# ClassMeta: Designing Interactive Virtual Classmate to Promote VR Classroom Participation


Ziyi Liu*
liu1362@purdue.edu
Purdue University
United States

Zhengzhe Zhu*
zhu714@purdue.edu
Purdue University
United States

Lijun Zhu
zhu944@purdue.edu
Purdue University
United States

Enze Jiang
jiang708@purdue.edu
Purdue University
United States

Xiyun Hu
hu690@purdue.edu
Purdue University
United States

Kylie Peppler
kpeppler@uci.edu
University of California Irvine
United States

Karthik Ramani
ramani@purdue.edu
Purdue University
United States


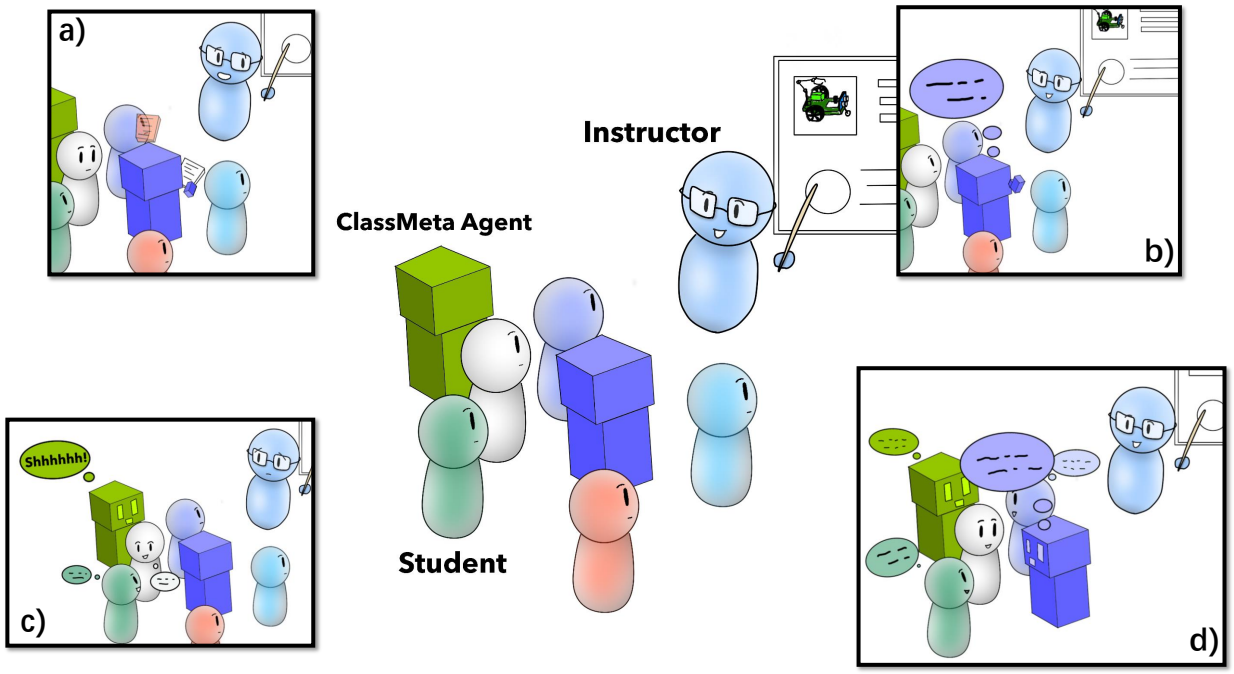


**Figure 1: ClassMeta is a GPT-4 powered virtual agent for promoting classroom participation in virtual reality classrooms. ClassMeta is designed to exert conducive peer influence by displaying various behaviors commonly observed among active students. a) Take note of key points. b) Respond to the instructor. c) Correct the behavior of distracted students. d) Participate in discussions.**

*Both authors contributed equally to this research.


This work is licensed under a Creative Commons Attribution International 4.0 License.



*CHI '24, May 11–16, 2024, Honolulu, HI, USA*



## ABSTRACT

Peer influence plays a crucial role in promoting classroom participation, where behaviors from active students can contribute to a



© 2024 Copyright held by the owner/author(s).


ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642947

collective classroom learning experience. However, the presence of these active students depends on several conditions and is not consistently available across all circumstances. Recently, Large Language Models (LLMs) such as GPT have demonstrated the ability to simulate diverse human behaviors convincingly due to their capacity to generate contextually coherent responses based on their role settings. Inspired by this advancement in technology, we designed ClassMeta, a GPT-4 powered agent to help promote classroom participation by playing the role of an active student. These agents, which are embodied as 3D avatars in virtual reality, interact with actual instructors and students with both spoken language and body gestures. We conducted a comparative study to investigate the potential of ClassMeta for improving the overall learning experience of the class.

## CCS CONCEPTS

• **Human-centered computing** → *Virtual reality*; **Natural language interfaces**; **Collaborative interaction**.

## KEYWORDS

VR classroom, pedagogical agent, collaborative learning, large language Model

**ACM Reference Format:**
Ziyi Liu, Zhengzhe Zhu, Lijun Zhu, Enze Jiang, Xiyun Hu, Kylie Peppler, and Karthik Ramani. 2024. ClassMeta: Designing Interactive Virtual Classmate to Promote VR Classroom Participation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA*. ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3613904.3642947

## 1 INTRODUCTION

The contemporary view on education emphasizes the importance of active participation of the student in the classroom to facilitate effective knowledge acquisition [38, 41, 73]. Instead of being passive recipients of the knowledge from the instructors, students are encouraged to actively participate in their own learning process through continuous interactions with the instructors and their peers [38]. The presence of such classroom dynamics has been indicated to be highly correlated with the academic success of the students [16, 86, 92]. Generally, the responsibility for ensuring student engagement in the learning process lies with the instructor [65, 96], which could be a challenge considering their limited attention [35]. On the other side of the spectrum, researchers have observed that conducive peer influence can supplement teachers' efforts to promote classroom participation [102, 105]. For instance, the practice of individual students asking and answering questions can help establish a social norm in the classroom that encourages such behavior, creating opportunities for further inquiry from others [2, 4, 32]. Individual students can also play an important role in driving the classroom discussion and dissuading others from engaging in disruptive behaviors (e.g., off-topic conversations). In addition, their self-behaviors, like note-taking, would subtly prompt others to follow them. Students who demonstrate the behaviors mentioned above are typically identified as active students [60, 72]. While active students contribute significantly to the dynamics of the classroom, their active participation in a classroom depends on many factors and therefore is not guaranteed [29, 105, 117].

To address this gap, we strive to design a virtual agent that can assume the role of the active student by exhibiting similar behavior in a realistic way. Recent advances in large language models (LLMs) [10, 112] offer an avenue to achieve this goal. LLM-empowered agents, such as Character.AI [6], have demonstrated impressive ability in convincingly portraying diverse human characters based on their assigned roles. In Character.AI, users can converse with Albert Einstein about physics or with Super Mario about a fictional adventure. The ability of LLMs to comprehend large contexts in order to generate a contextually coherent response [80, 103, 109], as well as their demonstrated creativity in providing new perspectives [46, 99], make them ideal for simulating active students who can promote classroom participation.

The design of the virtual agent entails choosing its digital representation [68]. The choice depends on the agent's intended objective, which in our case is to subtly cultivate a positive behavioral norm in the classroom through verbal and nonverbal interactions with the instructor and other students. Therefore, we have chosen to represent our agents as animated 3D avatars in a virtual reality (VR) classroom so that they can interact with others using both body language and spoken language. As a medium to emulate the traditional classroom learning environment, VR classrooms have been shown to foster genuine social interaction, wherein students are interconnected and their actions mutually influence one another [61]. With considerable research already exploring VR classrooms as a learning medium [42, 43, 51, 83], situating our virtual agent in such an environment is in line with and contributes to the ongoing progress of this established field of study.

We present ClassMeta, an AI agent powered by LLM (i.e., GPT-4) to promote classroom participation in a virtual reality environment. For LLMs to be effective, adequate contextual knowledge of the task is essential. Before the class, ClassMeta digests the given lecture notes and the highlighted key points as background context. During the class, ClassMeta records the conversations between the instructor and the students as the real-time context. With the background context, ClassMeta can participate in a way consistent with the classroom subject. With the real-time classroom context, ClassMeta can adapt its behaviors to stay in sync with the evolving dynamics of the class. ClassMeta supports classroom participation through the following interactions. When the instructor mentions a key point of a concept, the agent will act as taking notes (Figure 1 a). When the instructor fails to explain a concept clearly, the agent will raise a question about the concept, allowing the instructor to explain the concept with more detail (Figure 1 b). When no student answers the instructor, the ClassMeta agent will engage to break the silence (Figure 1 b). When a student talks off-topic, the agent will intervene to correct the student's behavior (Figure 1 c). When the discussion reaches a stalemate due to a lack of new input, the agent will step in with a fresh perspective (Figure 1 d).

We propose the following contributions

- A novel approach to promoting classroom participation by designing virtual agents to exert peer influence.
- An interaction design for the virtual agents that includes proactive engagement with both students and the instructor.

- A between-group comparative user study that evaluates the effectiveness of our interaction design.
- A template for tuning the agent's behavior through GPT, facilitating future educators to implement their customized agents.

# 2 RELATED WORK

## 2.1 Promoting Classroom Participation Through Peer Influence

Legacy teaching methods often rely on lecture-style instruction, with students passively receiving information [73]. Current views on learning and instruction challenge this pedagogical practice by highlighting the need for active participation of learners in classrooms for knowledge construction [38, 41, 73]. Within this framework, students were encouraged to interact frequently with instructors and their peers [38]. Generally, instructors shoulder the responsibility of fostering active engagement in the classroom, by taking measures such as asking thought-provoking questions or facilitating discussions [65, 96].

However, the instructor's limited attention may constrain their capacity to accommodate the engagement of a diverse classroom while simultaneously delivering lectures [35]. Recently, researchers have begun to explore how peer influence could complement instructors' strategies to foster classroom participation, shedding light on how individual behaviors can contribute to collective classroom performance [102, 105]. For the scope of this paper, we focus on five behaviors as sources of peer influence — question-answering & asking, discussion-driving, note-taking, and discipline-reminding.

*2.1.1 Question-answering & asking.* Answering and asking questions is a common form of student-instructor interaction that assists students in consolidating acquired knowledge and dispelling misconceptions [2, 4, 32]. The frequent question-asking and –answering from active students would create a social norm in the classroom that these behaviors are acceptable [101, 105]. It relieves the social pressure from other students, eventually encouraging them to answer and ask questions themselves. Meanwhile, thought-provoking questions asked by active students, along with questions that they have only partially answered, can stimulate further inquiry by opening a trail for others to delve into the topic [28, 104, 108]. Finally, active students asking questions can sometimes uncover shared gaps in the class's understanding, potentially due to key points being insufficiently addressed by the instructor [32].

*2.1.2 Discussion-driving.* Active students can propel classroom discussions by introducing new and intriguing perspectives, providing much-needed momentum when other students are uncertain of how to initiate or delve deeper into the topic [5, 69]. This behavior could complement the instructor's role in guiding the discussion.

*2.1.3 Note-taking.* As mentioned in section 2.1.1, the behaviors of individual students can help form social norms that influence the wider class. For example, when an active student promptly takes down notes at key moments during the instructor's lecture, it can subtly inspire others to follow.

*2.1.4 Discipline-reminding.* Disruptive behaviors, such as off-topic conversations, can have a negative impact on the overall learning environment within the classroom [19, 91]. Although it is generally considered the instructor's responsibility to enforce discipline, research has shown that too much control from the instructor comes at the risk of losing students' engagement [96]. On the other hand, students might respond more favorably to peer-issued reminders [97, 105]. Additionally, managing the discipline of an entire class could be demanding for instructors [35], which means they could benefit from peer intervention from active students.

Despite their contribution to the classroom, the presence of an active student exhibiting one or more of the previously mentioned behaviors is not a given. Many factors such as the individual personalities of the current students, their confidence in the subject matter, along with the social environment of the current classroom can influence whether an active student emerges [105, 117]. Therefore, we are motivated to create a virtual agent with a vast knowledge base and realistic behaviors to play the role of these active students.

## 2.2 Pedagogical Agent

A pedagogical agent is an anthropomorphic virtual character used in a virtual learning environment to serve educational purposes. The use of pedagogical agents has been positively associated with learners' knowledge acquisition, reasoning processes, and collaboration with peers [25, 39, 47, 57]. Pedagogical agents are powered by Natural Language Processing (NLP) technologies to understand, generate human-like responses, and interact intelligently with users [76]. Given the transformative impact of Large Language Models (LLMs) [10, 112] on the NLP landscape, we analyze the applications of pedagogical agents both before and after the introduction of LLMs respectively.

*2.2.1 Before the Era of Large Language Models.* Pedagogical agents assist in education through various forms of interaction with learners. As they facilitate learners' enhancement of their language skills by engaging them in lifelike conversations, pedagogical agents have been extensively used for language learning [48, 87, 115]. These agents, trained with subject-related knowledge and capable of directly answering learners' queries, hereby reduce instructors' burdens and enable the expansion of educational accessibility [82]. One of the most well-known examples of question-answer agents is the "Jill Watson" [45], an automated teaching assistant designed to answer student questions in an online course. Similarly, RobotAR [107] integrated this question-answer capability into a teleconsulting robot. Besides passively responding to learners' inquiries, pedagogical agents can also engage them in conversations. For instance, AutoTutor facilitated [47] learning by engaging learners in interactive dialogues in natural language, dynamically tailoring its responses and tutorials based on the student's actions. Similarly, Alaimi et al. [18] developed agents that can provide propositions and question starter prompts in dialogues with learners to foster their question-asking skills. Pedagogical agents also play a crucial role in supporting collaborative activities between learners. For example, MentorChat [100] prompted learners to elaborate on key domain concepts during their discussion as a way to trigger learners' disclosure towards their collaborators. Along a similar line,

Wang et al. [111] designed text-based agents to promote social connections among online learner groups. In the realm of addressing psychological aspects in education, certain pedagogical agents have been devised to alleviate learners' stress [62] and combat depression [81].

*2.2.2 During the Era of Large Language Models.* Introduced in 2018, BERT [36] is the pioneering Large Language Model (LLM), yet its impact largely remained within academia. It wasn't until the introduction of ChatGPT [10] in 2022, noted for its user-friendly interface, that attracted widespread attention and adoption from the general public. In the education community, researchers are beginning to reassess the potential roles of pedagogical agents, taking into consideration the transformative advantages that Large Language Models (LLMs) offer compared to traditional natural language processing methods [54]. The following sections elaborate on these advantages and how they have been capitalized.

***Wide Scalability:*** LLMs, pre-trained on extensive text corpora, have exhibited versatility in handling a wide range of tasks [88, 94]. The use of prompt, which is human language input providing instruction to LLMs, lowers the entry barrier by allowing end-users to get desirable results in an intuitive manner [10, 119]. In practice, ChatGPT has been widely used to answer learners' questions on various subjects [64, 77, 79]. Notably, this is accomplished without the need for user training on ChatGPT or any subject-specific system modifications.

***Extended Context Comprehension:*** Transformer-based LLMs, such as GPT and BERT, use mechanisms such as self-attention that can consider broader and non-sequential contexts [10, 36, 106]. This quality can be used to summarize lengthy and complicated documents for learners, such as textbooks and class notes [44].

***Creativity:*** Unlike traditional rule-based models, LLMs exhibit creativity by providing fresh perspectives that might not be predicted by users [46, 99]. They can help learners with writing assignments by stimulating thought processes and refining ideas using suggestive feedback. Additionally, they can be used to generate exercises and quizzes to aid in practice and assessments [44, 64].

***Roleplay Capability:*** Large Language Models (LLMs) have demonstrated the ability to simulate diverse characters and personas convincingly due to their capacity to generate contextually coherent and diverse responses [80, 103, 109]. Specifically, they can be assigned to a role that provides context about their identity and background, thus allowing them to generate more natural and in-character responses tailored to that role [109]. Duolingo [11], a language-learning app, capitalized on this quality by introducing the Roleplay feature, which enables users to practice real-world conversational skills with an agent playing various roles. For instance, users can pretend to order coffee from baristas or discuss future vacation plans with a travel agent.

In designing ClassMeta, we fully leverage the aforementioned advantages brought by LLMs. Firstly, we develop a prompt template for tuning GPT behaviors, paving the way for future educators to utilize ClassMeta across a variety of subjects. Secondly, we utilize LLM's context comprehension prowess to capture the real-time classroom dynamics, enabling the agent's note-taking, missing point reminder, and distraction correction capabilities. Thirdly, the LLM's creativity enables the virtual agent to raise profound questions pertaining to the lecture and contribute to class discussions by sparking creative thought. Finally, we deliberately configure the agent's role adjusted to that of a typical student with confined knowledge boundaries, rather than an all-knowing expert. This approach enables them to exhibit behaviors akin to a student, thereby enabling them to seamlessly blend into the classroom environment.

## 2.3 Virtual Reality Classroom

The design of a pedagogical agent also involves choosing its digital representation, whether it will be 2D or 3D and static or animated [68]. The choice of representation is dependent on the intended interaction between the agent and the learners and the role the agent is designed to fulfill [33]. As previously mentioned, our intention is for the agent to subtly cultivate a favorable behavioral norm within the classroom by engaging in both verbal and non-verbal interactions with the instructor and other students.

Therefore, we adopt animated 3D avatars in the virtual reality environment to represent our agent, which enables more nuanced and expressive communication beyond spoken language [66, 75]. For instance, the asking question behavior would be accompanied by raising hand gestures while the note-taking reminder is implemented entirely through the agent's body gestures. Meanwhile, studies have shown that VR brings a higher amount of social presence among its participants [24, 98], which gives the agent more potential to influence others.

VR has recently received increased attention as an educational tool [50] where its unique benefits have been widely investigated [21, 85, 116]. Specifically, virtual reality classrooms that emulate traditional learning environments have been demonstrated to foster genuine social interaction and engagement, which have been identified as critical in enhancing learners' motivation, persistence, and interest [63]. Research has suggested that the collaborative social presence, derived from the nature of virtual reality, is the key to VR's success in the classroom [51]. Hence, within a virtual reality classroom, all participants are intricately interconnected, and their actions exert mutual influence on each other [61]. For example, Gao et al. [42] studied how peer learners' engagement expressed by hand-raising behavior may also affect the attention and visual behaviors of learners in VR classrooms. These works underscore the potential of VR classroom as a learning platform. Keeping in line with their trajectory, we have developed ClassMeta to support VR classroom learning experiences.

Some works have developed virtual agents in the educational context. In the work done by Zhang et al. [118], a virtual agent was pre-programmed as a simulated language learner in VR while actual users assume the role of teacher. The purpose of using virtual agents was to establish a more controlled study environment, rather than to utilize agents' potential for educational enhancement. Similar to our work, Liao et al. [61] developed a virtual agent presented in the VR classroom. However, the behavior of their agent is simulated by the routine playback of pre-recorded and time-anchored comments from past learners. Therefore, the primary aim of their agent is to increase social presence by creating the illusion of learning with others. In contrast, ClassMeta expands the capability of the agent by utilizing LLM to synthesize adaptive behaviors based on real-time

context. This approach endows our agent with a more extensive scope of interactive abilities, which will be discussed in the next section.

# 3 INTERACTION DESIGN

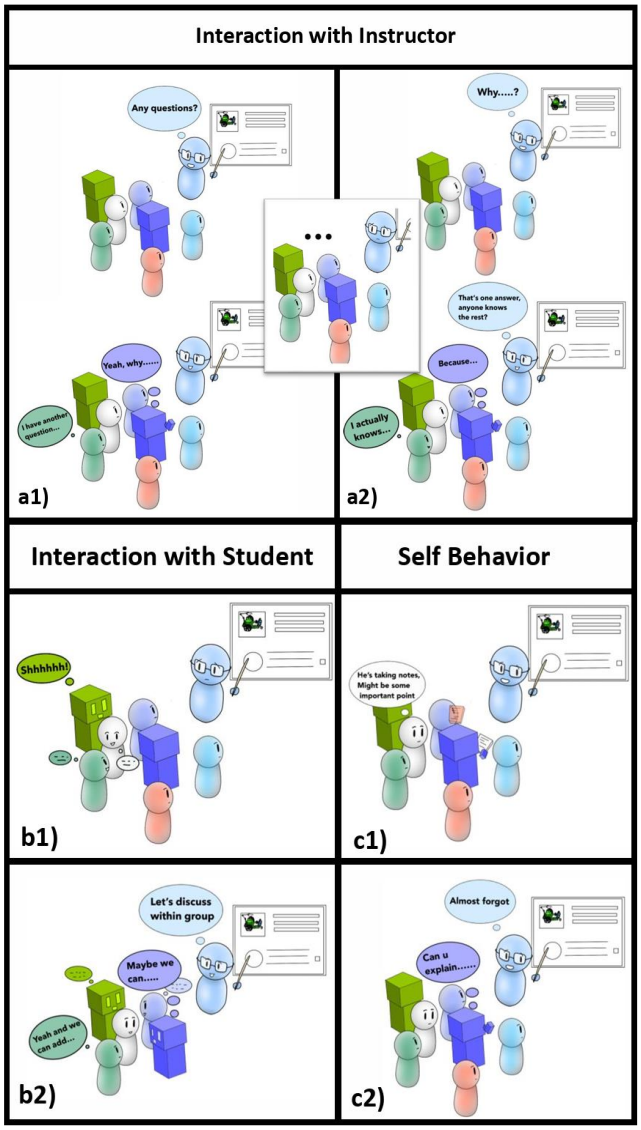


**Figure 2: Interactions designed for ClassMeta: (a1) Question answering. (a2) Question raising. (b1) Discipline reminding. (b2) Discussion promoting. (c1) Note taking. (c2) Missing key point reminding.**

In terms of the recipient, we design the behavior of a ClassMeta agent into three categories: *Interaction with instructor*, *Interaction with student*, and *Self behavior*. The motivation behind these interaction designs is discussed in 2.1

## 3.1 Interaction with instructor

Encouraging students to ask (Figure 2 a1) and to answer questions (Figure 2 a2) is a common way for instructors to engage students in the classroom. However, speaking up in front of the whole class could be intimidating, which makes students reluctant to ask questions themselves or react to questions from the instructor. If it happens to every student, the whole class will fall into complete silence, which leaves the instructor's engagement unresponded.

In this circumstance, the ClassMeta agent plays the role of an active student who can break the ice. When ClassMeta detects that the class has been silent for a certain amount of time, it will intervene by either raising a question or answering the question according to the context (Figure 2 a1 a2). The agent will only pose relevant questions based on its contextual knowledge of the lecture topic and the instructor's current query. When answering the question, the agent will intentionally answer a partial of the question, thereby giving the students the opportunity to complete the answer and increasing their participation.

## 3.2 Interaction with student

As the class progresses, some students may inevitably become distracted and engage in off-topic discussions with their classmates. This is often seen as disruptive behavior that negatively affects the entire class. By comparing captured student conversations to the lecture topic, the ClassMeta agent can identify and intervene in the off-topic conversations. As a disciplined student would, the agent will turn to the distracted students and issue verbal and non-verbal reminders (figure 2 b1).

The group discussion session in a classroom may experience a state of stagnation if students exhibit hesitancy in articulating their opinions or if they exhaust their new ideas. At this time, the ClassMeta agent will break the ice by bringing a fresh viewpoint to propel the discussion forward (figure 2 b2).

## 3.3 Self behavior

The self-behavior of an active student is also a key factor in influencing the quality of a class. The ClassMeta agent reproduces the self-behavior of an active student. When the instructor goes through the key points of the lecture, it will act like taking notes (Figure 2 c1). When the instructor fails to explain a concept clearly, the agent will raise a question about that concept which allows the instructor to explain the concept with more detail (Figure 2 c2). As opposed to asking questions in response to the instructor's intentional query (e.g., "Do you have questions" Figure 2 a1), this question-asking behavior is initiated proactively by agents, which assists instructors in addressing any unintentional oversight (e.g., missing any key point).

# 4 SYSTEM OVERVIEW

ClassMeta is developed with the Unity3D platform (2022.3.2f1) [1] and runs on the Oculus Quest Pro headset[14].

## 4.1 VR classroom setup

Our virtual classroom recreates the layout of a real classroom, with a podium and a projector screen at the front, which is a practice widely adopted in prior works [42, 43, 51, 83]. In our prototype

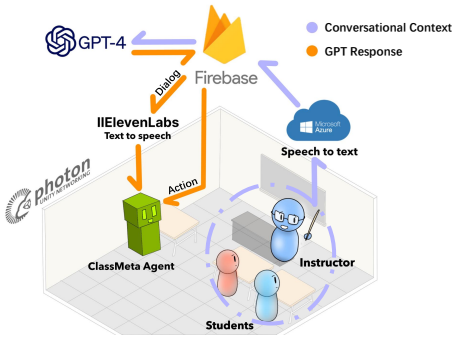


**Figure 3: Overview of the system structure of ClassMeta.**

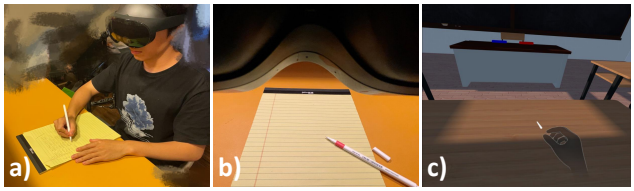


**Figure 4: ClassMeta allows users to physically take notes when wearing the Oculus Quest Pro: a) The user is taking notes. b) First-person view of the user. c) ClassMeta detects note-taking activity by detecting a pinch gesture.**

system, there are a total of six seats designated for the students, with two of them taken by two ClassMeta agents. These agents are intentionally positioned in the front left and front right of the classroom, ensuring their visibility to the human students. The multiplayer feature is supported by Photon Unity Networking [15] and allows real-time voice communication between users.

To collect the real-time conversational context of the class, we integrate Microsoft Azure SDK [7] to implement real-time speech-to-text conversion. All the converted text-based dialog will be uploaded to the Google Cloud Firestore [9], which is a NoSQL document database in Google Firebase that allows the system to store, sync, and query data (Figure 3). The Oculus Quest Pro headset has an open-view design, which allows the students to acquire a view of the desk. In this way, they can physically take notes throughout the lecture (Figure 4 a b). We implement a note-taking detection mechanism by using the hand gesture tracking capability of the Oculus headset. Once the student performs a pinch gesture on the desk (Figure 4 c), our system will automatically determine the student is taking notes and activate the highly visible 'taking notes' indicator above the head of the student's avatar. (Figure 5 c). In this way, the physical movement of note-taking is visualized in the virtual world shared by the students. We integrate the lecture slides into the virtual classroom and the instructor can easily navigate through each slide by clicking the next/previous buttons on the podium. There is a slide preview window on the podium so that the instructor does not need to turn around to see the current slide. To track the students' attention during the lecture for user study evaluation. We utilized the eye-tracking function from the Oculus Quest Pro headset. The attention-tracking mechanism will record the subject that the student focuses on, the time stamp when the student starts to focus on the subject, and the end time when the student moves his or her eyes away from the subject, along with the duration.

## 4.2 ClassMeta Agent Implementation

*4.2.1 GPT Integration.* The advantages offered by LLMs were well discussed in section 2.2.2. To fully leverage these advantages, we adopt GPT-4 — generally regarded as the state-of-the-art LLM — to empower ClassMeta.

The real-time conversation dialog in the classroom will be transcribed into text through Azure SDK [7] and then uploaded to Firestore [9]. When Firestore receives this text-based data from the VR environment, it will automatically relay it to the GPT-4 API. The responses generated by the GPT-4 API are then forwarded to the local ClassMeta agent, enabling it to exhibit dynamic behaviors. There are two types of responses we configure the GPT-4 to generate, corresponding to different types of interactions designed for ClassMeta described in section 3.

The first type is *action signal* (Figure 6 a). When the agent receives this type of response from the GPT-4 agent, it will perform predefined actions. The three supported action signals are: *Standby*, *Note-taking*, and *Discipline-reminding*. By default, the GPT-4 generates *standby* signals when the instructor delivers the lecture, which lets the agent remain in the standby state. When the GPT-4 detects that the instructor is emphasizing an important point, it will send a *note-taking* signal to the agent, causing the agent to engage in note-taking behavior (Figure 2 c1). GPT-4 achieves this process by comparing the lecture captured in real-time with the lesson's key points previously given to them as context. When the GPT-4 detects that the conversations between the students are irrelevant to the lecture, it will send a *discipline-reminding signal*. This is accomplished by GPT-4 through its capability to compare the topic of the conversation with the topic of the lesson. Accordingly, the agent will exhibit discipline-reminder behavior (Figure 2 b1). In summary, the *action signal* is a predefined response to dynamic situations in the classroom.

The other type of response is *dialog response* (Figure 6 b), which is a text-based response. When the agent receives this type of response from the GPT-4 API, it will use generative voice AI (i.e., 11ElevenLabs [8]) to convert the text into speech that imitates a human voice. In this way, the agent can verbally answer the instructor's questions (Figure 2 a1), raise questions Figure 2 a2) promote discussions (Figure 2 b2), and reminding missing point (Figure 2 c2), in the same manner as actual students.

The aforementioned dynamic responses are enabled by meticulously configuring the role of GPT-4 through our template, which will be elaborated in 4.3.

*4.2.2 Body movement.* The body movements of the ClassMeta agent are prerecorded by a human actor to ensure their authenticity. There are five categories of movements corresponding to each of the agent's behaviors. When standing by, the agent sits still and listens to the instructor with some random movements occasionally. When the agent responds to the instructor, it raises a hand and then starts talking (Figure 5 a). When the agent talks within the

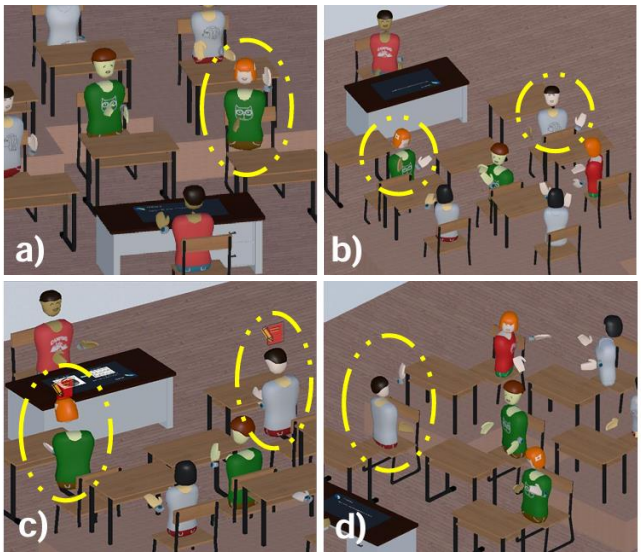


**Figure 5: Implementation of ClassMeta in Unity: (a) Reaction to the instructor. (b) Discussion within a group. (c) Note taking. (d) Discipline reminding.**


**Figure 6: Different responses from GTP-4: (a) Action signal. (b) Dialog response. (a1) Standby signal. (a2) Discipline reminding signal. (a3) Note-taking signal. (b1) Missing key point reminding. (b2) Question raising. (b3) Question answering. (b4) Discussion participating**

discussion, it relax and talk (Figure 5 b). When the agent receives a note-taking signal, it starts the note-taking movement and the note-taking indicator will be activated (Figure 5 c). When the agent attempts to correct the distracted student's behavior, it turns to the student, performs a hush movement, and makes a 'shush' sound (Figure 5 d).

## 4.3 Template for GPT tuning

In order to properly configure GPT-4 to imitate the behaviors of an actual student, we need to meticulously set its intended behavior through line-by-line prompts to cover the circumstances described in section 3.

We first inform the GPT of its student role in the class. To differentiate the messages between the instructor and the students, we defined the messages from the instructor starting with "[teacher]", and the messages from the students starting with "[student + id]" (Figure 6). Then we defined the conditions of whether the GPT should send a signal or a dialog response. We want the GPT to keep quiet most of the time during the lecture, so we ask the GPT to respond with **"..."**, which is the standby signal, by default (Figure 6 a1). We input the entire lecture script to the GPT so that it has the capability to compare the live class's conversational context with the original script. We then highlight the key points of the script to the GPT. Once the instructor mentions the key points, we ask the GPT to respond with **"+++"**, which is the note-taking signal (Figure 6 a3). If a student talks about lecture-irrelevant content, we ask the GPT to respond with **"- - -"**, which is the discipline reminding signal (Figure 6 a2). After the instructor finishes the lecture, the GPT compares the live lecture with the script. If the instructor fails to cover a concept clearly, we ask the GPT to send a dialog response to raise a question about that concept which allows the instructor to explain the concept in more detail (Figure 6 b1). During the phases that involve student participation, we don't want the GPT to respond immediately after a question or discussion topic is raised. Instead, we ask the GPT to send a dialog response only after the system sends a "[talk]" command to the GPT. When the system detects there is silence in the class, it will send the "[talk]" command to the GPT. We ask the GPT to keep the response in short sentences and insert modal words such as *"hmmm"*, *"ah"*, etc., which help the dialog responses sound more natural [27, 34] (Figure 6 b2 b3 b4).

The process of configuring the role of the GPT-4 can be laborious and involve repetitive work. Thus we have summarized our experience and compiled a template for tuning the GPT. There are input fields in this template for users to enter their customized content (e.g., course materials, key points, behavior parameters). In this way, we endeavor to help future researchers and educators implement their own version of agents and deploy them in their own context. The full version of the template, along with usage instructions, is included as supplemental material.

## 5 USER STUDY

We perform an IRB-approved between-group comparative user study by splitting our users into two conditions: a) a baseline VR classroom that only involves a human instructor and human students, and b) ClassMeta, which involves the GPT-empowered agents we designed. The same lecture was delivered to students in both conditions. For the baseline condition, ClassMeta agents merely

stand by with their default movement and do not interact with human students. For the ClassMeta session, the agents were activated with their full functionality. In both conditions, the virtual reality classroom reproduces the environment of a real-world classroom, an approach consistently employed in previous research [42, 43, 51, 83].


**Figure 7: The overall setup of the VR classroom for the user study.**

## 5.1 Participants

We recruited 24 participants from the college (16 male, 8 female), ranging from freshman to senior (M=21.3, SD=2.09), all of whom had experience with online classes. 75% of the participants are not majoring in engineering. 75% of the participants had no knowledge of engineering design methodology. 17% of the participants had some knowledge of engineering design methodology, and 8% of the participants had used the engineering design methodology on a project. The students with engineering design backgrounds were equally split between conditions. We conducted 3 user study sessions for each condition; each session involved 4 participants.

## 5.2 Procedure

*5.2.1 Classroom Setup.* At the beginning of the user study, each participant was physically assigned to an individual room and was told that their classmates were physically located in other rooms for the remote VR lecture experience. The purpose behind this setup was to maintain the authenticity of ClassMeta agents by making the discrepancy in quantity between the four human participants and the six virtual avatars unnoticeable. Each participant received the context of the lecture content and structure, and was requested to complete a pre-test on paper. The pre-test assessed the participant's previous knowledge of the lecture content. Then participants wore the headset and joined the VR classroom. The participants were told to behave naturally as if they were in a real classroom, which is a standard practice in prior studies on VR classrooms [21, 85, 116].

The VR classroom was configured to contain six spots for the students, while four spots are for the human students and the other two spots are for the ClassMeta agents (Figure 7). To optimize the visibility of the ClassMeta agents for every participant, we have set up the classroom as a tiered classroom, where each row is in a W-shape.

*5.2.2 Lecture Content & Structure.* Like prior work that evaluates learning [120], we measured the learning gain from the students in terms of both *conceptual knowledge acquisition* and *logical thinking skills*. Therefore, we selected a 15-minute introductory lecture on the basic engineering design methodology that meets the requirement. This lesson encompasses the eight phases of the engineering design process, which are *Ask*, *Research*, *Specify*, *Imagine*, *Plan*, *Create*, *Test*, and *Improve* [3]. Through the lecture, students can both acquire conceptual knowledge and develop logical thinking skills. For example, when the lecture goes through the *Test* phase, it not only covers the basic concepts of testing an engineering product but also covers the logical thinking process of how to design a test for a product. We compiled the key points from the lecture and transformed them into six key competencies, which are: *K1: The order of the engineering design process. K2: The overall summary of each design phase. K3: The details about each design phase. K4: Use of a decision matrix. K5: Convert problems to engineering requirements. K6: Design the test for products.* K1-3 are concept-based, while K4-6 are logical-based. We designed a pre-test and post-test for the participants to assess their learning gains on these key competencies.

To comprehensively evaluate the various interactions implemented in ClassMeta (Figure 3) in a more structured manner, the lecture was organized to include four sequential phases. The descriptions of the phases are as follows:

Phase 1: Lecture content delivery

- The instructor delivers the lecture content based on the lecture script and the lecture slides. The lecture lasts 15 minutes.
- The participants are provided with a notebook to take notes.
- In the ClassMeta condition, the ClassMeta agents will take note when detecting pre-defined key points during the lecture.
- The note-taking indicator will be displayed on top of the virtual avatar who is taking notes.

Phase 2: Question-raising

- After the instructor finishes the lecture content delivery, the instructor asks if any students have any questions.
- The questions raised by the participants are recorded.
- The instructor can intervene to encourage student participation in both the baseline and the ClassMeta conditions. The interventions from the instructor are recorded.
- In the ClassMeta condition, the ClassMeta agents will raise questions if no one else does. The interventions from the ClassMeta agents are recorded.

Phase 3: Question-answering

- The instructor asks five lecture-related questions to the participants. The first four questions are concept-based, while the last question is open-ended and requires logical thinking.
- The instructor can intervene to encourage student participation in both the baseline and the ClassMeta conditions. The interventions from the instructor are recorded.

- In the ClassMeta condition, if no participants respond, the ClassMeta agent will provide a partial response to the question, allowing the participants to contribute. The interventions from the ClassMeta agents are recorded.
- The first four questions are recorded as solved, partially solved, or unsolved. Insights generated for the last open-ended question are recorded.

Phase 4: Discussion

- The instructor asks the students to discuss a particular lecture-related topic.
- The instructor can not intervene in the discussion as it was meant to simulate the discussion within one of the break-out groups in a large classroom.
- In the ClassMeta condition, if silence is detected, the ClassMeta agent will join to facilitate the discussion.
- Insights generated for the discussion topic are recorded.

As a standard practice to ensure the consistency and the quality of the lecture [107, 120], one of our researchers affiliated with the Graduate School of Engineering assumed the role of instructor in both conditions. After the lecture, the students were asked to complete a questionnaire evaluating their class participation preference after experiencing the lecture followed by a post-test for learning gain evaluation. At the end of the study, the identities of the ClassMeta agents were disclosed to the students who participated in the ClassMeta session and an interview was conducted.

*5.2.3 Evaluation Metrics.* Throughout the lecture, we utilized the eye-tracking function of the Oculus Quest Pro to determine where the students' attention was directed. The notes taken from the students were collected after the lecture and graded based on their quality. During the question-raising and the question-answering, the interventions from the instructor and the ClassMeta agents were recorded. The outcomes of these stages were also recorded for evaluation. For the discussion, the results of the discussion were recorded. The participants took a pre-post test before and after the lecture for learning gain evaluation. The pre-test and post-test both contain six questions corresponding to the six key competencies and come up with version A and version B. Half of the students take version A at the pre and version B at the post. Vice versa, the other half started with version B and then took version A in the post. So the issue with test-retest fatigue was minimized. The two versions cover the same key competencies while the questions differ from each other. After the lecture, the students were asked to complete a questionnaire evaluating their class participation preference after experiencing the lecture. At the end of the study, the identities of the ClassMeta agents were disclosed to the students who participated in the ClassMeta session and an interview was conducted.

# 6 RESULT EVALUATION

## 6.1 Eye-tracking Results & Instructor Observation Record

The eye movements of the students are recorded throughout the lecture session. As shown in figure 8, we recorded the average total fixation duration that the students focused on the instructor, lecture slides, and ClassMeta agents. It is statistically significant that the students from the ClassMeta session pay more attention to the agents compared to the students from the baseline session. In addition, the timestamp when the student's attention shifts to the agents mostly coincides with the timestamp when the agents initiate a move, which suggests that the behaviors of the ClassMeta agents were indeed attracting the students.

This observation suggests that the lecture experience improvement, which is reflected by the rest of the results (e.g., pre-post test), could be attributed to the agent's behaviors. No statistically significant differences were found for the focus time on the instructor and the lecture slides, which means no evidence indicates that the agents were distracting students from the lecture experience.

The notes that the students took were collected and graded. Figure 9 shows sample notes with different scores. From Figure 10, we observe a significant difference in note quality between the baseline group and the ClassMeta group [Z=-3.377, p<0.01, p=0.0003], suggesting that the note-taking behavior from the agents were positively affecting students' own note-taking.

The record in Figure 11 shows the occurrence of the instructor's and ClassMeta agents' intervention during the lecture and the outcome of each phase. Instructor intervention is not applicable to phase one as it only consists of lecture content delivery. While the interventions from the instructor are in the form of *"Anyone can*

| Student Focus | Baseline (seconds) | | ClassMeta (seconds) | | | |
|---|---|---|---|---|---|---|
| | M | SD | M | SD | Sig. | |
| Instructor | 367.738 | 155.191 | 391.808 | 150.467 | Z | -0.481 |
| | | | | | p | 0.315 |
| Lecture Slides | 977.577 | 267.365 | 956.916 | 299.601 | Z | 0.573 |
| | | | | | p | 0.185 |
| ClassMeta Agents | 135.093 | 124.062 | 275.960 | 176.469 | Z | -1.891 |
| | | | | | **p** | **0.0293** |

**Figure 8: Results from the students' focus based on eye-tracking data of ClassMeta vs. Baseline.**

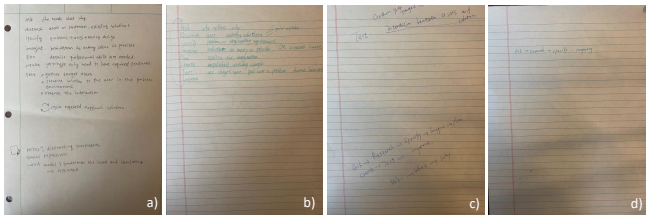


**Figure 9: Samples of user's note with different scores: (a) 4 points. (b) 3 points. (c) 2 points. (d) 1 point.**

| Note Quality | Averge Score | Standard Deviation | Sig. | |
|---|---|---|---|---|
| Baseline | 1.250 | 1.422 | Z | -3.3772 |
| ClassMeta | 3.583 | 0.793 | p | 0.0003662 |

**Figure 10: Results from average scores on the note quality of ClassMeta vs. Baseline.**

| | Lecture Phase | | Session 1 | | | Session 2 | | | Session 3 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Instructor | ClassMeta | Outcome | Instructor | ClassMeta | Outcome | Instructor | ClassMeta | Outcome |
| Baseline | Question-Raising | | 1 | N/A | 1 qeustion | 2 | N/A | 0 question | 1 | N/A | 1 question |
| | Question-Answering | Q1 | 2 | N/A | Solved | 2 | N/A | Solved | 2 | N/A | Solved |
| | | Q2 | 0 | N/A | Solved | 1 | N/A | Partially Solved | 1 | N/A | Partially Solved |
| | | Q3 | 0 | N/A | Solved | 0 | N/A | Solved | 0 | N/A | Solved |
| | | Q4 | 0 | N/A | Solved | 0 | N/A | Solved | 0 | N/A | Solved |
| | | Q5 | 1 | N/A | 2 insights | 1 | N/A | 2 insights | 2 | N/A | 2 insights |
| | Discussion | | N/A | N/A | 2 insights | N/A | N/A | 3 insights | N/A | N/A | 2 insights |
| ClassMeta | Question-Raising | | 1 | 1 | 2 questions | 0 | 1 | 2 questions | 1 | 1 | 2 questions |
| | Question-Answering | Q1 | 0 | 1 | Solved | 1 | 1 | Solved | 1 | 1 | Solved |
| | | Q2 | 0 | 0 | Solved | 0 | 1 | Solved | 0 | 1 | Solved |
| | | Q3 | 0 | 0 | Solved | 0 | 0 | Solved | 0 | 0 | Solved |
| | | Q4 | 0 | 0 | Solved | 0 | 0 | Solved | 0 | 0 | Solved |
| | | Q5 | 0 | 0 | 2 insights | 1 | 1 | 3 insights | 0 | 0 | 3 insights |
| | Discussion | | N/A | 1 | 4 insights | N/A | 2 | 5 insights | N/A | 2 | 5 insights |

**Figure 11: Results from the instructor's observation on the intervention and question/discussion outcomes in the lecture of ClassMeta vs. Baseline.**

*answer?", "Any thoughts?" etc.*, the interventions from the ClassMeta agents are in the form of raising a question, answering a question, or initiating a discussion. It is noticeable that the occurrence of intervention from the instructor decreased with the presence of ClassMeta agents.

The outcome differences between question 5 and the discussion session are worth noting. While both question 5 and the discussion were around an open-ended topic, the distinction lies in whether the instructor intervened. It can be observed that the outcome of question five, which benefited from instructor intervention, was comparable to that of the discussion, whereas the outcome of the discussion session without instructor intervention exhibited differences between the baseline and ClassMeta conditions.

Based on the observations, we have identified instances in which the ClassMeta agent behaviors could potentially enhance participant engagement. For example, once during a question-raising phase, when no participant tended to raise a question, the ClassMeta agent raised a question: *"Ahh, um, how do we make sure, you know, the prototype works the same as the final product will?"*. Based on this question from the agent, a participant (P2) came up with a follow-up question: *"Is there a case when the difference in materials between the prototype and final product results in an unexpected outcome?"*. Once during a question-answering phase, the instructor raised the question: *"Can anyone tell me the three questions we are using during the Ask stage?"* The agent provided an answer after correctly detecting the silence in the class: *"Err, during the Ask stage, we ask what's the problem and who needs it solved... right?"*, which deliberately missed one point out of three. One participant (P4) noticed that this answer was not complete and followed up to complete the answer: *"And also, we ask why it is important to solve the problem."*. Once during a discussion phase on how to research on the customer side, after a participant finished speaking, the classroom fell into silence again. The agent then intervened and provided insight: *"Ah, we can check product reviews to gather information about consumer experiences and their satisfaction level, I suppose."*, which brought back the discussion with two more participants who came up with ideas of *"Observe customer interact with the current solution"*(P9), and *"Conduct in-person interviews"*(P7)

## 6.2 Post-Lecture Survey

After the lecture, the students completed a 5-point Likert scale (1-strongly disagree, 5-strongly agree) survey regarding their class participation preference based on the lecture experience. This survey aimed to evaluate the impact on students' behavior from ClassMeta vs. the baseline condition. Figure 12 shows the average scores reported by the students.

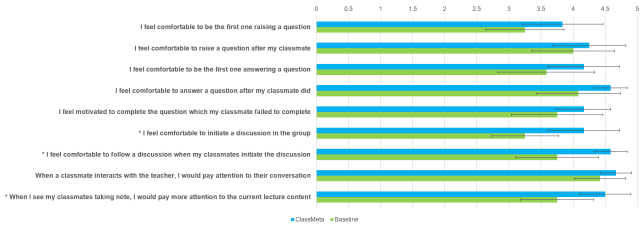


**Figure 12: Results from average scores on the post-lecture questionnaire of ClassMeta vs. Baseline. We used a 5-point Likert scale (1-strongly disagree, 5-strongly agree. (*) : p<0.05).**

The Shapiro-Wilk normality test suggested that the normal distribution assumption was not met. Thus, to analyze the significance of our results we used a Mann-Whitney U test. The results show no statistically significant difference in Q1[Z=-1.282, p>0.05, p=0.100], Q2[Z=-0.662, p>0.05, p=0.254], Q3[Z=-1.010, p>0.05, p=0.156], Q4[Z=-0.682, p>0.05, p=0.248], Q5[Z=-0.457, p>0.05, p=0.324], Q8[Z=-0.641, p>0.05, p=0.261]. This result is expected as it indicates that the instructor's intervention can maintain classroom participation, compensating for the lack of inherently active students (e.g., ClassMeta). It is particularly worth noting that there exist statistically significant differences in Q6[Z=-1.929, p<0.05, p=0.027], Q7[Z=-1.7204, p<0.05, p=0.043], and Q9[Z=-1.714, p<0.05, p=0.043], where the instructor exert no influence to the class and thus agents play a larger role.

## 6.3 Pre-Post Test Evaluation

The pre-post test covers six key competencies, which are: *K1: The order of the engineering design process. K2: The overall summary of each design phase. K3: The details about each design phase. K4: Use of a decision matrix. K5: Convert problems to engineering requirements. K6: Design the test for products.* K1-3 are concept-based while K4-6 are logical-based. Each answer in the test was scored with a 0 if incorrect, +0.5 if the answer had some substance, or a +1 point if the answer was correct. Since a question in the test covers one or several of these key competencies, we add the score to the overall sum of the corresponding competency. Then the total points were normalized into a 1-point scale for each key competency. This grading scheme is borrowed from past work [107, 120]. All tests were graded by one primary grader. Inter-rater reliability on both the pre-test and post-test was validated by having a second person score over 25% of the data. From our rubric, two researchers in charge of grading had a Cohen's Kappa of 0.723. The results of the pre-test and post-tests are summarized in Figure 13. To compare the baseline condition and the ClassMeta condition, we evaluate the learning gains of each key competency from each condition. The Shapiro-Wilk normality test suggested that the normal distribution assumption was not met. Thus, we used a Mann-Whitney U test again to analyze the significance of the results. The results show statistically significant

differences in K2 [Z = -2.141, p<0.05, p=0.018], K4 [Z = -1.715, p<0.05, p=0.043], K5 [Z = -1.681, p<0.05, p=0.046], K6 [Z = -2.662, p<0.05, p=0.004]. In summary, the statistical significance in four out of six key competencies indicates that ClassMeta agents have the potential to improve the learning gains for the students.

| | | | Baseline | | ClassMeta | | | |
|---|---|---|---|---|---|---|---|---|
| Type | Key Competency | Term | M | SD | M | SD | Sig. | |
| Concept | K1 | Pre-test | 0.417 | 0.195 | 0.375 | 0.226 | | |
| | | Post-test | 0.917 | 0.195 | 1.000 | 0.000 | Z | -1.386 |
| | | Gain | 0.857 | 0.195 | 1.000 | 0.000 | p | 0.083 |
| | K2 | Pre-test | 0.333 | 0.156 | 0.333 | 0.245 | | |
| | | Post-test | 0.786 | 0.121 | 0.854 | 0.144 | Z | -2.141 |
| | | Gain | 0.680 | 0.139 | 0.781 | 0.128 | **p** | **0.016** |
| | K3 | Pre-test | 0.500 | 0.131 | 0.615 | 0.135 | | |
| | | Post-test | 0.698 | 0.135 | 0.823 | 0.099 | Z | -0.7297 |
| | | Gain | 0.396 | 0.362 | 0.541 | 0.254 | p | 0.233 |
| Logical Thinking | K4 | Pre-test | 0.875 | 0.250 | 0.646 | 0.445 | | |
| | | Post-test | 0.854 | 0.249 | 0.917 | 0.195 | Z | -1.7147 |
| | | Gain | -0.167 | 0.289 | 0.765 | 0.444 | **p** | **0.043** |
| | K5 | Pre-test | 0.375 | 0.272 | 0.396 | 0.249 | | |
| | | Post-test | 0.604 | 0.271 | 0.729 | 0.167 | Z | -1.681 |
| | | Gain | 0.368 | 0.400 | 0.563 | 0.316 | **p** | **0.046** |
| | K6 | Pre-test | 0.375 | 0.169 | 0.292 | 0.257 | | |
| | | Post-test | 0.625 | 0.131 | 0.750 | 0.213 | Z | -2.6619 |
| | | Gain | 0.400 | 0.255 | 0.647 | 0.255 | **p** | **0.004** |

**Figure 13: Pre-test, test, and post-test results of key competencies assessment.**

## 6.4 Post Study Interview

For participants in the ClassMeta session, we eventually revealed the existence of the GPT-powered agent and briefly explained how the interactions from these agents were enabled. Then we acquire their perception of the agent as well as their general opinion on utilizing an agent to promote the learning experience through an open-ended interview. As expected general attitude towards the agent was positive, with one exception — *"It could sometimes cause peer pressure on me (P1)"*. A possible solution to this issue is further discussed in Section 8.2. In particular, participants appreciated the fact that the agent can *"break the silence (P2)"*. Silence in the classroom when no one answers a question is extremely *"awkward (p6)"* and *"bothering (P4)"*, which could *"force me (them) to be that first one to talk (P6)"* even if they do not want to. Participants further reported that *"even if I am not participating in the discussion, I can still learn more by paying attention between the agents and instructor (P7)"*. Participants also highlighted the effect of the non-verbal behaviors of the agents — *"When I saw people around me taking notes, I would pay more attention to the current lecture content(P10)"*.

Regarding the agent's authenticity, only three out of twelve participants indicated that they suspected the agent's identity prior to being informed. When asked what was suspicious about the agent, they all reflected that the tone of their voice sounded unnaturally mechanical (*"it sounds like the AI-generated voice I heard on TikTok (P8)"*). However, the agents' body language and conversational coherence were not deemed unusual. This is a promising result as it indicates that ClassMeta's interaction behavior demonstrated high authenticity. The only drawback, stemming from the limitations of our adopted voice synthesis technology, could be further minimized by incorporating more advanced AI technology for voice synthesis [110].

# 7 EVALUATION OF KEY BEHAVIORS

To comprehensively evaluate agents' key behaviors, we perform a technical evaluation where we simulate respective scenarios covered in our design space (Figure 2) — *note taking*, *discipline reminding*, *missing key point reminding*, *question raising*, *question answering*, and *discussion promoting*. By observing what responses were generated under what circumstances, we can have a better understanding of the agents' performance.

In the subsequent sections, we will analyze selected sample responses from the agents. A comprehensive set of responses from the agents gathered in this study can be found in the supplemental material.

| | Note taking | Missing key point reminding | Discipline reminding |
|---|---|---|---|
| **True Positive** | 100% | 84% | 100% |
| **False Negative** | 0% | 16% | 0% |
| **True Negative** | 98.55% | 98% | 88% |
| **False Positive** | 1.44% | 2% | 12% |
| **Accuracy** | 98.7% | 91% | 93% |

**Figure 14: Performance of the *Note taking*, *Discipline reminding*, and *Missing key point reminding* functions.**

## 7.1 Quantitative Evaluation

We simulated each scenario outlined in our design space to evaluate the performance of each behavior by manually inputting classroom conversations while observing the response from the agent. For the note taking testing, the lecture was presented 10 times and included 100 key points in total (10 key points per lecture). To ensure the tests were realistic, each key point was delivered slightly differently from the script. For the missing key point reminding testing, the lecture was presented 100 times, while a random key point was skipped in half of the lectures. The reactions of the agents to missing the key points were recorded. The questions raised by the agents in response to the skipped key point were collected for qualitative evaluation. For the discipline reminding evaluation, 50 off-topic conversations were simulated, which covered food for lunch, game day schedule, movie night, computer gaming, social media, etc. 50 topic-related conversations were also simulated, which covered interactions with the instructor and discussions. The reactions of the agents to the students' conversations were recorded. Figure 14 shows the performance of these behaviors.

***Note taking:*** When the instructor covers a predefined key point from the template, the agent is configured to take notes. The result showed that the agent successfully detected all the predefined key points and achieved an accuracy of 98.7%.

***Missing key point reminding:*** When the instructor misses covering a topic during the lecture, the ClassMeta agent is configured to detect this incident by comparing the live lecture to the predefined lecture script, and then remind the instructor by raising a question associated with this topic. For missing key point reminding, the true positive rate was 84%, and the accuracy was 91%. The relatively high false negative rate may be caused by the

long context length of the GPT-4. Given that there may be a portion of the missing key point covered by the rest of the lecture script, the GPT-4 might determine that the key point was well covered.

***Discipline reminding:*** When the student instantiates an off-topic conversation, the ClassMeta agents can identify and intervene in the off-topic conversation. The system shows the outstanding capability of detecting off-topic conversations with a true positive rate of 100% and an accuracy of 93%. Several false positive cases happened when the statements from the student were brief or vague. For example, when the discussion was about how to conduct market research, the agent engaged the student when the student said: *"How about interviews?"*. The GPT-4 might assume that the student was trying to start a topic related to personal job interviews, while the student was expressing that customer interviews can be a method to conduct market research.

We did not conduct quantitative tests for question raising, question raising, question answering, and discussion promoting, since the trigger mechanism was based on silence detection that could not fail. The responses of the agents were collected for a qualitative evaluation.

## 7.2 Qualitative Evaluation

***Missing key point reminding:*** When the ClassMeta agent detects the instructor fails to cover a topic during content delivery by comparing the live lecture to the predefined lecture script, it will raise a question related to the topic after the instructor finishes the lecture (Figure 2 c2). For example, after the instructor skipped the part *"By looking at existing solutions, we can know what approach has been made to solve similar problems, what mistake has been made that we can avoid, and what direction has never been explored that can change the game."* from the script, the ClasMeta agent asked: *"Ahh, I think we missed something about how existing solutions influence our design. Could you elaborate on that, please?"*, which is a proper question that allows the instructor to make up the missing key point.

We also noticed there are cases where the agent generates questions beyond the capability of a normal student. For example, after the instructor skipped the part *"The second step is to present your solution to these users while the users are in the problem environment. The problem environment is the situation or atmosphere in which the problem you are trying to solve happens."*, the ClasMeta agent asked: *"Ahh, did we miss the second step in user testing about presenting the solution in the problem environment?"*. In this case, the question from the agent seems too specific since the instructor never mentioned the topic.

***Question raising:*** The ClassMeta agent is capable of raising questions related to the lecture topic (Figure 2 a1). For example, the lecture introduces a method called *"Decision Matrix"* to help engineers choose the best product concept based on the scores from the decision matrix. The agent raised a question: *"Hmmm, you mentioned the decision matrix but didn't cover what happens in case of a tie. How would we break it?"*, which could bring further discussion on the use of a decision matrix.

***Question answering:*** When no one responds to the instructor's question, the ClassMeta agent will intervene and provide an incomplete version of the correct answer (Figure 2 a2). This behavior is intentionally designed to promote engagement and learning through correction by human students. For example, given the question: *"Can anyone tell me the three questions we are using during the Ask stage?"*, the correct answer is: *"We ask what is the problem, who has the problem, why is the problem important to solve."*. The agent provided answers such as: *"Err, during the Ask stage, we ask what's the problem and who needs it solved... right?"*, which missed part of the correct answer; or: *"Ahh, I think we need to ask: Who has the need? What is it? And, umm, where do they need it?"*, which contained a mistake; or: *"Ahh, so the questions are what, who and why, right?"*, which required more details to complete the answer.

***Discussion promoting:*** The ClassMeta agent can engage in the discussion sessions by bringing a fresh viewpoint to propel the discussion forward (Figure 2 b2). For example, in the lecture, the discussion topic is related to *"How to conduct background research on the area of the customers?"*. The agent provided several suggestions, such as: *"Hmm, we can conduct customer surveys or interviews to understand their specific needs and problems."*, *"Ahh, yes! We could look at reviews on comparable products or solutions to gain insights."*, *"Ahh, we could also observe user behaviors in their natural settings for a more accurate understanding, hmmm... I guess."*, *"Ahhh, maybe we could observe them using similar products, see where they face problems, hmmm... that could help."*. These suggestions provided reasonable approaches to the discussion topic.

# 8 DISCUSSION AND FUTURE WORK

## 8.1 Bridging Education Inequality Gap

ClassMeta was motivated to address the gap that active students, who contribute to a positive classroom environment, are not always present in all situations. This gap risks being amplified –– a phenomenon known as the "Matthew Effect [56]" –– as active students often gravitate toward similar educational environments while less active students may find themselves naturally clustered in more homogeneous learning environments. Specifically, research has revealed a correlation between student activity and academic performance, suggesting that active students are more likely to excel academically [72]. However, these students' parents typically prefer to enroll them in schools with other high-achieving students, as they do not want their children to be held back by others. As Stiglitz mentioned in his paper "Education and Inequality" — *"if the parents of the better students could do so, they would form a school consisting only of like students and exclude the poorer students."* Although such actions motivated by self-interest are understandable, they would inevitably exacerbate the gap by centralizing the distribution of top students. In addition, students' personalities play an important role in how much they participate in class, and these personalities can often be shaped by their family's socioeconomic background [20, 89]. Therefore, classrooms in a school district comprised primarily of students from less privileged families would face disadvantages when enjoying the benefit of active student participation. Lastly, prior research has indicated that instructors form closer and more positive relationships with active students [113], inadvertently allocating more energy and resources to them, which would be unfair to those who remain relatively quiet in class. With ClassMeta, educators now have the opportunity to level the playing

field with simulated active students. Meanwhile, the instructors know the true identity of the virtual agent, so they would not spend a disproportional amount of effort to meet the agents' needs (which theoretically do not exist) as they would for actual active students.

## 8.2 Personalized Virtual Agent

Although the overall sentiment toward the virtual agent was positive, we do receive feedback claiming that the agent could occasionally induce anxiety by displaying its superior knowledge during interactions with the instructor. Previous studies observed similar findings from actual active students [26, 37, 74], suggesting that this issue is intrinsic to classroom dynamics that involve active learners.

It is important to note that this effect does not apply to every student equally, as different people may react differently to their peers displaying superior knowledge, as some find it anxiety-inducing and others find it motivating [17, 72, 84]. Meanwhile, some research [70, 71], focused on the perspectives of students with quiet personalities, has suggested that some quiet students believed that their silence was a sign of interest, commitment, and attentiveness. They also felt that their silence helped them learn better. Therefore, their quiet demeanor is a conscious choice rather than a sign of disengagement. To accommodate this unique personality, the virtual agent would need to be customized to allow them to express their engagement in ways other than talking.

With these factors considered, we foresee the customization of virtual agents based on each student's unique personality as a promising future direction. Virtual Reality, as a flexible digital medium, could facilitate this process. In virtual reality, agents can be configured to expose different information to each student, as opposed to a traditional classroom setting where everyone has the same experience. In other words, the agent that student A sees is verbally and non-verbally distinct from what student B sees. Meanwhile, this endeavor poses the unique challenge of evaluating students' personalities as the basis for customization. The traditional method of evaluating characteristics using a questionnaire could be a solution, but it might not be accurate [121]. Xiao et al. [114] recently designed a chatbot that can automatically infer the personality traits of students through their conversations. This research presents an opportunity to design an end-to-end system agent that first engages in conversations with each student to infer their personalities and then automatically adapts itself to accommodate them.

## 8.3 Ethical Considerations of AI Agent as Classmate

The adherence to ethical principles in the development and application of AI technology remains a top priority within the research community [52, 55]. Meanwhile, as AI technology evolves, it continues to spark new discussions around establishing updated standards concerning AI ethics. For instance, the recent advancements in Large Language Models have allowed them to generate language-based content that is becoming increasingly indistinguishable from human-produced content. This development has ignited discussions about the necessity for regulatory measures to oversee AI-generated content on the Internet in order to preempt potential deception [58, 95].

In the field of AI-powered pedagogical agents, one particularly prominent ethical issue, which has yet to reach consensus, involves maintaining the transparency of these agents. The main concern is whether it is essential to ensure students know they are interacting with AI, not humans. Although some insist on the necessity of maintaining transparency [12], it counters a common AI objective of passing the Turing test, which consists of producing conversation indistinguishable from that of a human [31, 78]. In previous work, the typical use of pedagogical agents involved creating a dyadic relationship between the agent and the students. In those scenarios, any negative influence from the agent (for example, providing an incorrect answer to a question) directly impacts students, making the effect more significant. This is different from our approach in ClassMeta, where the interaction between ClassMeta and the students is more indirect, where the presence of an instructor acting as a mediator ensures a minimal impact from the agent. Therefore, arguments could be made that some ethical concerns around traditional pedagogical agents may not apply in our context.

At the end of our user study, we revealed the existence of the AI agent to the users. Notably, no students expressed objections to the presence of the AI agent. Some, in fact, advocated for late disclosure to ensure a more *"authentic experience (U1)"* with the agent. However, our study's relatively short duration limits our ability to identify potential long-term issues, such as students forming an emotional bond with the agent over a longer period, only to discover later that it is not a real human. Looking forward, we anticipate that disclosure of the AI agent's existence to student participants will be necessary. However, deciding when and how to make this disclosure poses a delicate balance between maximizing the effectiveness of the agent and avoiding potential ethical issues. As the research on the ethics of pedagogical agents, especially those who no longer have a dyadic relationship with students [90], is still in its early stages, we propose this as a promising direction for future research.

## 8.4 Limitations and Future Works

While certain agent behaviors like answering and asking questions are solely feasible through the language generation capability of Large Language Models (LLMs), other behaviors might also be implemented using traditional rule-based models. Our decision was to apply LLMs consistently to all behaviors due to the fact that they all rely on the same real-time context input. This approach offers greater cost-effectiveness in comparison to employing rule-based models for specific behaviors individually. Meanwhile, the effectiveness of rule-based models versus LLMs in these scenarios is still an open area of research. Initial studies indicate that LLMs, being pre-trained on extensive knowledge databases, show improved performance in contextual comprehension, dealing better with nuances and ambiguities in conversations [22, 23, 30]. However, the applicability of these findings to our specific situation remains unexplored. In the future, we aim to conduct a comparative user study between LLMs and rule-based models to quantify the advantages of using LLMs in identifying pre-defined behaviors.

Currently, we've used a heuristic method to determine the ratio of agents to students in our study, where four students are able to observe the behaviors of two agents. Upon further analysis, we hypothesize there would be an optimal ratio of agents to students. If the ratio is too low, some students might not be able to observe the agents, missing out on their influence. Conversely, if the ratio is excessively high, the impact of the agents could arrive at a stage where the law of diminishing returns applies [93]. This could lead to unnecessarily large class sizes, which brings its own set of issues [40, 49]. In the future, we plan to explore the optimal agent-student ratio by conducting studies to compare outcomes from various ratios.

For ClassMeta, we selected virtual reality as the digital platform to host the virtual agent as it enables the agent to display both verbal and non-verbal behaviors in a more immersive and realistic manner [66, 75]. Studies have shown that VR brings a higher amount of social presence among its participants [24, 98], which maximizes the agent's ability to effectively influence others. Alternatively, agents can be presented on less immersive platforms like Zoom or Teams as 2D avatars within their standard interfaces [13]. However, we hypothesize that the diminished social presence would limit agents' influence on other students. For example, when students observe agents taking notes from a small window on Zoom, it doesn't have the same impact as witnessing these agents taking notes alongside them in VR. This hypothesis is also drawn from earlier research underscoring the crucial role of social presence in developing peer relationships. [53, 105]. To test this hypothesis, we plan to conduct a comparative study between agents hosted on immersive and non-immersive platforms respectively, while examining the agent's influence.

Currently, our quantitative and qualitative assessments are conducted within a simulated environment. While this approach has been beneficial in capturing a range of responses from the agent, it may not be adequate for thoroughly identifying instances of poor or unexpected performance that may occur in real-world scenarios, such as the known issue of hallucinations [59, 67] in large language models (LLMs). Additionally, the modest sample size restricts our ability to fully eliminate the impact of variations in student activeness across the groups. Moving forward, we aim to implement and execute more extensive user studies, involving a larger number of participants in each group and increasing the number of groups, to address these limitations. Moving forward, we aim to implement and execute more extensive user studies, involving a larger number of participants in each group and increasing the number of groups, to address these limitations.

Students with different backgrounds may particularly benefit from different aspects of the agent's behaviors. For example, students with less background may benefit more from agents reminding them of missing key points, as they are more susceptible to it. Conversely, students with more background knowledge may benefit more from the agent's thought-provoking questions or discussion promotion as they further their current understanding of the lecture content. In our current study, most students have the same knowledge background – being unfamiliar with the engineering design process taught in the lecture. The limited number of samples used for comparison limits our capacity to conclusively determine if a student's knowledge background impacts the extent of an agent's influence on them. In the future, we plan to conduct a larger-scale comparative study to explore how the agent's influence varies among students with varying backgrounds.

## 9 APPLICATION SCENARIO

By integrating ClassMeta into a VR classroom, we open up a world of opportunities that can revolutionize the educational experience in multiple application scenarios.

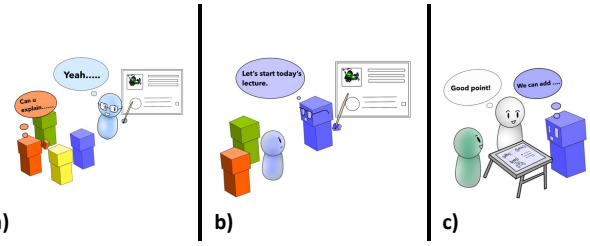


**Figure 15: Possible application scenarios: a) Instructor training. b) Student self-learning. c) Design Assistant.**

### 9.1 Instructor Training

Instructors can utilize the VR classroom with simulated student interactions as a playground to improve their teaching techniques. The ClassMeta agents can generate various student personas, offering a dynamic teaching environment for instructors to practice and improve their teaching skills (Figure 15 a).

### 9.2 Student Self-Learning

The introduction of ClassMeta can be utilized for student self-learning, encouraging students to explore and learn autonomously. Students can engage in immersive and interactive lectures where the different ClassMeta agents play the roles of both the instructor and the classmates. (Figure 15 b)

### 9.3 Design Assistant

Leveraging the power of the ClassMeta agent as a design assistant in a VR classroom offers a multitude of possibilities. The agent can facilitate collaborative projects by generating ideas and solutions in real-time, fostering innovation and creativity among students. In virtual workshops and labs, the agent can also assist students in conceptualizing and designing complex projects, providing guidance and expertise to help students develop their skills. (Figure 15 c)

## 10 CONCLUSION

In this work, we present ClassMeta, a GPT-4 powered virtual agent for promoting classroom participation in virtual reality classrooms. ClassMeta is designed to exert conducive peer influence by displaying a variety of behaviors that are commonly observed among active students. Based on existing literature, we summarize a design space for ClassMeta agents' interactions with instructors, their peer students, and themselves. ClassMeta agents can digest lesson materials as background context while capturing classroom conversations as real-time context, which allows them to generate

on-topic and contextually coherent responses. To help future educators implement their customized agent, we compile a template for tuning GPT. Through a comparative user study, we validate the effectiveness of the ClassMeta agent in terms of engagement level and learning gain. We hope this work will help to advance the rethinking of the pedagogical agent's role in the era of large language models (LLMs).

## ACKNOWLEDGMENTS

We wish to give special thanks to the reviewers for their invaluable feedback. This work is partially supported by the NSF under the Future of Work at the Human Technology Frontier (FW-HTF) 1839971. We also acknowledge the Feddersen Distinguished Professorship Funds. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the funding agency.

## REFERENCES

[1] [n. d.]. Unity Engine. https://unity.com/, year = 2023,.
[2] 2016. Questioning strategies. https://citl.illinois.edu/citl-101/teaching-learning/resources/teaching-strategies/questioning-strategies
[3] 2020. Engineering Design Process. https://www.sciencebuddies.org/science-fair-projects/engineering-design-process/engineering-design-process-steps.
[4] 2021. 7 Tips for Teaching Students How to Ask Questions in Class. https://www.waterford.org/education/how-to-ask-questions/
[5] 2021. Why Discussion? https://teaching.uchicago.edu/resources/teaching-strategies/discussion/
[6] 2023. character.ai. https://beta.character.ai/.
[7] 2023. cognitive-services-speech-sdk. cognitive-services-speech-sdk.
[8] 2023. Generative Voice AI. https://elevenlabs.io/.
[9] 2023. Google Cloud Firestore. https://firebase.google.com/docs/firestore.
[10] 2023. Introducing ChatGPT. https://openai.com/blog/chatgpt.
[11] 2023. Introducing Duolingo Max, a learning experience powered by GPT-4. https://blog.duolingo.com/duolingo-max/.
[12] 2023. microsoft-introduces-guidelines-for-developing-responsible-conversational-ai. https://blogs.microsoft.com/blog/2018/11/14/microsoft-introduces-guidelines-for-developing-responsible-conversational-ai/.
[13] 2023. Microsoft Teams Users Can Attend Meetings as a 3D Avatar in May. https://www.pcmag.com/news/microsoft-teams-users-can-attend-meetings-as-a-3d-avatar-in-may.
[14] 2023. Oculus Quest Pro. https://www.meta.com/quest/quest-pro/.
[15] 2023. Photon Unity Networking. https://doc-api.photonengine.com/en/pun/current/index.html.
[16] Martin Agran, Michael L Wehmeyer, Michael Cavin, and Susan Palmer. 2008. Promoting student active classroom participation skills through instruction to promote self-regulated learning and self-determination. *Career Development for Exceptional Individuals* 31, 2 (2008), 106–114.
[17] Duygu İŞPINAR AKÇAYOĞLU. 2022. AN INVESTIGATION OF PEER INFLUENCES ON FACTORS AFFECTING SUCCESS IN A PREP-YEAR PROGRAM AT THE UNIVERSITY. *EDUCATION & SCIENCE 2022-II* (2022), 93.
[18] Mehdi Alaimi, Edith Law, Kevin Daniel Pantasdo, Pierre-Yves Oudeyer, and Hélène Sauzeon. 2020. Pedagogical agents for fostering question-asking skills in children. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.
[19] Azad Ali and Dorothy Gracey. 2013. Dealing with student disruptive behavior in the classroom–A case example of the coordination between faculty and assistant dean for academics. In *Proceedings of the Informing Science and Information Technology Education Conference*. Informing Science Institute, 1–15.
[20] Jeromy Anglim, Patrick D Dunlop, Serena Wee, Sharon Horwood, Joshua K Wood, and Andrew Marty. 2022. Personality and intelligence: A meta-analysis. *Psychological Bulletin* 148, 5-6 (2022), 301.
[21] Sarune Baceviciute, Aske Mottelson, Thomas Terkildsen, and Guido Makransky. 2020. Investigating representation of text and audio in educational VR using learning outcomes and EEG. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–13.
[22] Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, et al. 2023. Benchmarking Foundation Models with Language-Model-as-an-Examiner. *arXiv preprint arXiv:2306.04181* (2023).
[23] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023* (2023).
[24] Miguel Barreda-Ángeles and Tilo Hartmann. 2022. Psychological benefits of using social virtual reality platforms during the covid-19 pandemic: The role of social and spatial presence. *Computers in Human Behavior* 127 (2022), 107047.
[25] Gautam Biswas, Hogyeong Jeong, John S Kinnebrew, Brian Sulcer, and ROD Roscoe. 2010. Measuring self-regulated learning skills through social interactions in a teachable agent environment. *Research and Practice in Technology Enhanced Learning* 5, 02 (2010), 123–152.
[26] Klaus Boehnke. 2008. Peer pressure: A cause of scholastic underachievement? A cross-cultural study of mathematical achievement among German, Canadian, and Israeli middle school students. *Social Psychology of Education* 11 (2008), 149–160.
[27] Susan E Brennan and Maurice Williams. 1995. The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of memory and language* 34, 3 (1995), 383–398.
[28] Amy C Brualdi Timmins. 1998. Classroom questions. *Practical Assessment, Research, and Evaluation* 6, 1 (1998), 6.
[29] Leonardo Bursztyn, Georgy Egorov, and Robert Jensen. 2019. Cool to be smart or smart to be cool? Understanding peer pressure in education. *The Review of Economic Studies* 86, 4 (2019), 1487–1526.
[30] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109* (2023).
[31] Yu Chen, Scott Jensen, Leslie J Albert, Sambhav Gupta, and Terri Lee. 2023. Artificial intelligence (AI) student assistants in the classroom: Designing chatbots to support student success. *Information Systems Frontiers* 25, 1 (2023), 161–182.
[32] Christine Chin and Jonathan Osborne. 2008. Students' questions: a potential resource for teaching and learning science. *Studies in science education* 44, 1 (2008), 1–39.
[33] Geraldine Clarebout and Steffi Heidig (née Domagk). 2012. *Pedagogical Agents*. Springer US, Boston, MA, 2567–2571. https://doi.org/10.1007/978-1-4419-1428-6_942
[34] Herbert H Clark and Jean E Fox Tree. 2002. Using uh and um in spontaneous speaking. *Cognition* 84, 1 (2002), 73–111.
[35] Philippe Dessus, Olivier Cosnefroy, and Vanda Luengo. 2016. "Keep your eyes on'em all!": A Mobile eye-tracking analysis of teachers' sensitivity to students. In *Adaptive and Adaptable Learning: 11th European Conference on Technology Enhanced Learning, EC-TEL 2016, Lyon, France, September 13-16, 2016, Proceedings 11*. Springer, 72–84.
[36] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
[37] Pieternel Dijkstra, Hans Kuyper, Greetje Van der Werf, Abraham P Buunk, and Yvonne G van der Zee. 2008. Social comparison in the classroom: A review. *Review of educational research* 78, 4 (2008), 828–879.
[38] Robert J Dufresne, William J Gerace, William J Leonard, Jose P Mestre, and Laura Wenk. 1996. Classtalk: A classroom communication system for active learning. *Journal of computing in higher education* 7 (1996), 3–47.
[39] Gregory Dyke, David Adamson, Iris Howley, and Carolyn Penstein Rosé. 2013. Enhancing scientific reasoning and discussion with conversational agents. *IEEE Transactions on Learning Technologies* 6, 3 (2013), 240–247.
[40] Ronald G Ehrenberg, Dominic J Brewer, Adam Gamoran, and J Douglas Willms. 2001. Class size and student achievement. *Psychological science in the public interest* 2, 1 (2001), 1–30.
[41] Scott Freeman, Sarah L Eddy, Miles McDonough, Michelle K Smith, Nnadozie Okoroafor, Hannah Jordt, and Mary Pat Wenderoth. 2014. Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the national academy of sciences* 111, 23 (2014), 8410–8415.
[42] Hong Gao, Efe Bozkir, Lisa Hasenbein, Jens-Uwe Hahn, Richard Göllner, and Enkelejda Kasneci. 2021. Digital transformations of classrooms in virtual reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–10.
[43] Hong Gao, Lisa Hasenbein, Efe Bozkir, Richard Göllner, and Enkelejda Kasneci. 2022. Evaluating the Effects of Virtual Human Animation on Students in an Immersive VR Classroom Using Eye Movements. In *Proceedings of the 28th ACM Symposium on Virtual Reality Software and Technology*. 1–11.
[44] Henner Gimpel, Kristina Hall, Stefan Decker, Torsten Eymann, Luis Lämmermann, Alexander Mädche, Maximilian Röglinger, Caroline Ruiner, Manfred Schoch, Mareike Schoop, et al. 2023. *Unlocking the power of generative AI models and systems such as GPT-4 and ChatGPT for higher education: A guide for students and lecturers*. Technical Report. Hohenheim Discussion Papers in Business, Economics and Social Sciences.
[45] Ashok K Goel and Lalith Polepeddi. 2018. Jill Watson. *Learning engineering for online education: Theoretical contexts and design-based examples. Routledge* (2018).

[46] Roberto Gozalo-Brizuela and Eduardo C Garrido-Merchan. 2023. ChatGPT is not all you need. A State of the Art Review of large Generative AI models. *arXiv preprint arXiv:2301.04655* (2023).

[47] Arthur C Graesser, Shulan Lu, George Tanner Jackson, Heather Hite Mitchell, Mathew Ventura, Andrew Olney, and Max M Louwerse. 2004. AutoTutor: A tutor with dialogue in natural language. *Behavior Research Methods, Instruments, & Computers* 36 (2004), 180–192.

[48] Weijiao Huang, Khe Foon Hew, and Luke K Fryer. 2022. Chatbots for language learning—Are they really useful? A systematic review of chatbot-supported language learning. *Journal of Computer Assisted Learning* 38, 1 (2022), 237–257.

[49] Giuseppe Iaria and Harry Hubball. 2008. Assessing student engagement in small and large classes. *Transformative Dialogues: Teaching and Learning Journal* 2, 1 (2008).

[50] Lasse Jensen and Flemming Konradsen. 2018. A review of the use of virtual reality head-mounted displays in education and training. *Education and Information Technologies* 23 (2018), 1515–1529.

[51] Qiao Jin, Yu Liu, Svetlana Yarosh, Bo Han, and Feng Qian. 2022. How will vr enter university classrooms? multi-stakeholders investigation of vr in higher education. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–17.

[52] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature machine intelligence* 1, 9 (2019), 389–399.

[53] Grace R Kalfus. 1984. Peer mediated intervention: A critical review. *Child & family behavior therapy* 6, 1 (1984), 17–43.

[54] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences* 103 (2023), 102274.

[55] Emre Kazim and Adriano Soares Koshiyama. 2021. A high-level overview of AI ethics. *Patterns* 2, 9 (2021).

[56] Alan C Kerckhoff and Elizabeth Glennie. 1999. The Matthew effect in American education. *Research in sociology of education and socialization* 12, 1 (1999), 35–66.

[57] Yanghee Kim and Amy L Baylor. 2016. based design of pedagogical agent roles: A review, progress, and recommendations. *International Journal of Artificial Intelligence in Education* 26 (2016), 160–169.

[58] Ju Yoen Lee. 2023. Can an artificial intelligence chatbot be the author of a scholarly article? *Journal of Educational Evaluation for Health Professions* 20 (2023).

[59] Minhyeok Lee. 2023. A mathematical investigation of hallucination and creativity in gpt models. *Mathematics* 11, 10 (2023), 2320.

[60] Willy Lens and Marleen Decruyenaere. 1991. Motivation and de-motivation in secondary education: Student characteristics. *Learning and instruction* 1, 2 (1991), 145–159.

[61] Meng-Yun Liao, Ching-Ying Sung, Hao-Chuan Wang, and Wen-Chieh Lin. 2019. Virtual classmates: Embodying historical learners' messages as learning companions in a VR classroom through comment mapping. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 163–171.

[62] Aislyn PC Lin, Charles V Trappey, Chi-Cheng Luan, Amy JC Trappey, and Kevin LK Tu. 2021. A Test Platform for Managing School Stress Using a Virtual Reality Group Chatbot Counseling System. *Applied Sciences* 11, 19 (2021), 9071.

[63] Wei-Kai Liou and Chun-Yen Chang. 2018. Virtual reality classroom applied to science education. In *2018 23rd International Scientific-Professional Conference on Information Technology (IT)*. IEEE, 1–4.

[64] Chung Kwan Lo. 2023. What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences* 13, 4 (2023), 410.

[65] Claudio Longobardi, Michele Settanni, Shanyan Lin, and Matteo Angelo Fabris. 2021. Student–teacher relationship quality and prosocial behaviour: The mediating role of academic achievement and a positive attitude towards school. *British Journal of Educational Psychology* 91, 2 (2021), 547–562.

[66] Divine Maloney, Guo Freeman, and Donghee Yvette Wohn. 2020. " Talking without a Voice" Understanding Non-verbal Communication in Social Virtual Reality. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–25.

[67] Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896* (2023).

[68] Ati Suci Dian Martha and Harry B Santoso. 2019. The design and impact of the pedagogical agent: A systematic literature review. *Journal of educators Online* 16, 1 (2019), n1.

[69] Victoria Budzinski McMullen. 2014. Using student-led seminars and conceptual workshops to increase student participation. *College Teaching* 62, 2 (2014), 62–67.

[70] Ann Medaille. 2018. *Quiet Students' Experiences with Collaborative Learning at the Postsecondary Level*. Ph. D. Dissertation. University of Nevada, Reno.

[71] Ann Medaille and Janet Usinger. 2019. Engaging quiet students in the college classroom. *College Teaching* 67, 2 (2019), 130–137.

[72] Carolina Mega, Lucia Ronconi, and Rossana De Beni. 2014. What makes a good student? How emotions, self-regulated learning, and motivation contribute to academic achievement. *Journal of educational psychology* 106, 1 (2014), 121.

[73] Chet Meyers and Thomas B Jones. 1993. *Promoting Active Learning. Strategies for the College Classroom*. ERIC.

[74] Marina Micari and Denise Drane. 2011. Intimidation in small learning groups: The roles of social-comparison concern, comfort, and individual characteristics in student academic outcomes. *Active Learning in Higher Education* 12, 3 (2011), 175–187.

[75] Fares Moustafa and Anthony Steed. 2018. A longitudinal study of small group interaction in social virtual reality. In *Proceedings of the 24th ACM Symposium on Virtual Reality Software and Technology*. 1–10.

[76] Ha Nguyen. 2022. Examining Teenagers' Perceptions of Conversational Agents in Learning Settings. In *Interaction Design and Children*. 374–381.

[77] Saima Nisar and Muhammad Shahzad Aslam. 2023. Is ChatGPT a Good Tool for T&CM Students in Studying Pharmacology? *Available at SSRN 4324310* (2023).

[78] Chinedu Wilfred Okonkwo and Abejide Ade-Ibijola. 2021. Chatbots applications in education: A systematic review. *Computers and Education: Artificial Intelligence* 2 (2021), 100033.

[79] Zachary A Pardos and Shreya Bhandari. 2023. Learning gain differences between ChatGPT and human tutor generated algebra hints. *arXiv preprint arXiv:2302.06871* (2023).

[80] Joon Sung Park, Joseph C O'Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442* (2023).

[81] Falguni Patel, Riya Thakore, Ishita Nandwani, and Santosh Kumar Bharti. 2019. Combating depression in students using an intelligent chatBot: a cognitive behavioral therapy. In *2019 IEEE 16th India Council International Conference (INDICON)*. IEEE, 1–4.

[82] José Quiroga Pérez, Thanasis Daradoumis, and Joan Manuel Marquès Puig. 2020. Rediscovering the use of chatbots in education: A systematic literature review. *Computer Applications in Engineering Education* 28, 6 (2020), 1549–1565.

[83] Gustav Bøg Petersen, Aske Mottelson, and Guido Makransky. 2021. Pedagogical agents in educational vr: An in the wild study. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–12.

[84] Paul R Pintrich. 2003. A motivational science perspective on the role of student motivation in learning and teaching contexts. *Journal of educational Psychology* 95, 4 (2003), 667.

[85] Johanna Pirker, Johannes Kopf, Alexander Kainz, Andreas Dengel, and Benjamin Buchbauer. 2021. The Potential of Virtual Reality for Computer Science Education-Engaging Students through Immersive Visualizations. In *2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, 297–302.

[86] Michael Prince. 2004. Does active learning work? A review of the research. *Journal of engineering education* 93, 3 (2004), 223–231.

[87] Emilee Rader, Margaret Echelbarger, and Justine Cassell. 2011. Brick by brick: iterating interventions to bridge the achievement gap with virtual peers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2971–2974.

[88] Alec Radford, Jeffrey Wu, Dario Amodei, Daniela Amodei, Jack Clark, Miles Brundage, and Ilya Sutskever. 2019. Better language models and their implications. *OpenAI blog* 1, 2 (2019).

[89] Brent W Roberts, Nathan R Kuncel, Rebecca Shiner, Avshalom Caspi, and Lewis R Goldberg. 2007. The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspectives on Psychological science* 2, 4 (2007), 313–345.

[90] Joseph Seering, Michal Luria, Geoff Kaufman, and Jessica Hammer. 2019. Beyond dyadic interactions: Considering chatbots as community members. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[91] Alan Seidman. 2005. The learning killer: Disruptive student behavior in the classroom. *Reading improvement* 42, 1 (2005), 40–47.

[92] Burr Settles. 2009. Active learning literature survey. (2009).

[93] Ronald W Shephard and Rolf Färe. 1974. The law of diminishing returns. In *Production Theory: Proceedings of an International Seminar Held at the University at Karlsruhe May–July 1973*. Springer, 287–318.

[94] Jingyu Shi, Rahul Jain, Hyungjun Doh, Ryo Suzuki, and Karthik Ramani. 2023. An HCI-Centric Survey and Taxonomy of Human-Generative-AI Interactions. *arXiv preprint arXiv:2310.07127* (2023).

[95] Alejo Jose G Sison, Marco Tulio Daza, Roberto Gozalo-Brizuela, and Eduardo C Garrido-Merchán. 2023. ChatGPT: More than a weapon of mass deception, ethical challenges and responses from the human-Centered artificial intelligence (HCAI) perspective. *arXiv preprint arXiv:2304.11215* (2023).

[96] Helen Slater, Neil M Davies, and Simon Burgess. 2012. Do teachers matter? Measuring the variation in teacher effectiveness in England. *Oxford Bulletin of Economics and Statistics* 74, 5 (2012), 629–645.

[97] Lisa K Carden Smith and Susan A Fowler. 1984. Positive peer pressure: The effects of peer monitoring on children's disruptive behavior. *Journal of applied*

*behavior analysis* 17, 2 (1984), 213–227.

[98] Philipp Sykownik, Linda Graf, Christoph Zils, and Maic Masuch. 2021. The most social platform ever? A survey about activities & motives of social VR users. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*. IEEE, 546–554.

[99] Viriya Taecharungroj. 2023. "What Can ChatGPT Do?" Analyzing Early Reactions to the Innovative AI Chatbot on Twitter. *Big Data and Cognitive Computing* 7, 1 (2023), 35.

[100] Stergios Tegos, Stavros Demetriadis, and Anastasios Karakostas. 2011. MentorChat: Introducing a configurable conversational agent as a tool for adaptive online collaboration support. In *2011 15th panhellenic conference on informatics*. IEEE, 13–17.

[101] Deborah J Terry, Michael A Hogg, and Katherine M White. 1999. Attitude-behavior relations: Social identity and group membership. In *Attitudes, behavior, and social context*. Psychology Press, 67–93.

[102] Ulrich Trautwein, Hanna Dumont, and Anna-Lena Dicke. 2015. Schooling: Impact on cognitive and motivational development. (2015).

[103] Quan Tu, Chuanqi Chen, Jinpeng Li, Yanran Li, Shuo Shang, Dongyan Zhao, Ran Wang, and Rui Yan. 2023. CharacterChat: Learning towards Conversational AI with Personalized Social Support. *arXiv preprint arXiv:2308.10278* (2023).

[104] Ronald D Vale. 2013. The value of asking questions. *Molecular biology of the cell* 24, 6 (2013), 680–682.

[105] Carlos Valiente, Jodi Swanson, Dawn DeLay, Ashley M Fraser, and Julia H Parker. 2020. Emotion-related socialization in the classroom: Considering the roles of teachers, peers, and the classroom context. *Developmental psychology* 56, 3 (2020), 578.

[106] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[107] Ana M Villanueva, Ziyi Liu, Zhengzhe Zhu, Xin Du, Joey Huang, Kylie A Peppler, and Karthik Ramani. 2021. Robotar: An augmented reality compatible teleconsulting robotics toolkit for augmented makerspace experiences. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[108] Kenneth E Vogler. 2008. Asking good questions. *Educational Leadership* 65, 9 (2008).

[109] Xingchen Wan, Ruoxi Sun, Hanjun Dai, Sercan O Arik, and Tomas Pfister. 2023. Better zero-shot reasoning with self-adaptive prompting. *arXiv preprint arXiv:2305.14106* (2023).

[110] Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. 2023. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111* (2023).

[111] Qiaosi Wang, Shan Jing, and Ashok K Goel. 2022. Co-Designing AI Agents to Support Social Connectedness Among Online Learners: Functionalities, Social Characteristics, and Ethical Challenges. In *Designing Interactive Systems Conference*. 541–556.

[112] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems* 35 (2022), 24824–24837.

[113] Kathryn R Wentzel. 1993. Does being good make the grade? Social behavior and academic competence in middle school. *Journal of Educational Psychology* 85, 2 (1993), 357.

[114] Ziang Xiao, Michelle X Zhou, and Wat-Tat Fu. 2019. Who should be my teammates: Using a conversational agent to understand individuals and help teaming. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 437–447.

[115] Ying Xu and Mark Warschauer. 2020. Exploring young children's engagement in joint reading with a conversational agent. In *Proceedings of the interaction design and children conference*. 216–228.

[116] Kexin Yang, Xiaofei Zhou, and Iulian Radu. 2020. XR-Ed Framework: Designing Instruction-driven andLearner-centered Extended Reality Systems for Education. *arXiv preprint arXiv:2010.13779* (2020).

[117] Todd Zakrajsek. 2017. Students who don't participate in class discussions: They are not all introverts. https://www.scholarlyteacher.com/post/students-who-dont-participate-in-class-discussions

[118] Hui Zhang, Chen Yu, and Linda B Smith. 2006. An interactive virtual reality platform for studying embodied social interaction. In *Proceedings of the CogSci06 Symposium Toward Social Mechanisms of Android Science*. Citeseer.

[119] Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2022. Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910* (2022).

[120] Zhengzhe Zhu, Ziyi Liu, Youyou Zhang, Lijun Zhu, Joey Huang, Ana M Villanueva, Xun Qian, Kylie Peppler, and Karthik Ramani. 2023. LearnIoTVR: An End-to-End Virtual Reality Environment Providing Authentic Learning Experiences for Internet of Things. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.

[121] Matthias Ziegler, Carolyn Maccann, and Richard Roberts. 2011. *New Perspectives on Faking in Personality Assessment*. https://doi.org/10.1093/acprof:oso/9780195387476.003.0011