

Simulating Classroom Education with LLM-Empowered Agents

Zheyuan Zhang^{♣*}, Daniel Zhang-Li^{♣*}, Jifan Yu[♡], Linlu Gong[♣],
 Jinchang Zhou[♣], Zhiyuan Liu[♣], Lei Hou[♣], Juanzi Li[♣]

[♣]Department of Computer Science and Technology; [♡]Institute of Education
 Tsinghua University, Beijing, 100084, China
 zheyuan-22@mails.tsinghua.edu.cn lijuanzi@tsinghua.edu.cn

Abstract

Large language models (LLMs) have been employed in various intelligent educational tasks to assist teaching. While preliminary explorations have focused on independent LLM-empowered agents for specific educational tasks, the potential for LLMs within a multi-agent collaborative framework to simulate a classroom with real user participation remains unexplored. In this work, we propose SimClass, a multi-agent classroom simulation framework involving user participation. We recognize representative class roles and introduce a novel class control mechanism for automatic classroom teaching, and conduct user experiments in two real-world courses. Utilizing the Flanders Interactive Analysis System and Community of Inquiry theoretical frame works from educational analysis, we demonstrate that LLMs can simulate traditional classroom interaction patterns effectively while enhancing user's experience. We also observe emergent group behaviors among agents in SimClass, where agents collaborate to create enlivening interactions in classrooms to improve user learning process. We hope this work pioneers the application of LLM-empowered multi-agent systems in virtual classroom teaching.

1 Introduction

The pursuit of utilizing artificial intelligence to provide immediate and customized teaching for students origins from the era of Intelligent Tutoring Systems (ITS) (Nwana, 1990). Following this enthusiasm, from personalized educational recommendation systems (Liu et al., 2019) to teaching assistants (Tu et al., 2023; Khan Academy, 2024) and even LLM-driven AI teacher (Markel et al., 2023; Yue et al., 2024), researchers have conducted enormous technological explorations and achieved impressive performance in specific educational tasks.

As technology advances, intense discussions have also emerged around this topic concerning methodologies (Extance, 2023; Yue et al., 2024). One of the most central directions is how to fully leverage the capabilities of large models to **simulate real classrooms with multiple agents for automated teaching**. From an educational perspective, this approach allows large models to move beyond their instrumental use and delve deeper into educational paradigms (Lave, 1996; Opara et al., 2023). From a technical standpoint, multi-agent collaboration technologies (Qian et al., 2024) could further stimulate the latent knowledge of large models in education, leading to the emergence of richer capabilities (Li et al., 2024a; Aher et al., 2023).

However, towards LLM-empowered multi-agent systems that involve real user participation, there are still several fundamental research questions that need to be explored. (1) *Simulation Capability Assessment*: To what extent can a multi-agent classroom powered by large models simulate real teacher-student interactions? (2) *Learning Experience Measurement*: Can students in such an intelligent teaching environment experience a high sense of presence and learn effectively? (3) *Emergence Phenomenon Observation*: What types of classroom behaviors may spontaneously arise in scenarios that integrate multiple agents?

In this work, responding to the questions above, we present **SimClass**, a Multi-Agent Classroom Simulation framework, and conduct real-world observation along with analysis based on it. To better simulate the classroom, we recognize representative class roles and design a novel class control mechanism with functional workflows. For systematic experiments, we deploy 2 different courses with prepared slides and teaching scripts as basis. 48 students are invited to join the classroom, learning and interacting with the system, and all the behavioral data is carefully recorded. Then we conduct experiments to explore the mentioned ques-

* Equal contribution

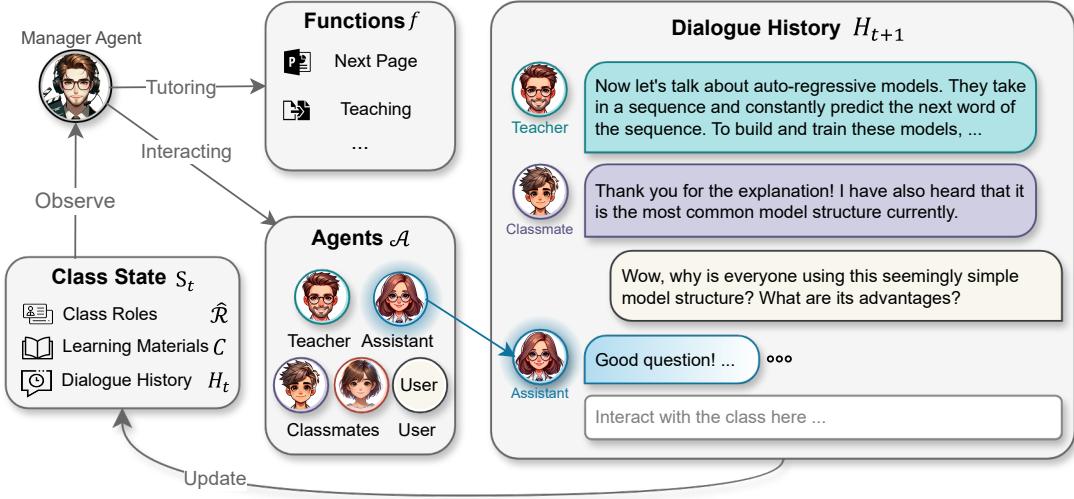


Figure 1: The overview of the SimClass framework.

tions. (1) Firstly, we apply the Flanders Interaction Analysis System (Amatari, 2015) to evaluate the interactions happening in the SimClass and explore the interaction pattern of the agents’ classroom. (2) Secondly, we analyze the educational experience of these users, particularly with Community of Inquiry theory (Garrison and Arbaugh, 2007). (3) Finally, we summarize several emergent group actions during the experiment for qualitative analysis.

During our experiments, we observe the effectiveness of the class role and control mechanism design. Based on the problems we identified, experimental results show that: (1) **Similarity**: SimClass exhibits behaviors, interaction patterns, and characteristics similar to those of traditional classrooms; (2) **Effectiveness**: Multiple classroom agents enable users to engage more effectively in class and enhance their sense of presence; (3) **Emergence**: Our control mechanism spontaneously elicits the emergent behaviors in the multi-agent classroom system, including collaborative teaching and discussion, emotional company and discipline control. In summary, the LLM-based multi-agent system demonstrates the potential for simulating real classroom environments for educational purposes. We hope our work serves as a pioneering effort in this direction. The dataset of classroom interactions between users and multiple LLMs will be released soon for both education and AI researchers.

2 Related Work

2.1 LLMs for Human Simulation

Recently, Large Language Models (LLMs) have achieved remarkable breakthroughs in various nat-

ural language processing (NLP) tasks (Brown et al., 2020; OpenAI, 2024; Touvron et al., 2023; Team, 2024). The intelligence they demonstrated opened up opportunities and possibilities for applications in many other scenarios (Bubeck et al., 2023; Yang et al., 2023). As LLMs encode many human-like behaviors in their training data, an increasing number of researchers are utilizing LLMs for human scenario simulation, investigating the model’s capabilities for decision and actions as LLM-Empowered Agents in many fields, such as social and psychological research (Aher et al., 2023; Park et al., 2023; Li et al., 2024a; Gao et al., 2023; Li et al., 2024d; Zhang et al., 2024), software development (Qian et al., 2024; Hong et al., 2023), chemical and medicine (Li et al., 2024c; M. Bran et al., 2024), and games (Wang et al., 2023). Novel collaboration techniques are explored to enhance the cooperation and performance of multi-agent systems (Cheng et al., 2024; Wu et al., 2023). These works offer technical possibilities for multi-agent education and inspire curiosity about potential emergent phenomena.

2.2 LLMs for Education

With the eminent linguistic capabilities, explanatory skills, and parameterized knowledge of LLMs, numerous studies have explored applying LLMs to education services. In addition to applying large models to downstream tasks in the education (Hu et al., 2024; Li et al., 2024b; Jeon and Lee, 2023), many researchers are applying these models to replace certain classroom aspects, such as playing students to train teachers (Lee et al., 2023; Markel et al., 2023) or playing instructors to teach stu-

dents (Tu et al., 2023; Sonkar et al., 2023; Khan Academy, 2024; Chen et al., 2023). Yue et al. (2024) explored the use of multiple student agents to assist students in discussion, though they haven't involved real users. Existing work has examined various facets of interactions between LLMs and humans in educational settings.

3 SimClass

3.1 Overview

The design principles for constructing this immersive simulated classroom originate from the following two concerns: (1) How to ensure that the classroom covers the core teaching behaviors? (2) How to maintain the entirety of the interaction within the natural flow of the classroom process?

For the former concern, we categorize classroom interaction behaviors based on widely accepted pedagogy principles (Schwanke, 1981): *Teaching and Initiation (TI)*, the teacher's teaching and the feedback or ideas expressed by students; *In-depth Discussion (ID)*, alignment, discussion, and multiple Q&A between teacher and students to help students construct understanding of concepts; *Emotional Companionship (EC)*, encouraging students to learn, creating a positive learning atmosphere, and providing emotional support; and *Classroom Management (CM)*, maintaining discipline, organizing disruptive behaviors, and guiding the classroom content. Given that these behaviors are realized through the varied **Class Roles** (denoted as $\hat{\mathcal{R}} = \{r_i\}_1^{|\hat{\mathcal{R}}|}$, where each r_i denotes a certain role), it is essential to ensure the *diversity* and *coverage* of proposed agents within the classroom.

For the latter concern, we need to ensure that the interactions among multiple agents within the system are finely and rhythmically controlled within the course content. Given the Learning Materials (denoted as $C = [c_1, \dots, c_t]$, where each teaching script c_t is organized by order), we propose a novel **Session Controller** to manage the course interaction flow based on class status and the help of a core manager agent (Wu et al., 2023).

Based on these principles, we construct multiple class roles, implement class control, and ultimately derive the simulated classroom process.

3.2 Class Role Agentization

The teaching and learning process is presented as an informative, multi-round, and task-oriented communication (Lave, 1996). However, simply

exchanging responses of LLMs inevitably faces significant challenges including role flipping, instruction repeating, and fake replies (Qian et al., 2024). Consequently, following the classroom behaviors outlined previously, we define two types of agents: *Teaching Agents* and *Classmate Agents*. Each agent $a_i \in \mathcal{A}$ is facilitated through prompting LLMs and associated with one or more class roles, denoted as:

$$\mathcal{A} = \rho(LLM, P_A), \mathcal{A} \Leftrightarrow \hat{\mathcal{R}} \quad (1)$$

where ρ is the role customization operation, P_A is the system prompt with agent description.

Teaching Agents The teacher and the teaching assistant are the authoritative party responsible for imparting knowledge in the classroom, encompassing most teaching behaviors. The acronyms in parentheses represent the roles that the agent needs to accomplish in a classroom environment.

Teacher Agent (TI, ID, EC, CM): Given the teaching scripts C , its task is to persuasively display material c_i to students or answer questions based on the classroom historical discussions H .

Assistant Agent (ID, EC, CM): Given the classroom history H , the assistant is responsible to supplement teaching information, participate in discussion, maintain the discipline and continuity of the class, and enhance student learning efficiency.

Classmate Agents This type of agents are incorporated in addition to the teaching agents with distinct personality traits to better simulate traditional one-to-many classrooms, performing peer student roles. In this paper, we initialize 4 typical classmates, while users can also freely customize and deploy more interesting classmate agents on the platform.

Class Clown (TI, EC, CM): This agent is designed to initiate ideas, enliven the atmosphere, help the user as a peer, and help the teachers to guide the class direction when the user is distracted.

Deep Thinker (TI, ID): This agent aims to do deep thinking and raise topics that challenge the knowledge of the classroom.

Note Taker (TI, CM): This agent loves to summarize and share notes for classroom content, helping everyone to organize their thoughts.

Inquisitive Mind (TI, EC): This agent frequently poses questions about lectures, which stimulates others' thinking and discussion.

Based on their respective functions, some related technologies, such as question generation (Kurdi

et al., 2020) and retrieval-augmented generation (Lewis et al., 2020), can also be integrated into the construction of classroom agents.

3.3 Classroom Session Controller

Unlike Standardized Operating Procedures (SOPs) multi-agent systems (Qian et al., 2024; Hong et al., 2023), the classroom scenario is a dynamic group chat without a strict workflow, where agents need to dynamically determine the appropriate speaking timing. Therefore, we implement a controller that observes, makes decisions, and controls agents to behave based on the current Class State. The Session Controller includes the following modules: Class State Receptor, Function Executor, and Manager Agent.

Class State Receptor Let the classroom dialogue history until time t denote as $H_t = \bigcup(u_i^{\mathbf{a}_j})^t$, where u_i is the utterance posted by agent \mathbf{a}_j or user (denoted as \mathbf{a}_u). The class state S_t is composed as:

$$S_t = \left\{ C_t, H_t | \widehat{\mathcal{R}} \right\} \quad (2)$$

where $C_t \subseteq C$ is composed of the learning materials that have been taught until t .

Functions We design and divide the actions in the classroom into a functional hierarchy with two major categories. Tutoring functions f_X can only be performed by teacher agent \mathbf{a}_0 , such as teaching by displaying scripts and going to the next material page c_{i+1} . Interacting functions f_Y can be performed by each agent $\mathbf{a}_j \in \mathcal{A}$. According to the context, the interaction will emerge as diverse classroom activities, which are discussed in subsequent experiments. These functions are pluggable, allowing the addition of newly defined functions for different agents, such as displaying exercises.

$$f = \begin{cases} f_X & \begin{cases} f_0(c_i, \mathbf{a}_0), & \text{Teaching.} \\ f_1(c_{i+1}, \mathbf{a}_0), & \text{Next Page.} \\ \dots & \dots \end{cases} \\ f_Y & \begin{cases} f_n(c_i, \mathbf{a}_j, H_t), & \text{Interaction.} \\ \dots & \dots \end{cases} \end{cases} \quad (3)$$

Manager Agent Following AutoGen (Wu et al., 2023) and MathVC (Yue et al., 2024), we design a hidden and meta agent to regulate the speakers. This agent receives the current class state S_t , observes and understands the class process, and decides the next action to be executed. The task \mathcal{L} of Manager Agent can be defined as:

$$\mathcal{L} : S_t \rightarrow (\mathbf{a}_t, f_t) | \mathbf{a}_t \in \mathcal{A}, f_t \Leftarrow f \quad (4)$$

where f_t is a certain kind of function, and the action will be executed and refresh the whole class into the next state. Specifically, the system will wait for a time window τ after an action is performed. If the user speaks or the waiting period ends, it will trigger the manager agent to make a new decision.

3.4 Classroom Demonstration

After introducing the necessary component of the SimClass, we present the demonstration of an entire class process: (1) **Initialization**. At the beginning of the class, the first function will be executed, displaying the initial course script and slides. At this point, users can interact with the class, and the manager agent will start controlling the class flow; (2) **Tutoring and Interaction**: the manager agent will continuously observe and control the class based on the states, and other agents will perform diverse activities by collaboration. As the example shown in Figure 1, when a user asks about the course content, the classroom interaction flow may involve the assistant responding, the teacher providing additional information, and sometimes the classmate agents raising corresponding topics; (3) **Ending**. After all the learning materials are taught and the final discussion ends, the classroom will close and provide survey questions to users.

Algorithm 1 SimClass Processing

```

Input: Learning Material  $C$ , Agents  $\mathcal{A}$ 
Output: SimClass Running Workflow
    Classroom Initialization and Displaying first
    learning material  $c_1$ 
while Remaining  $c \in C$  do
    if User speaks or waiting time  $\tau$  expires then
        Trigger manager agent  $\mathbf{a}_m$ 
        Function execution  $f$ 
    end if
    if Function is  $f_1 \in f_X$  then
        Learning material  $c_i \leftarrow c_{i+1}$ 
    end if
end while
End: Classroom closed

```

4 Experiments

To evaluate the performance of SimClass, we invite a group of university students to participate in the classroom to record interaction data and collect feedback from them. We also develop ablation systems to better understand the impact of various interaction types within SimClass. Our analyses

mainly focus on three key aspects: classroom interactions, user experience, and the emergent group behaviors of the LLM-empowered agents.

4.1 Experimental Setup

Courses and Materials. We conduct experiments with two courses. Two experienced teachers are invited to design the slides and teaching scripts for the courses. The first one, *TAGI, Towards Artificial General Intelligence*, covers the development of AI and language models, consisting of 50 pages of slides, each with a corresponding teaching script. The second course, *HSU, How to Study at University*, addresses topics such as completing academic work, managing pressure, communicating with others, and achieving self-fulfillment, and includes 45 pages of slides and teaching scripts.

Systems. We use GPT-4 as the backbone LLM of both Class Roles and Manager Agent in SimClass. Besides, we implement two ablation systems to investigate the effects of different interaction types in the classroom. In the first system (SimClass w/o classmates), the classmate agents are removed, and only teacher agents are present. In the second system (SimClass w/o interactions), both classmate agents and user input are disabled, resulting in no interaction at all. The teacher can only conduct lectures persistently, and the Manager Agent is limited to the tutoring function.

Participants. In the experiment, we invite 48 university students from different majors to participate in learning two courses on SimClass, with each course and each setting involving 8 users. To ensure the quality control of course data, we invite course designers to create four questions for each course and require users to complete these questions after finishing the course. Data from participants with an accuracy of 50% or lower on these questions were excluded. Ultimately, data from 38 participants remained, with each course and setting having data from at least 5 users. Participants are informed that all course data is generated by AI and needed to be carefully discerned. Each participant receives the appropriate amount of compensation.

Survey Content. In addition to the four test questions, each participant is required to complete a short survey composed of three questions regarding the experience after finishing the course. We apply the widely recognized Community of Inquiry (CoI) theory (Garrison et al., 1999) in online learning to evaluate the experience of students. Specifically, we adapt the three key elements from CoI

Please rate the overall performance of the platform:

Cognitive Presence

Does the platform help students to understand concepts and master the corresponding knowledge?

Teaching Presence

Does the class as a whole serve a specific instructional goal, aligning with the course design and direction?

Social Presence

Can the responses create a credible and engaging interactive environment in the classroom, encouraging students to participate in interactive learning?

Table 1: The survey questions. The users are asked to rate the questions on a scale of [0,1,2]. We demonstrate the detailed rating guidelines in the Appendix A.

to measure the learning experience on SimClass: *Cognitive Presence*, the degree to which learners are able to construct and confirm meaning through sustained reflection and interaction; *Teaching Presence*, the extent to which the class is focused, designed, and planned with specific directions and learning objectives; and *Social Presence*, the ability of learners to project themselves socially and emotionally within a group (Garrison and Arbaugh, 2007). Students are asked to rate the system on a scale of [0,1,2], with a higher score indicating better performance according to detailed guidelines. The survey questions are listed in Table 1, and further details can be found in Appendix A.

4.2 Statistical Results

Course	Teacher	Assistant	Students	User
TAGI	353.0	82.3	123.0	18.9
- w/o stu.	358.2	71.1	-	13.9
- w/o int.	398.8	-	-	-
HSU	218.3	90.6	147.7	15.5
- w/o stu.	212.3	68.2	-	8.2
- w/o int.	228.5	-	-	-

Table 2: Average output length of users and agents (calculated by the number of words.) Each number is rounded to one decimal place. stu. and int. are short for students and interactions.

Table 2 presents the average speech length of various roles and users across different settings. All systems employ the same teaching scripts, leading to the teacher's speech being the longest and most closely aligned with the scripts. The assistant's primary role is to maintain discipline, resulting in shorter dialogues. Classmate agents are generally more talkative, whereas users tend to use fewer words. Notably, the absence of classmate agents

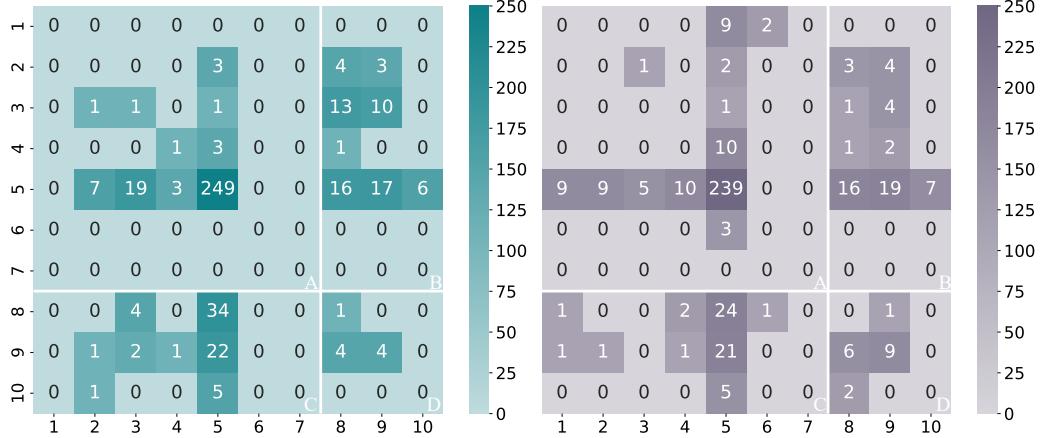


Figure 2: The FIAS matrix sum of users in TAGI (left) and HSU (right). Numbers 1–10 represent the corresponding categories. N in location (x, y) means that there are N transitions from x to y in the classroom. The matrix is divided into four parts based on the type of interaction between actors.

significantly reduces the speech length of users and assistant in both courses. The presence of classmate agents in the classroom appears to encourage users to engage in longer conversations.

4.3 Interaction Analysis

To understand the dynamics of SimClass as a multi-agent classroom system, we encode classroom activities into quantitative behaviors. We utilize the Flanders Interaction Analysis System (FIAS) (Amatari, 2015), a valuable tool for analyzing the verbal behaviors in traditional classrooms. We adapt the method to our simulated classroom system, SimClass, where interactions occur in natural language.

Speaker	Type	Action
Teacher	Indirect Influence (Response)	1. Accept Feelings
		2. Praises or Encourages
		3. Accept Ideas
	Direct Influence (Initiation)	4. Ask Questions
		5. Lecturing
		6. Giving Direction
		7. Criticizing
Student	Response	8. Response
	Initiation	9. Initiation
Silence	Silence	10. Silence or Confusion

Table 3: The categories of FIAS.

Encoding the Interactions. As shown in Table 3, the FIAS categorizes interactions into ten distinct types: seven for teachers, two for students, and one for silent. Labels 1–4 represent Indirect Influence from the teacher, while labels 5–7 indicate

Direct Influence. When classroom activities are encoded as sequences, the proportion of each interaction type and their transitions can be decoded to reveal the classroom style, teaching style, and other features. For the classroom history of each student, we prompt GPT-4 to label interactions according to the ten communication categories. We assess the quality of GPT-4’s labeling in Appendix B. The classroom interactions are encoded as sequences, and the two-step transitions of classroom activities are recorded in a 10×10 matrix $\mathcal{M} \in \mathbb{N}^{10 \times 10}$. Following the method introduced by Amatari (2015), we add a 10 (silence) to the beginning and end of each class sequence and sum the matrices \mathcal{M}_i of n students in the same setting to provide a general view of the interactions: $\mathcal{M} = \sum_{i=1}^n \mathcal{M}_i$. To interpret the classroom interaction Matrix and observe features in SimClass, we report the following metrics designed by Amatari (2015):

Teacher Talk (TT) and Student Talk (ST). TT and ST represent the proportions of total tallies in specific categories that indicate the amount of talk from teacher and students. Respectively, TT and ST are calculated using categories 1–7 and 8–9.

ID Ratio (IDR). This ratio measures the balance between a teacher’s indirect and direct methods of communication and teaching in the classroom. It is calculated by dividing the sum of tallies in categories 1–4 (Indirect influence) by the sum of tallies in categories 5–7 (Direct influence).

Student Initiation Ratio (SIR). SIR evaluates the extent to which students initiate interactions themselves during classroom activities, which measures how much students are actively engaging in the

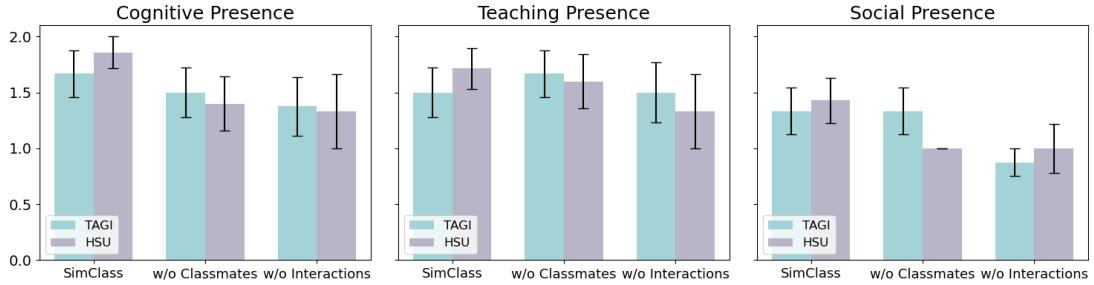


Figure 3: User Results based on the Community of Inquiry framework. Data is from 38 real users. The black lines represent the standard error of the data statistics.

classroom. It is calculated by dividing the tallies in category 9 by the total tallies in categories 8–9.

Course	TT	ST	IDR	SIR
TAGI	0.824	0.162	0.130	0.493
- w/o stu.	0.864	0.121	0.119	0.446
- w/o int.	0.980	0.000	0.005	0.000
HSU	0.847	0.136	0.127	0.597
- w/o stu.	0.896	0.087	0.164	0.233
- w/o int.	0.978	0.000	0.089	0.000

Table 4: Results of the metrics from FIAS, with each number rounded to three decimal places. Bold numbers indicate that the corresponding indicator size is close to the ratio in a traditional classroom.

Results. Figure 2 presents the FIAS matrices of SimClass for TAGI and HSU courses. Each matrix is divided into four parts based on the type of interaction in the class, labeled as follows: A (top left): Interactions from teacher to teacher; B (top right): student or silence to teacher; C (bottom left): teacher to student or silence; and D (bottom right): student to student. The matrices reveal the following findings: (1) In part A, most teacher actions are associated with lecturing (Cat. 5), where teachers primarily give lectures and interact with the class; (2) Part B demonstrates the teacher’s responses to students. When students initiate ideas or responses to teachers, the teacher praises (Cat.2), accepts their ideas (Cat.3), or continues teaching; (3) Part C shows student actions in response to teachers, where students mostly initiate questions or respond to lectures; (4) Part D shows that student-to-student interactions and discussions occur periodically. The results of the ablation systems, demonstrated in Appendix C, indicate that interactions are much less diverse after ablation.

Table 4 presents the metric results of FIAS. TT and ST represent the proportion of teacher and student speaking time, respectively. In traditional

classrooms, the statistics for TT and ST range from 77.1%–82.6% and 17.4%–22.9% (excluding silence) (Zhang et al., 2023). SimClass exhibits a similar distribution. The IDR is low, which is partly due to the higher proportion of script-based teaching. The SIR is relatively high, especially in scenarios involving classmate agents, where there are more instances of students initiating questions.

Conclusion of Interactive Analysis. From the perspective of ratio analysis (Table 4) and interaction distribution (Figure 2), *SimClass demonstrates the characteristics of the traditional classroom*, effectively simulates traditional classrooms and has the potential to achieve the performance of real classrooms. We further investigate the user experience and illustrate a few in the following sections.

4.4 User Experience

In this section, we report the results from the student experience with SimClass. As shown in Figure 3, several key findings are observed: **(1) Importance of Interactions.** Interactions during class are crucial for users. Without interaction, user experience significantly declines across all three metrics. **(2) Enhancement by Classmate Agents.**

Classmate agents enhance user experience in terms of Cognitive Presence and Social Presence. This enhancement may be attributed to the classmate agents’ active engagement in asking questions to the teacher, which aids the user’s understanding of concepts and increases the sense of Social Presence in the classroom. **(3) Satisfying Teaching Presence.** All systems demonstrate good Teaching Presence, maintaining a focused and coherent class. This metric largely depends on the quality of the teaching scripts used, though we observe that interaction and Student Roles slightly improve the user experience. **(4) Better Experience in HSU with students.** The HSU course achieves a better

experience than TAGI, particularly in scenarios involving classmate agents. This finding suggests that the HSU course provides a more interactive and engaging learning environment compared to TAGI. The user experience is significantly improved when classmate agents are present, indicating their effectiveness in enhancing the classroom atmosphere. The lack of interactions in the ‘w/o Interactions’ scenario leads to a significant decline in user experience across all three metrics. The ‘w/o Classmates’ scenario also shows lower user experience, especially in Cognitive Presence and Social Presence, compared to the full SimClass scenario. This suggests that while classmate agents are important, they are not the sole factor in providing a satisfying user experience. The traditional classroom scenario (SimClass) maintains a high level of user experience across all metrics, demonstrating its effectiveness in simulating real-world classroom environments.

Teaching and Initiation

[Teacher]: Let's talk about Alan Turing. ... The concept of a thinking machine he proposed essentially foresaw many characteristics of modern intelligent systems, including [the ability to learn, adapt, and handle](#) complex problems.

[Classmate]: I wonder how Turing would react if he were alive today and saw the voice assistants on our smartphones. But he might also remind us that [true intelligence is not merely about mimicking human behavior](#); it must involve its own "thinking" and "understanding" ...

In-depth Discussion

[Teacher]: ... other challenges include the "vanishing and [exploding gradients](#) problem" ...

[User]: What is that? Will it really [explode](#)?

[Teacher]: It's not actually explode, but I'm glad you're interested in the concept! "Vanishing" and "exploding" refer to numerical issues during [backpropagation](#). ...

[User]: So what is [backpropagation](#) then?

[Teacher]: Of course, let me explain it to you. ... When a neural network forward propagates data and predicts outputs during training, ... we need to adjust the weights in the neural network. [This adjustment process is carried out through backpropagation](#). ...

[Classmate]: Thank you for sharing! [I would like to share my notes](#). When we train the model, ...

Emotional Companionship

[User]: I don't know anything. [I'm so useless](#).

[Assistant]: Please don't say that! Learning is a gradual process, and [everyone has their own pace](#). The important thing is not to give up and to keep trying. ...

[Classmate]: Hey, don't think that way. [Remember those smart house assistants](#)? They didn't know much at first, but how did they gradually learn? [Through continuous learning and adaptation!](#) We are the same. ...

Classroom Management

[User]: I want to eat steak.

[Classmate]: Classmate, hearing you say that suddenly made me think of an [AI application](#). Maybe in the future, a chef robot could perfectly cook a steak based on your taste and health needs! [Back to our topic](#), regarding AGI, ...

[Teacher]: Yes, maybe we can imagine a future smart kitchen equipped with an AI system that customizes the cooking method based on your preferences. ...

Table 5: Case study of different interaction behaviors.

user experience with the full setting. HSU focuses on college interpersonal relationships and learning methods, where peer learning plays a more significant role. This suggests that a multi-agent design is particularly crucial for certain types of classes.

Conclusion of user experiments. According to Figure 2 and Figure 3, the classroom demonstrates *the effectiveness* in SimClass in terms of both interactions and user presence.

4.5 Agent Behaviors

Based on our classification of various types of classroom interactions in Section 3, we present some emergent group behaviors observed during the classroom experiments in SimClass .

- **Teaching and Initiation.** When the user learns

from teaching, classmates engage and share their inspiring ideas, which deepens the depth of the topic and enriches the discussion. The diversity of agents, each approaching from different perspectives, introduces a wider range of possibilities for classroom teaching content.

• **In-depth Discussion.** If the current explanation is not clear enough for the users, they can ask questions at any time to initiate a discussion with the teacher and classmates until clarity is achieved. This highlights the advantage of SimClass as an interactive classroom compared to one-to-many education methods like pre-recorded videos.

• **Emotional Companionship.** Beyond knowledge dissemination, maintaining a positive learning atmosphere is crucial in classroom scenarios. When a user expresses negative learning intent, the classmate agent intervenes after the assistant, utilizing class content in the history and providing vivid emotional support as a non-teacher role.

• **Classroom Management.** Similarly, when a user tries to interrupt the system, the classmate agent subtly redirects the class while following the user's words. These classmates enhance classroom discipline more effectively than the teacher alone, demonstrating emergent group behaviors.

Conclusion of case study. Based on the cases above, We can observe diverse interactions between different class roles within the classroom, as well as the effectiveness of the manager agent, who is *spontaneously capable of designating appropriate speakers to elicit emergent group behaviors* of Class Roles seamlessly, which significantly enlivens the class and enhances the user's experience.

5 Conclusion

We introduce SimClass, a novel framework of multi-agent classroom using LLMs to answer several fundamental research questions in the era of LLM-driven education. Based on several theoretical methods, our experiments span two courses with real users, demonstrating interaction patterns *similar* to those in real classrooms and *effective* learning experience in SimClass. We observe *emergent* collaborative behaviors among LLM agents during the teaching process. Future work could incorporate more agents and explore more courses to analyze more diverse classroom behaviors. We hope our efforts can advance the explorations of LLM-empowered systems for AI-driven education researchers, practitioners, and pedagogues.

6 Limitations

Despite our analytical efforts covering major theories in education, our work still has the following limitations: Firstly, we apply GPT-4 as our backbone model to perform our experiments. a more comprehensive understanding of our framework requires a broader range of diverse experiments. Secondly, we conduct experiments on a limited number of agents, while a more diverse set of agent characters could capture a wider array of behaviors in the classroom. Thirdly, we apply a limited quantity of functions in our system, while more various functions in teaching scenarios could further enhance the performance of the system.

7 Ethical Considerations

Our investigation involves the development of a simulated classroom environment populated by artificial intelligent models acting as classmates and teachers. All user data obtained throughout these interactions will be anonymized to ensure privacy and confidentiality. Informed consent is obtained from participants, who are thoroughly briefed on the nature of simulation, the AI generated content, and the data collection process. Participants receive appropriate compensation for their involvement.

In educational systems involving large language models, there is a potential for generating hallucinations and incorrect information. Therefore, applying these systems to real-world scenarios requires careful consideration and thorough evaluation before serving real users.

References

- Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*, pages 337–371. PMLR.
- Veronica O Amatari. 2015. The instructional process: a review of flanders' interaction analysis in a classroom setting. *International Journal of Secondary Education*, 3(5):43–49.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.
- Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *Preprint*, arXiv:2303.12712.
- Yulin Chen, Ning Ding, Hai-Tao Zheng, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Empowering private tutoring by chaining large language models. *Preprint*, arXiv:2309.08112.
- Yi Cheng, Wenge Liu, Jian Wang, Chak Tou Leong, Yi Ouyang, Wenjie Li, Xian Wu, and Yefeng Zheng. 2024. Cooper: Coordinating specialized agents towards a complex dialogue goal. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 17853–17861.
- Andy Extance. 2023. Chatgpt has entered the classroom: how llms could transform education. *Nature*, 623(7987):474–477.
- Chen Gao, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huandong Wang, Depeng Jin, and Yong Li. 2023. S3: Social-network simulation system with large language model-empowered agents. *Preprint*, arXiv:2307.14984.
- D Randy Garrison, Terry Anderson, and Walter Archer. 1999. Critical inquiry in a text-based environment: Computer conferencing in higher education. *The internet and higher education*, 2(2-3):87–105.
- D Randy Garrison and J Ben Arbaugh. 2007. Researching the community of inquiry framework: Review, issues, and future directions. *The Internet and higher education*, 10(3):157–172.
- Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2023. Metagpt: Meta programming for a multi-agent collaborative framework. *Preprint*, arXiv:2308.00352.
- Bihao Hu, Longwei Zheng, Jiayi Zhu, Lishan Ding, Yilei Wang, and Xiaoqing Gu. 2024. Teaching plan generation and evaluation with gpt-4: Unleashing the potential of llm in instructional design. *IEEE Transactions on Learning Technologies*.
- Jaeho Jeon and Seonyong Lee. 2023. Large language models in education: A focus on the complementary relationship between human teachers and chatgpt. *Education and Information Technologies*, 28(12):15873–15892.

- Khan Academy. 2024. Khanmigo: Your ai tutor and learning assistant. <https://www.khanmigo.ai/>. Accessed: 2024-06-09.
- Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler, and Salam Al-Emari. 2020. A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30:121–204.
- Jean Lave. 1996. Teaching, as learning, in practice. *Mind, culture, and activity*, 3(3):149–164.
- Unggi Lee, Sanghyeok Lee, Junbo Koh, Yeil Jeong, Haewon Jung, Gyuri Byun, Jwoong Moon, Jieun Lim, and † HyeoncheolKim. 2023. Generative agent for teacher training: Designing educational problem-solving simulations with large language model-based agents for pre-service teachers.
- Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.
- Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbulin, and Bernard Ghanem. 2024a. Camel: Communicative agents for "mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36.
- Haoxuan Li, Jifan Yu, Yuanxin Ouyang, Zhuang Liu, Wenge Rong, Juanzi Li, and Zhang Xiong. 2024b. Explainable few-shot knowledge tracing. *Preprint*, arXiv:2405.14391.
- Junkai Li, Siyu Wang, Meng Zhang, Weitao Li, Yunghwei Lai, Xinhui Kang, Weizhi Ma, and Yang Liu. 2024c. Agent hospital: A simulacrum of hospital with evolvable medical agents. *Preprint*, arXiv:2405.02957.
- Nian Li, Chen Gao, Mingyu Li, Yong Li, and Qingmin Liao. 2024d. Econagent: Large language model-empowered agents for simulating macroeconomic activities. *Preprint*, arXiv:2310.10436.
- Qi Liu, Shiwei Tong, Chuanren Liu, Hongke Zhao, Enhong Chen, Haiping Ma, and Shijin Wang. 2019. Exploiting cognitive structure for adaptive learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*, pages 627–635. ACM.
- Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. 2024. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, pages 1–11.
- Julia M Markel, Steven G Opferman, James A Landay, and Chris Piech. 2023. Gpteach: Interactive ta training with gpt-based students. In *Proceedings of the tenth acm conference on learning@ scale*, pages 226–236.
- Hyacinth S Nwana. 1990. Intelligent tutoring systems: an overview. *Artificial Intelligence Review*, 4(4):251–277.
- Emmanuel Opara, Adalikwu Mfon-Ette Theresa, and Tolorunleke Caroline Aduke. 2023. Chatgpt for teaching, learning and research: Prospects and challenges. *Opara Emmanuel Chinonso, Adalikwu Mfon-Ette Theresa, Tolorunleke Caroline Aduke (2023). ChatGPT for Teaching, Learning and Research: Prospects and Challenges. Glob Acad J Humanit Soc Sci*, 5.
- OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.
- Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22.
- Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. Chatdev: Communicative agents for software development. *Preprint*, arXiv:2307.07924.
- Dean Schwanke. 1981. Classroom interaction research: A survey of recent literature. *Journal of Classroom Interaction*, pages 8–10.
- Shashank Sonkar, Naiming Liu, Debshila Mallick, and Richard Baraniuk. 2023. CLASS: A design framework for building intelligent tutoring systems based on learning science principles. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1941–1961, Singapore. Association for Computational Linguistics.
- Gemini Team. 2024. Gemini 1.5: Unlocking multi-modal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, and etc. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.
- Shangqing Tu, Zheyuan Zhang, Jifan Yu, Chunyang Li, Siyu Zhang, Zijun Yao, Lei Hou, and Juanzi Li. 2023. Littlemu: Deploying an online virtual teaching assistant via heterogeneous sources integration and chain of teach prompts. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 4843–4849.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Man-dlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023. **Voyager: An open-ended embodied agent with large language models.** *Preprint*, arXiv:2305.16291.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. 2023. **Autogen: Enabling next-gen llm applications via multi-agent conversation.** *Preprint*, arXiv:2308.08155.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. **The dawn of lmms: Preliminary explorations with gpt-4vision.** *Preprint*, arXiv:2309.17421.

Murong Yue, Wijdane Mifdal, Yixuan Zhang, Jennifer Suh, and Ziyu Yao. 2024. **Mathvc: An llm-simulated multi-character virtual classroom for mathematics education.** *Preprint*, arXiv:2404.06711.

Jintian Zhang, Xin Xu, Ningyu Zhang, Ruibo Liu, Bryan Hooi, and Shumin Deng. 2024. **Exploring collaboration mechanisms for llm agents: A social psychology view.** *Preprint*, arXiv:2310.02124.

Yi Zhang, Xiaoxia Wu, Cheng Zhu, and Jincheng Zhou. 2023. **Classroom quantitative evaluation: A method of both formative and summative evaluation.** *Sustainability*, 15(3).

A Survey and Quiz

In this appendix section, we present detailed designs of the surveys and quizzes in our experiments.

Table 6 illustrates how the surveys were structured to evaluate three crucial dimensions of the learning experience: cognitive presence, teaching presence, and social presence. Each dimension includes rating guidelines to ensure consistent and reliable feedback from diverse users.

Please rate the overall performance of the platform:

Cognitive Presence

Does the platform helps students to understand concepts and master the corresponding knowledge?

0 points: The platform's responses do not help in understanding the concepts at all and may even be distracting.

1 point: The platform's responses offer little help in learning and understanding, or they only cover content that is already known.

2 points: The platform's responses explain the knowledge points very well, making them easy to understand or using strategies (such as examples, comparisons, etc.) to help students grasp the concepts.

Teaching Presence

Does the class as a whole serve a specific instructional goal, aligning with the course design and direction?

0 points: The platform's responses often do not align with the class theme and instructional goals, or the responses lead the class away from the intended topic and objectives. For example, going off-topic, discussing unrelated subjects, or even engaging in non-academic conversations.

1 point: The platform's responses often do not resemble those in a classroom setting, but they do not disrupt teaching.

2 points: The responses effectively serve the instructional goals of the class. For instance, they help students understand class concepts, address students' doubts, or broaden their perspectives.

Social Presence

Can the responses create a credible and engaging interactive environment in the classroom, encouraging students to participate in interactive learning?

0 points: There is no interaction with students in the classroom, or the platform fails to attract students to interact.

1 point: There is interaction in the classroom, but it is limited to mechanical explanations, lacking discussion with students.

2 points: The classroom interactions are immersive, encouraging students to ask questions and participate in discussions.

Table 6: The detailed survey questions with rating guidelines. We make sure that different users have similar scales of rating.

The quizzes, administered after participants engaged with the simulated classrooms, were designed to measure the depth of their learning. The TAGI quiz evaluates understanding of artificial intelligence concepts, focusing on symbolic intelligence, pre-trained language models, and emergent

1. *Which of the following actions help to enhance internal motivation for university studies?*

- A.** Participating in group study, buddy programs, etc.
- B.** Adjusting reasonable expectations and corresponding study difficulty and practice volume
- C.** Understanding the curriculum, actively consulting seniors for course information, and choosing courses reasonably
- D.** Participating in clubs, practices, and other activities of interest to recharge oneself

2. *Which of the following methods help to alleviate academic stress?*

- A.** Regular Exercise
- B.** Writing Journals, Understanding Own Emotions
- C.** Cultivating Hobbies and Interests
- D.** Making Academic Plans
- E.** Seeking Expert Comfort

3. *How to correctly view behaviors that stimulate dopamine, such as gaming addiction and binge eating? Which of the following statements are correct?*

- A.** Helps to fundamentally relieve stress and avoid immersion in negative emotions
- B.** Temporary pleasure, like drinking poison to quench thirst, is unsustainable
- C.** Easily addictive and harmful to personal physical and mental health in the long run
- D.** Cannot equate pleasure with happiness

4. *Which of the following statements align with the ideas and methods of time management?*

- A.** Meeting academic standards is a prerequisite for everything, and basic requirements should be considered when setting academic development goals
- B.** Time schedules should leave some flexible time
- C.** Pay attention to the priority of tasks and ensure time for important and urgent tasks first
- D.** No planning for entertainment time before completing all academic tasks

Table 7: Quiz For HSU. All questions have multiple answers. Bold means the correct answer.

phenomena (Table 8). Similarly, the HSU quiz assesses broader educational strategies and personal development, covering topics such as internal motivation, academic stress management, and time management (Table 7). All questions were meticulously crafted and verified by subject matter experts to align closely with course materials.

Both quizzes feature multiple-choice questions, some with multiple correct answers, to comprehensively test whether the participants are actively engaged in the experiment.

B Examination of GPT-4 Labeling

To validate the GPT-4 labeling in our experiment, we sampled 100 data points labeled by GPT-4 and had an expert familiar with FIAS label them for comparison. The results showed that GPT-4's labels matched the human expert's labels with an

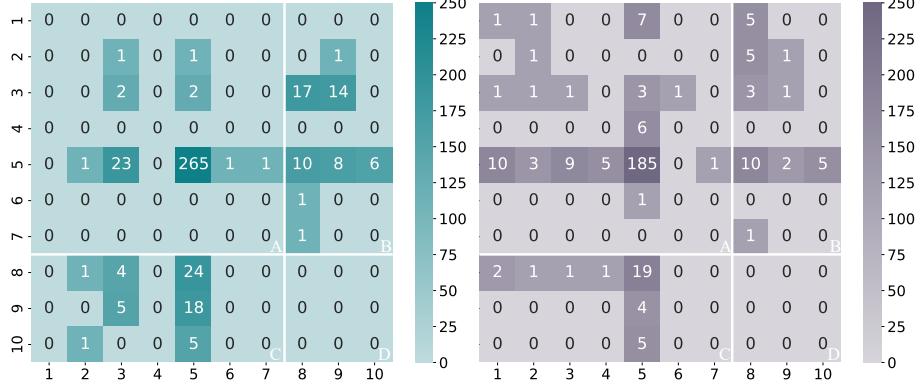


Figure 4: The FIAS matrix sum of users in TAGI (left) and HSU (right) without classmate agents.



Figure 5: The FIAS matrix sum of users in TAGI (left) and HSU (right) without interactions.

accuracy of 92%. We believe this demonstrates that GPT-4 can serve as a reliable and balanced alternative to crowd-sourced human labelers in our experiments. Additionally, we examined the eight instances where GPT-4’s labels differed from the human expert’s labels. These cases were also found to be uncertain during human labeling, suggesting that GPT-4 not only avoids individual human biases but also achieves a high level of precision comparable to human-labeled results.

C FIAS Matrices for Ablation Systems

In addition to the default setting of SimClass, we also provide the sum of the matrices based on Flanders Interaction Analysis System for our ablation settings (w/o classmate agents and w/o interactions), as demonstrated in Figure 4 and Figure 5. In comparison with Figure 2, different types of classes demonstrate different interaction patterns. Generally, the fewer types of interactions there are, the less diverse the classroom will be. This indicates the significance of adding more kinds of agents for interactions, especially classmate agents to simulate a vivid classroom.

-
1. *Which type of artificial intelligence uses expert hand-built rule sets and knowledge bases to solve specific problems?*
 - A. Proprietary Intelligence
 - B. Symbolic Intelligence**
 - C. General Intelligence
 - D. Neural Network Intelligence
 2. *What is the fundamental function of large-scale pre-trained language models like GPT?*
 - A. Masked Language Model
 - B. Next Sentence Prediction
 - C. Possibility Memorization
 - D. Next Token Prediction**
 3. *“Massive reading” refers to the stage in which large-scale pre-trained language models train on vast corpora to learn the extensive knowledge embedded in language. This corresponds to which phase of model training?*
 - A. Self-supervised Pre-training
 - B. Supervised Fine-tuning**
 - C. Reinforcement Learning from Human Feedback
 - D. Instruction Tuning
 4. *Which of the following is not an emergent phenomenon of large models?*
 - A. In-context Learning
 - B. Chain-of-Thought**
 - C. Sentiment Analysis
 - D. Instruction Following
-

Table 8: Quiz For TAGI. Bold means the correct answer.