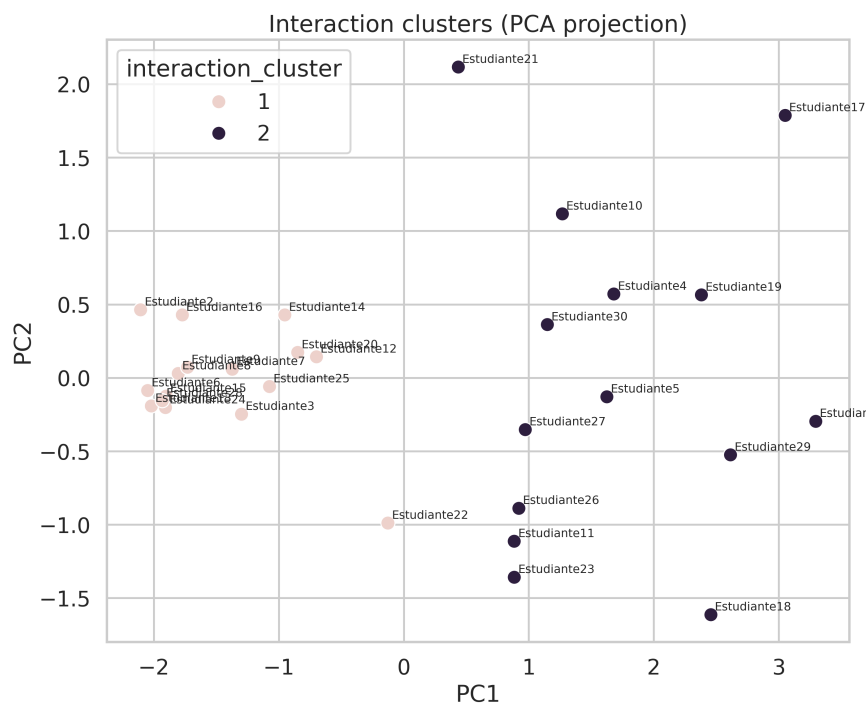


Figure 5: Principal-component projection of the retained two-cluster interaction solution.

about 71%, compared with only 19% in the lower-interaction profile. In practical terms, this was not a marginal difference: students who sustained richer interaction with the agent were much more likely to complete the full two-task sequence successfully. The odds ratio of 10.83 means that, within this pilot sample, the odds of fully solving the two-task sequence were about eleven times higher for students in the higher-interaction profile than for students in the lower-interaction profile. The wide confidence interval reflects the small sample size and should be interpreted accordingly; replication with a larger sample is needed before drawing definitive conclusions about the magnitude of the effect.

We then asked what the new CPS metrics add beyond simple dialogue volume. A sequence of nested logistic models showed that adding the primary global CPS index did *not* materially improve the baseline-plus-volume model, but adding selected CPS subdimensions did. Table 6 summarizes the model comparison.

Table 6: Nested logistic models predicting total achievement.

Model	Predictors	AIC	McFadden R^2	Accuracy	AUC
M1	Self-efficacy only	44.48	0.014	0.60	0.568
M2	Self-efficacy + message volume	43.23	0.093	0.70	0.686
M3	M2 + primary global CPS index	45.14	0.095	0.70	0.710
M4	M2 + take appropriate action (2) share + monitor and reflect (D) share	44.31	0.164	0.80	0.765

Note. M2 uses message volume as the aggregate dialogue-volume control. M4 shows that selected CPS subdimensions improved classification more than the primary global CPS index in this pilot sample.

The coefficients in the best-performing model (M4) suggested that the share of take appropriate action (2) moves was more promising than the global CPS index as an explanatory variable, even though the coefficient did not reach conventional significance in this pilot sample ($B = 0.72$, $p = .100$). The overall pattern is theoretically important: the global index captures meaningful variation, but some of its predictive value overlaps heavily with interaction volume. More interpretable subdimensions may therefore be more informative than the total score alone. In practical terms, the best model was the one that combined simple message volume with a small number of theoretically meaningful CPS subdimensions. Its 0.80 accuracy means that it classified about 8 out of 10 students correctly in this sample, and its AUC of 0.765 indicates a noticeably better-than-chance ability to separate students who solved both tasks from those who did not.

This interpretation is reinforced by the linear models. Interaction-cluster membership

explained 53.0% of the variance in the primary CPS index ($B = 10.09$, $p < .001$). When total achievement sum was modeled as a function of interaction-cluster membership and calculus self-efficacy, the cluster effect remained significant ($B = 0.653$, $p = .004$), while self-efficacy did not. Together, these analyses suggest that the higher-interaction profile remains the strongest quantitative organizing pattern in the dataset, while the new CPS metrics add explanatory depth primarily through selected subdimensions rather than through a single global score. Substantively, this means that profile membership marked a large gap in observed collaborative performance: the primary CPS index was roughly 10 points higher on average in the higher-interaction profile, and cluster membership alone accounted for more than half of the between-student variation in that index. Likewise, once profile membership was known, calculus self-efficacy added little explanatory value for total achievement sum in this pilot.

We also inspected exploratory CPS-only cluster solutions in parallel with the retained two-cluster interaction solution. Although a five-cluster CPS solution slightly improved some internal fit indices, it produced smaller and less stable groups that were difficult to interpret substantively in this pilot sample. We therefore did not retain it as a primary quantitative result. Instead, for qualitative sampling and interpretation only, we used an exploratory four-cluster CPS solution because it offered a more interpretable theory-guided spread of cases without over-fragmenting the sample. Accordingly, the two-cluster interaction solution remained the primary inferential clustering result, whereas the four-cluster CPS solution served only as an interpretive aid for qualitative case selection.

6.4 RQ3: Qualitative participation patterns in eight conversations

To complement the quantitative analyses, we conducted a thematic case analysis of eight representative conversations. The cases were selected through cluster-informed purposive sampling anchored in the retained two-cluster interaction solution from Section 6.3, while preserving maximum variation in total achievement (solved, partial, failed) and in the emerging CPS composition patterns. This qualitative sample was therefore explanatory rather than representative: it was designed to clarify the mechanisms behind the quantitative interaction profiles with information-rich cases, not to estimate new frequencies. The qualitative analysis did not replace the PISA-grounded coding; rather, it interpreted how those coded behaviors were enacted in practice, focusing on power relations with the AI, responses to Socratic scaffolding, mathematical autonomy, and the visibility of students' reasoning. Accordingly, the *interaction profiles* reported in Section 6.3 refer to quantitative student-level groupings derived from platform traces, whereas the *qualitative participation patterns* developed here refer to recurrent case-based ways in which students enacted collaboration with the LLM peer. In other words, the cases were used to explain the quantitative results in Sections 6.2–6.3, not to stand apart from them. All quoted excerpts in this section are translated from Spanish for readability.

Table 7 makes this bridge explicit by showing, for each case, its interaction-cluster membership and total achievement alongside the qualitative macro-/micro-profile labels.

Table 7: Representative qualitative cases used to interpret the quantitative profiles.

Student	Interaction profile	Achievement	Macro-family	Micro-profile	Interpretive summary
E4	High (27 msgs)	Solved (2.0)	High-engagement collaborative regulator	The Co-Regulator	Highly engaged, mathematically productive; actively maintains shared understanding and coordination.
E5	High (22 msgs)	Solved (2.0)	Engaged but initially misaligned	The Recalibrating Collaborator	Starts active but misaligned, then adapts and uses the agent as a complementary partner.
E11	High (31 msgs)	Solved (2.0)	Low-effort minimally elaborative	The Low-Effort Guesser	Stays active but avoids deeper algebraic restructuring; relies on arithmetic trial-and-error.
E18	High (50 msgs)	Solved (2.0)	High-engagement delegation-oriented	The Delegating Director	Highly engaged while increasingly delegating procedural work to the agent, retaining oversight.
E25	Low (12 msgs)	Partial (1.0)	High-dependence misaligned	The Misaligned Follower	Depends heavily on the agent yet fails to build a correct shared representation.
E2	Low (4 msgs)	Solved (2.0)	Low-engagement answer-dependent	The Low-Engagement Answer-Seeker	Minimally elaborative; repeatedly seeks direct answers without engaging the scaffolding prompts.
E6	Low (6 msgs)	Failed (0.0)	Self-reliant low-externalization	The Efficient Knowledge Seeker	Uses the agent for targeted knowledge but externalizes too little reasoning for reciprocal collaboration.
E28	Low (6 msgs)	Solved (2.0)	Low-effort non-collaborative	The Early Disengager	Little sustained effort or collaboration; quickly abandons meaningful uptake of the scaffolding.

Note. Interaction profile reports cluster membership (Section 6.3); message count links qualitative cases to aggregate traces. Achievement: 2.0 = solved both tasks; 1.0 = partial; 0.0 = failed both. Macro-families summarize broad collaboration orientations; micro-profiles name the specific interaction role.

Rather than forming a simple progression from “weak” to “strong” collaboration, the cases revealed three cross-case tensions that qualify the quantitative results.

First, successful task completion did not guarantee collaborative engagement with the AI. E2 and E28 both belonged to the lower-interaction profile yet still solved both tasks. In E2, the student repeatedly asked for direct guidance — “*Guide me*” and later “*What is the answer?*” — with little reciprocal uptake of the agent’s prompts. In E28, the interaction moved quickly from “*Give me more information*” to “*I have no idea*” and then to minimal follow-up. These cases suggest that successful outcomes can occur even when relatively little CPS becomes visible in the dialogue, whether because part of the mathematical work happens off-chat or because the task can be completed with very limited interaction. If we looked only at the final score, both students would look equally successful, yet their dialogues show almost no sustained collaborative work with the agent. The qualitative evidence therefore helps explain why achievement cannot be treated as a direct substitute for visible CPS.

Second, high interaction was not equivalent to high-quality collaboration. E4 and E11 both belonged to the higher-interaction profile and both solved the tasks, but they did so very differently. E4 repeatedly externalized reasoning and helped keep the shared problem state aligned through utterances such as “*I can see a rectangle with base A and height B*” and later “*I was wrong, there are two critical points*”. E11, by contrast, remained in the activity while relying on low-effort arithmetic trial-and-error, for example “*I propose a height of 2.5 and a base of 2.5*” and “*You give me the values*”. The qualitative memos therefore clarify that the volume-based interaction cluster captures an important participation difference, but not the semantic or regulatory quality of that participation.

Third, the cases point to concrete design opportunities for future LLM peers. E18

remained highly engaged but increasingly offloaded execution through imperative turns such as “*You do it*” and “*Check the calculations yourself*”, whereas E25 and E28 illustrate how the interaction can drift into misalignment or disengagement when common ground is weak. These contrasts suggest that future agents could benefit from lightweight real-time profiling: over-delegation might trigger prompts that require the student to externalize the next formula or reasoning step, while early disengagement or persistent misalignment might trigger motivational or common-ground-repair prompts instead of further task exposition. In practical design terms, the cases suggest that one-size-fits-all support is unlikely to be enough. The next best move for an over-delegating student is different from the next best move for a disengaged student or a misaligned student.

Taken together, the qualitative layer clarifies why a single global CPS index is useful but insufficient on its own. The cases show that the same quantitative profile can contain very different collaboration mechanisms, and that some students with successful outcomes nevertheless make very little collaboration visible. Conversely, some students who remain highly engaged do not necessarily collaborate symmetrically. This is precisely why the mixed-methods design matters: the quantitative analyses identify broad participation patterns, while the memos explain the interactional logic behind those patterns.

7 Discussion

This study set out to examine how undergraduate students interact with a generative AI peer during open-ended collaborative problem solving, and what conditions distinguish productive from unproductive human–GenAI collaboration in higher-education mathematics. The study contributes three layers of evidence: a behaviorally grounded typology of eight participation roles, quantitative evidence linking interaction patterns to achievement, and initial validation of an automated scoring pipeline that makes process-based measurement of human–GenAI interaction feasible at scale.

Three main conclusions emerge. First, human–GenAI collaboration is not monolithic. The eight participation roles identified here span a wide behavioral space — from students who co-regulate genuinely, maintaining a shared cognitive state with the AI and externalizing reasoning at each step, to those who delegate procedural work while retaining oversight, to those who disengage or interact at a purely transactional level. This typology addresses a key gap in the literature on human–GenAI interaction in higher education: simply knowing that a student “used an AI” conveys little about whether that interaction constituted educationally valuable collaboration. Second, the new scorer shows encouraging but bounded validity evidence. Agreement with humans is strongest at the broader collaborative-competence and cognitive-process levels and more modest at the finest 12-code level, which suggests that the scoring pipeline is currently better suited to detecting *types of collaborative/cognitive moves* than to making psychometrically strong

claims about every fine-grained code boundary. Third, interaction quality still matters greatly: the most robust quantitative pattern in the study remains the distinction between low- and higher-interaction profiles, with the latter associated with substantially better achievement ($OR \approx 10.83$) and higher CPS indices. Critically, this is not simply a story about interaction volume: the qualitative cases show that the productive mechanisms involve keeping reasoning visible, maintaining common ground, and distributing cognitive responsibility in ways that allow the AI to respond contingently.

These findings can be situated within recent work on co-regulation, socially shared regulation, and group awareness in collaborative learning. Reviews of classroom assessment as co-regulated learning emphasize that productive assessment environments depend not only on what is measured, but also on whether interaction makes thinking visible enough for reciprocal regulation to occur (Andrade et al., 2021). Similarly, group-awareness and regulation research shows that collaboration benefits when partners can keep a shared model of progress, misunderstanding, and next steps visible to one another (Author et al., 2022; Schnaubert & Bodemer, 2022). Our qualitative cases extend this logic to human–LLM collaboration: the strongest dialogues were not merely longer, but more successful at externalizing assumptions, progress, and uncertainty in ways that enabled the agent to act as a meaningful peer rather than as a mere answer provider.

7.1 Why does interaction still matter so much?

The new analyses sharpen the interpretation of the higher-interaction pattern documented in Section 6.3. The strongest quantitative signal is not simply that more messages correlate with more CPS. Rather, students in the higher-interaction profile showed more mathematically elaborated dialogue, more monitoring/reflection activity associated with monitor and reflect (D), and a larger share of contributions associated with take appropriate action (2). In other words, their higher activity was accompanied by interaction forms that are more consistent with productive problem solving. The strongest quantitative signal is therefore not “chat more and you will succeed,” but something narrower: students who kept the interaction going also tended, on average, to externalize more mathematically actionable work and more monitoring/reflection, which are precisely the kinds of moves that help a partner respond contingently.

The study cautions against two overly simple readings. First, higher achievement does not guarantee stronger CPS: E2 and E28 solved both tasks while showing weak uptake of the AI’s scaffolding and little visible reciprocal reasoning, a pattern consistent with outcome success coexisting with shallow support-system engagement (Baker et al., 2004; Muldner et al., 2011). Second, higher interaction does not guarantee higher-quality collaboration: E4 and E11 were both high-interaction solvers, but E4 co-regulated continuously while E11 stayed at an arithmetic, trial-and-error level. Neither achievement nor message volume is a sufficient proxy for CPS quality. Process-based evidence is needed

to distinguish solving *with* the peer from solving *around* the peer.

These tensions align with profile-oriented CSCL research showing that learner clusters differ in regulatory quality, not only in participation volume (De Backer et al., 2022). Our cases sharpen that finding for human–LLM contexts: E4 exemplifies co-regulated learning (externalizing equations, monitoring the joint state, updating after errors); E6 and E18 reduced the agent’s regulatory capacity by internalizing reasoning or over-delegating execution. The eight participation roles thus represent meaningfully distinct collaboration mechanisms, not a simple engagement continuum.

7.2 What does the multi-agent scorer add?

The scorer’s value lies in enabling subdimensional profiling rather than a single total score. The primary global CPS index was meaningful (it differed by achievement group and cluster) but correlated strongly with message volume. In nested logistic models, adding the global index over self-efficacy and dialogue volume changed little; adding selected subdimensions (especially the take appropriate action (2) share, and to a lesser extent the monitor and reflect (D) share) improved the model substantially (Table 6). This aligns with work arguing that CPS is best understood as patterned combinations of social and cognitive activity, not undifferentiated totals (Lu et al., 2022). The scorer’s most useful role is therefore as an intermediate analytic layer that distinguishes broad CPS profiles — exploring vs. executing vs. monitoring vs. delegating — between raw turn volume and rich qualitative interpretation.

7.3 Implications for formative assessment and LLM-peer design

The results reported across the three substantive findings sections (Sections 6.1–6.4) have two design implications. From an assessment perspective, LLM peers can do more than host interaction: when combined with a transparent coding pipeline, they can generate multiple layers of CPS evidence — message-level, task-level, student-level, and profile-level — that can support formative feedback. From a task-design perspective, the comparison between Task 1 and Task 2 shows that different information-split structures elicit different forms of CPS. This is encouraging because it means the platform can be used not only to score collaboration, but also to design tasks that selectively surface different parts of the CPS construct. In practical terms, this means the platform can support richer teacher-facing interpretations than “correct/incorrect” or “used the chatbot a lot.” It can begin to answer questions such as: Did the student mainly try to coordinate, to execute mathematics, to monitor, to delegate, or to disengage?

The qualitative case analysis in Section 6.4 adds a more concrete design implication. Feedback should probably not be identical for all “low-CPS” students. Some students need help building shared understanding; others need support externalizing their reason-

ing; others need nudges away from over-delegation; and some need early re-engagement prompts. A practical next step would be to add a lightweight intent-classification layer that monitors rolling windows of recent student turns and flags shifts toward unproductive participation patterns. For example, if the system detects repeated imperative delegation moves consistent with E18’s *Delegating Director* profile (e.g., translated turns such as “*You do it*”), the peer could withhold the next numerical step and instead require the student to provide the next formula or justification before the interaction continues. If it detects patterns closer to E28’s *Early Disengager* or E25’s *Misaligned Follower*, the peer could switch from content delivery to a shorter common-ground repair or motivational re-entry move. In that sense, the most promising future direction is not simply more accurate scoring, but *profile-sensitive support* built on top of the scoring layer.

For higher-education instructors, the eight participation roles offer a concrete vocabulary for interpreting AI-peer activity logs and designing role-differentiated pedagogical interventions. An instructor noticing Co-Regulator patterns — frequent externalization of intermediate steps, self-initiated monitoring, and collaborative hypothesis formation — can provide relatively minimal additional scaffolding; the interaction is already doing productive epistemic work. By contrast, students showing Delegating Director patterns may benefit from structured prompts that require them to produce the next derivation step before the AI continues, thereby interrupting the delegation cycle. Students matching Early Disengager or Misaligned Follower profiles are the most urgent intervention targets: the former needs re-entry prompts that restore motivational engagement before reasoning can begin, while the latter needs explicit common-ground repair activities that surface and correct the misaligned shared representation. Low-Engagement Answer-Seekers may respond better to task redesigns that raise the cost of pure answer-seeking — for example, requiring the student to justify why the AI’s proposed step is correct before the interaction advances. This role-based framing moves beyond the binary of “high-engagement vs. low-engagement” and gives instructors actionable differentiation criteria tied to process evidence rather than merely to time-on-task.

This profile-sensitive view also resonates with recent studies of LLM peers. [Oppenheimer et al. \(2025\)](#) show that students can treat LLM partners as genuine interlocutors for refining arguments, but our cases suggest that this potential is highly contingent on how the student positions the agent: as a co-regulator, as a targeted knowledge source, as a delegated executor, or as a minimally used interface. A similar lesson appears in broader authentic collaborative settings. In an interdisciplinary hackathon study, [Law et al. \(2025\)](#) found that generative AI contributed most productively when teams used it within shared communicative and knowledge-building routines rather than as a stand-alone solution engine. Although our setting is dyadic and domain-specific rather than team-based and interdisciplinary, the same pattern appears here: the value of the LLM peer depended not simply on model capability, but on whether the interaction made reasoning visible,

preserved common ground, and distributed cognitive responsibility in workable ways. Designing future peers therefore requires attention not only to helpfulness and safety, but also to how the agent’s generosity, prompting style, and task framing distribute cognitive responsibility across the interaction. In this respect, the scoring layer described in Section 4 is not only an analytic component; it can also be understood as the basis for future adaptive orchestration rules that close the loop between trace interpretation and differentiated support (Author et al., 2023).

7.4 Limitations and next steps

The study is a single-site pilot ($n = 30$, two tasks), limiting statistical power and generalizability. The scorer was validated on 76 messages from four conversations (initial, not definitive evidence); fine-grained 12-code agreement was moderate ($\kappa = .60$), and some message-level distinctions remain difficult even for human raters. The LLM peer was not systematically audited for mathematical correctness; informal inspection found no frequent failures, but systematic error logging is needed in future iterations to estimate hallucination rates and their effects on CPS. Qualitative participation patterns are PISA-grounded but analytic constructs specific to this setting, not canonical CPS categories. Achievement was coded with low-range outcomes (0/1 per task), potentially obscuring partial progress.

Future work should therefore focus on three fronts. First, the scoring pipeline should be tested across larger and more diverse samples, additional tasks, and more extensive human-coded benchmarks. Second, the strongest qualitative and quantitative distinctions identified here — such as co-regulation, delegation, misalignment, and low-externalization problem solving — should be linked to profile-sensitive scaffolds and tested experimentally. Third, future models should explore richer learning outcomes and stronger task-level designs so that the relationship between collaboration quality and mathematical learning can be estimated with more precision.

8 Conclusion

This paper presented an LLM peer environment for collaborative calculus problem solving and demonstrated that the resulting interaction can be interpreted at multiple levels: dialogue traces, coded CPS moves, task-level differences, interaction profiles, and qualitative participation patterns. The core finding is that human–GenAI collaboration in higher education is not monolithic. The eight participation roles identified here — from Co-Regulator to Early Disengager — capture meaningfully distinct mechanisms that neither task scores nor interaction volume alone can distinguish. Students who externalized their reasoning, maintained a shared cognitive state, and distributed epistemic responsibility were substantially more likely to succeed ($OR \approx 10.83$), while students with identical

achievement profiles sometimes enacted collaboration very differently.

These findings suggest that LLM peers can function as *measurement-enabled collaborative partners* that both elicit CPS evidence and support its automated interpretation in authentic STEM tasks. The most useful measurement signal comes from combining subdimensional CPS profiles, interaction patterns, and qualitative interpretation rather than relying on a single global score — a lesson with direct implications for how AI-peer tools in higher education should be designed, deployed, and evaluated.

Declarations

Ethics approval and consent to participate

The study protocol (objectives, instruments, and procedures) was reviewed and approved by the institutional ethics committee within the university’s Faculty of Engineering and Sciences. All participants provided informed consent prior to registration, including acknowledgment that participation was voluntary and that they could withdraw at any time without consequences.

Consent for publication

All participants provided informed consent, which included consent for the anonymized publication of interaction data and findings derived from the study.

Availability of data and materials

The anonymized interaction logs and scoring pipeline code supporting the findings of this study are available upon reasonable request to the corresponding author.

Competing interests

The authors declare that they have no competing interests.

Funding

Funding information has been withheld to support double-blind peer review and will be provided upon acceptance.

Generative AI and AI-assisted technologies in the production of this work

Generative AI was used in three capacities in this work. First, as the object of study: the LLM peer environment deployed in the experiment uses two large-language-model-powered agents—a Responding Agent and a Provocateur Agent—built on GPT-4o mini. These agents are the collaborative interlocutors whose behavior students engaged with,

and are fully described in Section 3; they are not tools of the research team. Second, as an automated CPS scorer: a GPT-4o mini judge assigns PISA 2015 Collaborative Problem Solving codes to student messages as part of the scoring pipeline. This constitutes a core methodological component, reported in full in Sections 4 and 5.5 with validation against human-coded benchmarks. Third, large language model tools were used to assist with drafting, paraphrasing, and editing portions of the manuscript text. All substantive scientific content, analytical decisions, interpretations, and conclusions were produced and verified by the authors, who take full responsibility for the integrity of the work.

References

- Andrade, H. L., Brookhart, S. M., & Yu, E. C. (2021). Classroom assessment as co-regulated learning: A systematic review. *Frontiers in Education*, 6, 751168.
- Andrews-Todd, J., & Forsyth, C. M. (2020). Exploring social and cognitive dimensions of collaborative problem solving in an open online simulation-based task. *Computers in Human Behavior*, 104, 105759.
- Baker, R. S., Corbett, A. T., Koedinger, K. R., & Wagner, A. Z. (2004). Off-task behavior in the cognitive tutor classroom: When students “game the system.” In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 383–390). ACM.
- Chai, H., Hu, T., & Wu, L. (2024). Computer-based assessment of collaborative problem solving skills: A systematic review of empirical research. *Educational Research Review*, 43, 100591.
- Chu, Z., Wang, S., Xie, J., Zhu, T., Yan, Y., Ye, J., ... & Wen, Q. (2025). LLM agents for education: Advances and applications. *arXiv preprint arXiv:2503.11733*.
- Creswell, J. W., & Plano Clark, V. L. (2017). *Designing and conducting mixed methods research* (3rd ed.). SAGE Publications.
- De Backer, L., Van Keer, H., De Smedt, F., Merchie, E., & Valcke, M. (2022). Identifying regulation profiles during computer-supported collaborative learning and examining their relation with students’ performance, motivation, and self-efficacy for learning. *Computers & Education*, 179, 104421.
- Emirtekin, E. (2025). Large language model-powered automated assessment: A systematic review. *Applied Sciences*, 15(10), 5683.
- Felmer, P. (2023). Collaborative problem-solving in mathematics. *Current Opinion in Behavioral Sciences*, 52, 101296.

- Fiore, S. M., Graesser, A. C., & Greiff, S. (2018). Collaborative problem-solving education for the twenty-first-century workforce. *Nature Human Behaviour*, 2(6), 367–369.
- Flor, M., Yoon, S. Y., Hao, J., Liu, L., & von Davier, A. (2016). Automated classification of collaborative problem solving interactions in simulated science tasks. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 31–41). Association for Computational Linguistics.
- Forsyth, C. M., Zapata-Rivera, D., Graf, E. A., & Jiang, Y. (2024). Complex conversations: LLM vs. knowledge engineering conversation-based assessment. In B. Paaßen & C. D. Epp (Eds.), *Proceedings of the 17th International Conference on Educational Data Mining* (pp. 868–871). International Educational Data Mining Society. <https://doi.org/10.5281/zenodo.12729976>
- Graesser, A. C., Fiore, S. M., Greiff, S., Andrews-Todd, J., Foltz, P. W., & Hesse, F. W. (2018). Advancing the science of collaborative problem solving. *Psychological Science in the Public Interest*, 19(2), 59–92.
- Graesser, A. C., Greiff, S., Stadler, M., & Shubeck, K. T. (2020). Collaboration in the 21st century: The theory, assessment, and teaching of collaborative problem solving. *Computers in Human Behavior*, 104, 106134.
- Griffin, P., & Care, E. (Eds.). (2014). *Assessment and teaching of 21st century skills: Methods and approach*. Springer.
- Hao, J., Chen, L., Flor, M., Liu, L., & von Davier, A. A. (2017). CPS-Rater: Automated sequential annotation for conversations in collaborative problem-solving activities. *ETS Research Report Series*, 2017(1), 1–9.
- Hao, J., Cui, W., Kyllonen, P., Kerzabi, E., Liu, L., & Flor, M. (2025). Automated coding of communications in collaborative problem-solving tasks using ChatGPT. *Journal of Educational Measurement*, 62(4), 809–837.
- Herborn, K., Stadler, M., Mustafić, M., & Greiff, S. (2020). The assessment of collaborative problem solving in PISA 2015: Can computer agents replace humans? *Computers in Human Behavior*, 104, 105624.
- Hesse, F., Care, E., Buder, J., Sassenberg, K., & Griffin, P. (2015). A framework for teachable collaborative problem solving skills. In P. Griffin & E. Care (Eds.), *Assessment and teaching of 21st century skills* (pp. 37–56). Springer.
- Hou, X., Forsyth, C. M., Andrews-Todd, J., Rice, J., Cai, Z., Jiang, Y., Zapata-Rivera, D., & Graesser, A. C. (2025). An LLM-enhanced multi-agent architecture for conversation-

based assessment. In *Proceedings of the 26th International Conference on Artificial Intelligence in Education*.

Law, N., Wang, N., Ma, M., Liu, Z., Lei, L., Feng, S., Hu, X., & Tsao, J. (2025). The role of generative AI in collaborative problem-solving of authentic challenges. *British Journal of Educational Technology*, 56, 1–21. <https://doi.org/10.1111/bjet.70010>

Li, H., Dong, Q., Chen, J., Su, H., Zhou, Y., Ai, Q., ... & Liu, Y. (2024). LLMs-as-Judges: A comprehensive survey on LLM-based evaluation methods. *arXiv preprint arXiv:2412.05579*.

Lu, J., Tuo, P., Feng, R., Stephens, M., Zhang, M., & Shen, Z. (2022). Visualizing commognitive responsibility shift in collaborative problem-solving during computer-supported one-to-one math tutoring. *Frontiers in Psychology*, 13, 815625.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1(1), 3–62.

Muldner, K., Burleson, W., Van De Sande, B., & VanLehn, K. (2011). An analysis of students' gaming behaviors in an intelligent tutoring system: Predictors and impacts. *User Modeling and User-Adapted Interaction*, 21(1–2), 99–135.

OECD. (2017). *PISA 2015 assessment and analytical framework: Science, reading, mathematical, financial literacy and collaborative problem solving*. OECD Publishing.

Oppenheimer, D. M., Cash, T. N., & Connell Pensky, A. E. (2025). You've got AI friend in me: LLMs as collaborative learning partners. *International Journal of Artificial Intelligence in Education*, 35, 3896–3921.

Author et al. (2021). [Reference withheld for double-blind review.]

Author et al. (2022). [Reference withheld for double-blind review.]

Author et al. (2023). [Reference withheld for double-blind review.]

Author et al. (2025). [Reference withheld for double-blind review.]

Rosen, Y. (2015). Computer-based assessment of collaborative problem solving: Exploring the feasibility of the human-to-agent approach. *International Journal of Artificial Intelligence in Education*, 25(3), 380–406.

Sætra, H. S. (2022). Scaffolding human champions: AI as a more competent other. *Human Arenas*, 8, 56–78.

- Schnaubert, L., & Bodemer, D. (2022). Group awareness and regulation in computer-supported collaborative learning. *International Journal of Computer-Supported Collaborative Learning*, 17, 11–38.
- Scoular, C., & Care, E. (2020). Monitoring patterns of social and cognitive student behaviors in online collaborative problem solving assessments. *Computers in Human Behavior*, 104, 105874.
- Scoular, C., Care, E., & Hesse, F. W. (2017). Designs for operationalizing collaborative problem solving for automated assessment. *Journal of Educational Measurement*, 54(1), 12–35.
- Steinbach, M., Bhandari, S., Meyer, J., & Pardos, Z. A. (2025). When LLMs hallucinate: Examining the effects of erroneous feedback in math tutoring systems. In *Proceedings of the Twelfth ACM Conference on Learning @ Scale* (pp. 139–150). ACM.
- Sun, C., Shute, V. J., Stewart, A., Yonehiro, J., Duran, N., & D’Mello, S. (2020). Towards a generalized competency model of collaborative problem solving. *Computers & Education*, 143, 103672.
- Yue, M., Lyu, W., Mifdal, W., Suh, J., Zhang, Y., & Yao, Z. (2024). MathVC: An LLM-simulated multi-character virtual classroom for mathematics education. *arXiv preprint arXiv:2404.06711*.
- Zhang, L., Zhai, X., Lin, J., Kleiman, J., Zapata-Rivera, D., Forsyth, C. M., Jiang, Y., Hu, X., & Graesser, A. C. (2025). Exploring communication strategies for collaborative LLM agents in mathematical problem-solving. *arXiv preprint arXiv:2507.17753*.

A PISA-grounded message-level CPS codebook

Table 8 summarizes the message-level codebook used in the new scoring pipeline. The codes preserve the PISA 2015 matrix structure while adapting it to open-ended calculus dialogue.

Quality rubric. Each coded message receives a quality score from 1 to 3. Quality 1 indicates weak/minimal evidence; quality 2 indicates clear and sufficient evidence; quality 3 indicates strong, explicit, or especially well-developed evidence.

Tie-breaking rules. The scorer uses cumulative prior context to resolve short or ambiguous messages. In particular, short affirmations are differentiated depending on whether they primarily maintain establish and maintain shared understanding (1) (e.g., agreeing with a mathematical statement), advance coordinated action associated with establish and maintain team organization (3) (e.g., agreeing to continue), or simply acknowledge

Table 8: Message-level CPS codebook used in the automated scoring pipeline.

Code	Cognitive process	Collaborative competence	Operational meaning in dialogue
A1	Explore & understand (A)	Shared understanding (1)	Describes what is visible, identifies relevant features, or makes the current problem state explicit for both partners.
A2	Explore & understand (A)	Take action (2)	Explores by trying a mathematically relevant move or test.
A3	Explore & understand (A)	Team organization (3)	Asks how to start, what the task requires, or how work should be organized at the outset.
B1	Represent & formulate (B)	Shared understanding (1)	Proposes or negotiates a conceptual or mathematical relation in a way that supports shared meaning.
B2	Represent & formulate (B)	Take action (2)	Introduces an actionable formula, equivalence, substitution, or symbolic representation that advances the solution.
B3	Represent & formulate (B)	Team organization (3)	Requests help with a formula or representation, or orients the partner toward a representational move.
C1	Plan & execute (C)	Shared understanding (1)	Communicates or confirms the current action or intermediate step so both partners remain aligned.
C2	Plan & execute (C)	Take action (2)	Performs a calculation, substitution, derivative, algebraic manipulation, or reports an actionable result.
C3	Plan & execute (C)	Team organization (3)	Delegates execution, asks the partner to perform the next step, or coordinates who carries out the procedure.
D1	Monitor & reflect (D)	Shared understanding (1)	Detects and corrects an error or updates the shared interpretation of what has happened in the solution.
D2	Monitor & reflect (D)	Take action (2)	Validates whether a result makes sense, satisfies a constraint, or is mathematically consistent.
D3	Monitor & reflect (D)	Team organization (3)	Closes a subtask, changes direction, or reorganizes the collaboration after monitoring joint progress.

Note. Codes preserve the PISA 2015 matrix while adapting it to open-ended calculus dialogue. In the paper body, full process and competence names are used first, with letter/number notation in parentheses for compactness.

the partner. Similarly, isolated numbers or symbolic expressions are interpreted as contributions associated with represent and formulate (B) or plan and execute (C) depending on the preceding dialogue.