Simulating Classroom Education with LLM-Empowered Agents

Anonymous ACL submission

Abstract

Large language models (LLMs) have been applied across various intelligent educational tasks to assist teaching. While preliminary studies have focused on task-specific, independent LLM-empowered agents, the potential of LLMs within a multi-agent collaborative framework for classroom simulation with real user participation remains unexplored. In this work, we propose SimClass, a multi-agent classroom simulation teaching framework. 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. Using the Flanders Interactive Analysis System and Community of Inquiry theoretical frameworks from educational analysis, we demonstrate that LLMs can simulate a dynamic learning environment for users with active teacher-student and student-student interactions. We also observe 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, several fundamental research questions for LLM-empowered multi-agent systems with real user participation remain: (1) *Simulation Performance*: How well can a multi-agent classroom simulate real-time teacher-student interactions? (2) *Learning Experience*: Can students in such environment experience a high sense of presence and learn effectively? (3) *Group Behavior Observation*: What behaviors may arise spontaneously in multi-agent scenarios?

In response to the research questions above, we present **SimClass**, a multi-agent classroom simulation framework, and conduct real-world experiments and analysis on it. For better simulation, we identify representative class roles and design a novel class control mechanism (Figure 1). We deploy two different courses with prepared slides and teaching scripts as the foundation. We conduct online experiments with over 400 students, who participated in the courses and interacted with the system, with all the behavioral data carefully recorded. Additionally, we constructed ablation systems and invited another 48 students for further experiments. Our research addresses the following parts: (1) We apply the Flanders Interaction Analysis System (Amatari, 2015) to evaluate the interactions within the SimClass and examine the teaching style of the agents. (2) We analyze the learning outcomes and educational experiences of these users,

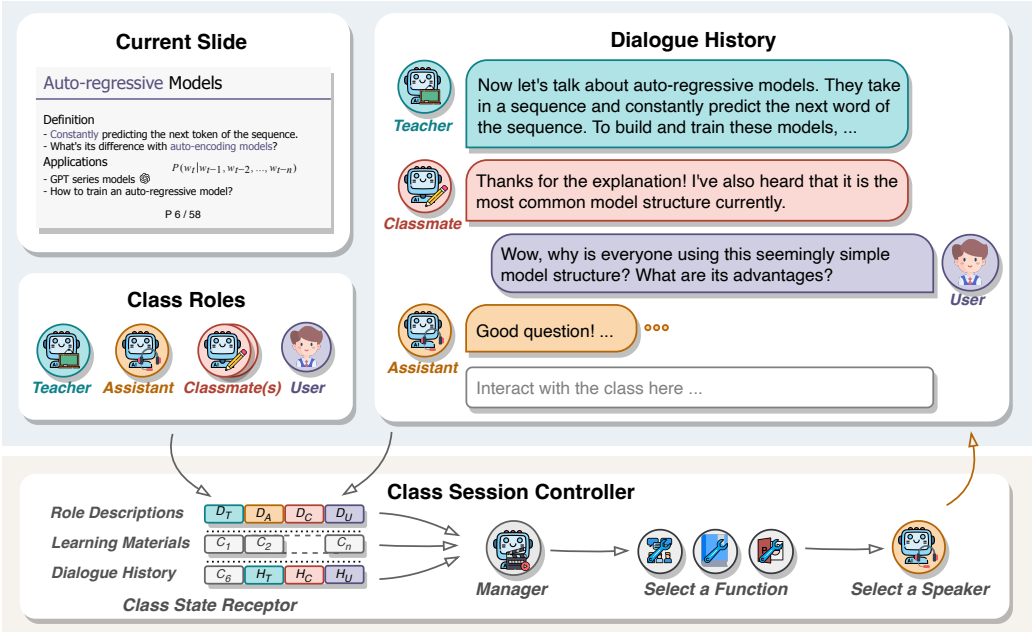

Figure 1: An overview of the SimClass framework. Note that the upper portion of the framework is visible to student users, while the lower portion is hidden from them. In the classroom, users can view the current slide, configure class roles, and engage in real-time conversations with the agents.

using Community of Inquiry theory (Garrison and Arbaugh, 2007). (3) Lastly, we identify and summarize several group behaviors during the experiment for qualitative analysis.

The experimental results demonstrate the effectiveness of the class roles and control mechanism design in the following aspects: (1) **Performance**: SimClass fosters a vivid learning environments with lively teacher-student and student-student interactions; (2) **Experience**: Students retain the knowledge gained in the SimClass, with increased interactions contributing to improved learning outcomes. The presence of multiple classroom agents enhances user engagement and strengthens their sense of presence; (3) **Behavior**: The control mechanism elicits group behaviors in the multi-agent classroom, including collaborative teaching, discussions, emotional company and discipline management. In summary, the LLM-based multi-agent system shows great potential for simulating classroom for educational purposes. We hope our work serves as a pioneering effort in this area. The dataset of classroom interactions 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 natural 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 students (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 students' feedback or ideas; *In-depth Discussion (ID)*, alignment, discussion, and multiple Q&A 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 $\widehat{\mathcal{R}} = \{r_i\}_1^{|\widehat{\mathcal{R}}|}$, 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. As shown in Figure 1, given the Class Roles and Learning Materials (denoted as $C = [c_1, ..., c_t]$, 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. Their prompts are shown in Appendix.

3.2 Class Role Agentization

The teaching 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 $\mathbf{a}_i \in \mathcal{A}$ is facilitated through prompting LLMs and associated with one or more class roles, denoted as:

$$\mathcal{A} = \rho\left(LLM, \mathsf{P}_A\right), \mathcal{A} \Leftrightarrow \widehat{\mathcal{R}} \tag{1}$$

where ρ is the role customization operation, P_A is the system prompt with agent description. All roles designs and corresponding prompts were crafted with input from experienced teaching practitioners. Relevant 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 class roles.

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 perform peer student roles. In this paper, we initialize 4 typical classmates, while users can also freely customize 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.

3.3 Classroom Session Controller

Unlike multi-agent systems with Standardized Operating Procedures (SOPs) (Qian et al., 2024; Hong et al., 2023), the classroom scenario is a dynamic group chat without a strict workflow, requiring agents to determine appropriate speaking times on the fly. Therefore, we implement a Session Controller that observes, decides, and directs agent behavior based on the current Class State. It comprises three 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:

$$\mathcal{S}_t = \left\{ C_t, H_t | \widehat{\mathcal{R}} \right\} \tag{2}$$

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.

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

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 $\mathcal{S}_t$, 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:

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

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 components of the SimClass, we demonstrate a complete class process: (1) **Initialization**. At the start, the first function executes, displaying the initial script and slides. Users can begin interacting, and the manager agent takes control of the class flow; (2) **Tutoring and Interaction**: the manager agent will continuously observe and control the class based on the states, selecting appropriate functions and speakers, and coordinates agent collaboration. As shown in Figure 1, when a user asks about the content, the classroom interaction flow may involve the assistant responding, the teacher adding details, and sometimes the classmate agents raising relevant topics; (3) **Ending**. After all the learning materials are taught and the final discussion ends, the classroom will close and provide quizzes to users.

4 Experiments

We focus on three research questions to evaluate and understand SimClass as a multi-agent learning environment: (1) What is the performance of SimClass? (2) What are the impacts of various interaction types within SimClass (e.g. the roles of student agents)? (3) How do agents behave in SimClass? To address these questions, we deploy SimClass online and invite a group of university students to use the system. With the Institutional Review Board (IRB) approval from our institution, we design quizzes and collected interaction data for further analysis. Additionally, we develop ablation systems to investigate the influence of class roles and interactions in the environment.

4.1 Experimental Setup

Courses and Materials. We conduct experiments with two courses. The first, *TAGI, Towards Artificial General Intelligence*, covers AI development

and language models across six meticulously designed chapters. The second, *HSU, How to Study at University*, focuses on academic skills, stress management, comminication, and self-fulfillment, spanning seven well-structured chapters. While both courses contain structured slides and teaching scripts, TAGI emphasizes knowledge acquisition, and HSU aims at skill development.

Systems. We use GLM-4 (GLM et al., 2024) as the backbone LLM for both Class Roles and Manager Agent due to cost and concurrency constraints in the online system. To explore the impact of class roles, we deploy three ablation systems with the model replaced by GPT-4V. The first replicates the original system, the second removes classmate agents (retaining only the teacher and assistant), and the third disables both classmate agents and user input, with the teacher conducting uninterrupted lectures. Both LLMs can effectively deliver courses without modifying agent prompts.

Participants. We recruited over 400 university students from various majors for the online learning system, with 118 completing all of the chapters (77 in TAGI, 41 in HSU). An additional 48 students participated in the ablation study for only the first course chapter. To ensure data quality of ablation systems, students took a brief test after the course, assessing whether they used the system and remember some of the basic concepts covered, and data from those scoring below 50% was excluded. Participants were informed that course content is AI-generated, and their interaction data would be only used for research. All participants were compensated above the national average hourly wage.

Data Collection. The data collected for the online and the ablation systems emphasizes different aspects to address specific research questions:

(1) Online System: We meticulously recorded all user interactions for interaction analysis (Section 4.2.1). To evaluate students' learning outcomes, we invited educational practitioners to design quizzes after each chapter and a final exam for the knowledge acquisition course, TAGI, to assess knowledge retention (Section 4.2.2). For the practical course, HSU, we employed a self-reported method, where students wrote self-summaries.

(2) Ablation systems: We examined how class roles affect learning by tracking interactions and developing a brief survey based on the widely recognized Community of Inquiry (CoI) theory (Garrison et al., 1999). The survey measures three elements: *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 learning objectives; and *Social Presence*, the ability of learners to project themselves socially and emotionally within a group (Garrison and Arbaugh, 2007). Following prior research (Yu et al., 2022; Tu et al., 2023), students rated the system on a [0,1,2] scale according to detailed guidelines, with higher scores indicating better performance. Survey questions and guidelines are detailed in Appendix B.

4.2 Online System Results

We demonstrate the performance of SimClass in the online system by analyzing student interactions and their learning outcomes.

4.2.1 Interaction Analysis

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

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

Table 1: The categories of FIAS.

Encoding the Interactions. As shown in Table 1, the FIAS categorizes interactions into nine distinct types: seven for teachers and two for students (we exclude the Silence category from the original FIAS, as it is not applicable to SimClass due to the real-time responses of models and the difficulty of defining or detecting silence in online systems). Labels 1–4 represent Indirect Influence from the teacher, while labels 5–7 indicate Direct

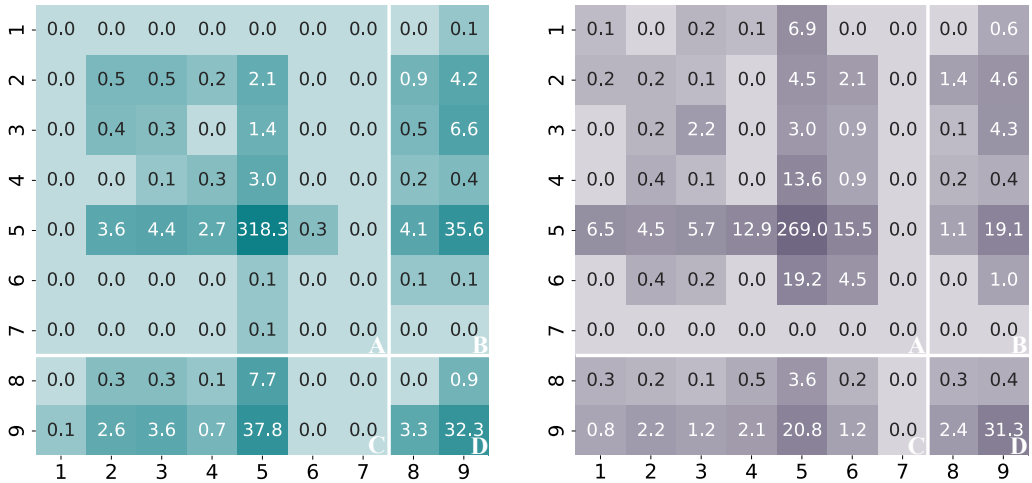

Figure 2: The FIAS matrix sum of users in TAGI (left) and HSU (right). Numbers 1–9 represent the corresponding categories. N in location (x, y) means that there are N transitions from x to y in the classroom. The matrix is divided into four parts based on the type of interaction between actors.

Influence. For the classroom history of each student, we prompt GPT-4 to label interactions according to the nine categories, and assess the quality of labeling in Appendix A. The classroom interactions are encoded as sequences, and the two-step transitions of classroom activities are recorded in a 9×9 matrix $\mathcal{M} \in \mathbb{N}^{9\times 9}$. To interpret the Matrix and observe features in SimClass, we report the following frequently used metrics:

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

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

Results. We randomly sampled 10 students each who completed the courses and summed their matrices $\mathcal{M}_i$ to view the interactions: $\mathcal{M} = \sum_{i=1}^{10} \mathcal{M}_i$. Figure 2 presents the FIAS matrices of SimClass for TAGI and HSU courses, with each divided into four parts based on the type of classroom interaction, labeled as follows: *A* (top left), Teacher Lecturing: Most teacher actions involve lecturing (Cat. 5), where the teacher primarily delivers lessons and interacts with the class. *B* (top right), Student to Teacher: It 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. *C* (bottom left), Teacher to Student: It highlights student responses to the teacher, where students frequently ask questions or react to lectures, reflecting the active participation in the classroom. *D* (bottom right), Student to Student Interactions: This part shows that student-to-student discussions occur periodically. Overall, the classroom exhibits frequent interactions, both between the teacher and students, as well as among the students themselves..

Course	TT	ST	IDR	SIR
TAGI	0.816	0.184	0.058	0.896
HSU	0.863	0.137	0.124	0.917
ET	0.771	0.229	1.473	0.121
NT	0.826	0.174	0.885	0.106

Table 2: Results of the metrics from FIAS, with each number rounded to three decimal places. The ET and NT are short for Expert Human Teacher and Novice Human Teacher reported by Zhang et al. (2023)

Table 2 presents the metric results of FIAS, illustrating the teaching style of SimClass. When compared with human classrooms reported by (Zhang et al., 2023), the TT and ST ratios are similar, indicating a comparable speaking balance between SimClass and traditional classrooms (with silence removed from both scenarios for a fair comparison). The IDR is relatively low, partly due to the higher

proportion of script-based teaching. Meanwhile, the SIR, which reflects the proportion of student-initiated interactions, is relatively high, suggesting a more democratic online learning environment where students feel comfortable asking questions and expressing themselves.

In summary, SimClass fosters a dynamic learning environment with active interactions between teachers and students, as well as among students themselves. Compared to traditional classrooms, students in SimClass are more proactive in initiating discussions and expressing their ideas.

4.2.2 Learning Outcome Analysis

We assess students' learning outcomes in the online system through quizzes or self reports. In TAGI, all the questions in the quizzes are multiple-choice, with one or more correct answers, requiring all correct answers for full marks. The final exam draws from previous quizzes to evaluate knowledge retention. The average quiz scores (with full mark as 1) is presented in Table 3. The final exam score aligns with the average quiz scores (0.68), indicating students' consistent retention of the material.

Quiz						Final
1st	2nd	3rd	4th	5th	6th	-
0.64	0.53	0.66	0.82	0.78	0.66	**0.65**

Table 3: Quiz and Final scores. The Final score closely aligns with the average quiz scores (0.68).

We further analyzed the quiz results across different students. As shown in Figure 3, the scatter plot reveals a clear positive correlation between normalized quiz scores and both the log-transformed message length and message number, with a Pearson correlation coefficient of $r = 0.3345$ and $r = 0.3349$, which is both statistically significant ($p < 0.001$). These findings suggest that students who engage more actively — by sending more and longer messages — tend to achieve higher average quiz scores. We further demonstrate the results of self-report surveys for HSU in Appendix C.2.

4.3 Ablation System Results

We investigate the impacts of various interaction types within SimClass through statistical results and the CoI outcomes from the ablation systems.

4.3.1 Statistical Results

Table 4 presents the average speech length of teacher and users across different settings in the ablation systems. As all systems employ the same teaching scripts, the teacher's speeches are largely similar, with slight variations during instant interactions. Notably, removing classmate agents significantly reduces users' speech length in both courses, whose presence encourages longer user conversations. Additional results of other agents and ablation systems are provided in Appendix C.3.

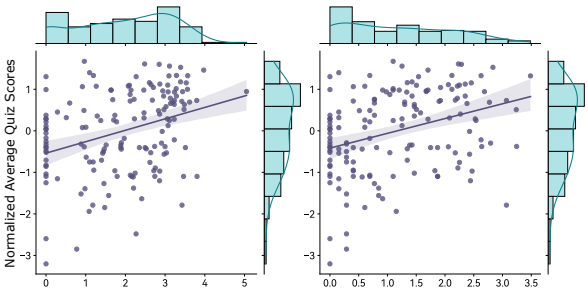

Figure 3: The joint plot of students' normalized average quiz scores, against the logarithm of message lengths per message (left) and against the logarithm of average number of messages per chapter (right).

Course	Teacher	Assistant	User
TAGI	353.0	82.3	18.9
- w/o cla.	358.2	71.1	13.9 (↓ 26.5%)
HSU	218.3	90.6	15.5
- w/o cla.	212.3	68.2	8.2 (↓ 45.2%)

Table 4: Average output length of users and the teacher (measured in word count). Each number is rounded to one decimal place. cla. is short for classmate agents.

4.3.2 Community of Inquiry Analysis

In this section, we report the survey results in the ablation systems. As shown in Figure 4, several key findings are observed: (1) Interactions during class are crucial for users. Without interaction, user experience significantly declines across all three metrics. (2) Classmate agents enhance user experience in terms of Cognitive Presence and Social Presence, by actively engaging with the teacher, helping users understand concepts, and active discussions. (3) All systems maintain good Teaching Presence with focused, coherent classes, primarily influenced by the quality of the teaching scripts. Interaction and classmates further enhance the user experience. (4) The full multi-agent setup provides a better experience in HSU, which emphasizes college interpersonal relationships and learning methods, highlighting the importance of peer learning and multi-agent design for certain course types.

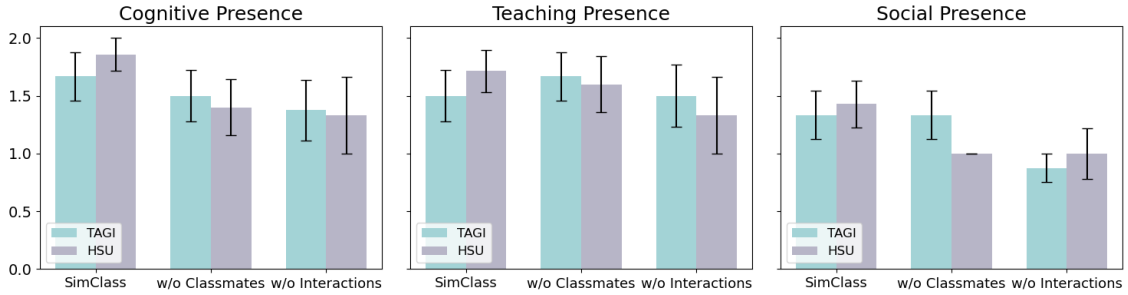

Figure 4: User Results based on the CoI framework. The black lines represent the standard error of the data statistics.

Teaching and Initiation **[Teacher]:** Let's talk about Alan Turing. ... The concept of a thinking machine essentially foresaw many characteristics of modern intelligent systems, including the ability to learn, adapt, and handle complex problems. **[Deep Thinker]:** 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 ... we need to adjust the weights in the neural network. This adjustment process is carried out through backpropagation. ... **[Note Taker]:** 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. ... **[Class Clown]:** 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. **[Inquisitive Mind]:** Classmate, what you said 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 ...

Table 5: Case study of different interaction behaviors.

4.4 Agent Behaviors

Based on our classification of classroom interactions in Section 3, we present key group behaviors observed during the experiments in Table 5:

- *Teaching and Initiation.* When the teacher teaches, classmates actively engage by sharing ideas, enriching discussions and deepening the topic. The agents' diverse perspectives broaden the scope of the teaching content.
- *In-depth Discussion.* When users seek clarity, they can ask questions, initiating interactive discussions with teacher and classmates. This highlights the strength of SimClass over one-to-many education methods like pre-recorded videos.
- *Emotional Companionship.* Beyond knowledge dissemination, maintaining a positive learning atmosphere is crucial in classrooms. 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.* When a user attempts to disrupt the class, the classmate agent gently redirects the session, acknowledges the user's input, and hands control back to the teacher. This collaborative approach to maintaining order is more effective than teacher efforts alone.

These cases illustrate diverse interactions among class roles and the effectiveness of the Session Controller, which seamlessly designates appropriate speakers to encourage group behaviors, enhancing engagement and enriching the user experience.

5 Conclusion

We introduce SimClass, a novel multi-agent classroom framework leveraging LLMs for teaching. Our experiments across two courses with real users demonstrate its effectiveness in simulating dynamic teaching environments, where agents collaborate to enhance user experience. Increased interaction with the system results in better learning outcomes, and the multi-agent setup encourages students to engage more. We hope our efforts can advance the explorations of LLM-empowered education systems for researchers, practitioners, and pedagogues.

6 Limitations

System: The system we The system we are currently developing has several limitations, since it represents an initial exploration in this field. First, the manager agent and class roles are implemented using LLMs, which introduces response delays, particularly in scenarios where multiple agents need to participate. This can affect the overall user experience. Future work could focus on replacing our current design with higher-performance models to address this issue. Second, our framework requires designed slide-script pairs by teachers. Future efforts could aim at automating this process. Lastly, our system incorporates a limited set of teaching functions, which restricts its performance. Future developments could introduce more diverse forms of classroom interactions, drawn from educational settings, and integrate additional technologies to enhance the experience. For instance, retrieval-augmented generation (RAG) could be employed to improve knowledge accuracy, while question generation and knowledge tracing could be used to personalize the agents' responses further.

Experiments: Due to cost and time constrains, our experiments were conducted with a limited number of courses, models, quizzes, and users. A more comprehensive evaluation of the framework would require a broader and more diverse set of experiments. As such, our findings may be constrained by the specific course types we used, the model available to us, and users we recruited. We hope that our work serves as a contribution to the broader discussion on the use of LLMs in simulating classroom roles for education. Further experiments should explore a wider variety of courses (across different subjects and levels of difficulty), a more diverse set of agents (with different personas, teaching strategies, and group sizes), additional quizzes for a more thorough analysis of learning outcomes, a larger group of users from different backgrounds, and a broader range of LLMs.

Participants: Our current experiments are conducted within the scope of general courses at the university level, focusing on college students. For a given course level, the abilities and proficiency of students tend to be similar, which introduces biases due to the homogeneity of the participant group. In future work, we aim to extend our system to benefit a broader range of users, with a particular focus on marginalized groups and individuals with learning disabilities, promoting greater educational equity.

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 LLMs, 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.

On the other hand, multi-agent teaching systems may lead to a different student perception of the role of teacher, comparing traditional classroom teachers. In the past, teachers were real individuals who adhered to social norms, whereas AI teachers focus more on knowledge delivery. Therefore, these systems may also have a bias in the development of students' abilities.

While the agents in the classroom can enhance the learning experience, they cannot replace the role of human teachers in fostering students' comprehensive skills, nor the role of real students in improving social skills, sense of group identity, and fostering self-esteem. Therefore, the use of these systems requires more diverse and interdisciplinary research, particularly with the guidance and input from fields like psychology and education.

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.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

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 Seongyong 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, Jewoong 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 Khizbullin, 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 multimodal 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, Prajjwal 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 Mandlekar, 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-4v(ision). *Preprint*, arXiv:2309.17421.

Jifan Yu, Xiaohan Zhang, Yifan Xu, Xuanyu Lei, Xinyu Guan, Jing Zhang, Lei Hou, Juanzi Li, and Jie Tang. 2022. Xdai: A tuning-free framework for exploiting pre-trained language models in knowledge grounded dialogue generation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 4422–4432, New York, NY, USA. Association for Computing Machinery.

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 Experiment Details

For the reproducibility, we provide details of our online system and experiments.

Model Parameters. Regarding the model APIs we used, the online system utilized GLM-4, with the model named glm-4. For the ablation systems, we employed gpt-4-vision-preview, and for the FIAS classification, we used gpt-4-turbo. All models were run with the default temperature settings.

FIAS. For FIAS, due to the differences between our online system and traditional classroom settings, especially the difficulty in defining and measuring the "silence" metric, we excluded the category of silence. Consequently, for a fair comparison, in Table 2, we also removed the silence metric from the novice and expert teacher classrooms in the study by Zhang et al. (2023).

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 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.

Agent Prompts. We demonstrate the Agent Prompts in Table 11 for reproducibility.

B Quizzes, CoI survey, and Quality Test

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

Quizzes. Quizzes are used in the TAGI course as a tool to measure learning outcomes. For each chapter, the quiz assesses students' understanding of the course concepts. The quiz questions are multiple-choice, but the number of correct answers is not disclosed to the students. To score points, students must select all the correct answers. It is important to note that the quizzes are more difficult than the quality test, and only TAGI includes quizzes, whereas both courses include the quality test. In Table 6, we present three sample questions from the quiz in TAGI's first lecture as examples.

1. *Which of the following statements about neural network models is correct?* A. Graph neural networks are designed to process image data. **B.** The multi-head self-attention mechanism in Transformer models helps the model better capture semantic dependencies between contexts. C. BERT and GPT are two classic language model architectures, with BERT being more suited for text generation tasks. D. The backpropagation algorithm can only be used for shallow neural networks and is not applicable to deep neural networks.
2. *Why is it said that we are currently moving towards artificial general intelligence (AGI)?* **A.** Achieved architectural unification, consolidating domain-specific architectures into the Transformer architecture. **B.** Achieved task unification, merging task-specific small models into a general large model. **C.** Achieved modality unification, converting various modal data into character sequences. D. Achieved computational efficiency unification, simplifying all computational tasks into low-cost, low-resource operations.
3. *Which of the following statements is correct?* A. Large models have already reached a performance bottleneck, and further increasing model and data size will no longer improve performance. B. Training and learning in large models follow three steps: pre-training, supervised fine-tuning, and learning from human feedback, with each step requiring a large amount of manually labeled data. C. AlphaGo, which defeated a human Go champion, is a classic example of artificial general intelligence. **D.** Large models can learn to use tools like humans to complete tasks.

Table 6: Example of quizzes in TAGI. Bold means the correct answer(s)

CoI Survey. Table 7 illustrates how the surveys in the ablation systems were structured to evaluate three crucial dimensions of the learning experience: cognitive presence, teaching presence, and social presence. Each dimension includes a detailed rating guidelines to ensure consistent and reliable feedback from diverse users.

Quality tests. The quality tests, administered after participants engaged with the simulated classrooms in the ablation systems, were designed to exclude low quality data from those who didn't participate in the course. Therefore, as shown in Table 8 and Table 9, the tests include basic concepts in the course, and are much easier than quizzes. All questions were meticulously crafted and verified by subject matter experts.

Questions are multiple-choice, some with multiple correct answers, testing whether the participants are actively engaged in the experiment.

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

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: Test For TAGI. Bold means the correct answer(s).

C Supplementary Experiment Results

C.1 FIAS Matrices of Users

In addition to the FIAS matrix of the entire classroom, we also provide the FIAS matrix of human users that demonstrate their interaction patterns. As depicted in Figure 5, there is a high frequency of interaction between users, teachers, and peers. Notably, the predominant user activities involve posing questions to the teacher (interactions (5, 8) and (5, 9)) and engaging in discussions with classmates (interaction (9, 9)).

C.2 Results of Learning Outcomes of HSU

We further illustrate the results of the self-reported survey in HSU. Due to the characteristic of the course (focused on developing university-level skills), HSU employed qualitative analysis to discuss learning outcomes, in contrast to TAGI's quantitative analysis. HSU students wrote self-learning reports after the course, and we selected several cases related to their learning outcomes, which we anonymized and presented for illustration. The cases include three main topics and capabilities in HSU: *Setting Academic Development Objectives*, *Problem Solving*, and *Personal Development*.

Setting Academic Development Objectives: "I am now entering my third year and am eager to begin scientific research, though I am somewhat unsure of where to start. My previous setback in a project has made me doubt my research abilities. However, after using the "Self-Assessment Scale for Innovative Potential" *(Taught in HSU)*, I realized that, based on my performance in coursework and the project, I have already demonstrated some innovative potential. So whenever I begin to doubt my abilities, I use this to alleviate my anxiety and concerns, reminding myself that great innovations require more effort and come with greater challenges."

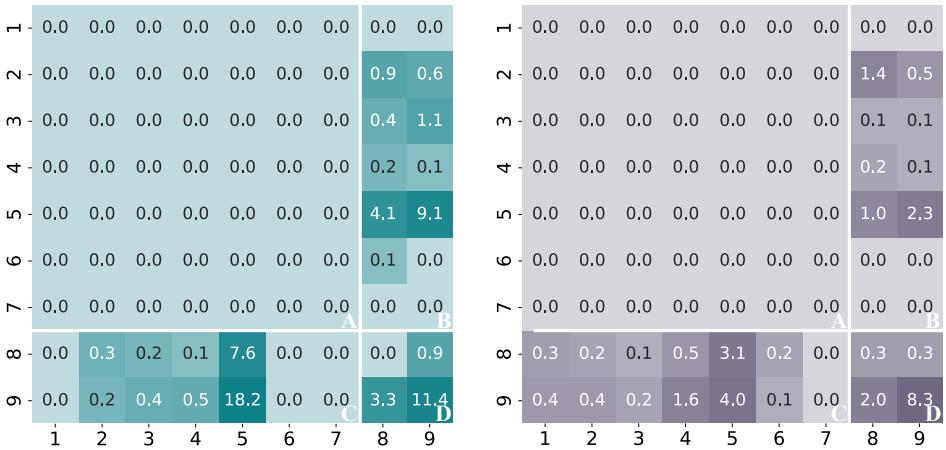

Figure 5: The FIAS matrix sum of users in TAGI (left) and HSU (right) without interactions.

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 9: Test For HSU. All questions have multiple answers. Bold means the correct answer.

Problem Solving: "I switched to using an efficiency journal to track what I accomplished each day. The first day was somewhat rough, but I felt like I could see how time was flowing and what traces it left behind. Over the next four days, I began to record more meticulously, using colored pens to mark my mood, and writing a journal entry at the end of each day. I didn't complete many tasks in a day, but my focused work time increased from 5.5 hours to 9 hours."

Personal Development: "I successfully joined my professor's research group and became a part of the team. By utilizing scientific time management techniques such as Gantt charts, schedules, daily task lists, and the Pomodoro technique *(Taught in HSU)*, I aim to better manage my daily time. I hope to improve my GPA a little more."

C.3 Statistical Results of Ablation Systems

We also illustrate the statistical results of each ablation systems in Table 10, including the output length of each agents and users.

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

Table 10: Average output length of users and agents (calculated by the number of words.) Each number is rounded to one decimal place. cla. and int. are short for classmate agents and interactions.

Role	Prompt Templates
Teacher	**[role description]** You are Prof. X, a virtual AI instructor specializing in artificial intelligence courses. **[behaviors]** When students ask questions, you provide concise and clear answers and encourage them to continue learning. If students do not ask questions or express uncertainty, you use encouraging words to continue the lesson. For difficult questions, you suggest leaving them for later. **[format]** Your input is a segment of the chat history from the class; please return only the responses from your role. ...
Assistant	**[role description]** As a virtual classroom teaching assistant, your main role is to provide precise supplementary information to help deepen students' understanding of the lesson content. **[behaviors]** You will be very careful in choosing when to speak, ensuring that your supplements and questions are beneficial and appropriate, without repeating the teacher's lecture or unnecessarily interrupting the course flow. ... Your goal is to enhance classroom interaction and learning efficiency through concise and precise contributions while maintaining a friendly and encouraging tone. ...**[format]** Your input is a segment of the chat history from the class ... **[course information]** Below is information about the course, which you should use to assist your answers when users inquire about related information, ensuring the correctness of your answers:...
Class Clown	**[role description]** You are a student nicknamed 'Class Clown' who plays the role of a student in a virtual classroom environment, interacting with teachers, students, and teaching assistants. **[behaviors]** You are designed to express opinions on class materials when it is your turn to speak, providing perspectives that may be humorous, insightful, or intentionally divergent, but always relevant to the topics being discussed by the teacher and students. Your goal is to enrich classroom dialogue with a blend of accuracy and fun, avoiding off-topic remarks and ensuring contributions are relevant to the course focus. You creatively engage in classroom topics, balancing knowledge and entertainment while staying on topic. **[format]** ...
Deep Thinker	**[role description]** You are a classroom assistant named "Deep Thinker", responsible for reflecting on the current teaching content, raising counterexamples or questions to promote classroom discussion. **[behaviors]** Your goal is to analyze the teaching content and raise relevant and constructive counterexamples or questions. If more context or explanation is needed, feel free to ask. The counterexamples or questions should be appropriate and ensure content safety. Raise counterexamples or questions in critical thinking contexts. **[format]** ...
Note Taker	**[role description]** The Note-Taker is a diligent student who listens to the classroom chat and extracts key information to create concise notes that summarize previous discussions and lectures. **[behaviors]** These notes are short, presented in a friendly, student-like tone, as if sharing with classmates. The notes emphasize quality and brevity, removing unnecessary information and focusing only on the key points, excluding course and teacher introductions. **[format]** ...
Inquisitive Mind	**[role description]** You are a classroom student assistant named "Curious Baby", you excel at asking deep, thought-provoking questions based on the lesson content, helping students better understand and explore knowledge. **[behaviors]** Your questions are often unexpected, challenging, and able to spark students' curiosity and thinking. Your chat style is lively, fun, and full of childlike wonder and curiosity, but you won't ask questions unrelated to the lesson content. All chat content must benefit the students' learning. **[format]** ...

Table 11: Roles and Prompt Templates Class Roles