RESEARCH-ARTICLE

Piecing Together Teamwork: A Responsible Approach to an LLM-based Educational Jigsaw Agent

Published: 26 April 2025

CHI 2025: CHI Conference on Human Factors in Computing Systems
April 26 - May 1, 2025
Yokohama, Japan

Conference Sponsors:
SIGCHI

CHI '25: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (April 2025)
https://doi.org/10.1145/3706598.3713349
ISBN: 9798400713941

Piecing Together Teamwork: A Responsible Approach to an LLM-based Educational Jigsaw Agent

Emily Doherty
University of Colorado Boulder
Boulder, Colorado, USA
emily.doherty@colorado.edu

E. Margaret Perkoff
University of Colorado Boulder
Boulder, Colorado, USA
margaret.perkoff@colorado.edu

Sean von Bayern
University of Colorado Boulder
Boulder, Colorado, USA
sean.vonbayern@colorado.edu

Rui Zhang
University of Colorado Boulder
Boulder, Colorado, USA
ruzh7356@colorado.edu

Indrani Dey
University of Wisconsin-Madison
Madison, Wisconsin, USA
idey2@wisc.edu

Michal Bodzianowski
University of Colorado Boulder
Boulder, Colorado, USA
michal.bodzianowski@colorado.edu

Sadhana Puntambekar
University of Wisconsin-Madison
Madison, Wisconsin, USA
puntambekar@education.wisc.edu

Leanne Hirshfield
University of Colorado Boulder
Boulder, Colorado, USA
leanne.hirshfield@colorado.edu

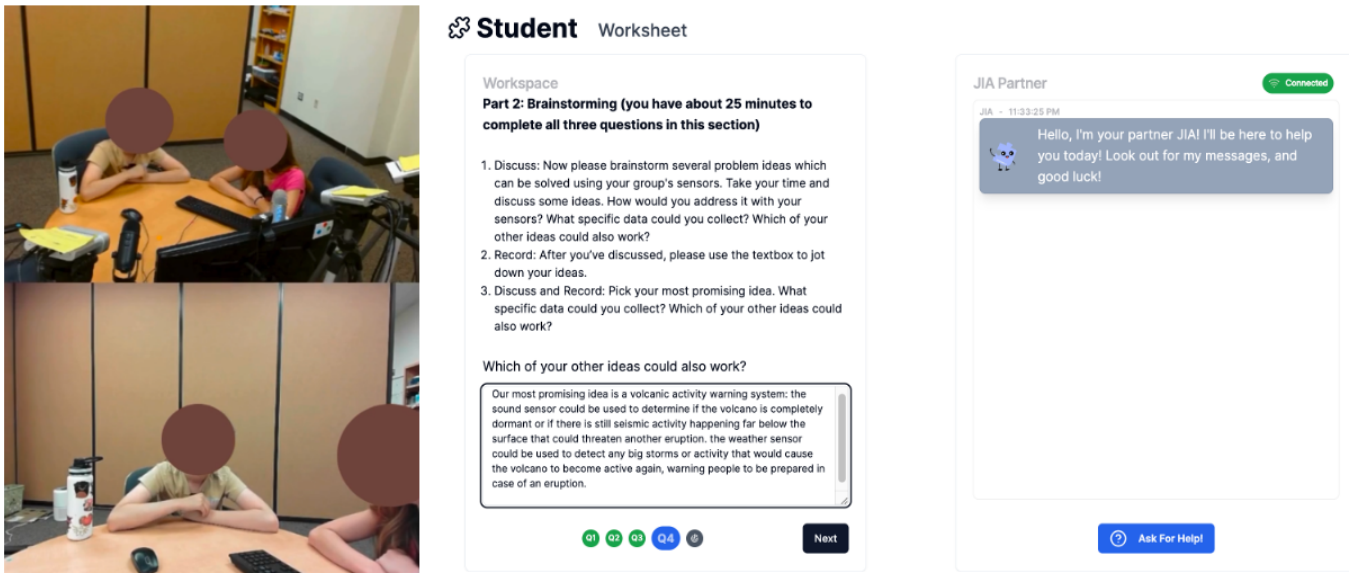

Figure 1: Left: a dyad collaborates during the jigsaw activity. Right: The 'Student' view of the JIA web app interface, where the group answers the jigsaw questions and receives interventions from JIA.

Abstract

Conversational agents have been used to support student learning for some time, but the emergence of Large Language Models (LLMs) poses a novel opportunity to enhance their capabilities in collaborative settings. LLM-powered agents can provide timely interventions in collaborative conversations when a teacher is unable to assist the students. However, the use of LLMs in such tools raises many ethical questions and concerns, especially for use with young, impressionable populations. In this work, we present the human-centered design and evaluation of an LLM-based agent aimed to facilitate small group collaboration in middle- and high-school classrooms. Fifty-eight groups of dyads and triads (145 participants), aged 12-17, collaborated in a jigsaw activity and were assigned to be assisted by our agent or not. The results showed decreased self-reported ratings of social loafing and increased use of language related to respectful collaboration in interactions with the agent compared to those without.

CHI '25, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3713349

CCS Concepts

• **Human-centered computing → Collaborative and social computing design and evaluation methods**; • **Applied computing → Interactive learning environments**; • **Computing methodologies → Natural language processing**.

Keywords

Conversational Agent, Jigsaw, System Design, Artificial Intelligence, Education

ACM Reference Format:

Emily Doherty, E. Margaret Perkoff, Sean von Bayern, Rui Zhang, Indrani Dey, Michal Bodzianowski, Sadhana Puntambekar, and Leanne Hirshfield. 2025. Piecing Together Teamwork: A Responsible Approach to an LLM-based Educational Jigsaw Agent. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3706598.3713349

1 Introduction

Whether it's surgeons in an operating room, a construction crew fixing a breached dam following a storm, or teachers developing new curriculum units for a K-12 classroom, the workforce of the future will be increasingly driven by teams working together to solve complex problems. Yet, many employers have noted that workers exhibit deficits and difficulties collaborating in the workplace [93]. In fact, the PISA 2015 international collaborative problem solving (CPS) assessment among 15-year-old students across 52 economies found that less than 30% of students demonstrated success on even low complexity problems [60], compelling experts to proclaim a "global deficit" in collaboration skills [28, 33]. Researchers have suggested that collaboration deficits partially stem from a lack of adequate training on collaboration skills [28]. However, assigning students to work in groups does not necessarily ensure they will work collaboratively [52] and simply asking students to communicate and interact does not lead to deep, meaningful learning [58, 86]. These challenges are exacerbated by the significant demands placed on teachers in classroom settings, where they must juggle several tasks at once in hectic and noisy learning environments. They monitor student progress during collaborative activities, provide student groups with personalized guidance, and support important knowledge-building group conversations. Moreover, they must do so across multiple groups of students and at the same time [8, 50, 91]. Under such conditions, it can be difficult for teachers to track when a group struggles to construct shared knowledge through collaboration. We propose a solution to this problem: an AI agent that monitors small group collaborations and intervenes at moments when it is necessary to support knowledge sharing and collaboration.

Specifically, our Jigsaw Interactive Agent (JIA), shown in Figure 1, supports diverse student groups as they collaborate on Jigsaw classroom activities, a type of curriculum activity designed to foster group collaboration and knowledge sharing. In the jigsaw method, students are put into small groups, and each is assigned a different but related topic to study [2]. Each student independently becomes an "expert" in their assigned topic. These "experts" are then regrouped so that each new group includes one expert from each topic. They then share their knowledge, allowing the group to work collaboratively and deepen their understanding of the subject matter. In the current study, we use the jigsaw method as it is an established pedagogical practice and is a prime example of a collaborative problem-solving activity deployed in both middle- and high- school settings.

The design and implementation of supportive agents are hampered by several interdependent challenges. Although the field of Computer Supported Collaborative Learning (CSCL) has long envisioned the future of collaborative learning using technology-based support systems like in [22, 26, 77], many studies investigating AI applications focus on cognitive outcomes and not facilitating the knowledge sharing or collaboration processes crucial to learning [84]. This is due in large part to a disconnect between the developers of AI systems and the domain experts with a deep understanding of key values, user goals, ethical considerations, and targeted outcomes for an AI partner for collaborative learning. The use of AI in K-12 classroom settings is a polarizing topic, with ethical AI design and responsible innovation considerations being crucial. Key concerns include the privacy of personal data [14, 98], fears of future dependency on AI [42], and the lack of a universal set of regulations for its use [80]. Thus, the question of how to design a system that can be trusted for use in sensitive contexts, such as a classroom, is one of the most prominent challenges facing researchers and designers today. Consider the code of conduct that both teachers and students abide by, such as harboring an appropriate, inclusive, and respectful learning environment. It is difficult to instill this tacit knowledge into AI systems [10] as "human reasoning is embodied, situated, in a social context and involves actions, often improvised, in the world, the complexity of which formal models cannot replicate" [11].

With the growing concern regarding AI's potential to exploit and mislead, there has been a recent shift towards Human-Centered AI (HCAI), which emphasizes the importance of creating AI tools centered around human expertise and feedback [11, 74]. Additionally, with the rapid advancements in AI, it is essential to innovate responsibly by not only taking a human-centered approach with key stakeholders and domain experts included, but also to have these design teams carefully consider the unintended consequences of such research [79]. More recently, Capel and colleagues (2023) articulated this point:

> "It is undeniable that AI and HCI need each other and that HCAI research can benefit from stronger collaborations across fields and efforts to understand each other's work and values... domain experts can highlight considerations around values and potential consequences that may not be obvious to AI designers, giving them greater influence in HCAI focused design. Given the multiplicity of interests, varieties of users, business interests and domain interests at play in any situation, this is no small challenge." [11]

Aligning with this sentiment, in this work, we take a responsible, interdisciplinary, and human-centered approach to the design and evaluation of our large language model (LLM)-based conversational agent. Guiding this work are the key principles of Stilgoe et al.'s responsible innovation framework: reflexivity, anticipation, inclusion, and responsiveness [79]. Throughout each stage of the design

process, we carefully incorporate the feedback and teachings of experts in Natural Language Processing (NLP), Human-Computer Interaction (HCI), CSCL, and learning sciences who play the role of subject matter experts (SMEs) due to their direct experience working with students in classrooms on the jigsaw activities within which JIA operates.

To achieve our high-level goal, we first created a configurable, web-based environment where participants in small groups work on a jigsaw activity that was adapted from an existing STEM curriculum, called Sensor Immersion [72]. This curriculum has been successfully implemented in diverse classrooms across school districts nationwide. To understand when and how a partner should intervene, we conducted studies using SMEs (i.e., CSCL researchers and learning scientists with direct experience working with students in classrooms on Sensor Immersion) in a Wizard-of-Oz (WoZ) paradigm. These Wizard Subject Matter Experts (WoZ-SMEs) delivered real-time interventions to the group as JIA.

Using video data from those studies, we applied a pedagogical annotation schema evaluating each moment that support was offered by the WoZ. We annotated participants' behaviors before an intervention to learn more about the collaborative, team-level state that prompted the WoZ to intervene, as well as what type of intervention was offered (e.g., validation, task-related support). Using these annotations, we conducted factor analysis grouping co-occurring annotations into three groups, which were then labeled by our CSCL experts into three collaborative states that occurred before an intervention was sent (i.e., Parallel or Limited Interaction, Contributing to Shared Problem Space, and Unproductive Perseverance). We also transcribed and analyzed the participants' discourse before interventions were sent, as well as at randomly selected times when no intervention was sent, using several automated discourse classification models relating to collaboration. The discourse model results were used as feature vectors, each labeled with a collaborative state, to train a decision tree model to predict these states from participants' dialogue. The resulting decision tree was used to create transparent rules that form the basis of JIA's dialogue policy which detects a state from the discourse. These states and some recommended actions are ultimately fed as part of a prompt to an LLM that outputs a real-time intervention to the students as they collaborate. Incorporating these types of dialogue action constraints to an LLM has been proven to increase its adherence to a set of guidelines and the helpfulness of the model in a teacher-like interaction [63, 67, 82]. Following creation of the dialogue policy, a human-in-the-loop capability (HITL) was integrated into JIA's architecture, enabling human experts to review JIA's suggested prompts, before sending the prompts on to participants.

Finally, we ran an evaluation study investigating the efficacy of our agent at promoting knowledge sharing and group collaboration as well as its effect on user experiences working with the agent. We performed a between-groups study with comparisons made between JIA with human-in-the-loop capability (JIA-LLM-HITL), a control condition with no intervention/agent (Control), and the previously described WoZ-SME condition.

Evaluation results show that the JIA-LLM-HITL condition fostered more thoughtful, respectful, and engaged communication between participants when compared with participants in the WoZ-SME condition who displayed more emotional expression but less analytic and respectful interactions. Participants in the Control condition used more 1st person singular language (I, me, my, mine) than JIA-LLM-HITL and WoZ-SME groups, indicating less collaborative efforts in the absence of JIA. The self-report survey results also showed that participants in the WoZ-SME group had significantly less social loafing compared to the Control group. However, there were no significant differences between the Control and JIA-LLM-HITL groups, or between the JIA-LLM-HITL and WoZ-SME groups for other self-report metrics including psychological safety and trust in agent.

The remainder of this paper is organized as follows: Section 2 describes responsible innovation and HCAI approaches guiding our primary research goals and key activities, as well as related work in collaborative learning through problem solving and pedagogical agents to support these activities. Section 3 describes the human-centered activities taken to design and develop JIA, with Section 4 presenting the LLM and prompting methods used. Section 5 describes the software architecture of JIA. Section 6 describes the evaluation study results and interpretation. We discuss our findings in Section 7, followed by limitations in Section 8, and conclusions and future work in Section 9.

2 Related Work

2.1 Responsible Innovation

While the idea of responsible innovation (RI) is not a novel concept, Stilgoe et al.'s framework, an effort funded by the UK Research Councils, is a scoping yet succinct list of guidelines that apply to all fields of research [79]. The authors outline the four driving principles of RI as anticipation, reflexivity, inclusion, and responsiveness. To innovate responsibly, scientists and researchers must 1) anticipate the consequences and gains of the ever-evolving technological progress, 2) be reflexive and consider the moral responsibilities of our work, 3) be inclusive and promote research that extends out to the wider public, and 4) be responsive and adapt to the fluctuation of public views and consistent growth of science and discovery. A more recently proposed framework specifically addresses the profound implications of research in AI and highlights the need for transparency in this ever-evolving landscape of innovation [9]. While RI calls for open discourse and reflection, AI challenges this through its opacity—poor transparency, explainability, and accountability. Thus, there is a heightened need for both prospective and retrospective transparency throughout all stages of innovation. We leveraged both frameworks to guide the design and evaluation of the conversational agent, and we revisit these choices in the discussion (Section 7).

2.2 Human-Centered Design & Evaluation of AI

Concerns over the potential societal impacts of AI have caused researchers to take more human-centered approaches toward the design and evaluation of AI systems. In their recent review of HCAI, Capel & Brereton explore the claim of human-centeredness and how it affects the interaction between the human and AI and the resulting impacts [11]. Specifically, they identify several areas of emerging and overlapping HCAI research, including *Interaction with AI* and *Ethical AI*. Interaction with AI enables humans to

directly engage with the AI learning process. This includes the sub-area of *Contestable AI*, where humans may contest the decision of AI to augment its learning process. This sub-area requires multidisciplinary skilled teams that draw together diverse technical, design, and domain-specific skills. Capel and Brereton note that these interactions with AI have been slow to emerge but have great potential for HCAI. Ethical AI seeks accountability regarding fundamental human values and rights, and advocates for more transparent design of AI. Its overarching claim to human-centeredness is that it considers the rights and values of the people who are working with the AI or are impacted by the AI, particularly within sensitive contexts [11]. In their review, the authors identify several design and evaluation methods used in the current HCAI literature including data-enabled design as well as Wizard-of-Oz and Human-in-the-loop paradigms. Data-enabled design utilizes data in the early stages of design to better understand context and user needs [30]. Further, design is an iterative practice of using data to inform the construction of an intelligent system, testing it in the real world, and adapting the design until the process becomes self-sustaining [47, 59]. For example, a form of data-enabled design includes engaging stakeholders and subject matter experts early in the design process, as encouraged by many in the HCI community [3, 30]. One such way to engage these outside entities is techniques like Wizard-of-Oz (WoZ), which has been deployed in wider human-computer interaction research for decades [17]. In a WoZ paradigm, participants believe they are working with a computer-based entity, which is actually controlled by a human confederate [44]. More recently and related to our study, WoZ paradigms have been used in the design and evaluation of pedagogical conversational agents [43, 73].

Another human-centered approach includes Human-in-the-Loop (HITL) experimental designs. HITL is an umbrella term for a myriad of methods where human feedback is used to inform, train, and improve learning of AI agents [12, 18]. One of the earliest and most frequently used forms of human advice involves the use of expert-generated rules as the backbone of intelligent systems. Rule-based systems (also known as expert systems) are one of the simplest forms of AI, using rules as the representation for knowledge coded into the system [34]. For example, SMEs can develop a set of if-then-else rules to be programmed into intelligent systems [34]. Although these rule-based systems are simpler in nature than other learning algorithms (e.g., deep learning), they have the added benefits of transparency and explainability. Agent decisions can be aligned with best practices through the theoretically-grounded, empirically-validated work in the domain represented by the SMEs. In this work, we take a human-centered approach by employing data-enabled design as well as WoZ and HITL paradigms in the design and testing of JIA.

2.3 Collaborative Learning through Problem Solving

CSCL research aims to understand how students engage in meaning-making, construct knowledge, and solve problems together by participating in a joint activity [22, 39]. This activity is usually mediated by technology and may be supported through pedagogical practices, such as the jigsaw method. CSCL environments are grounded in the sociocultural perspective, which posits that learning is facilitated through social interactions [95]. Therefore, examining how students interact during a joint activity is critical to understanding their collaboration. This includes the examination of students' creation of a shared problem space [4, 69], how they contribute to this space, and build on each other's ideas [71, 83], to solve problems together [21]. As noted previously, this work is largely focused on the specific CSCL activity of collaborative problem solving (CPS), as assessments have shown deficiencies in these skills worldwide [60].

However, assigning students to work in groups does not always ensure they will work collaboratively [52], and asking students to simply communicate and interact doesn't lead to deep, meaningful learning [58, 86]. While dialogue is informative and can reveal cognitive processes like consensus building and conflict resolution, many of these processes may not occur without guidance [23]. Even with guidance, students may not interact with prompts unless they elicit a cognitive or metacognitive response [99]. Thus, it is imperative that an agent designed to facilitate collaboration is able to 1) understand the group's cognitive state via their dialogue and 2) deliver an intervention that is truly encourages groups to collaborate and learn from one another.

2.4 Team-Level State Detection via Dialogue

Dialogue is arguably the richest indicator of collaboration and cognitive processing at the team level (i.e., team cognition) [15]. Many of the cognitive processes implicated in collaborative learning including shared knowledge building and problem solving can studied through social discourse [36, 81]. With the growth of NLP tools to evaluate discourse, we can deduce information about a team's cognitive processing through trained discourse models. For example, Sun et al.'s framework of collaborative problem skills (CPS) [81] been used to annotate utterances and train a BERT-based classification model to classify these skills during live discourse [65, 78]. Similarly, Breideband et al. developed a RoBERTa-based NLP model to evaluate discourse in relation to shared norms (i.e., being respectful) between teachers and students to guide classroom collaboration [8]. In addition to these CSCL-based discourse models, we can also explore discourse more generally with text analysis tools. For example, many text corpora have been used to develop dictionaries that count words in psychologically meaningful categories. A notable example is Linguistic Inquiry and Word Count (LIWC), a dictionary featuring over 100 word categories including social, affective, and cognitive processes [6, 85]. Using more robust tools like LIWC allows for more generalizable insights that complement the findings from the CSCL-specific models, deepening our understanding of the collaborative interactions.

For a conversational agent to support collaborative learning, it must be able to deduce information about the team's cognitive *state*, to know when and how to intervene accordingly. While NLP models and techniques can provide a wealth of information about cognitive processes that occur during collaboration, inferring the collective cognitive *state* at the team level remains a more complex challenge. This difficulty arises because team-level states often emerge from nuanced interactions and shared dynamics that are not easily captured through speech alone [16]. Thus, in the development of JIA's

dialogue policy, we used both annotated data of team behaviors and NLP indicators to model states (with the help of CSCL experts) in which JIA should intervene. In the next section, we discuss several examples of conversational agents developed to intervene during collaboration to augment student performance and learning.

2.5 Pedagogical Conversational Agents to Support Collaborative Learning

Recent advancements in NLP, particularly the emergence of many publicly available LLMs, have made the development of realistic and helpful conversational agents possible. Pedagogical conversational agents [35] encompass a wide variety of systems in which the user interacts with an agent in a learning environment. Conversational agents [45] refer to the class of NLP systems that engage with a user in a back-and-forth dialogue. Prior work in this space has shown that agent interactions can improve students' motivation, engagement levels, and in some cases individual and group learning outcomes [25, 75, 87]. Recent surveys on pedagogical agents have distinguished these systems from one another based on implementation goals, the impact of the system on learning outcomes, and pedagogical roles of the agent [48]. Frequently, the agents are deployed in an online learning setting to interact with students via a web-based interface [89]. Some systems are designed to engage with students like a one-on-one tutor [97], whereas others are built to facilitate conversations between groups [88]. These studies have demonstrated the ability of pedagogical agents to improve students' self-regulation skills, understanding of subject matter, and collaboration [49]. Only a handful of studies have specifically focused on assessing CPS skills and providing interventions during remote collaborative gameplay [24, 78]. There is much work to be done in this space, particularly with the recent boom of LLMs, provided these innovations are developed ethically and responsibly.

3 Human-Centered Design of JIA

Our high-level goal is to create and implement a dialogue policy that is i) tightly aligned with theoretical and empirical work in collaborative learning and content support, ii) has a transparent, explainable, and justifiable mapping between the higher-level dialogue policy states and actions and the actual underlying measures/models that are inferring a given state and recommending an action. Figure 2 provides a high-level view of the design activities undertaken to develop JIA with an eye toward transparent, responsible innovation.

3.1 Data Collection in Support of JIA development

Key activities include 1) a data collection where WoZ-SMEs observed participants collaborating on a jigsaw activity (Figure 3), intervening in ways to support group knowledge building and collaboration, 2) annotated the data to deduce collaborative states of participants occurring prior to WoZ-SME interventions, and paired those annotations with, 3) NLP-based indicators of collaborative states using the Multimodal Intelligent Analyzer (MMIA), 4) used data from steps 3 and 4 to train a decision tree model to classify group's collaborative states, which ultimately is used to generate a rule-based dialogue policy, which is used to prompt the Mistral LLM.

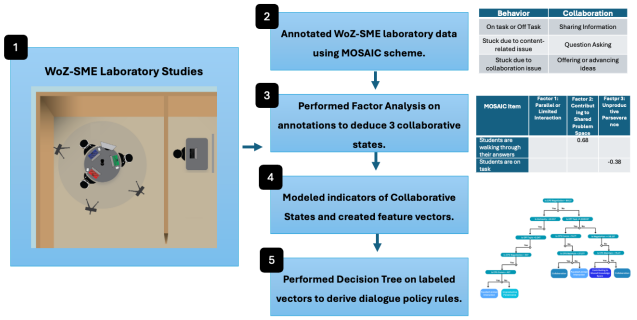

Figure 2: This figure depicts the main activities that were undertaken to design JIA at a high level. This includes annotation of WoZ-SME study data, factor analysis of those annotations to deduce collaborative states, and deriving dialogue policy rules via a trained decision tree using dialogue features which we describe in detail next.

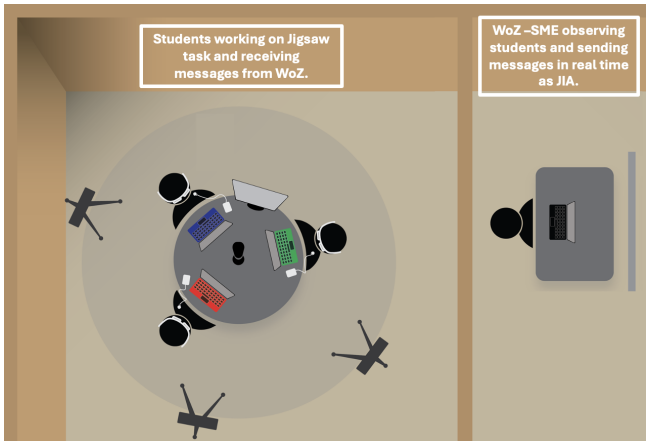

Figure 3: While students collaborate on the jigsaw activity in groups of 2-3 students, a WoZ-SME observes the collaboration remotely via video conferencing and sends messages in real-time as JIA.

3.2 Jigsaw Activity

In this study, participants completed a jigsaw activity based on the Sensor Immersion curriculum unit which has been implemented with 54 teachers and more than 5000 students across three school districts. The unit has five lessons, which can take 5 or more classes to implement, depending on the teacher. In the curriculum unit, students learn about basic programming concepts, using a drag and drop block programming environment in MakeCode, where they learn how to wire several hardware sensors and how to design a program (in MakeCode) to take in data from each sensor and show information on a LED display using programmatic logic. One of the lessons within Sensor Immersion is a jigsaw activity, where each student becomes an expert in a specific MakeCode sensor (environmental, moisture, or sound) after completing individual

tutorials on how to code and test their sensor. After developing this expertise, students are grouped together to fill in a paper worksheet, which includes questions about each sensor (knowledge sharing), as well as questions that ask students to brainstorm scenarios where they might combine all three sensors to help 'solve' a real-world problem (brainstorming/collaboration).

We adapted the jigsaw portion of the Sensor Immersion curriculum for investigation in these lab data collections. For the jigsaw activity, we designed a flexible, web-based app with a separate page for each question and editable text box. Behavioral data such as amount of time spent on each question and written text were logged using Amazon DynamoDB [76]. In the 'Student' view interface, participants collaborate on a jigsaw activity which includes three questions relating to knowledge sharing (e.g., "What data can your sensor collect?") and a final question relating to brainstorming (i.e., "Brainstorm several problem ideas which can be solved using your group's sensors"). The interface has a textbox to type in their answers (Figure 1), as well as a chatbot interface in which WoZ-SMEs would send messages as JIA. Participants could select a button labeled 'Ask JIA for help', in which a message saying, 'Help me JIA!' would be sent to the WoZ-SME. In the 'Wizard' view, WoZ-SMEs viewed the question and text written by participants in real time, as well as a chat interface in which they could send messages as JIA. More details on the software architecture connecting the web app, dialogue data, and JIA are in Section 5.

3.3 Annotation of WoZ Support Movements

To better understand the collaborative state which prompted a WoZ-SME to send a message, we annotated video data from the WoZ-SME sessions using a pedagogical schema, MOSAIC-AI. The original MOSAIC coding protocol was developed to evaluate the behaviors before and after a supportive intervention provided by teachers or peers during collaborative activities in the classroom [20, 37, 38]. The protocol examines student behaviors one minute before and after the intervention, to provide more context about the interaction, i.e., why the support was provided, who initiated and provided the support, and how the support was taken up. For the minute before and after an intervention, annotators watch the video data and label each behavior they observe depending on how long they observed it for, on a scale from 1 (none of the time) to 5 (all of the time).

The original MOSAIC protocol also allows researchers to identify the type of supportive intervention delivered (e.g., encouragement/validation, explanation about the task). Specifically, we used an adapted protocol, MOSAIC-AI, that more accurately captures human-AI interactions, rather than student-teacher or peer-to-peer interactions. This adapted protocol was revised from the original MOSAIC schema to 1) evaluate the support offered by an AI agent, rather than a teacher or peer, 2) include behaviors observed in the video data and exclude those not observed in the controlled lab setting (i.e., announcement by teacher), 3) include students' collaborative state by examining their dialogue, and 4) provide more information about the appropriateness of the intervention and students' reactions to it.

For every intervention message sent by the WoZ-SME, the moment before and after was annotated using the MOSAIC-AI schema. Across the 20 WoZ-SME sessions, there were 143 interventions delivered, amounting to approximately 286 minutes of video data that were annotated. The WoZ-SMEs annotated a group of sessions that were randomly assigned to them, excluding any sessions where they were delivering interventions. All WoZ-SMEs were trained in annotating the support moments following the MOSAIC-AI schema. The first session was annotated by all WoZ-SMEs and discussed together to remove any bias and disagreements on interpreting the MOSAIC-AI schema.

3.4 Factor Analysis of MOSAIC-AI Annotations

To understand the pre-intervention behavior, we used all MOSAIC-AI annotations from the minute before an intervention delivered by the WoZ-SME. The possible annotations regarded the students' general and collaboration behaviors (See Table 1), dialogue states, and negotiation status. Dialogue state was a binary variable that annotators rated, 0 being productive discussion, 1 being unproductive discussion. Negotiation status similarly was coded 0 for no negotiation and 1 for negotiation present. For each annotation label (e.g., 'Students are walking through their answers'), the corresponding scores from 1-5 were used to performed factor analysis, in order to cluster co-occurring behaviors together. The factor analysis resulted in three factors that combined students' general and collaboration behaviors, dialogue states, and negotiation status. The Bartlett's test results show that $p < 0.001$, Chi-square = 305.07. The value of Kaiser-Meyer-Olkin is 0.56, larger than 0.5 indicating that the data is acceptable for factor analysis [41]. All the variables used to conduct the factor analysis derive from the adapted MOSAIC annotations of 20 experimental WoZ-SME sessions. We considered data from all times when the WoZ-SME intervened to support the group. We consulted with CSCL experts who qualitatively categorized the three factors into 'collaborative states' based on the characteristics of the variable loadings, drawing from CSCL literature to find similar states. This resulted in a mapping from MOSAIC item loadings to factors that were labeled with CSCL states, as detailed in Table 2.

The first factor is labeled as the collaborative state of Parallel or Limited Interaction. The variables *students being off task* (0.51), *no collaboration in group* (0.75), and the Group's *dialogue state* (0.64) loaded positively on this component, and *Students offering or advancing ideas* (-0.42), and *students build off others' ideas or paraphrase* (-0.35) loaded negatively on this component. Together, these describe a state where students are likely not collaborating well, i.e., they are not working on the task together, and there is a lack of interaction with other group members. Similar states have also been described in other studies, such as, limited verbal interaction between group members, students rejecting others' ideas without further discussion [5], or not building on each others' ideas [68]. These group members are likely not working well together [52] and thus are not collaborating well.

The second factor to be labeled as a collaborative state is: Contributing to Shared Problem Space. The variables *Students are walking through their answers* (0.68), and *Students are sharing information* (0.89) loaded positively on this component, and *No collaboration in group* (-0.39) loaded negatively on this component. These indicate that students are sharing their knowledge with other team members

Behavior	Collaboration
On task or Off Task	Sharing Information
Stuck due to content-related issue	Question Asking
Stuck due to collaboration issue	Offering or advancing ideas
Stuck due to procedural issue	Building off others' ideas or paraphrasing
Walking through answers	Talking about how they work together
Express need for direction	No collaboration

Table 1: Example of behavior and collaboration annotation categories from the MOSAIC-AI schema. Each behavior and collaboration category was rated on a Likert scale from 1 (observed none of the time) to 5 (observed all of the time for the moment preceding an intervention by JIA).

Factor	State 1: Parallel or Limited Interaction	State 2: Contributing to Shared Problem Space	State 3: Unproductive Perseverance
Students are walking through their answers		0.68	
Students are on task			-0.38
Students are working and then get stuck			0.64
Students are off task	0.51		
Students are sharing information		0.89	
Students ask questions			0.37
Students are offering or advancing ideas	-0.42		
Students build off others' ideas or paraphrase	-0.35		
No collaboration in group	0.75	-0.39	
Group's dialogue state	0.64		
Group's negotiation status			0.39

Table 2: States, factors, and loadings yielded by factor analysis. Only loadings $\geq |0.3|$ are displayed. The scores of the MOSAIC-AI annotation categories were used as factors to group collaborative states, which were then labeled by our CSCL experts.

and contributing to the collaborative task and aligns with characteristics of successful collaborations as demonstrated in previous research.

Unproductive Perseverance, as the third factor and collaborative state, had positive loadings from *Students are working and then get stuck* (0.64), *Students ask questions* (0.37), and the Group's *negotiation status* (0.39) and negative loading from *Students are on task* (-0.38). This state describes scenarios where students may be stuck and trying to overcome their difficulties or challenges (possibly without success) by asking questions of each other and negotiating, as observed in other studies [32, 56, 90]. If students are stuck for too long, it may be challenging for them to stay on task, as was observed in our study.

3.5 Modeling Indicators of Collaborative States and Creation of Feature Vectors

We collected discourse data via microphones that was then transcribed and analyzed in real time using a Multimodal Intelligent Analyzer (MMIA), following the work of [8]. The MMIA was designed specifically to evaluate collaboration in small groups. The MMIA allows flexible integration of individual analysis models including Automatic Speech Recognition (ASR), diarization, and a suite of automatic classification models including Off-Topic/Task [31], Collaborative Problem Solving (CPS) skills [65], and Community Agreement models (respect, committed to community, moving thinking forward) [8], which we describe next. Data is processed through the MMIA in 10s chunks.

For this study, data was transcribed using OpenAI's Whisper medium.en model [66] and diarized using an XVector model as implemented in the SincTDNN class in pyannote.audio library [7, 64] to extract speaker embeddings. Off-Topic and -Task are both binary classification models of utterances relating to a classroom-specific topic or activity [31]. These models result in a probability score from 0 (on-topic/task) to 1 (off-topic/task) per 10s chunk. The CPS classification model is based on a theoretically grounded, empirically validated CPS framework [81], which consists of three facets: shared knowledge construction, negotiation/coordination, and maintaining team function. The BERT-based model assigns probability scores from 0 to 1 according to each facet per utterance [65]. Importantly, this BERT model of CPS was trained on a wide array of classroom audio data of students working with Sensor Immersion, as well as other curricula. Finally, the community agreements model labels utterances with scores related to being respectful, showing commitment to community, and moving the group's thinking forward. For each agreement, a separate pre-trained RoBERTa model (also trained on Sensor Immersion classroom data) outputs a probability in the range [0, 1] for each 10s chunk of data. The model assigns scores to utterances during the 10s that may be considered an example of one of the community agreements, with probabilities greater than 0.5 signaling a positive match [8].

For each minute (six 10-second chunks) preceding an intervention message from the WoZ-SME, MMIA speech data is extracted and labeled with a state using the above factor loadings. To do so, each 10s chunk was given three 'collaborative state scores' that were computed using factor loadings (from Table 2) * MOSAIC item score[1]. Whichever collaborative state score was greatest dictated the collaboration state label for that chunk. Additionally, we extracted chunks of data from each session during times that an intervention was not delivered, as it is important to develop a set of rules that suggest when to intervene and when to not interrupt the collaboration. These chunks of data were labeled as 'Collaboration' and these were times where dialogue was productive, not warranting any intervention.

This resulted in four target labels of collaborative state: 1) Parallel or Limited Interaction, 2) Contributing to Shared Problem Space, 3) Unproductive Perseverance, and 4) Collaboration (where no interventions were sent). For every 10s chunk, a feature vector was created including the classification model scores and the resulting collaborative state label. Next, the feature vectors were used to train a decision tree to derive dialogue policy rules for each collaborative state, as described next.

3.6 Decision Tree Classification to Derive JIA's Dialogue Policy

We wished to design a dialogue policy using a data-enabled approach using the dialogue features, combined with labeled expert human feedback. Thus, a decision tree was trained using the MATLAB Classification Learner app. We employed five-fold cross-validation and 10% of the data was set aside for testing. The feature set included three CPS scores, three community agreement scores, off-task and off-topic scores, and verbosity (number of words spoken) for each 10-second window of time, with four classes predicted (the three collaborative states via factor analysis & the additional "Collaboration" state).

The number of feature vectors for each class were imbalanced, with 59% of features being labeled "Collaboration", 35% of features labeled "Contributing to Shared Problem Space", 4% of features labeled "Parallel or Limited Interaction" , and 2% of features labeled "Unproductive Perseverance". In our study, the latter two states occurred less often. There were more instances of when JIA did not need to intervene or did so by encouraging participants while they contributed to the shared problem space. To address class imbalance, the training data was balanced using the Synthetic Minority Over-sampling Technique (SMOTE) [13]. SMOTE generates synthetic samples for the minority class by interpolating between existing minority instances, thereby enhancing the model's ability to generalize and perform well on imbalanced datasets. "Parallel or Limited Interaction" and "Unproductive Perseverance" were treated as the minority classes and oversampled because, although rare, these are critical states for JIA to recognize and address. The decision tree was configured as an optimizable tree with a maximum of 30 splits, and grid search was used as the optimizer within the Matlab Learner application. We report the results of the decision tree in Table 3 and the test confusion matrix in Figure 4.

Metric	Result
Training Accuracy	62.8%
Test Accuracy	62%
Parallel or Limited Interaction AUC	0.8324
Contributing to Shared Problem Space AUC	0.6818
Unproductive Perseverance AUC	0.8891
Collaboration (No Intervention) AUC	0.7345

Table 3: Decision Tree Results. This table summarizes the performance metrics of the trained decision tree model used to derive JIA's dialogue policy. Training and test accuracy indicate the model's ability to classify collaborative states based on dialogue features. AUC values for each class reflect the model's discrimination ability across the collaborative states." The relatively high AUC for minority classes demonstrates the effectiveness of oversampling using SMOTE, despite their infrequent occurrence. These results support the model's capability to identify critical collaborative states for intervention.

When we consider that we have four class values to predict, random guess would achieve 25% performance. However, the test accuracy of 62% and relatively high AUC values suggest that the model can distinguish between the classes, especially "Unproductive Perseverance" (AUC = 0.8891) and "Parallel or Limited Interaction" (AUC = 0.8324). These two states are particularly critical, as they represent moments when students were struggling significantly and required intervention from JIA. Prioritizing the accurate detection of these states ensures that the system can intervene effectively to support students during these challenging moments.

[1]For example, the state score for Contributing to the Shared Problem Space would be calculated as follows: (Students are walking through their answers Score * 0.68) + (Students are sharing information Score * 0.89) + (No Collaboration Score * -0.39).

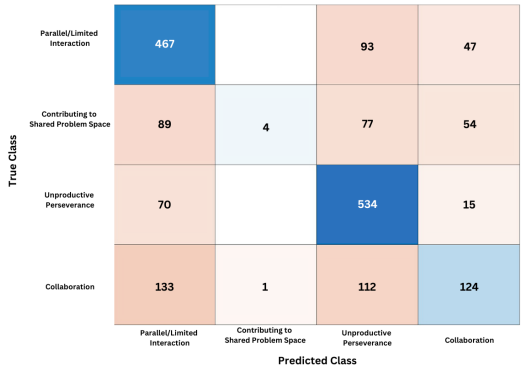

Figure 4: Test Confusion Matrix. The matrix illustrates the model's performance on the test set, which comprised 10% of the dataset held out during training. The distribution of true labels versus predicted labels indicates that "Contributing to the Shared Problem Space" was the most frequently misclassified category.

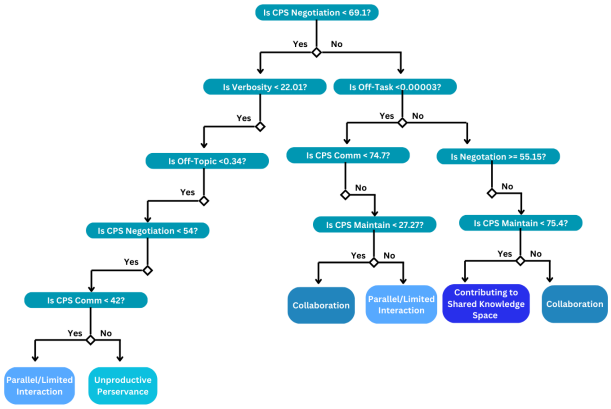

Figure 5: A pruned version of the decision tree. The tree classifies collaborative states based on dialogue features such as CPS metrics, CoBi agreements, off-topic/task scores, and verbosity. The decision tree was configured as an optimizable tree with a maximum of 30 splits, with the pruned version highlighting key decision points that inform JIA's interventions. The complete decision tree is provided in Appendix ??.

While the AUC for "Contributing to Shared Problem Space" was lower (0.6818), this may reflect an overlap in feature patterns between this class and others, leading to misclassifications. As show in the test confusion matrix (Figure 4), the model frequently predicted "Contributing to Shared Problem Space" as "Parallel/Limited Interaction" (89 instances), "Unproductive Perseverance" (77 instances), or "Collaboration" (54 instances). This suggests significant feature overlap among these states, particularly in verbosity and other shared collaboration metrics. We discuss how future work could work on refining features to better distinguish between states in Section 8.

3.7 Dialogue Policy Rule Implementation

To extract dialogue policy rules from the resulting tree, we begin at the root node and create rules using the following decision and leaf nodes. Due to paper space restrictions, we show a pruned version displaying only 5 splits in Figure 5 and the full decision tree is in Appendix **??**.

3.7.1 Implementation of a Timekeeper Function. Our rule-based dialogue policy derived from the decision tree outputs an updated state every 10 seconds. Even with the promising accuracies reported in Table 3, a state prediction every 10 seconds will produce many misclassified false positives, causing too many interventions to be sent, resulting in detrimental interruptions. Thus, we implemented an additional timekeeper function to act as a 'gatekeeper', to ensure that too many interventions are not sent. To guide the timekeeper function development, we looked at the timings of interventions made by our WoZ-SMEs. In the WoZ-SME studies, a total of 143 interventions were sent throughout 20 sessions. Most of the support was provided during the brainstorming activity (Q4) (51.05%), followed by knowledge sharing questions: Q2 (23.78%), Q1 (16.78%), and Q3 (8.39%). The average time between every two interventions sent by the WoZ-SME across all sessions was 3 minutes 25 seconds

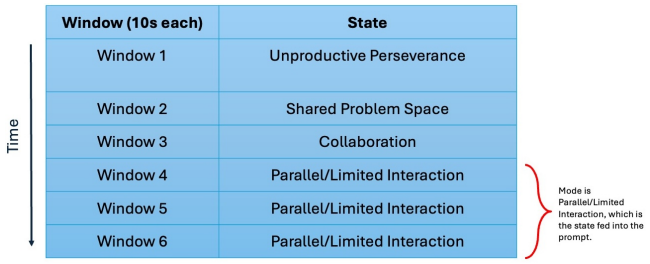

Figure 6: Timekeeper of states to ensure misclassifications by the decision tree models do not cause excessive interventions. For every one minute of utterances (six 10s chunks), the most common state in the queue is what is prompted to Mistral.

(SD=00:01:21). On average, around 7 interventions (Mean=7.15, SD=3.56) were sent to the students per session.

Based on the analysis of intervention timing from the WoZ-SME studies, we implemented a timekeeper function so that while states are generated every 10s, the most common state over the last minute is considered in the Mistral prompt (Figure 6). The timekeeper is an n-length queue (where n=6) that stores the last *n* states. Then, we take the mode of that queue; and, if there are multiple modes, we take whichever is most recent. Currently, the timekeeper is updated every 10s when a new chunk is added, and the oldest chunk is discarded. Mistral is prompted every 10s using the mode of the timekeeper.

4 Controllable Response Generation and Mistral Prompt Design

Within NLP, response generation is a complex task that involves taking the most recent utterance in an interaction, the surrounding

context, and providing an appropriate response. LLMs that have been trained on large amounts of conversational data have shown great promise in this space, able to engage users in interactive dialogues online. However, they are by nature trained to generalize well to a variety of conversation topics– which can make them ill-suited to domains that require highly specific responses, including the classroom setting. To better fit conversational models to this context, it is necessary to constrain the output based on a particular set of criteria or to leverage controllable response generation. In the case of the JIA system, the model is constrained by the dialogue policy. The flow of the dialogue system is as follows: every ten seconds the dialogue agent receives a chunk of data to process and respond to, this chunk includes the most recent student utterances and the dialogue state. This data is included in a prompt that is then sent to the LLM, Mistral [40]. The output from the model is then returned to the user via a web interface as detailed in Figure 7.

The conversational agent is backed by Mistral AI's 7B instruction-finetuned language model [40]. In a 2023 shared task focused on generating teacher-like responses, prompt-based models were notably high performers [82]. The winning system, NAISTeacher [94], was backed by GPT-3.5 Turbo [61]. However, the GPT family of models is not suitable for this use case for several reasons, the least of which being that it is a paid service. OpenAI explicitly states in Section 6 of their privacy policy that their service is not intended for children under the age of 13; if the service itself is not available to the target demographic, then it would be inappropriate to use the service in a context wherein they would interact with [62]. Furthermore, the data collected in this study is protected under the Institute Review Board, making it ineligible for use with a service that may be collecting prompt data. In addition to these ethical concerns, there is also a constraint on the compute resources since the system needs to be able to run on a laptop to be practical for classroom use. For these reasons, we sought out a high-performing open-source language model that would not stress the limited computational power. Mistral-7B outperforms other LLMs of comparable size, including Meta's Llama 2 [92], on a variety of language task benchmarks such as Hellaswag for commonsense reasoning and GSM8K for solving math word problems [96, 100]. Our choice to use Mistral over tools such as GPT-3.5 Turbo is another example of prioritizing responsible innovation over sheer computing or algorithmic power.

The two main techniques we use are dynamic prompt segments [63], where we can turn segments of the prompt on and off according to available data, and prompting templates [53], where we insert our own calculated features (e.g., states) directly into the prompt at execution. Please see Appendices **??** and **??** for the full text of our prompt templates. The active segments of our prompts are as follows:

(1) Preamble: Assigns the model an identity and primes it for further instructions.
(2) Setting & Role: Describes the classroom environment, lists behavior appropriate for an educational assistant, and provides current question text.
(3) Formatting: Restricts the length and content of output string.
(4) Context: Presents a recent conversation history and instructions for how to use it.

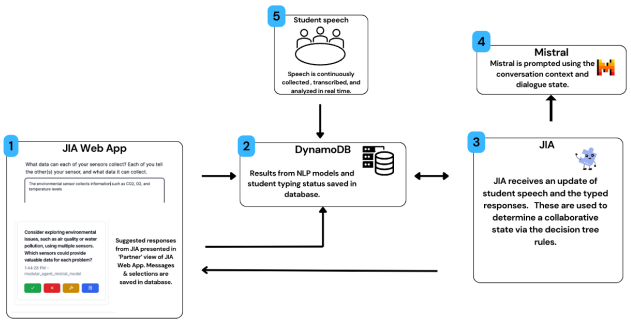

Figure 7: System Architecture. (1) Students type their response into the textbox of the JIA web app. (2) Data from the JIA web app and real-time speech analytics are saved to DynamoDB. (3) Dialogue state is determined from the decision tree results. (4) The dialogue state is used to prompt Mistral for a suggested message, which is then sent to the 'Partner' view of the JIA web app.

(5) Assignment: Assigns the model its task, including which state action to take.
(6) State Action: State actions are presented with three components: (1) a description of the current state, (2) the consequences of allowing that state to persist, and (3) suggested interventions for advancing the conversation.

Within the state action, we provide a suggested intervention based on the current state. To suggest an appropriate intervention type according to the collaborate state detected, we used both the collaborative state scores (described in Section 3.5) and MOSAIC annotations about the intervention type. We performed Pearson correlations between these state scores and intervention types (e.g., validation, explanation about the task) from the MOSAIC-AI schema to see which intervention type was most correlated to each collaborative state (full correlation matrix in Appendix **??**). For the state of limited/parallel interaction, the suggested intervention was to *direct their participation to the task* (0.14, $p<0.005$). While students were contributing to the shared problem space, the suggested response to was to *connect their discussion to a higher-level goal* (0.14, $p<0.005$) or to *ask a question* (0.19, $p<0.005$). Finally, when in unproductive perseverance, the suggested response was to *give an explanation or direction on task* (0.30, $p<0.005$) or *ask a question of the group* (0.21, $p<0.005$).

5 The JIA LLM-Based Agent and Addition of a Human-in-the-Loop Component

Real-time communication between the students and JIA was facilitated by a WebSocket layer built on top of AWS Lambda, using a DynamoDB database [76] to keep track of active connections and participants. Likewise, data and state updates from either the student or the partner, including answers, messages, and the current question, were recorded in real time and written to a DynamoDB database (Figure 7).

The system architecture shown in Figure 7 depicts our fully functional automated JIA. However, with our commitment to using

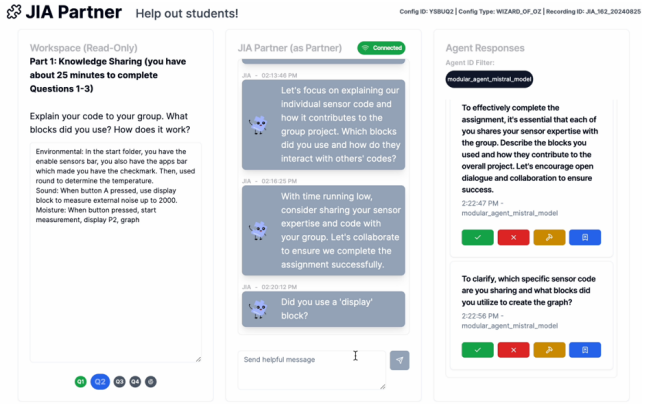

Figure 8: The 'Partner' view of the JIA interface. Left: Human observes student's writing in real time. Middle: The chatbot interface where messages are sent between students and JIA. Right: Suggested messages from Mistral populate the queue with options for the SME (i.e., 'human' in the loop) to accept, reject, modify, or ignore each message.

humans-in-the-loop (HITL) and responsible innovation [79], we opted to build out the capability for a human expert to moderate the interventions being sent to the group from JIA. The JIA-LLM-HITL interface is presented in Figure 8. For each suggested message from the JIA LLM-based agent, the SME has the choice of accepting, rejecting, modifying, or ignoring messages. 'Accept' would directly send the message to participants, while 'Reject' would delete the message from the queue and mark it as 'reject'. Modify would allow the SME to modify the message before sending and Ignore would remove the message and mark it as 'ignore'.

6 Evaluation of the JIA LLM Agent with Human-in-the-Loop

We know that the "gold standard" for an agent that supports group knowledge sharing and collaboration in a jigsaw activity is likely to be the subject matter expert (WoZ-SME), who has experience supporting students in classrooms on that jigsaw activity. A key goal is to develop a JIA-LLM based agent with HITL (JIA-LLM-HITL) that will achieve similar utility; both in its effectiveness and in its impact on user experiences, to the WoZ-SME agent. We expect that both human and LLM versions of JIA will outperform a control condition with no support provided to students.

We hypothesize that when compared against the Control condition, our Jigsaw Agent (JIA-LLM-HITL) and our WoZ-SME agent will lead to

- H1a) improved communication patterns in the group discourse
- H1b) improved user experiences, with a focus on more positive affect and perception of team processes (team cohesion, trust in AI).

Furthermore, we posit that the WoZ-SME condition will outperform the JIA-LLM-HITL on most, if not all, of these measures, with the WoZ-SME condition yielding

- H2a) improved communication patterns in the group discourse
- H2b) improved user experiences, with a focus on more positive affect and perception of team processes (team cohesion, trust in AI).

6.1 Experimental Design

We designed an experiment to investigate the hypotheses above. The experiment is a between-groups design with three conditions (Control, WoZ-SME, and JIA-LLM-HITL). Participants were children aged 12-17 recruited from Boulder, CO and the surrounding areas via recruitment flyers posted locally and on social media. Studies took place in a university research laboratory. All children and parents signed assent and consent documents approved by the university's institutional review board.

Participants were recruited to participate in the lab study as dyads and triads. After signing informed consent forms, they filled out self-report surveys reporting their demographics and computer literacy levels. Next, they worked together on the jigsaw activity described in Section 3.1, which included participants each learning about a specific sensor (sound, moisture, environmental) via individualized tutorials before working as a group on a shared computer to answer the open-ended jigsaw questions. The total study (including consent, assent, pre- and post-experiment surveys, and the jigsaw activity) took about 2 hours to complete.

Conditions included:

(1) Control condition ('Control'): Participants worked on the jigsaw activities without any supports provided.
(2) WoZ subject matter expert as JIA condition ('WoZ-SME'): This condition included the human SMEs playing the role of JIA (WoZ-SME), as first described in Section 3.1.
(3) JIA LLM Agent with HITL condition ('JIA-LLM-HITL'): Participants worked with JIA developed in Section 5, which includes the HITL capability (Figure 8) where our SMEs played the role of the *human* in the loop. The HITL was advised to 'Reject' messages that they would not send, because it didn't make sense, wasn't appropriate timing, or any other reason. 'Ignore' was instructed to be used when the message was appropriate and could have been sent, but the timing was not appropriate. The HITL could also 'Accept' messages to send them directly to the students or 'Modify' the message before sending it.

Fifty-eight groups of dyads and triads (145 participants), aged 12-17, participated in this study. The self-report survey data for groups in each condition are in Table 4. We collected data from 10 dyads and 11 triads in the control condition, 21 groups total (32 males, 19 females, 1 non-binary, 1 prefer not to say), and for the JIA-LLM-HITL condition, we collected data from 9 dyads and 8 triads, 17 groups total (28 males, 14 females). For the WoZ-SME condition, we collected data from 10 dyads and 10 triads, 20 groups total (26 males, 22 females, 2 non-binary). Across the 17 JIA-LLM-HITL sessions, there was an average of 6.29 accepted, 5.53 rejected, 3.29 ignored, and 1 modified response from the SME (which we discuss in Section 7).

6.1.1 Dependent Measures. The dependent measures included self-report surveys administered before and after the study and speech

Category	Details	Control (n=53)	HITL (n=42)	WoZ (n=50)
Group Type	Dyads	10	9	10
	Triads	11	8	10
	Total Groups	21	17	20
Gender	Male	32	28	26
	Female	19	14	22
	Non-binary	1	–	2
	Prefer not to say	1	–	–
Grade	6th	4	2	7
	7th	11	8	11
	8th	12	11	20
	9th	7	6	6
	10th	4	9	3
	11th	10	4	2
	12th	4	2	1
	Prefer not to say	1	–	–
Ethnicity	White	41	32	38
	Asian/Asian American or Pacific Islander	6	5	8
	Latin@/Hispanic	2	2	–
	White & Latin@/Hispanic	1	2	2
	White & Asian/Asian American or Pacific Islander	3	1	2
Main Language at Home	English Only	38	37	45
	English and Other*	9	5	4
	Non-English**	5	–	1
	Prefer not to say	1	–	–
Computer Literacy	Novice computer programmer	19	24	22
	Never computer programmed	17	12	20
	Computer programmer	16	6	8
	Expert computer programmer	1	–	–

Table 4: Demographics, Language, Ethnicity, and Computer Literacy Across Conditions. *Includes English combined with Spanish, Russian, Thai, Mandarin, Bulgarian, French, Hindi, or Arabic. **Includes Catalan, Mandarin, Russian, Tamil, and Tulu.

transcripts. Self-report measures included pre- and post-experiment survey data. Post-experiment survey measures included psychological safety [27], trust in teammate(s) and agent [57], positive group interaction and social loafing [51] and team processes [54, 55]. Audio was collected with Yeti microphones for the group and lapel microphones for individualized audio. For reduced latency, this data was transcribed using faster-whisper [2] (medium), which was derived from OpenAI's Whisper and operates 4x faster for the same accuracy while using less memory. We used ECAPA diarization [19] to further improve accuracy.

6.2 Discourse Results Comparison

Several of our hypotheses are investigated through analysis of the collaborative discourse using automated text analysis tools, which we report here. For all fifty-eight sessions, the transcripts were truncated to only include data from the collaborative jigsaw activity. Based on how much each group discussed, the length of these transcripts varied. On the truncated transcripts, we applied LIWC-22, which is a word-counting dictionary providing the percentage of words associated with a given psychologically relevant construct (e.g., affect-, authentic-, and analytic-related language) [6]. All LIWC category scores were averaged over each session. We also averaged the results of automatic discourse classification models (CPS, on/off topic, and CoBi agreements).

We conducted a one-way ANOVA to compare the effects of three conditions (Control, JIA-LLM-HITL, WoZ-SME) on all LIWC and discourse classification results. Significant differences were found in LIWC variables such as affect, authentic, analytic, positive tone, use of first-person singular pronouns, and all or none thinking, as well as community agreements of respectful collaboration (CoBi-Respect), and pushing our thinking forward (CoBi-Thinking).

6.2.1 LIWC Results. More specifically, we report the significant results from post-hoc Tukey HSD tests using LIWC on dialogue: First, the WoZ-SME condition demonstrated significantly higher levels of affect-related language compared to JIA-LLM-HITL (mean difference = 3.17, $p = 0.0417$), indicating that participants in the WoZ-SME condition exhibited more emotional language or sentiment.

[2]https://github.com/SYSTRAN/faster-whisper

Further, the WoZ-SME condition demonstrated significantly higher levels of positive tone language compared to JIA-LLM-HITL (mean difference = 3.53, p = 0.0218). Regarding analytic thinking, defined as a metric of logical, formal thinking, participants in WoZ-SME used significantly more analytic language than those in JIA-LLM-HITL (mean difference = 8.49, p = 0.007), reflecting deeper cognitive processing and confirming H2a of improved communication in WoZ-SME groups.

The JIA-LLM-HITL condition showed interesting differences of language use in comparison to the other conditions. JIA-LLM-HITL groups used significantly more all or none language ("all", "none", "never", "always") than WoZ-SME groups (mean difference = 0.72, p =0.0298). JIA-LLM-HITL groups used more words per sentence (WPS) than WoZ-SME and Control groups. JIA-LLM-HITL participants were also more authentic in their communication than both Control (mean difference = 5.98, p = 0.0228) and WoZ-SME groups (mean difference = -6.92, p = 0.0078), suggesting a more genuine or spontaneous style in the JIA-LLM-HITL groups. Finally, Control groups used 1st person singular pronouns ("me", "myself", "I") significantly more than JIA-LLM-HITL groups (mean difference = 2.40, p = 0.0005), indicating more individualized efforts rather than collaborative.

6.2.2 CoBi Discourse Classification Results. Regarding respectful collaboration (CoBi-Respect), the JIA-LLM-HITL groups were found to exhibit significantly higher levels of respect compared to both Control (mean difference = 0.0599, p = 0.0033) and WoZ-SME (mean difference = -0.0641, p = 0.0018) groups. Pushing our thinking forward (CoBi-Thinking) was also notably higher in JIA-LLM-HITL groups compared to Control groups (mean difference = 0.0438, p < 0.001) and WoZ-SME groups (mean difference = -0.0456, p < 0.001), indicating more collaborative and thoughtful dialogue in JIA-LLM-HITL groups.

These detailed post-hoc statistics further emphasize that the JIA-LLM-HITL condition fostered more thoughtful, respectful, and engaged communication, with the WoZ-SME condition displaying higher emotional expression but less authentic and respectful interactions.

6.3 Self-Report Survey Results

To investigate self-report survey differences across the three conditions (Control, JIA-LLM-HITL, and WoZ-SME), we first averaged self-report scores of social loafing, psychological safety, positive interaction, team processes, and trust in teammate for each group. We used the Shapiro-Wilk test to assess whether the data was normally distributed. Social loafing, psychological safety and team processes were confirmed to be normally distributed with p-values exceeding the 0.05 threshold.

We proceeded with a one-way ANOVA with the normally distributed measures to examine whether there were statistically significant differences in the group means across the three conditions (see Table 5). The analysis revealed a statistically significant difference between the groups for the measure of social loafing (F = 3.69, p = 0.031). However, no significant differences were found for the other measures: psychological safety (F = 0.36, p = 0.701) and team processes (F = 0.74, p = 0.482).

Measure	H-statistic	p-value
Social Loafing	3.69	**0.031***
Psychological Safety	0.36	0.701
Team Processes	0.74	0.482

Table 5: One-Way ANOVA Results for the self-reported survey measures of social loafing, psychological safety, and team processes.

There was only a significant difference found for social loafing (p = 0.031), suggesting that this measure varies significantly across the three conditions. To further explore the significant finding for social loafing, we conducted a Tukey HSD post-hoc test to determine which specific groups differed from each other. The results showed that there is a statistically significant difference in the means between the Control and WoZ-SME groups (p = 0.0369), with the WoZ-SME group having less social loafing compared to the Control group. This somewhat confirms H1b, with less indication of social loafing in the WoZ-SME condition. However, there are no significant differences between the Control and JIA-LLM-HITL groups or between the JIA-LLM-HITL and WoZ-SME groups.

For survey measures that were not normally distributed (positive interaction, trust in teammate), we conducted a Kruskal-Wallis H test to compare the differences across experimental conditions. The results revealed no significant differences between the experimental conditions for positive interaction average (H(2) = 0.72, p = 0.699) or trust in teammate (H(2) = 0.21, p = 0.899).

To investigate participants' trust in the agent, we were only able to compare the JIA-LLM-HITL and WoZ-SME conditions as the Control condition did not have an agent, thus results are presented for those two conditions. For both the WoZ-SME and JIA-LLM-HITL conditions, participants rated their trust in the agent. Because these measures were not normally distributed, we used the Mann-Whitney U test to compare trust ratings between the two conditions. The test revealed that there was no significant difference in trust in agent ratings between the JIA-LLM-HITL and WoZ-SME conditions (U-statistic = 917.5, p-value = 0.282). This indicates that participants' trust in the agent was similar across both conditions, rejecting H2b that the WoZ-SME would result in greater trust in the agent than JIA-LLM-HITL.

7 Discussion

The proliferation of LLMs in society poses a new opportunity to promote productive collaboration by designing pedagogical agents to support students in the classroom. However, the use of LLMs in this context raises serious ethical questions and concerns. The goal of this work was to explore how to create a responsible and effective pedagogical agent by incorporating expert human feedback. In this section, we discuss the results from our evaluation study, review the key activities rooted in responsible innovation and HCAI undertaken, and further discuss how incorporating LLMs in future work could utilize the same HCAI approaches and framework.

Our results show that the JIA-LLM-HITL condition fostered more thoughtful, respectful, and engaged communication than those interacting with a WoZ-SME. Specifically, LIWC analyses on the

groups' communication showed lower emotional expression but more authentic and respectful interactions in their communication patterns when they were supported by the JIA-LLM-HITL agent. This is certainly a promising sign, indicating that the JIA-LLM-HITL agent produced quality dialogue that was both authentic and respectful. These differences in CoBi and LIWC results may be due to the formal responses offered by the LLM vs. the WoZ-SME. For example, the WoZ-SMEs more often encouraged participants when they collaborated effectively, whereas the JIA-LLM-HITL agent more often offered detailed, task-related information to move the groups' thinking forward. To more closely mimic the WoZ-SME interventions, future work could revise the prompt for each collaborative state and also instruct the model to give succinct responses when possible. While the discourse between conditions differed, participants' self-reported trust in the agent was similar for the WoZ-SME and the JIA-LLM-HITL conditions, indicating that both the human-written and LLM-generated responses (with human input) were trusted by participants.

Notably, participants in the WoZ-SME group reported significantly less social loafing compared to the Control group, but there were no significant differences between the Control and JIA-LLM-HITL groups or between the JIA-LLM-HITL and WoZ-SME groups. This suggests that the WoZ-SME likely picked up on social loafing nuances in the group that the JIA-LLM-HITL did not detect. While the SME observed the group over video conferencing during the task for both JIA-LLM-HITL and WoZ-SME conditions, in the WoZ-SME condition, they were able to quickly type a response if they picked up on a nonverbal behavior indicating social loafing (e.g., slumped back in chair, not engaged). On the other hand, the LLM did not have this visual information to consider and thus did not pick up on social loafing as quickly or efficiently relying solely on speech. In the future work section, we revisit these findings to discuss future work in longitudinal classroom environments that may be better suited to see effects where our one-session lab experiment did not.

Throughout the JIA-LLM-HITL condition, the SME most often accepted the suggested messages rather than reject, modify, or ignore. However, they rejected an average of 5.53 messages per session. We found that the SMEs disagreed with some of the LLM-generated responses because they made assumptions (hallucinations) about the task that were not true (e.g., the LLM would talk about light sensors, however this was not a sensor that students learned about during the task) or relevant (e.g., suggest that the students look back at the written material, which was not an actual option during the study). Only 1 message, on average, was modified by the SME before being sent, suggesting that most suggestions required little adjustment or were either fully accepted or dismissed. The most common reason for modification was to shorten the LLM-generated response, as they often were superfluous and failed to mimic a human's succinctness.

Regarding responsible innovation [79] and the need to include prospective and retrospective transparency in AI development [9], we take several important steps. We first review our process using the four dimensions of anticipation, reflexivity, inclusion, and responsiveness by taking a human-centered approach. We adhered to the dimensions of anticipation (anticipating the consequences and gains of the ever-evolving technological progress), reflexivity (to be reflexive and consider the moral responsibilities of our work), inclusivity (to promote research that extends out to the wider public), and responsiveness (being responsive and adapting to the fluctuation of public views and consistent growth of science and discovery) via our HCAI approach to the development of JIA. Specifically, we anticipated the consequences of introducing an AI partner in classrooms by carefully selecting an appropriate LLM and testing several prompting templates before finalizing the one presented here. Throughout the design process of JIA, we continuously considered the moral responsibilities of designing a pedagogical agent including ensuring inclusivity and accessibility for diverse learners and fostering a respective learning environment. Finally, we were responsive to public views towards AI and LLMs by carefully choosing a transparent LLM with safeguards in place.

By way of our HCAI approach, we ensured that SMEs engaged directly in the development of our dialogue policy and LLM response generation, for the means of both prospective and retrospective transparency. The SMEs first played the role of the WoZ-SMEs in our early development studies, and employed MOSAIC annotation schema to identify groups' collaborative states for the AI agent to monitor that were meaningful to teachers and learning scientists. Once the collaborative states were labeled, the SMEs also helped to craft the prompt template for each state based on CSCL literature. This choice to keep our dialogue policy and LLM prompts transparent and explainable to learning scientists is a key aspect showcasing this commitment to responsible innovation. However, it important to note the tradeoff between balancing the decision tree's accuracy with deriving collaborative states and subsequent features using a pedagogical schema (MOSAIC). It was important to involve the SMEs in several steps of the design process and to transparently label the states so that the final model is accessible to several disciplines, but this came at the cost of lower model accuracy. We note this as a limitation and discuss how the model could be improved in Section 8.

This study demonstrates an approach of designing a LLM-based conversational agent for use with children, that is dedicated to responsibility and ethics. With the recent boom of LLMs and mixed feelings from the public about them, it is vital for researchers to be transparent when developing AI tools with such models. Transparency not only builds trust but also promotes greater accessibility and accountability in the adoption of these technologies. This work highlights how the responsible innovation framework can guide the ethical and transparent integration of LLMs, setting a precedent for future applications in educational and collaborative AI tools.

8 Limitations

We note several limitations in our work regarding the experimental design and interpretability of the decision tree model. First, these studies were conducted in a single two-hour long laboratory session per group, while the jigsaw activity from the sensor immersion curriculum is usually included in 5 class lessons. Our single session design may have limited our ability to consider key learning outcomes, knowledge sharing, and collaboration that would occur more naturally between student groups and their AI supports over the course of time in classroom settings. Future longitudinal studies

should be considered to better assess the effects of JIA's support in longitudinal settings.

Second, our data collections occurred with one group at a time, in a controlled lab setting, with students of different ages. While this approach allowed us to reach a higher sample size, it does not guarantee transfer to the noisiness of the naturalistic classroom environment, nor the traditional classroom context where students are typically around the same age. However, the 2017 NEAP Report on assessing CPS skills discusses how heterogeneous naturally teams naturally in classrooms with a wide variation in background knowledge, culture, maturity, and social skills [29]. Further, the 2015 PISA framework of CPS states that students must be prepared to work effectively within heterogeneous groups of familiar and unfamiliar members in real life [60]. We aimed to design JIA to aid these heterogeneous groups build their CPS skills through knowledge sharing and learning. Future work should move beyond the lab environment and test these tools in authentic classroom settings, where factors such as different group dynamics may influence the outcomes. Additionally, future work could separately study groups of middle-school or high-school students to better understand how JIA extends to each population and adapts accordingly.

Third, our HITL capability was only used for the human SMEs to provide oversight over the LLM-based JIA agent. There is a missed opportunity here to utilize the HITL evaluative feedback (e.g., accept, reject) as evaluative feedback that could be directly incorporated into the learning algorithms, opening the door for forms of interactive machine learning interactive reinforcement learning [1, 46]. Indeed, the HITL evaluative feedback interface was designed with human evaluative feedback in mind, and we will be incorporating that important component into future work.

We also note several limitations regarding the ML model performance. While decision trees are inherently interpretable and well-suited for our initial exploration, they are not the only interpretable machine learning models available. The test accuracy of 62% is relatively low for practical applications, especially in the context of JIA's real-time intervention requirements. The imbalanced training dataset, coupled with feature overlap among classes, contributed to a lower AUC for certain classes, such as "Contributing to Shared Problem Space" (AUC = 0.6818). To address these issues, future work could collect more labeled data for the minority classes and refine feature engineering to reduce overlap among classes. Future work could explore more complex yet interpretable models (e.g., Random Forests, Logistic Regression) to improve model performance, therefore increasing the reliability and trustworthiness of JIA's dialogue system.

9 Conclusions & Ethical Considerations

In this work, we presented the human-centered design and evaluation of an LLM-based agent to facilitate small group collaboration in middle- and high-school classrooms. We evaluated our agent with one-hundred and forty five participants aged 12-17, grouped into fifty-eight groups of dyads and triads, and placed in groups representing four conditions. These studies yielded promising results showing that when students interacted with a an LLM-based agent combined with a human-in-the-loop they had highly engaged and thoughtful conversations, more so than when they interacted with the WoZ-SME. The results also showed (via self-report surveys) that participants in the WoZ-SME group had significantly less social loafing compared to the Control group. This indicates that participants' trust in the agent was similar for the WoZ-SME and the JIA-LLM-HITL conditions. These results show the potential of designing AI supports for small groups in classrooms through responsible innovation and HCAI processes that value and integrate input from key domain experts from classroom contexts. The need for these transparent and meaningful collaborations between AI developers and key domain experts has been noted, and identified as non-trivial. Our work provides a roadmap to showcase how design processes can indeed be undertaken that adhere to these visionary goals, resulting in effective, ethical, and transparent AI.

Ethical considerations must continue to be carefully reviewed in future work. In particular, researchers should take into account how LLMs are used in classrooms where student privacy is paramount and LLM hallucinations can cause potential harm. With this in mind, our own future work will consider the use of conjecture mapping to further instantiate key tenets of the responsible innovation framework. Conjecture mapping is a tool that can be leveraged to implement the responsible innovation framework, by outlining a set of potential uses and misuses of AI partners that are driven by theory. Within the learning sciences, conjecture maps [70] are an established method for representing explicit linkages between proposed interventions, mediating processes, and intended outcomes. The path forward for LLM-based conversational agents to support students' to develop key collaboration skills is an exciting one, but with great power comes great responsibility. The work presented here is intended to continue the important work in the HCAI community to pave a path toward responsible innovation of LLM-based conversational agents in classrooms, to support students to develop the crucial collaboration skills necessary in the 21st century workforce.

Acknowledgments

This research was supported by the NSF National AI Institute for Student-AI Teaming (iSAT) under grant DRL 2019805. The opinions expressed are those of the authors and do not represent views of NSF.

References

[1] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *AI magazine* 35, 4 (2014), 105–120.

[2] Elliot Aronson. 2002. Building Empathy, Compassion, and Achievement in the Jigsaw Classroom. In *Improving Academic Achievement*, Joshua Aronson (Ed.). Academic Press, San Diego, 209–225.

[3] Amid Ayobi, Jacob Hughes, Christopher J. Duckworth, Jakub J. Dylag, Sam James, Paul Marshall, Matthew Guy, et al. 2023. Computational Notebooks as Co-Design Tools: Engaging Young Adults Living with Diabetes, Family Carers, and Clinicians with Machine Learning Models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–20.

[4] Michael J. Baker. 2015. Collaboration in Collaborative Learning. *Interaction Studies* 16, 3 (2015), 451–473.

[5] Brigid Barron and David A. Sears. 2002. Advancing Understanding of Learning in Interaction: How Ways of Participating Can Influence Joint Performance and Learning. In *International Conference on Computer Supported Collaborative Learning*. 593–594.

[6] Ryan L Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. 2022. The development and psychometric properties of LIWC-22. (2022), 1–47.

[7] Hervé Bredin. 2023. Pyannote.auDio 2.1 Speaker Diarization Pipeline: Principle, Benchmark, and Recipe. In *Proc. INTERSPEECH*.

[8] Thomas Breideband, Jeffrey Bush, Chelsea Chandler, Michael Chang, Rachel Dickler, Peter Foltz, Ananya Ganesh, et al. 2023. The Community Builder (CoBi): Helping Students to Develop Better Small Group Collaborative Learning Skills. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*. Association for Computing Machinery, New York, NY, USA, 376–380.

[9] Alexander Buhmann and Christian Fieseler. 2021. Towards a Deliberative Framework for Responsible Innovation in Artificial Intelligence. *Technology in Society* 64 (2021), 101475.

[10] Jenna Burrell. 2016. How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms. *Big Data & Society* 3, 1 (2016), 205395171562251.

[11] Tara Capel and Margot Brereton. 2023. What Is Human-Centered about Human-Centered AI? A Map of the Research Landscape. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–23.

[12] Liwei Chan, Yi-Chi Liao, George B. Mo, John J. Dudley, Chun-Lien Cheng, Per Ola Kristensson, and Antti Oulasvirta. 2022. Investigating Positive and Negative Qualities of Human-in-the-Loop Optimization for Designing Interaction Techniques. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA. https://doi.org/10.1145/3491102.3501850

[13] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. SMOTE: Synthetic Minority over-Sampling Technique. *The Journal of Artificial Intelligence Research* 16 (2002), 321–357.

[14] Thomas K. F. Chiu. 2024. Future Research Recommendations for Transforming Higher Education with Generative AI. *Computers and Education: Artificial Intelligence* 6 (2024), 100197.

[15] Nancy J Cooke. 2015. Team Cognition as Interaction. *Curr. Dir. Psychol. Sci.* 24, 6 (1 Dec. 2015), 415–419.

[16] Nancy J Cooke, Myke C Cohen, Walter C Fazio, Laura H Inderberg, Craig J Johnson, Glenn J Lematta, Matthew Peel, and Aaron Teo. 2024. From teams to teamness: Future directions in the science of team cognition. *Human Factors* 66, 6 (2024), 1669–1680.

[17] N. Dahlbäck, A. Jönsson, and L. Ahrenberg. 1993. Wizard Oz Studies: Why How?. In *Proceedings of the 1st International Conference on Intelligent User Interfaces*. New York, NY, USA, 193–200.

[18] Kerstin Dautenhahn. 1998. The Art of Designing Socially Intelligent Agents: Science, Fiction, and the Human in the Loop. *Applied Artificial Intelligence* 12, 7–8 (1998), 573–617.

[19] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. http://arxiv.org/abs/2005.07143. *ArXiv* 2005.07143 (2020).

[20] I. Dey, N. Hoang, and J. B. Bush. 2024. Analyzing Support Moments During Small Group Work [Poster Presentation]. In *American Educational Research Association*.

[21] Pierre Dillenbourg. 1999. What Do You Mean by Collaborative Learning. In *Collaborative-Learning: Cognitive and Computational Approaches*. Elsevier Science, Oxford, England, 1–19.

[22] Pierre Dillenbourg, Sanna Järvelä, and Frank Fischer. 2009. The Evolution of Research on Computer-Supported Collaborative Learning. In *Technology-Enhanced Learning: Principles and Products*, Nicolas Balacheff, Sten Ludvigsen, Ton de Jong, Ard Lazonder, and Sally Barnes (Eds.). Springer Netherlands, Dordrecht, 3–19.

[23] P. Dillenbourg and P. Tchounikine. 2007. Flexibility in Macro-scripts for Computer-supported Collaborative Learning: Flexibility in Macro-Scripts for CSCL. *Journal of Computer Assisted Learning* 23, 1 (2007), 1–13.

[24] Sidney K D'Mello, Nicholas Duran, Amanda Michaels, and Angela E B Stewart. 2024. Improving collaborative problem-solving skills via automated feedback and scaffolding: a quasi-experimental study with CPSCoach 2.0. *User Model. User-adapt Interact.* (14 Feb. 2024).

[25] Melissa C. Duffy and Roger Azevedo. 2015. Motivation Matters: Interactions between Achievement Goals and Agent Scaffolding for Self-Regulated Learning within an Intelligent Tutoring System. *Computers in Human Behavior* 52 (November 2015), 338–348.

[26] Sidney K. D'Mello, Quentin Biddy, Thomas Breideband, Jeffrey Bush, Michael Chang, Arturo Cortez, Jeffrey Flanigan, et al. 2024. From Learning Optimization to Learner Flourishing: Reimagining AI in Education at the Institute for Student-AI Teaming (ISAT). *AI Magazine* 45, 1 (2024), 61–68.

[27] Amy Edmondson. 1999. Psychological Safety and Learning Behavior in Work Teams. *Administrative Science Quarterly* 44, 2 (1999), 350–383.

[28] Stephen M. Fiore, Arthur Graesser, and Samuel Greiff. 2018. Collaborative Problem-Solving Education for the Twenty-First-Century Workforce. *Nature Human Behaviour* 2, 6 (2018), 367–369.

[29] S M Fiore, A Graesser, S Greiff, P Griffin, B Gong, and others. 2017. Collaborative problem solving: Considerations for the national assessment of educational progress. (2017).

[30] Mathias Funk, Peter Lovei, and Renee Noortman. 2024. Designing with Data, Data-Enabled and Data-Driven Design. In *Handbook of Human Computer Interaction*. Springer International Publishing, Cham, 1–32.

[31] Ananya Ganesh, Michael Alan Chang, Rachel Dickler, Michael Regan, Jon Cai, Kristin Wright-Bettner, James Pustejovsky, et al. 2023. Navigating Wanderland: Highlighting Off-Task Discussions in Classrooms. In *Artificial Intelligence in Education*. Springer Nature Switzerland, 727–732.

[32] Gloriana González and Anna F DeJarnette. 2015. Teachers' and students' negotiation moves when teachers scaffold group work. *Cognition and Instruction* 33, 1 (2015), 1–45.

[33] Arthur C. Graesser, Stephen M. Fiore, Samuel Greiff, Jessica Andrews-Todd, Peter W. Foltz, and Friedrich W. Hesse. 2018. Advancing the Science of Collaborative Problem Solving. *Psychological Science in the Public Interest: A Journal of the American Psychological Society* 19, 2 (2018), 59–92.

[34] Crina Grosan and Ajith Abraham. 2011. Rule-Based Expert Systems. In *Intelligent Systems*, Janusz Kacprzyk and Lakhmi C. Jain (Eds.). Intelligent Systems Reference Library, Vol. 17. Springer Berlin Heidelberg, Berlin, Heidelberg, 149–185.

[35] Agneta Gulz, Magnus Haake, Annika Silvervarg, Björn Sjödén, and George Veletsianos. 2011. Building a Social Conversational Pedagogical Agent: Design Challenges and Methodological Approaches. In *Conversational Agents and Natural Language Interaction*. IGI Global, 128–155.

[36] Cindy E. Hmelo-Silver and Howard S. Barrows. 2008. Facilitating Collaborative Knowledge Building. *Cognition and Instruction* 26, 1 (2008), 48–94.

[37] N. Hoang, J. B. Bush, and I. Dey. 2024. Support Patterns in Classrooms Implementing a Computer Science and Physical Computing Curriculum. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 2*. 1674–1675.

[38] Nga Hoang, Jeffrey B. Bush, Indrani Dey, Emily Watts, Charis Clevenger, and William R. Penuel. 2024. MOSAIC Protocol: Analyzing Small Group Work to Gain Insights into Collaboration Support for Middle School STEM Classrooms. In *Proceedings of the International Conference on Computer-Supported Collaborative Learning*. International Society of the Learning Sciences. https://doi.org/10.22318/cscl2024.344145

[39] Heisawn Jeong, Cindy E. Hmelo-Silver, and Kihyun Jo. 2019. Ten Years of Computer-Supported Collaborative Learning: A Meta-Analysis of CSCL in STEM Education during 2005–2014. *Educational Research Review* 28 (2019), 100284.

[40] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, et al. 2023. Mistral 7B. *ArXiv [Cs.CL]* (2023). http://arxiv.org/abs/2310.06825

[41] M. O. Kaiser. 1974. Kaiser-Meyer-Olkin Measure for Identity Correlation Matrix. *Journal of the Royal Statistical Society* 52, 1 (1974), 296–298.

[42] Sheer Karny, Lukas William Mayer, Jackie Ayoub, Miao Song, Haotian Su, Danyang Tian, Ehsan Moradi-Pari, and Mark Steyvers. 2024. Learning with AI Assistance: A Path to Better Task Performance or Dependence?. In *Proceedings of the ACM Collective Intelligence Conference*. ACM, New York, NY, USA, 10–17.

[43] Reet Kasepalu, Luis P. Prieto, Tobias Ley, and Pankaj Chejara. 2022. Teacher Artificial Intelligence-Supported Pedagogical Actions in Collaborative Learning Coregulation: A Wizard-of-Oz Study. *Frontiers in Education* 7 (2022). https://doi.org/10.3389/feduc.2022.736194

[44] J. F. Kelley. 1984. An Iterative Design Methodology for User-Friendly Natural Language Office Information Applications. *ACM Transactions on Information Systems* 2, 1 (1984), 26–41.

[45] R. Ellis Kerly and S. Bull. 2009. Conversational Agents E-Learning. In *Proceedings AI 2008*. 169–182.

[46] W Bradley Knox and Peter Stone. 2008. Tamer: Training an agent manually via evaluative reinforcement. In *2008 7th IEEE international conference on development and learning*. IEEE, 292–297.

[47] Jv Kollenburg and S. Bogers. 2019. Data-Enabled Design.

[48] Mohammad Amin Kuhail, Nazik Alturki, Salwa Alramlawi, and Kholood Alhejori. 2023. Interacting with Educational Chatbots: A Systematic Review. *Education and Information Technologies* 28, 1 (2023), 973–1018.

[49] R. Kumar and Carolyn P. Rose. 2011. Architecture for Building Conversational Agents That Support Collaborative Learning. *IEEE Transactions on Learning Technologies* 4, 1 (2011), 21–34.

[50] Vishesh Kumar and Mike Tissenbaum. 2022. Supporting Collaborative Classroom Networks through Technology: An Actor Network Theory Approach to Understanding Social Behaviours and Design. *British Journal of Educational Technology: Journal of the Council for Educational Technology* 53, 6 (2022), 1549–1570.

[51] Lisa Linnenbrink-Garcia, Toni Kempler Rogat, and Kristin LK Koskey. 2011. Affect and engagement during small group instruction. *Contemporary Educational Psychology* 36, 1 (2011), 13–24.

[52] K. Littleton and N. Mercer. 2010. The Significance of Educational Dialogues between Primary School Children. In *Educational Dialogues*. https://doi.org/10.4324/9780203863510-29/significance-educational-dialogues-primary-school-children-karen-littleton-neil-mercer

[53] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *Comput. Surveys* 55, 9 (2023), 1–35.

[54] Margaret M. Luciano, Leslie A. DeChurch, and John E. Mathieu. 2018. Multiteam Systems: A Structural Framework and Meso-Theory of System Functioning. *Journal of Management* 44, 3 (2018), 1065–1096.

[55] Michelle A. Marks, John E. Mathieu, and Stephen J. Zaccaro. 2001. A Temporally Based Framework and Taxonomy of Team Processes. *Academy of Management Review* 26, 3 (2001), 356–376.

[56] Dean B. McFarlin, Roy F. Baumeister, and Jim Blascovich. 1984. On Knowing When to Quit: Task Failure, Self-esteem, Advice, and Nonproductive Persistence. *Journal of Personality* 52, 2 (1984), 138–155.

[57] Stephanie M. Merritt. 2011. Affective Processes in Human–Automation Interactions. *Human Factors* 53, 4 (2011), 356–370.

[58] S. Michaels, M. C. O'Connor, M. W. Hall, and L. B. Resnick. 2010. Accountable Talk Sourcebook: For Classroom That Works (v.3.1). https://www.fredhutch.org/content/dam/www/about-us/education/sep/2020/AT-Sourcebook.pdf

[59] Renee Noortman, Peter Lovei, Mathias Funk, Eva Deckers, Stephan Wensveen, and Berry Eggen. 2022. Breaking up Data-Enabled Design: Expanding Scaling up Clinical Context. *Artificial Intelligence Engineering Design* 36, 1 (2022), 1–13.

[60] OECD. 2017. PISA 2015 Results: Collaborative Problem Solving.

[61] OpenAI. 2023. GPT-3.5 Turbo Fine-Tuning and API Updates. https://openai.com/index/gpt-3-5-turbo-fine-tuning-and-api-updates

[62] OpenAI. 2024. Privacy Policy. https://openai.com/policies/privacy-policy

[63] E. Margaret Perkoff, Angela Maria Ramirez, Sean von Bayern, Marilyn A. Walker, and James Martin. 2024. 'Keep up the Good Work!': Using Constraints in Zero Shot Prompting to Generate Supportive Teacher Responses. In *25th Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 24)*. Association for Computational Linguistics.

[64] Alexis Plaquet and Hervé Bredin. 2023. Powerset multi-class cross entropy loss for neural speaker diarization. In *Proc. INTERSPEECH 2023*.

[65] Samuel L. Pugh, Shree Krishna Subburaj, Arjun Ramesh Rao, Angela E. B. Stewart, Jessica Andrews-Todd, and Sidney K. D'Mello. 2022. Say What? Automatic Modeling of Collaborative Problem Solving Skills from Student Speech in the Wild. https://files.eric.ed.gov/fulltext/ED615653.pdf

[66] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. *ArXiv [Eess.AS]* (2022). http://arxiv.org/abs/2212.04356

[67] Angela Ramirez, Karik Agarwal, Juraj Juraska, Utkarsh Garg, and Marilyn A. Walker. 2023. Controllable Generation of Dialogue Acts for Dialogue Systems via Few-Shot Response Generation and Ranking. *arXiv [Cs.CL]* (2023). http://arxiv.org/abs/2307.14440

[68] Jeremy Roschelle. 2013. Special Issue on CSCL: Discussion. *Educational Psychologist* 48, 1 (2013), 67–70.

[69] Jeremy Roschelle and Stephanie D. Teasley. 1995. The Construction of Shared Knowledge in Collaborative Problem Solving. In *Computer Supported Collaborative Learning*. Springer Berlin Heidelberg, Berlin, Heidelberg, 69–97.

[70] William Sandoval. 2014. Conjecture Mapping: An Approach to Systematic Educational Design Research. *Journal of the Learning Sciences* 23, 1 (2014), 18–36.

[71] Marlene Scardamalia and Carl Bereiter. 2006. Knowledge Building: Theory, Pedagogy, and Technology. In *Cambridge Handbook of the Learning Sciences*, K. Sawyer (Ed.). Cambridge University Press, 97–118.

[72] SchoolWide Labs, University of Colorado Boulder. 2023. Sensor Immersion. https://www.colorado.edu/program/schoolwide-labs/sensor-immersion

[73] Sandra Schulz, Bruce M. McLaren, and Niels Pinkwart. 2022. Towards a Tutoring System to Support Robotics Activities in Classrooms – Two Wizard-of-Oz Studies. *International Journal of Artificial Intelligence in Education* (2022). https://doi.org/10.1007/s40593-022-00305-2

[74] Ben Shneiderman. 2022. *Human-centered AI.* Oxford University Press, London, England.

[75] Pieta Sikström, Chiara Valentini, Anu Sivunen, and Tommi Kärkkäinen. 2022. How Pedagogical Agents Communicate with Students: A Two-Phase Systematic Review. *Computers & Education* 188 (2022), 104564.

[76] Swaminathan Sivasubramanian. 2012. Amazon dynamoDB: a seamlessly scalable non-relational database service. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*. 729–730.

[77] Gerry Stahl and Kai Hakkarainen. 2021. Theories of CSCL. In *International Handbook of Computer-Supported Collaborative Learning*, Ulrike Cress, Carolyn Rosé, Alyssa Friend Wise, and Jun Oshima (Eds.). Springer International Publishing, Cham, 23–43.

[78] Angela E. B. Stewart, Arjun Rao, Amanda Michaels, Chen Sun, Nicholas D. Duran, Valerie J. Shute, and Sidney K. D'Mello. 2023. CPSCoach: The Design and Implementation of Intelligent Collaborative Problem Solving Feedback. In *Artificial Intelligence in Education*. Springer Nature Switzerland, 695–700.

[79] Jack Stilgoe, Richard Owen, and Phil Macnaghten. 2013. Developing a Framework for Responsible Innovation. *Research Policy* 42, 9 (2013), 1568–1580.

[80] Christian M. Stracke, Irene-Angelica Chounta, Vania Dimitrova, Beth Havinga, and Wayne Homes. 2024. Ethical AI and Education: The Need for International Regulation to Foster Human Rights, Democracy and the Rule of Law: 25th International Conference, AIED 2024, Recife, Brazil, July 8–12, 2024, Proceedings, Part II. In *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky (Communications in Computer and Information Science, Vol. 2151)*, Andrew M. Olney, Irene-Angelica Chounta, Zitao Liu, Olga C. Santos, and Ig Ibert Bittencourt (Eds.). Springer Nature Switzerland, Cham, 439–445.

[81] Chen Sun, Valerie J. Shute, Angela Stewart, Jade Yonehiro, Nicholas Duran, and Sidney D'Mello. 2020. Towards a Generalized Competency Model of Collaborative Problem Solving. *Computers & Education* 143 (2020), 103672.

[82] Anaïs Tack, Ekaterina Kochmar, Zheng Yuan, Serge Bibauw, and Chris Piech. 2023. The BEA 2023 Shared Task on Generating AI Teacher Responses in Educational Dialogues. *ArXiv [Cs.CL]* (2023). http://arxiv.org/abs/2306.06941

[83] Seng Chee Tan, Carol Chan, Katerine Bielaczyc, Leanne Ma, Marlene Scardamalia, and Carl Bereiter. 2021. Knowledge Building: Aligning Education with Needs for Knowledge Creation in the Digital Age. *Educational Technology Research and Development: ETR & D* 69, 4 (2021), 2243–2266.

[84] Jingwan Tang, Xiaofei Zhou, Xiaoyu Wan, and Fan Ouyang. 2022. A Systematic Review of AI Applications in Computer-Supported Collaborative Learning in STEM Education. In *Artificial Intelligence in STEM Education*. 333–358.

[85] Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology* 29, 1 (2010), 24–54.

[86] Stergios Tegos, Stavros Demetriadis, and Anastasios Karakostas. 2015. Promoting Academically Productive Talk with Conversational Agent Interventions in Collaborative Learning Settings. *Computers & Education* 87 (2015), 309–325.

[87] Stergios Tegos, Stavros Demetriadis, Pantelis M. Papadopoulos, and Armin Weinberger. 2016. Conversational Agents for Academically Productive Talk: A Comparison of Directed and Undirected Agent Interventions. *International Journal of Computer-Supported Collaborative Learning* 11, 4 (2016), 417–440.

[88] Stergios Tegos, Stavros Demetriadis, and Thrasyvoulos Tsiatsos. 2014. A Configurable Conversational Agent to Trigger Students' Productive Dialogue: A Pilot Study in the CALL Domain. *International Journal of Artificial Intelligence in Education* 24, 1 (2014), 62–91.

[89] Stergios Tegos, Georgios Psathas, Thrasyvoulos Tsiatsos, Christos Katsanos, Anastasios Karakostas, Costas Tsibanis, and Stavros Demetriadis. 2020. Enriching Synchronous Collaboration in Online Courses with Configurable Conversational Agents. In *Intelligent Tutoring Systems*. Springer International Publishing, Cham, 284–294.

[90] Michael Tissenbaum, M. Berland, and Vishesh Kumar. 2016. Modeling Visitor Behavior in a Game-Based Engineering Museum Exhibit with Hidden Markov Models. In *International Educational Data Mining Society*. 517–522.

[91] Mike Tissenbaum and James D. Slotta. 2015. Scripting and Orchestration of Learning across Contexts: A Role for Intelligent Agents and Data Mining. In *Seamless Learning in the Age of Mobile Connectivity*. Springer Singapore, Singapore, 223–257.

[92] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).

[93] Hasanuzzaman Tushar and Nanta Sooraksa. 2023. Global Employability Skills in the 21st Century Workplace: A Semi-Systematic Literature Review. *Heliyon* 9, 11 (2023), e21023.

[94] Justin Vasselli, Christopher Vasselli, Adam Nohejl, and Taro Watanabe. 2023. NAISTeacher: A Prompt and Rerank Approach to Generating Teacher Utterances in Educational Dialogues. In *Workshop on Innovative Use of NLP for Building Educational Applications*. 772–784.

[95] L. S. Vygotsky and Michael Cole. 1978. *Mind in Society: Development of Higher Psychological Processes*. Harvard University Press.

[96] S. Walker. 2024. GSM8K Benchmark. https://klu.ai/glossary/GSM8K-eval

[97] Thiemo Wambsganss, Tobias Kueng, Matthias Soellner, and Jan Marco Leimeister. 2021. ArgueTutor: An Adaptive Dialog-Based Learning System for Argumentation Skills. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21 683)*. Association for Computing Machinery, New York, NY, USA, 1–13.

[98] Maximiliane Windl, Verena Winterhalter, Albrecht Schmidt, and Sven Mayer. 2023. Understanding and Mitigating Technology-Facilitated Privacy Violations in the Physical World. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (ACM)*. New York, NY, USA, 5:1–16.

[99] G. E. Xun and Susan M. Land. 2004. A Conceptual Framework for Scaffolding III-Structured Problem-Solving Processes Using Question Prompts and Peer Interactions. *Educational Technology Research and Development: ETR & D* 52, 2 (2004), 5–22.

[100] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a Machine Really Finish Your Sentence? *ArXiv [Cs.CL]* (2019). http://arxiv.org/abs/1905.07830