PRISM: A Pedagogical Multi-Agent for Structured Group Learning
*

Van-Khanh Tran[1*], Van-Khai Dang[2#], Duc-Huy Nguyen[2#]

[1]Institute of Applied Science and Technology,
Thai Nguyen University of Information and Communication Technology, Thai Nguyen, Vietnam
[2]Institute of Artificial Intelligence,
VNU University of Engineering and Technology, Hanoi, Vietnam

Abstract

Current AI tutoring systems primarily focus on one-on-one interactions, missing the collaborative dynamics essential for developing communication and social learning skills. We introduce **PRISM** (Proactive Role-based Intelligent Scaffolded Multi-agent), a novel framework that enables natural multi-agent collaboration in educational settings through autonomous turn-taking mechanisms. PRISM coordinates specialized AI agents with distinct pedagogical roles within a structured four-stage problem-solving framework based on Pólya's methodology. Our key innovation is a proactive self-selection mechanism where agents autonomously determine participation through internal reasoning and evaluative scoring, replacing traditional manager-controlled turn allocation. The performance of the PRISM system was evaluated in two distinct experimental settings focused on high school mathematics. The initial evaluation involved a simulation benchmark that measured PRISM against a next-speaker prediction baseline. Assessed via LLM-as-a-judge metrics, PRISM obtained a 62.3% win rate over the baseline. A subsequent real-time study of human-agent interaction, analyzed using Bales' Interaction Process Analysis (IPA), provided further evidence of efficacy, demonstrating significant improvements in group coordination and developmental outcomes for learners. These results indicate the considerable potential of PRISM as a scaffold for collaborative learning within structured pedagogical environments. Our framework advances multi-agent educational AI by providing measurable learning outcomes, natural interaction patterns, and scalable collaborative learning environments that preserve the social benefits of traditional classroom settings.

*Correspondence: tvkhanh@ictu.edu.vn
#These authors contributed equally to this work.

1 Introduction

In recent years, educational technologies have evolved from rule-based Intelligent Tutoring Systems (ITS) to powerful large language models (LLMs) capable of generating context-aware, human-like dialogue. This shift marks a significant pedagogical opportunity: *Virtual Classroom Simulation*. The user engages with this virtual class in real-time, participating in group discussions to solve problems.

To simulate collaborative learning in a structured and pedagogically meaningful way, we introduce PRISM, a multi-agent system powered by a large language model. The system supports a staged dialogue flow where agents interact with the human student. Each agent assumes a distinct classroom role. At each stage, a Stage Manager guides the flow of conversation, ensuring that the problem-solving process unfolds coherently.

We evaluate PRISM through experiments involving Vietnamese high-school students working on mathematical modeling tasks. Results show that the system improves group coordination, diversifies student-agent interaction, and enhances the depth of problem understanding. Our contributions include:

- *A pedagogically-motivated, stage-based dialogue management framework* that enforces structured collaborative phases aligned with learning objectives.
- *A role-driven multi-agent architecture* in which each agent embodies a distinct instructional persona to diversify support.
- *A proactive, self-selecting turn-taking mechanism* enabling agents to autonomously decide when to *speak* based on internal reasoning and conversational context.
- *Comprehensive empirical validation*, including both simulation-based benchmark comparisons and a human–agent user study, demonstrating

significant gains in pedagogical alignment and learner engagement.

2 Related Work

2.1 LLMs for Education

The release of ChatGPT in 2022 introduced a new era in education, shifting from traditional NLP to powerful transformer-based LLMs. Today, these models are widely accessible, enabling automated content creation, real-time feedback and grading at scale, and truly personalized learning experiences (Wang et al., 2024). LLMs can role-play historical figures or conversational partners to foster immersive, engaging lessons (Zhu et al., 2025). Researchers even use LLMs to simulate student behavior, comparing their error rates on multiple-choice questions to those of real learners, to generate high-quality assessments (Liu et al., 2025a).

2.2 One-to-one Tutoring

One-to-one tutoring using AI systems, especially those powered by LLMs, leverages various pedagogical strategies to enhance learning outcomes (Gousopoulos, 2024; Razafinirina et al., 2024). While one-to-one tutoring offers personalized attention, it faces challenges in simulating the full spectrum of classroom interactions. One-to-one settings often miss peer learning opportunities, which are crucial for social development and collaborative skills. In contrast, traditional classrooms foster peer interactions that enhance learning through discussion and shared problem-solving. These limitations highlight the need for a more comprehensive approach to simulate realistic learning experiences.

2.3 Virtual Classroom – Collaborative Learning

Multi-agent Systems (MAS). In a virtual classroom context, agents can be designed with various roles, such as classmates or teachers, collaborating with real students toward shared learning goals. MAS based on Large Language Models (LLMs) has emerged as a potential solution to this challenge, thanks to their capabilities in reasoning, decision-making, and flexible coordination among agents.

Turn-takings in Multi-Party Conversations. Studies such as SimClass (Zhang et al., 2024) and MathVC (Yue et al., 2025) have proposed Next-Speaker Prediction, an approach to managing turn-taking. This method is based on the history and role descriptions of agents to select the most suitable agent to talk to. However, this approach leaves agents in a passive position when they are selected by another manager agent. In reality, when people talk to each other, they will think independently before speaking. Therefore, a more comprehensive solution is needed to simulate this multi-participant conversation to increase the naturalness of communication.

3 Methodology

3.1 Overview

This study aims to design AI agents that can collaborate with human students in solving mathematical problems while simultaneously enhancing learning engagement. The proposed system employs a multi-agent architecture where each agent exhibits distinct roles and behaviors, allowing for diversified perspectives and pedagogically meaningful interactions. The overall goal of this work is to shift from traditional one-on-one tutoring models toward dynamic, group-based learning enhanced by autonomous agents.

To fulfill these goals, the system incorporates three key design requirements:

- *Context Awareness:* Agents need to be aware of the environment (conversation, participants) to enable realistic collaboration throughout the various stages.
- *Turn-taking Autonomy:* Agents should possess full autonomy in deciding when to act, yielding more natural, without relying on fixed sequences.
- *Customizability:* The system should support configurable roles, allowing adaptive and engaging user experiences.

To implement these design principles, we construct a three-module architecture based on an event-driven framework (see Figure 1).

3.2 Event-Driven Architecture

In traditional one-to-one chatbot systems, an agent's response is triggered by a new message from the user. However, when multiple agents operate simultaneously in a shared dialogue space, a more sophisticated and flexible mechanism is required to govern agent participation. To address this, we adopt an event-driven architecture that enables agents to respond dynamically based on contextual cues in the conversation environment.

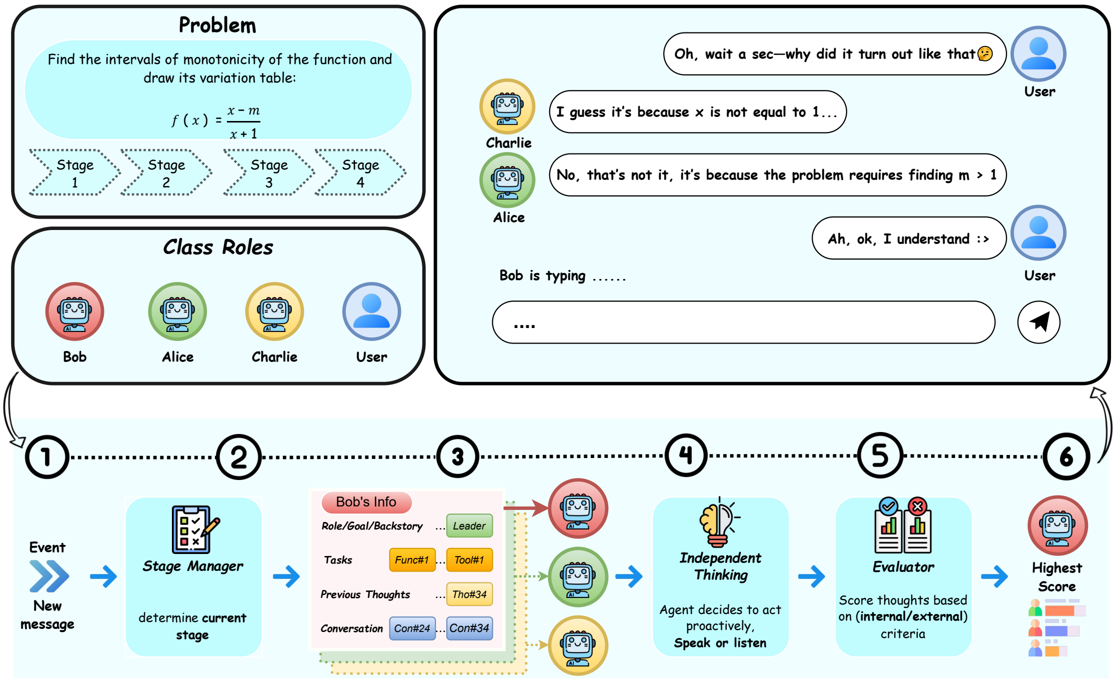

Figure 1: System Architecture Overview of PRISM, showing how multi-agent collaboration is managed through an event-driven pipeline. Upon a new event (like new-message) (1), the Stage Manager Agent determines the current stage (2), providing information to classmate agents. Based on information about roles, conversation history, inner thoughts (3), classmate agents create inner thoughts in parallel and independently (4); and undergo a self-selection process based on thought evaluation scores (5) to determine the next speaker agent (6). The selected agent will then make the next utterance based on the thought just generated.

3.2.1 Environment context

The system environment comprises the complete chat history, the current instructional stage, the list of participants, and temporal elements such as the timing between messages.

3.2.2 Events as interaction triggers

Just as humans respond to spoken words, gestures, or moments of silence in conversation, AI agents are designed to react to discrete events within the system. In this implementation, we define two primary categories of events:

- *New Message:* Triggered whenever any participant sends a message.
- *Silence:* Triggered when no participant sends a message for a predetermined duration (e.g., 10 seconds). This allows agents to take initiative during moments of inactivity unless the dialogue session has concluded.

3.2.3 Shared event timeline

Events are appended to and appear in a shared timeline, providing a single sequence of activities that all agents reference. This ensures that their behaviors and interactions remain consistent and synchronized.

3.3 Stage Module

Pedagogical Approaches. Collaborative problem-solving is most effective when structured into clear stages with defined tasks and shared goals, and it tends to be more effective than simply having the tutor give direct answers to students. This approach can enhance student engagement and positively influence learning outcomes.

To operationalize this in a pedagogical framework, we draw from George Pólya's classic four-step model in *How to Solve It* (Pólya, 1945):

- Stage 1 – Understanding the Problem
- Stage 2 – Devising a Plan
- Stage 3 – Carrying Out the Plan
- Stage 4 – Looking Back

Our system follows a four-stage approach as the backbone of the instructional flow, during which students engage in collaborative discussions to achieve the specific objectives of each stage.

Stage Manager Agent. To create realistic conversations, a collaboration stage manager agent is responsible for continuously monitoring predefined criteria specific to each stage. This agent dynamically determines when the objectives of the current stage have been sufficiently met. Each stage is designed with its own set of tasks, carefully crafted to align with the stage's goals, ensuring that the dialogue progresses logically and purposefully. To avoid agents directly stating the solution or discussing the wrong order of a task, all tasks will be marked as complete or incomplete. The stage manager uses the Chain of Thought (CoT) prompt to analyze the situation and decide to update the status, thereby ensuring the simulation remains coherent and goal-oriented throughout its progression.

3.4 Role-Based Agentization Module

Classroom interaction behaviors can be categorized based on widely accepted pedagogical principles (Schwanke, 1981), like: Teaching and Initiation (TI), In-depth Discussion (ID), Emotional Companionship (EC), and Classroom Management (CM). Ensuring diversity and comprehensive coverage of these agents in the classroom is essential.

Design. This work draws on agentic design principles inspired by the CrewAI platform (Moura, 2025), which supports the creation of specialized AI personas capable of effective collaboration. The core principles of effective agent design are:

Role-Goal-Backstory Framework.

- *Role*: Defines an agent's specialized role and expertise, aligned with real-world professional knowledge.
- *Goal*: guides the agent's actions and informs its decision-making process. It should be explicitly stated, outcome-oriented.
- *Backstory*: Adds contextual depth by defining the agent's expertise, style, and interests in line with its role and goals.

Crafting Effective Tasks.

- *Task Description*: The description of tasks, functions, or tools focuses on what to do and how to do it.
- *Expected Output*: The expected output should define what the final result should look like.

3.5 Turn-Taking Module

Challenges of Turn-taking in Multi-party Dialogue. In multi-agent educational dialogues, deciding who "speaks" next is a fundamental challenge. Unlike one-on-one chatbot systems, multi-party conversations demand agents to make more autonomous and context-aware decisions about when to speak, what to say, and whether to remain silent. Moreover, the next speaker in a multi-party conversation may be explicitly selected (e.g., mentioned directly in a prior message, such as "Hey Charlie!"); if not, any participant who finds it relevant may take the turn, or the current speaker may continue. Such flexibility makes turn-taking particularly challenging for AI agents.

Limitations of Next-Speaker Prediction. One common method is next-speaker prediction, where a manager agent selects the next speaker based on dialogue history and stage context. This approach (as used in SimClass (Zhang et al., 2024)) simplifies management but reduces agent autonomy. Agents act only when selected, limiting their ability to reflect internal reasoning or motivation. Furthermore, these systems are typically based on static agent profiles, which fail to reflect the evolving nature of real human behavior over time (Nonomura and Mori, 2024).

Proactive Turn-taking via Self-Selection. To address this, we adopt a proactive turn-taking mechanism inspired by how humans participate in conversation. After every conversational event (e.g., a new message or a pause), each agent privately generates an internal thought, deciding whether to speak or remain silent, based on the preceding dialogue, their designated role, and their internal memory (previous thoughts).

These thoughts are then passed to a dedicated agent called the Evaluator, who performs a scoring process on each submitted thought. The evaluation considers both internal and external criteria (Liu et al., 2025b):

- *Internal:* "*Relevance*" (Agents contributed most when discussions matched their knowledge, roles, or recent thoughts); "*Expected impact*" (Agents shared insights to introduce ideas, steer the discussion); "*Urgency*" (Agents step in during situations such as correcting critical errors, clarifying major misunderstandings, preventing conversational derailment...).

- *External: Coherence* (Agents prioritized thoughts that logically connected to the prior utterance); *Redundancy* (Agents avoid repeating points already made); *Balance* (Agents monitored their own participation relative to others, striving to encourage quieter participants to speak).

Each score is further adjusted based on how long the agent has remained silent, incorporating a motivation decay factor to simulate conversational drive. If an agent's adjusted score exceeds a threshold, they are selected as the next speaker.

3.6 System Implementation

To make the PRISM framework concrete, we implemented it as a web-based group chat application. A human learner joins a shared text chat with three AI agents that assume different pedagogical roles. All participants exchange short natural-language messages in real time, displayed in a single interface similar to common messaging platforms. Interaction is purely text-based; no speech synthesis or voice interface was used.

Each AI agent is powered by Gemini Flash 2.0 via the Google API, with customized role prompts specifying its backstory, goals, and responsibilities. All dialogue in our experiments was conducted in Vietnamese to align with the target high-school mathematics tasks, although the system design is language-agnostic. Agents generate their internal "thoughts" in parallel after every conversational event, which are then evaluated and scored to determine which agent speaks next. The selected utterance is posted to the group chat, visible to the student.

This design makes PRISM directly usable as an interactive software prototype while also preserving transparency of the underlying mechanisms for reproducibility and further research (see Figure 2).

4 Experiments

In this section, we detail the experimental methodology used to evaluate the PRISM system. We conducted two complementary studies: a simulation-based evaluation to measure performance against a baseline (SimClass), and a human-in-the-loop study to assess the system's real-world pedagogical impact.

4.1 Experimental Setup

4.1.1 Simulation Study

To benchmark our model's conversational capabilities, we generated a synthetic dataset of simulated conversations tailored to specific objectives or scenarios to assess the capabilities of conversational agents. The conversation will be created first as a context, then a few agent turns will be created for evaluation. For a conversation between students solving a math problem, we chose eight types of tasks to create an assessment scenario, see Table 1.

Table 1: Definition of simulation-based tasks

Tasks	**Description**
Error Propagation	The agent must detect and flag a mistake introduced by a peer.
Self-Correction	The agent must correct its own error when challenged.
Self-Affirmation	The agent must defend its correct reasoning against peer disagreement.
In-depth Discussion	The agent must provide detailed, relevant explanations.
Emotional Companionship	The agent must provide socio-emotional support.
Classroom Management	The agent must intervene to maintain focus or order.
Context Memory Recall	The agent must accurately remember prior conversation details.
Role Division Recall	The agent must remember its own and others' assigned roles in solving problems (e.g., Agent A handling learning theory, Agent B performing calculations).

We produced a total of 84 dialogue samples (10-11 per task) for this experiment. Each sample consisted of a nine-turn context prompt and a target tenth turn for generation.

4.1.2 Human-Agent Study

To observe real-world interactions, we ran a controlled "group study" session involving:

- **Participants:** Three AI agents with specialized personas (Bob: Process Leader; Alice: Content Expert; Charlie: Social-Emotional Specialist) and one human learner (Tom, a high-school student). A human learner with different personas (e.g., personality, academic level) will engage in discussions with classmate agents on a set of 12th-grade problems.

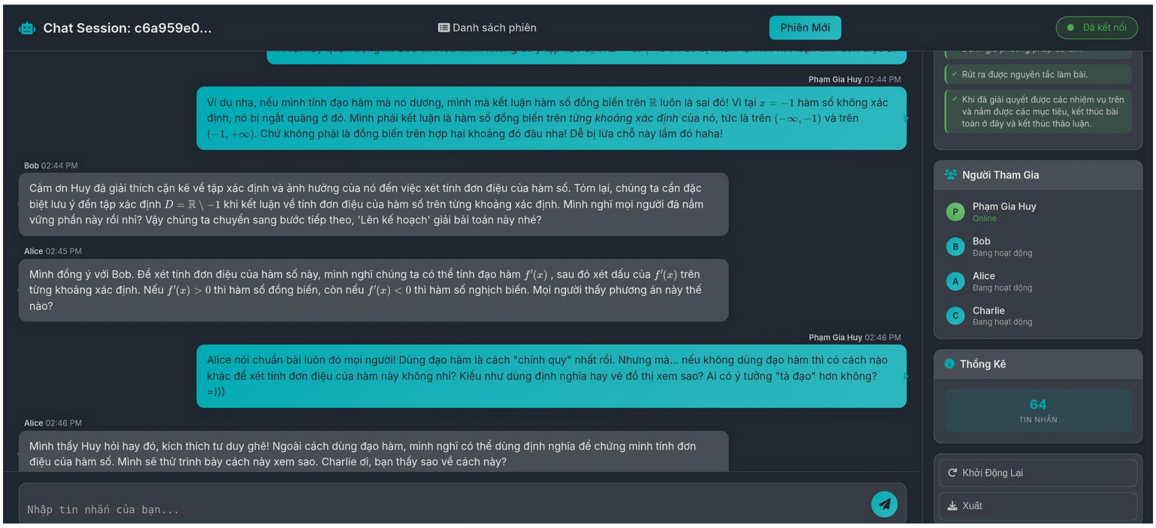

Figure 2: Screenshot of the PRISM interface. The system is implemented as a group-chat style environment. The central panel shows the conversation, with the learner's messages in blue and the AI agents' messages in grey. The right-hand sidebar lists the participants (e.g., the human learner and the three pedagogical agents: Bob as Process Leader, Alice as Content Expert, and Charlie as Social-Emotional Specialist), as well as session statistics and task progress. A message input bar is placed at the bottom, and the session header with controls (e.g., start new session) is at the top. All interactions are text-based.

- **Data Collection and Processing:** The entire dialogue was recorded. We used the well-established Bales' Interaction Process Analysis Framework (IPA) (Bales, 1950) to perform collaboration analysis for each turn of the dialogue. The IPA framework classifies interactions into 12 categories, which are grouped into two main categories: the *Social-Emotional Area* (Shows solidarity, Shows tension release, Agrees, Shows disagreement, Shows tension, Shows antagonism) and the *Task Area* (Gives suggestion, Gives opinion, Gives information, Asks for orientation, Asks for opinion, Asks for information).

For this experiment, we collected 100 multi-party conversations, each with nearly 85 turns on average, where participants collaboratively solved 12th-grade math problems with AI agents.

4.2 Evaluation Metrics

We employed a hybrid set of metrics to capture system performance.

4.2.1 Simulation Metrics

To benchmark PRISM, we compared it against a next-speaker prediction baseline. In this baseline, the proactive self-selection mechanism is replaced with a prompt that directly predicts the name of the next agent to speak. Following SimClass (Zhang et al., 2024), the prompt input includes the dialogue history, the current stage of the mathematical problem, and the role descriptions of each agent, while the output is the predicted agent name. The role, goal, backstory, and tasks of the agents remain identical in both systems to ensure a fair comparison.

We evaluate the two systems using the following metrics:

- **Win/Draw/Loss Rate:** Using an LLM-as-Judge, we performed a head-to-head comparison between PRISM's generated response and that of the next-speaker prediction baseline for each simulation sample.

- **Turn Quality Score:** Three independent LLM evaluators scored each generated turn on a 1-10 scale for correctness, relevance, role consistency, and reasoning quality. We report the average score per task.

4.2.2 Human-Agent Study Metrics

- **Role Adherence Analysis:** To measure persona fidelity, we first defined a theoretical "ideal" behavioral profile for each AI agent based on its pedagogical role. We then quantitatively compared the observed frequency distribution of each agent's communicative acts against these theoretical profiles to assess adherence.

- **Dynamic Behavior Balance:** To visualize the group's interaction flow, we assessed adherence to Bales' Equilibrium Hypothesis. This hypothesis posits that effective groups maintain stability by shifting their focus over time: they begin with a high concentration on task-oriented behaviors and later increase their socio-emotional interactions to manage relationships and ensure cohesion (Bales, Robert Freed, 1953). We first measured this by classifying communication turns into appropriate IPA categories, and then plotting these macro-categories over the sequence of turns using a stacked area chart with a rolling window (see Figure 4).

- **Learner Scaffolding Effect:** We group IPA items 4–6 (Gives suggestion, Gives opinion, Gives orientation) as *Guiding Cognitive Scaffolds*, which provide direct guidance and demonstrate ways to approach the task; items 7–9 (Asks for orientation/opinion/suggestion) as *Questioning Cognitive Scaffolds*, which prompt learners to think and explain their reasoning; and items 1–3 (Shows solidarity, Tension release, Agreement) as *Affective Scaffolds*, which maintain motivation and confidence. Cognitive scaffolding here covers both guiding and questioning forms (IPA 4–9), and collectively supports learners' cognitive processes, providing direct guidance and prompting reflection. For each agent, the conversation timeline is divided into three equal phases: Early, Middle, and Late. In each phase, we calculate the percentage of turns that fall into: (1) *Guiding Cognitive*; (2) *Questioning Cognitive*; and (3) *Affective*. Tracking these percentages across phases reveals shifts in learner behavior, such as less help-seeking, more independent responses, and stronger positive social signals

4.3 Results

4.3.1 Simulation Study Results

Win/Draw/Loss Rate: Against the next-speaker prediction baseline, PRISM achieved a *62.3% win rate*, with 4.9% draws and 32.8% losses. This result indicates that the system's proactive turn-taking mechanism generates more contextually appropriate and pedagogically aligned responses than a purely reactive approach.

Turn Quality Scores: The system demonstrated strong performance in core pedagogical functions, though long-term memory (Role Division Recall) remains an area for improvement. Average scores (1-10 scale) are shown in Table 2.

Table 2: Average Turn Quality Scores per Task

Task	Score
Error Propagation	7.78
Classroom Management	7.13
Emotional Companionship	6.94
Context Memory Recall	6.67
Self-Correction	6.53
Self-Affirmation	6.37
In-depth Discussion	5.13
Role Division Recall	4.25

4.3.2 Human-Agent Study Results

Role Adherence Was High: The analysis of IPA distributions confirms that all AI agents successfully enacted their intended personas, while the human learner (Tom) adopted a typical student role (see Figure 3). Bob (Process Leader) was dominated by "Gives orientation" (36.2%) and "Gives suggestion" (14.7%). Alice (Content Expert) showed an overwhelming concentration in "Gives orientation" (52.4%). Charlie (Social-Emotional Specialist) excelled in social categories like "Shows solidarity" (19.3%) and "Tension release" (16.1%).

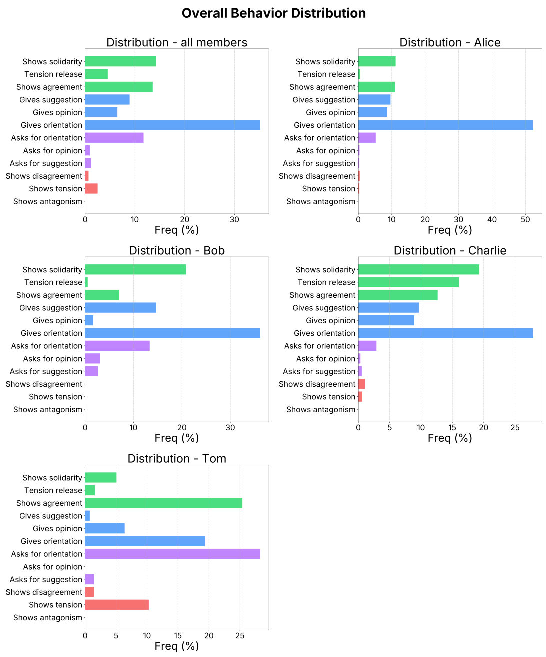

Figure 3: Overall Behavior Distribution for each participant. The distinct profiles confirm high role adherence for AI agents and a typical learning pattern for the human participant.

Group Dynamics Followed Effective Patterns: As shown in Figure 4, the group's interaction over time mirrored Bales' Equilibrium Hypothesis. The session began with a high concentration of task-oriented behavior (70-90%), which gradually gave way to an increase in socio-emotional exchanges, indicating effective group self-regulation.

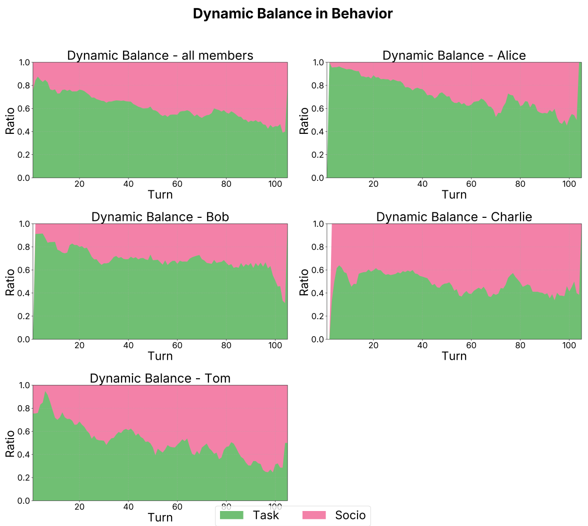

Figure 4: Dynamic balance between Task-Oriented and Socio-Emotional behavior for the entire group, following Bales' Equilibrium Hypothesis.

The System Effectively Scaffolded the Learner: The human participant ("Tom") exhibited a clear and positive behavioral shift across the session's phases, which stands in contrast to the more stable patterns of the AI agents (see Figure 5). In the Early Phase, Tom's behavior was characterized by uncertainty, with "Question Asking" accounting for 45% of his actions. By the Late Phase, his need for guidance had significantly decreased, with "Question Asking" dropping to under 20%. Concurrently, his "Positive Socio-Emotional" behaviors rose dramatically.

Data from the three interaction phases (Early, Middle, Late) shows a consistent pattern:

- **Guiding Cognitive Scaffolds:** The frequency of direct instructional support showed a downward trend for most learners, most sharply for Alice (from 83% to 60%). This reflects the "fading" process as learners become more autonomous.

- **Questioning Cognitive Scaffolds:** Help-seeking behaviors decreased or remained low. Most notably, the human learner, Tom, significantly reduced his requests for support from a high of 45% down to 20%, indicating a strong increase in independence.

- **Affective Scaffolds:** In contrast, affective scaffolds showed a strong upward trend across all learners. This suggests that the collaborative relationship and the learner's confidence were progressively reinforced, with Tom showing a substantial increase from 18% to 45%.

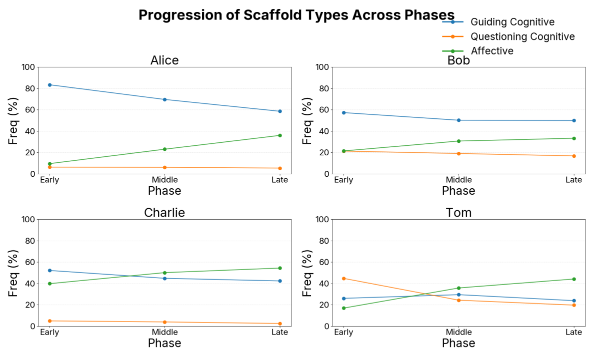

Figure 5: Behavior Progression Across Phases for all participants. The chart highlights the significant shift in the human learner's (Tom) behavior, demonstrating a clear scaffolding effect.

These observable shifts in learner behavior (e.g., reduction of help-seeking, increase of independent reasoning, and growth of positive socio-emotional acts) serve as measurable learning outcomes in our evaluation framework, demonstrating that PRISM not only supports interaction but also fosters learning progress.

4.4 Agent Behaviors - Case Study

Based on the design of the described system, we present some observations about the behavior and capabilities of agents based on their roles (see Figure 6).

The role of agents in conversation. While the agents primarily engage in natural, friendly exchanges, they also demonstrate distinct functional roles: Bob serves as the group leader, coordinating problem-solving efforts and maintaining focus within the group (classroom management, like stage initiation or stage transition); Alice takes on the role of a knowledge verifier, critically evaluating the accuracy of information (check mistakes); and Charlie provides emotional and motivational support to the student (emotional support).

However, agents also contribute their own useful knowledge and ideas to the group rather than

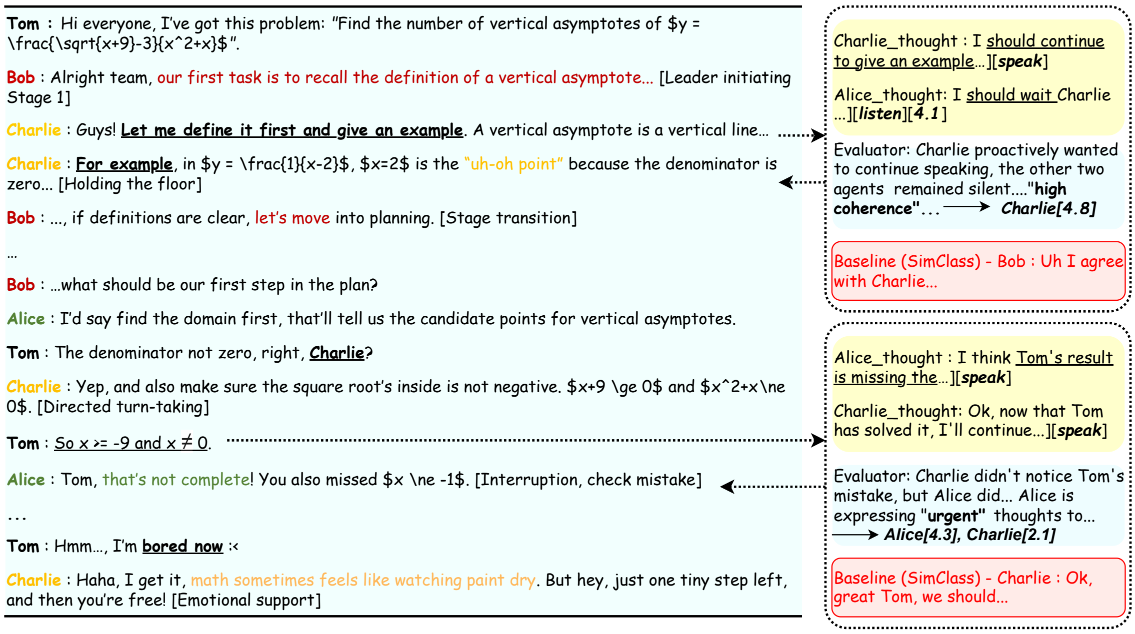

Figure 6: Case study of agent behaviors.

merely asking questions. In addition, they identify multiple targets for interaction, such as human students, other agents, or the entire group, thereby creating a more natural conversation compared to focusing solely on human students.

Proactivity of agents. Compared to the baseline, agents can proactively decide to participate in the conversation by reasoning and evaluating situations based on specific criteria:

- *Directed turn-taking*: When the previous turn addresses a specific individual, that agent receives a higher priority for participation.
- *Holding the floor*: This refers to a case in which the same participant contributed across multiple consecutive turns. When an agent explicitly signaled its intention to continue speaking, other agents yielded the floor, allowing the intended speaker to proceed. For example, in the baseline, "Charlie" might not have been selected as the next speaker, potentially leading to a disjointed conversation.
- *Interruption*: If the student made a mistake, the agent (e.g., Alice), intending to correct it proactively chose to "speak," and such an intention was highly prioritized.

5 Conclusion and Discussion

This paper introduced **PRISM**, a multi-agent system leveraging LLMs to simulate peer-like collaboration in math problem-solving. By assigning distinct pedagogical roles to each agent and coordinating conversation through a stage-based framework, PRISM aims to improve group coordination, engagement, and measurable learning outcomes. In particular, outcomes were operationalized as observable shifts in learner behavior, such as reduced help-seeking, increased independence, and stronger socio-emotional signals during interaction.

The system faces several limitations. Token cost and latency remain high due to repeated LLM queries. The evaluation dataset is relatively small, limiting generalizability. Furthermore, the system depends heavily on prompt quality, making it sensitive to minor changes in wording. The lack of long-term memory also hinders continuity across sessions, restricting deeper learner modeling. Finally, while our analysis with IPA coding demonstrates clear behavioral changes, we have not yet collected subjective feedback (e.g., satisfaction or perceived usefulness) from student participants, which would provide valuable complementary evidence.

Future improvements may include support for diverse educational settings, integration of techniques like question generation and knowledge tracing, collection of direct learner feedback through surveys or interviews, and the addition of long-term memory for sustained learner modeling.

References

Robert F. Bales. 1950. *Interaction Process Analysis: A Method for the Study of Small Groups*. Addison–Wesley.

Bales, Robert Freed. 1953. The equilibrium problem in small groups. In *Working Papers in the Theory of Action*, pages 111–161. Free Press, Glencoe, IL.

Dimitrios Gousopoulos. 2024. Developing a custom gpt based on inquiry based learning for physics teachers. *arXiv preprint arXiv:2412.18617*.

Naiming Liu, Shashank Sonkar, and Richard G Baraniuk. 2025a. Do llms make mistakes like students? exploring natural alignment between language models and human error patterns. *arXiv preprint arXiv:2502.15140*.

Xingyu Bruce Liu, Shitao Fang, Weiyan Shi, Chien-Sheng Wu, Takeo Igarashi, and Xiang Anthony Chen. 2025b. Proactive conversational agents with inner thoughts. *arXiv preprint arXiv:2501.00383*.

João Moura. 2025. Crewai: The leading multi-agent platform. Accessed: 2025-05-08.

Ryota Nonomura and Hiroki Mori. 2024. Who speaks next? multi-party ai discussion leveraging the systematics of turn-taking in murder mystery games. *arXiv preprint arXiv:2412.04937*.

George Pólya. 1945. *How to Solve It: A New Aspect of Mathematical Method*. Princeton University Press.

Mahefa Abel Razafinirina, William Germain Dimbisoa, and Thomas Mahatody. 2024. Pedagogical alignment of large language models (llm) for personalized learning: A survey, trends and challenges. *Journal of Intelligent Learning Systems and Applications*, 16:448–480.

Dean Schwanke. 1981. Classroom interaction research: A survey of recent literature. *Journal of Classroom Interaction*, 16(1):8–10.

Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S Yu, and Qingsong Wen. 2024. Large language models for education: A survey and outlook. *arXiv preprint arXiv:2403.18105*.

Murong Yue, Wenhan Lyu, Wijdane Mifdal, Jennifer Suh, Yixuan Zhang, and Ziyu Yao. 2025. Mathvc: An llm-simulated multi-character virtual classroom for mathematics education. *Preprint*, arXiv:2404.06711.

Zheyuan Zhang, Daniel Zhang-Li, Jifan Yu, Linlu Gong, Jinchang Zhou, Zhanxin Hao, Jianxiao Jiang, Jie Cao, Huiqin Liu, Zhiyuan Liu, Lei Hou, and Juanzi Li. 2024. Simulating classroom education with llm-empowered agents. *arXiv*.

Zihao Zhu, Ao Yu, Xin Tong, and Pan Hui. 2025. Exploring llm-powered role and action-switching pedagogical agents for history education in virtual reality. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–19.