# Foundations of PEERS: Assessing LLM Role Performance in Educational Simulations

**Anonymous ACL submission**

## Abstract

In education, peer instruction (PI) is widely recognized as an effective active learning strategy. However, real-world evaluations of PI are often limited by logistical constraints and variability in classroom settings. This paper introduces PEERS (Peer Enhanced Educational Realistic Simulation), a simulation framework that integrates Agent-Based Modeling (ABM), Large Language Models (LLMs), and Bayesian Knowledge Tracing (BKT) to emulate student learning dynamics. As an initial step, this study focuses on evaluating whether LLM-powered agents can effectively assume the roles of teachers and students within the simulation. Human evaluations and topic-based metrics show that LLMs can generate role-consistent and contextually appropriate classroom dialogues. These results serve as a foundational milestone toward building realistic, AI-driven educational simulations. Future work will include simulating the complete PEERS framework and validating its accuracy through actual classroom-based PI sessions. This research aims to contribute a scalable, cost-effective methodology for studying instructional strategies in controlled yet realistic environments.

## 1 Introduction

Classroom learning is an intricate process influenced by various variables such as student participation, peer interactions, and instructional strategies. Active learning, where students actively participate in the learning process, has gained popularity due to its effectiveness inside the classroom (Martella and Schneider, 2024). One notable strategy in active learning is Peer Instruction (PI), a pedagogical approach that promotes student interaction.

PI facilitates critical thinking, improves retention, and improves problem solving skills by encouraging collaborative dialogue and shared understanding (Garrison and Vaughan, 2008). For example, a decade-long study at Harvard demonstrated

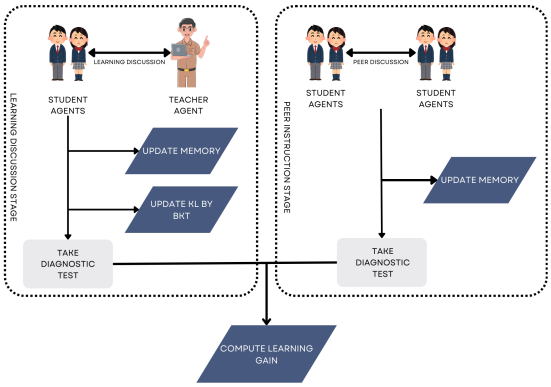


Figure 1: **PEERS Flowchart.** PEERS has 2 parts in order to deliver Peer Instruction. The Learning Discussion Stage shown is where the Student Agent gains a base knowledge regarding the topic by updating its memory and knowledge by BKT. The Peer Discussion stage reflects the knowledge from the previous stage, and then student agents discuss and give feedback on it. Learning gains are computed from pre-and post-test.

the efficacy of PI over traditional lectures, showing significant improvements in both conceptual reasoning and quantitative problem solving performance (Crouch and Mazur, 2001). This method has become a vital component of modern educational practices in disciplines such as physics, biology, and chemistry (Vickrey et al., 2015).

Although PI has been shown to provide substantial benefits, evaluating its effectiveness in authentic classroom environments presents significant challenges. Factors such as variability in student participation, personality types, dynamics of peer relationships, and external pressures frequently obscure the impact of instructional strategies (Black and Wiliam, 1998). Furthermore, logistical constraints and resource-intensive requirements limit the feasibility of conducting large-scale classroom experiments to fully investigate broader learning dynamics (Bieda et al., 2020). Although a previous

work (Elendu et al., 2024) shows that simulation-based studies provide an alternative by allowing precise control over variables and exploration of emerging learning behaviors, these models often rely on assumptions that may not fully capture the complexities of real-world interactions. This limitation underscores the need for methodologies that combine realism, scalability, and cost-effectiveness to thoroughly investigate the dynamics of PI.

To address these challenges, this thesis proposal introduces PEERS (Peer Enhanced Educational Realistic Simulation), a novel Agent-Based Modeling (ABM) framework augmented by Large Language Models (LLMs) and Bayesian Knowledge Tracing (BKT). Adopting ABM allows for the modeling of individual students as agents with distinct and evolving traits, such as knowledge level, engagement, and interaction frequency, allowing for the capture of emergent behaviors that reveal how individual and group dynamics contribute to learning outcomes. These behaviors, which are difficult to observe in real-life scenarios, provide valuable insights into the mechanisms underlying collaborative learning. To enhance the realism of these simulations, we used LLMs to generate nuanced, contextually relevant dialogues that emulate human-like classroom discussions, making the simulation results more applicable to real-world settings. Furthermore, we dynamically track the knowledge progression of each agent based on participation and quiz performance by BKT, offering a probabilistic mechanism to quantify learning outcomes during instructional activities. Unlike conventional pre- and post-test evaluations, this integrated approach provides granular insights, such as access to the peer conversations themselves, as well as a more direct observation of the impact of PI, enabling a more comprehensive understanding of its effectiveness.

The present work focuses on the first phase of this broader research agenda: Validating the ability of LLMs to assume distinct classroom roles (e.g., teacher, average student, below-average student) and engage in realistic, role-appropriate dialogues. Initial experiments evaluate LLM consistency and believability through human- and topic-based assessments.

The following objectives structure the overall direction of this research:

- Validate the ability of LLMs to assume classroom roles through human- and metric-based evaluation (current work).
- Simulate the full PEERS framework, integrating BKT and memory modeling to analyze learning dynamics (future work).
- Conduct actual classroom-based PI sessions to validate and calibrate the simulation framework (future work).

## 2 Related Work

PI fosters active learning by encouraging structured peer discussions, improving conceptual understanding, and problem-solving skills across disciplines (Mazur, 1997). Theoretical foundations include cultural evolutionary theory (Lew-Levy et al., 2023), collaborative learning (Yang, 2023), and cognitive constructivism (Keerthirathne and Keerthirathne, 2020). PI is widely implemented at all levels of education (Wang and Gao, 2021), (Arthur et al., 2022), with research showing that peer discussions and instructor explanations improve learning gains (Smith et al., 2011). However, social dynamics, time constraints, and logistical issues hinder its large-scale evaluation (Themeli, 2023), (Knight et al., 2013). To address these challenges, PEERS provides a scalable and controlled simulation framework that enables the systematic analysis of PI interactions without the constraints of traditional classroom settings. ABM enables the simulation of complex learning environments, providing insight into the optimization of instructional strategies (Vulic et al., 2024), (Ormazábal et al., 2021). ABM models human decision-making and social interactions, making it valuable for education research An (2012). However, it struggles to replicate the dynamics of a real classroom (Chopra et al., 2024). Integrating AI can improve ABM realism, particularly by using LLMs to generate human-like discussions that capture peer interactions (Chen et al., 2024).PEERS enhances ABM-based simulations by integrating LLMs, allowing for dynamic peer discussions that better reflect real classroom interactions. Artificial intelligence (AI), particularly LLM, has been widely used in education (Wang et al., 2024). LLMs can simulate classroom discussions by generating realistic dialogues, allowing for emergent behaviors that enhance learning (Zhang et al., 2024). Tools such as CodeAid provide LLM-driven personalized guidance (Kazemitabaar et al., 2024). However, the modeling of student behavior remains challenging (Nguyen et al., 2024). With

this, PEERS leverages LLMs to simulate student-driven dialogues and peer discussions, capturing emergent learning patterns that traditional models struggle to reproduce. BKT helps track and quantify knowledge progression, refining the realism of AI-driven classroom simulations (Corbett and Anderson, 1994). Despite progress in using ABM, LLMs, and BKT separately, little research has explored their combined application in PI environments. By integrating ABM, LLMs, and BKT, PEERS creates a novel framework for evaluating peer learning, enabling the continuous tracking of student knowledge states and interactions in a scalable, data-driven manner.

## 3 Methodology

### 3.1 Simulation Framework

The simulation framework consists of two primary agent roles: Teacher and Student agent. Each agent interacts in a simulated classroom environment using a set of predefined parameters. The simulation framework, illustrated in Figure 2, comprises two primary stages: the Learning Discussion Stage and the Peer Instruction Stage.

Each agent $i$ is defined by a set of basic attributes that determine its role $R$ and behavior. These attributes are further enhanced by the output generated from LLMs. In this simulation, there are two primary roles, teacher and student roles.

*Teacher Agent.* The teacher agent is characterized by three core components: the Teacher Script ($T$), the Test Set ($Q_t$) and the LLM Prompt ($P_t$). Hence, we can define the teacher agent's roles as

$$R_T = \{T, Q_t, P_t\}, \tag{1}$$

where

- $T$ is the teacher script that serves as the basic outline of the lecture that the teacher agent follows throughout the simulation. It provides structure to the class discussion, highlights key points, and determines where the discussion ends.
- $Q_t$ is the test set that the teacher agent will administer after the discussion. It assesses the student's learning and retention, and the results are used to compute the student's learning gain.
- $P_t$ is the LLM prompt to generate the teacher agent responses in the simulation. It defines the interaction style and depth of the responses, enabling the teacher agent to respond naturally and contextually based on the discussion.

*Student Agent.* The student agent is defined by a set of personalized attributes that model individual learning behaviors, which are implemented as behavioral parameters in the agent-based simulation. These attributes are encoded directly in the simulation code to guide the student agent's actions and responses. The student role is described as

$$R_S = \{K_i(t), F_i(t), E_i(t), Q_i(t), M_i(t), P_i\}, \tag{2}$$

where

- $K_i(t)$ is the Knowledge Level (KL) parameter that represents the student's understanding of the subject at time $t$. This parameter influences the agent's uncertainty, calculated as $1 - K_i(t)$. The knowledge level also affects the student's memory capacity,

$$MC = 5 + \exp(4K_i(t)), \tag{3}$$

  following Miller's Law ((Miller, 1956)).
- $F_i(t)$ is the Interaction Frequency (IF) parameter. This parameter triggers whether the agent actively participates (e.g. asks a question) or passively listens during discussion.
- $E_i(t)$ is the Engagement Level (EL) parameter that affects the complexity of the questions posed by the agent. Higher EL results in more detailed or in-depth questions.
- $Q_i(t)$ means Question Trigger (QT) which determines the threshold for the agent to ask questions influenced by uncertainty. The student will ask a question if $Uncertainty > Q_i(t)$. It shows that the student agents with higher uncertainty are more likely to seek clarification.
- $M_i(t)$ serves as the student's memory. It is the student agent's knowledge repository, where learned information is stored and accessed for future discussions and tests. The memory capacity is determined on the basis of Miller's law.
- $P_i$ is the LLM parameter prompt that describes how the student agent responds in class, from asking questions to participating in peer discussions. It customizes the tone, detail, and style of student response in the simulation, making each student's behavior more realistic and varied.

This student agent model enables the simulation to capture both individual learning dynamics and

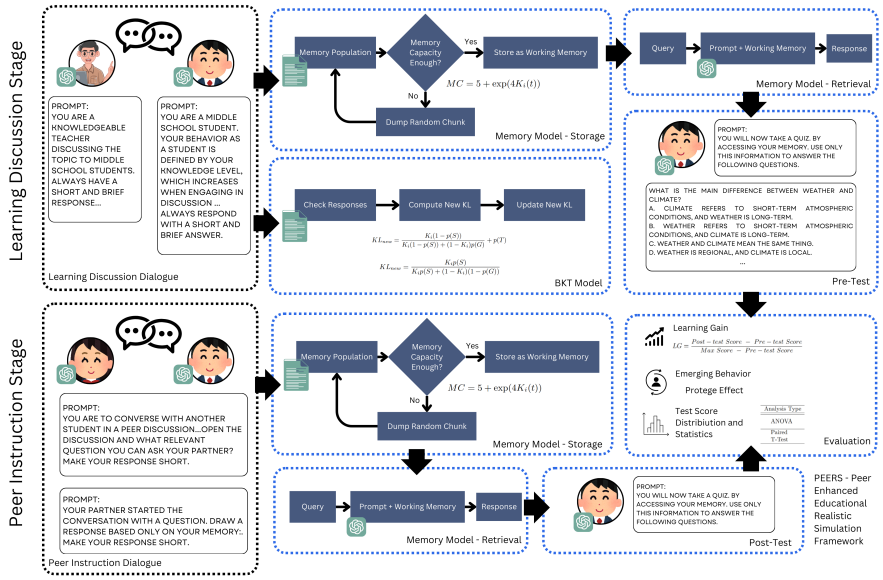


Figure 2: **PEERS Framework for Learning Discussion (upper) and Peer Instruction Stage (lower).** Every time agents engage in conversation, chunks of information are stored in their memory. The student agent's base knowledge is updated by BKT during the learning discussion stage. When the student agents take a test, they retrieve the information stored in their memory. PEERS will be able to capture the learning gain from the pre- and post-test.

group interactions, making it possible to measure the impact of peer instruction on student knowledge.

*Memory Model.* The memory model for student agents represents student learning. The model consists of two parts: storage and retrieval, as shown in Figure 2. This model adopts a straightforward approach, focusing on Miller's number to determine how many chunks of information can be stored in working memory. The information comes from the conversations during the discussions. In this case, the chunks are extracted from the conversation dialogue and stored in the form of textual information. As such, chunks are groups of keywords extracted from the discussion. This interprets the things a student agent remembers when in a discussion; they remember not all of it but key parts of the conversation (Stafford and Daly, 1984). For this method, we use NLP to extract the key words from the conversation. In the storage model, when new information arrives, the system first checks whether there is sufficient storage space. If space is available, the model stores the new information. However, if no space is available, the model randomly removes a memory chunk to accommodate the new information. This memory erasure mechanism implies that students tend to remember new information more than older information.

### 3.2 Session Structure

As shown in Figure 2, the PEERS framework consists of two stages: the Learning Discussion Stage and the Peer Instruction Stage. These stages mimic real-world classroom teaching strategies, where the teacher first discusses a topic, and peer discussions reinforce the learning from the covered material.

#### 3.2.1 Learning Discussion Stage

The Learning Discussion Stage is designed to mimic a conventional classroom environment in which the teacher agent presents a lecture and the student agents participate. In this stage, the teacher agent follows the script $T$ and discusses the material. In this paper, we demonstrate our framework using a simulation with climate change as the discussion topic. The student agents interact according to their parameters. The discussion flows naturally until all the points in the teacher script $T$ are covered. After completing the script, the teacher agent would ask each student agent questions regarding the topic. This simulates the question strategies used in classrooms to encourage critical thinking and analysis. After a student agent

answers a question, the teacher agent would provide feedback and a brief explanation of the answer. This response will serve as an input to BKT.

The BKT method updates the KL of a student dynamically based on their correct or incorrect responses to questions. For correct response, the formula to use for the KL update is

$$KL_{new} = \frac{K_i(1-p(S))}{K_i(1-p(S)) + (1-K_i)p(G)} + p(T), \quad (4)$$

and for an incorrect response, we have

$$KL_{new} = \frac{K_i p(S)}{K_i p(S) + (1-K_i)(1-p(G))}, \quad (5)$$

where $KL_{new}$ is the new KL after update, $K_i$ is the current KL of the student agent, $p(S)$ is the probability of answering incorrectly despite knowing, $p(G)$ is the probability of guessing the answer correctly, and $p(T)$ is the learning rate. Using the BKT process, the simulation offers a quantitative and dynamic method to monitor each student agent's learning progress. In addition, the student agents store information in their memory $M_i$ throughout the discussion.

#### 3.2.2 Peer Instruction Stage

In the Peer Instruction stage, student agents engage in peer instruction within a simulated row-column classroom layout. The PI occurs in two rounds: In the first round, each student pairs with the seatmate to their right. If no rightward partner exists, they pair with the student directly behind them. In the second round, students pair with their seatmates to the left. During PI, the student agents will discuss what they learned in the previous stage. The students access their memory to contribute to the discussion. Agents expand or reinforce their memory during PI based on their interaction with their peers. New knowledge and insights shared by peers are stored as memory entries, enhancing student learning.

#### 3.2.3 Simulation Parameters

The teacher and student agents are initialized to implement the simulation framework employing varied roles and behavioral parameters. The teacher agent receives a curated script on the topic of climate change, derived from widely available lectures, which serves as the basis for discussion. In addition, a set of diagnostic test questions was extracted from the script to assess the knowledge of the student agents at different stages.

The simulation features 20 student agents categorized into three distinct groups to represent a realistic middle school classroom. These groups include 10 average (Student _A), 4 above average (Student _AA), and 6 below average (Student _BA) students. The categorization was based on ranges of key behavioral parameters such as KL, EL, IF, and QT, as shown in Table 1.

The LLM used for both the student and the teacher agents, OpenAI GPT-4, was configured with a temperature setting of 0.1 to ensure relevant and deterministic responses. It was estimated that a single run uses 350k tokens at 12 USD.

| Parameter | Above Average | Average | Below Average |
|---|---|---|---|
| Knowledge Level | 0.35 - 0.5 | 0.2 - 0.35 | 0.1 - 0.2 |
| Engagement Level | 0.25 -0.4 | 0.1 - 0.25 | 0.05 - 0.1 |
| Interaction Frequency | 0.6 - 1.0 | 0.4 - 0.6 | 0.1 - 0.4 |
| Question Trigger | 0.2 - 0.3 | 0.1 - 0.2 | 0.05 - 0.1 |

Table 1: **Student Agent Parameters.** These values were randomly assigned within their respective ranges to introduce diversity in learning behaviors.

### 3.3 Actual PI Implementation

To evaluate the effectiveness of the PEERS framework, we carried out a practical implementation in a classroom setting. We observed two separate classrooms: one designated as the control group without any PI and the other implementing PI. Both classrooms were provided with identical course materials for discussion. Observers were stationed in each classroom to assess the interactions occurring there. Interaction metrics included monitoring the frequency of questions posed by both the teacher and students, analyzing the depth and frequency of student responses, and observing active listening through visual cues. The observers documented these interactions for potential replication in PEERS. Each classroom also participated in a diagnostic exam to gauge their understanding of the subject matter. Classroom 1, with no PI, was given a short test following the discussion, while Classroom 2, which utilized PI, took the test after both the discussion and the implementation of PI. Learning gains were evaluated using Hake's formula to assess student progress. The observed classroom interactions will be inputted into PEERS for comparison with the learning gain outputs. Figure represents the framework for the actual PI implementation.

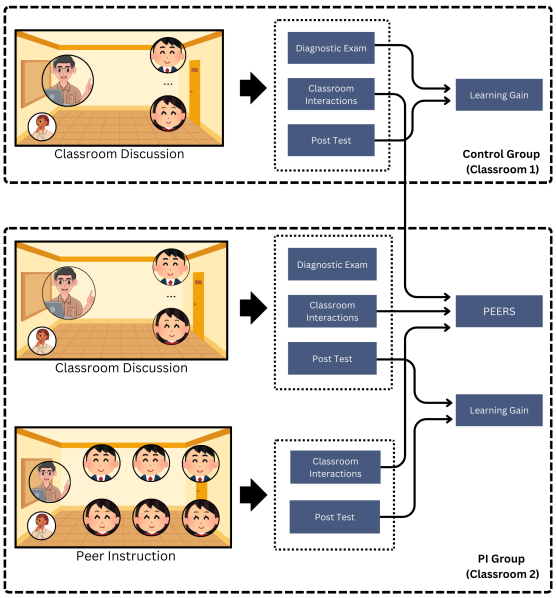


Figure 3: **Actual PI Implementation**Two classrooms were observed to obtain realistic PI results. Classroom 1, which did not implement PI, served as the control group, while Classroom 2 included PI. The resulting metric measurements were inputted into PEERs, and the learning gains were compared.

### 3.4 Evaluation Metrics

We evaluated how closely our simulation matches the classroom experience in the real world by assessing (1) how well the agents mimicked their assigned roles and (2) whether student agents actually learned, as measured by the learning gains and phenomena observed in a real classroom.

#### 3.4.1 LLM Role Evaluation

To ensure that the LLM agents effectively assumed their roles in the simulation, we evaluated them using both human evaluation and metric-based evaluation.

For the human evaluation, we took the transcript of the dialogues produced by the simulation and had them assessed by four human evaluators. The evaluators were randomly selected, and before participation, the details of the study were thoroughly explained to them. They were informed that their task was to identify roles in a dialogue within a given context. Additionally, they provided explicit consent, acknowledging that no compensation would be given and that their evaluations would be used solely for research purposes. Their responses were anonymized to ensure compliance with ethical guidelines on data privacy and confidentiality, as outlined in Annex A.

For the metric-based evaluation, we conducted a topic-based analysis to assess the consistency of the LLM agents in maintaining their assigned roles throughout the simulation. The topic-based analysis allowed us to determine whether the agents stayed focused on their assigned discussion topics rather than deviating into unrelated areas, a common issue with LLMs. Furthermore, evaluating the behavior of the student agents based on their defined behavioral parameters ensured that they behaved in alignment with their initial settings.

#### 3.4.2 Learning Gain

The effectiveness of this simulation in fostering knowledge acquisition through PI is quantified using learning gain, a widely recognized metric for evaluating educational interventions ((Evans et al., 2018)). By comparing pre-test and post-test scores, the learning gain provides a normalized measure of the improvement in knowledge achieved by the student agents through PI. The formula for calculating Learning Gain is based on Hake's model ((Hake, 2002)):

$$LG = \frac{Post - test\ Score\ -\ Pre - test\ Score}{Max\ Score\ -\ Pre - test\ Score} \tag{6}$$

This formula normalizes the gain by accounting for the student agent's initial level of knowledge, allowing comparisons across a heterogeneous population of agents with varying prior knowledge and engagement levels.

#### 3.4.3 Statistical Analysis

*T-test and ANOVA.* We use paired t-test and ANOVA on the learning gaining values to determine whether the student agents did learn. The paired t-test is used to determine whether there is a significant difference between pre-test and post-test scores, indicating the effectiveness of peer instruction. The null hypothesis $H_O$, is that there is no significant difference between pre-test and post-test scores, implying that the peer instruction framework does not significantly impact student learning. ANOVA will be used to determine whether there is a significant difference in learning gains across multiple simulation trials. The null hypothesis $H_O$, is that there is no significant difference in learning gains among the different trials i.e., the mean learning gains across trials are equal. Rejecting $H_O$ would confirm the effectiveness of peer learning and the framework reliably produces similar

learning outcomes across different runs.

#### 3.4.4 Emergent Behavior

For this simulation, one of the key advantages of employing an ABM framework is the ability to observe emergent behaviors: complex, collective phenomena arising from the interactions of individual agents. In this study, the interplay between teacher and student agents, governed by their parameters and decision-making rules, leads to several emergent outcomes that provide valuable insight into classroom dynamics. During the PI stage, collaboration among agents fosters discussions and knowledge exchange based on their stored memory. These interactions can result in scenarios where students with higher levels of knowledge reinforce the understanding of their less knowledgeable peers by sharing accurate information during discussions.

## 4 Initial Results and Discussion

### 4.1 LLM Role Experiments

#### 4.1.1 Human Evaluation

We asked human evaluators to review the transcript of the dialogues between the teacher and student agents. These dialogues were extracted from the Learning Discussion stage, where agents interacted in the environment. We selected three unique dialogues for evaluation. Their task was to analyze the dialogue and identify the speaker's role based on their perception and understanding of the script. They classified speakers as teachers or students and further classified students as below average, average, or above average. To avoid bias, we did not inform evaluators that an LLM generated the dialogue.

We selected four respondents as evaluators: two professors, one student, and one staff member. The evaluator's answers are compared with the true values. We evaluated accuracy using f1-score and Fleiss' Kappa. The f1-score measures the balance of precision and recall, particularly since below-average students rarely participates in class. We also used Fleiss' Kappa to assess the reliability of agreement among the evaluators.

Table 2 presents the measured f1-score and Fleiss' Kappa values. The results show that human evaluators successfully identified the teacher and student roles in the dialogues, with scores close to 1.0. However, the f1-scores for student categorization were lower, indicating that evaluators struggled to distinguish between student categories based only on dialogue. This challenge is reflected in the overall Fleiss' Kappa score of 0.53, suggesting moderate agreement among the respondents in identifying roles. Despite this limitation, LLM agents successfully generated a role-distinct dialogue with minor deviations in student classification.

| Dialogue | Role | f1-score | Fleiss' Kappa |
|---|---|---|---|
| 1 | Teacher | 0.9925 | 0.52 |
| | Student (Overall) | 0.99 | |
| | Below Average | – | |
| | Average | 0.35 | |
| | Above Average | 0.09 | |
| 2 | Teacher | 0.995 | 0.52 |
| | Student (Overall) | 0.9925 | |
| | Below Average | – | |
| | Average | 0.42 | |
| | Above Average | 0.09 | |
| 3 | Teacher | 1.00 | 0.55 |
| | Student (Overall) | 1.00 | |
| | Below Average | 0.31 | |
| | Average | 0.44 | |
| | Above Average | 0.15 | |
| | | AVERAGE | 0.55 |

Table 2: **Human Evaluation Result.** Human evaluators were able to capture the teacher ans student roles in the dialogues, however had difficulty assessing the student categorization. Dialogues 1 and 2 don't have any true value for Below Average student because no one in that group participated in the discussion.

#### 4.1.2 Metric-Based Evaluation

To further assess whether the LLM agents assumed their roles correctly, we conducted a metric evaluation for the student agents.

*Topic-Based Analysis.* We evaluated whether the teacher agent effectively discussed its assigned topic using topic modeling techniques. Specifically, we applied Latent Dirichlet Allocation (LDA) to extract key discussion topics from the dialogues. These topics served as representations of the main points discussed by the teacher agent. Table 3 presents the top topics extracted by LDA.

The results indicate that the top topics across the seven dialogues align with the intended topic of climate change. Topics 1 and 2 prominently feature terms like "climate," "gases," and "heat," demonstrating the teacher agent's focus on climate change. Additionally, the LLM appears to extend

the discussion by covering biodiversity, habitats, and species, likely in response to student questions. This suggests that the teacher agent dynamically guided the discussion based on student input, making the lesson more informative and interactive. Interestingly, the final extracted topic appears more educational in nature, indicating that the teacher agent assumed a classroom-oriented role by structuring discussions and responding effectively to student inquiries.

| Topic No. | Associated Words |
|---|---|
| 1 | "greenhouse", "climate", "change", "gases", "heat" |
| 2 | changes", "biodiversity", "species", "climate", "habitat" |
| 3 | "student", "answer", "climate", "weather", "aligns" |

Table 3: **Topic extraction from LDA.** The topics adheres with the topic assigned to the teacher agent to discuss which is climate change.

*Role Consistency in Behavior.* To verify whether student agents behaved according to their assigned roles, we analyzed four key metrics. First is Student Engagement that is measured engagement by counting how often each student participated in the dialogue and dividing it by the total number of dialogues. Then, Question Trigger calculated by how frequently each student asked questions by determining their proportion of total questions in the discussion. Third, Interaction Frequency where we analyzed how often each student performed an action by counting their dialogue entries and dividing by the total number of actions. And lastly Knowledge Level it was measured in the final part of the discussion, when the teacher asked a question, we counted how many correct responses each student provided to evaluate their base knowledge level.

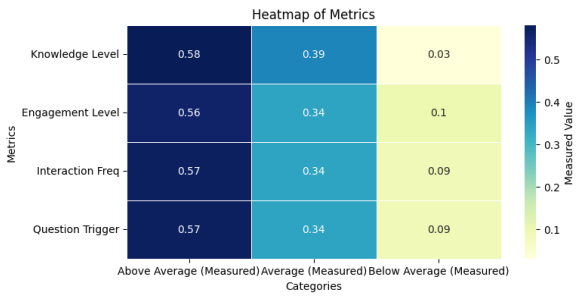


Figure 4: **Heatmap of Measured Metrics.** The figure shows a distinct differences (colors) in the student categorization within the four metrics.

Figure 4 presents a heatmap of the measured values across dialogues. The results indicate that almost no overlap exists between the student agent categories, meaning their behavior aligned with their assigned roles. Additionally, while some values deviated slightly, they remained within the predefined parameter ranges for each student category. This confirms that student agents effectively captured their assigned roles and behaved accordingly in the discussion.

## 5 Conclusion

This thesis presents initial findings from the PEERS framework, focusing on evaluating the effectiveness of LLMs in assuming teacher and student roles during simulated classroom interactions. Through human evaluation and topic modeling, the study demonstrates that LLM agents are capable of producing role-consistent, contextually appropriate dialogues. These results validate the feasibility of using LLMs as agent surrogates in educational simulations and mark an important step toward modeling more complex classroom dynamics.

While the broader PEERS framework incorporates memory modeling, Bayesian Knowledge Tracing (BKT), and agent-based learning simulations, these components remain outside the scope of the current study and are reserved for future work. The next steps include:

- Simulating the complete PEERS framework with learning discussions and peer instruction stages.
- Validating simulation accuracy through actual classroom PI implementations.

By establishing the role fidelity of LLM agents, this work lays the groundwork for future investigations into how AI-driven simulations can enhance our understanding of collaborative learning, offering a scalable alternative to traditional classroom research.

## 6 Limitations

This study has several limitations that future research can address. First, it does not explicitly categorize student behavior into predefined types; instead, it models learning dynamics through various parameters. The parameters of the student agent are assumed in this study. The literature lacks a definitive categorization of students. Additionally, the framework does not focus on modeling long-term memory retention in LLM agents,

since the memory system primarily functions as a knowledge-recall mechanism. The peer instruction dynamics in this study is structured and sequential and assesses immediate learning gains but does not track long-term retention, which could be addressed through delayed post-tests or longitudinal simulations. Addressing these limitations will enhance the realism, scalability, and cognitive modeling of AI-driven classroom simulations.

## 7 Ethical Considerations

This study involved human annotators to evaluate the dialogues produced by the LLM-powered student agents. The annotators evaluated the dialogue produced by the agents to validate that the LLM assumes their role. Since the study did not involve real human subjects providing personal data or performing experimental interventions, the institutional ethics review board deemed it exempted it from formal ethics review.

To uphold ethical research standards, all annotators were informed of their roles and responsibilities prior to participation. They gave their consent to evaluate the generated dialogues and were instructed to assess them objectively. No personally identifiable information was collected or processed during the evaluation, and all data used were generated in a controlled simulation environment.

## References


Li An. 2012. Modeling human decisions in coupled human and natural systems: Review of agent-based models. *Ecological Modelling*, 229:25–36. Modeling Human Decisions.

Yarhands Dissou Arthur, Simon Kojo Appiah, Kwadwo Amo-Asante, and Bright Asare. 2022. Modeling student's interest in mathematics: Role of history of mathematics, peer-assisted learning, and student's perception. *Eurasia J. Math. Sci. Technol. Educ.*, 18(10):em2168.

Kristen N. Bieda, Serena J. Salloum, Sihua Hu, Shannon Sweeny, John Lane, and Kaitlin Torphy. 2020. Issues with, and insights for, large-scale studies of classroom mathematical instruction. *The Journal of Classroom Interaction*, 55(1):41–63.

Paul Black and Dylan Wiliam. 1998. Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, 5(1):7–74.

John Chen, Xi Lu, Yuzhou Du, Michael Rejtig, Ruth Bagley, Mike Horn, and Uri Wilensky. 2024. Learning agent-based modeling with llm companions: Experiences of novices and experts using chatgpt & netlogo chat. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Ayush Chopra, Shashank Kumar, Nurullah Giray-Kuru, Ramesh Raskar, and Arnau Quera-Bofarull. 2024. On the limits of agency in agent-based models. *Preprint*, arXiv:2409.10568.

Albert T. Corbett and John R. Anderson. 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4(4):253–278.

Catherine H. Crouch and Eric Mazur. 2001. Peer instruction: Ten years of experience and results. *American Journal of Physics*, 69(9):970–977.

Chukwuka Elendu, Dependable C Amaechi, Alexander U Okatta, Emmanuel C Amaechi, Tochi C Elendu, Chiamaka P Ezeh, and Ijeoma D Elendu. 2024. The impact of simulation-based training in medical education: A review. *Medicine (Baltimore)*, 103(27):e38813.

C Evans, C Kandiko Howson, and A Forsythe. 2018. Making sense of learning gain in higher education. *High. Educ. Pedagog.*, 3(1):1–45.

D. Garrison and Norman Vaughan. 2008. *Blended Learning in Higher Education: Framework, Principles, and Guidelines*.

Richard R. Hake. 2002. Relationship of individual student normalized learning gains in mechanics with gender , high-school physics , and pretest scores on mathematics and spatial visualization.

Majeed Kazemitabaar, Runlong Ye, Xiaoning Wang, Austin Zachary Henley, Paul Denny, Michelle Craig, and Tovi Grossman. 2024. Codeaid: Evaluating a classroom deployment of an llm-based programming assistant that balances student and educator needs. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

W Keerthirathne and Dr Keerthirathne. 2020. Peer learning: an overview. *International Journal of Scientific Engineering and Science*, 4(11):1–6.

Jennifer K Knight, Sarah B Wise, and Katelyn M Southard. 2013. Understanding clicker discussions: student reasoning and the impact of instructional cues. *CBE Life Sci. Educ.*, 12(4):645–654.

Sheina Lew-Levy, Wouter van den Bos, Kathleen Corriveau, Natália Dutra, Emma Flynn, Eoin O'Sullivan, Sarah Pope-Caldwell, Bruce Rawlings, Marco Smolla, Jing Xu, and Lara Wood. 2023. Peer learning and cultural evolution. *Child Development Perspectives*, 17(2):97–105.

Amédee Martella and Darryl Schneider. 2024. A reflection on the current state of active learning research. *Journal of the Scholarship of Teaching and Learning*, 24:119–136.

E. Mazur. 1997. *Peer Instruction: A User's Manual*. Series in Educational Innovation. Prentice Hall.

G A Miller. 1956. The magical number seven plus or minus two: some limits on our capacity for processing information. *Psychol. Rev.*, 63(2):81–97.

Manh Hung Nguyen, Sebastian Tschiatschek, and Adish Singla. 2024. Large language models for in-context student modeling: Synthesizing student's behavior in visual programming. *Preprint*, arXiv:2310.10690.

Ignacio Ormazábal, Félix A. Borotto, and Hernán F. Astudillo. 2021. An agent-based model for teaching–learning processes. *Physica A: Statistical Mechanics and its Applications*, 565:125563.

M K Smith, W B Wood, K Krauter, and J K Knight. 2011. Combining peer discussion with instructor explanation increases student learning from in-class concept questions. *CBE Life Sci. Educ.*, 10(1):55–63.

Laura Stafford and John Daly. 1984. Conversational memory: The effects of recall mode and memory expectancies on remembrances of natural conversations. *Human Communication Research*, 10.

Chryssa Themeli. 2023. *Inclusive Peer Learning Augmented Reality in Higher Education: A Technology-Enhanced Learning (TEL) Perspective*. Power Learning Solutions.

Trisha Vickrey, Kaitlyn Rosploch, Reihaneh Rahmanian, Matthew Pilarz, and Marilyne Stains. 2015. Research-based implementation of peer instruction: A literature review. *CBE—Life Sciences Education*, 14(1):es3. PMID: 25713095.

John Vulic, Michael J. Jacobson, and James A. Levin. 2024. Exploring education as a complex system: Computational educational research with multi-level agent-based modeling. *Education Sciences*, 14(5).

Camilla Wang and Jian Gao. 2021. Peer teaching as an effective method: A case study at st university in china. *Journal of Higher Education Theory and Practice*, 21(6).

Shan Wang, Fang Wang, Zhen Zhu, Jingxuan Wang, Tam Tran, and Zhao Du. 2024. Artificial intelligence in education: A systematic literature review. *Expert Systems with Applications*, 252:124167.

Xigui Yang. 2023. A historical review of collaborative learning and cooperative learning. *TechTrends*, 67(4):718–728.

Zheyuan Zhang, Daniel Zhang-Li, Jifan Yu, Linlu Gong, Jinchang Zhou, Zhanxin Hao, Jianxiao Jiang, Jie Cao, Huiqin Liu, Zhiyuan Liu, Lei Hou, and Juanzi Li. 2024. Simulating classroom education with llm-empowered agents. *Preprint*, arXiv:2406.19226.

# A Sample Human Evaluator's Guide

This is the guide given to the annotators for the LLM role evaluation.

## EVALUATION FORM FOR CLASSROOM DIALOGUE

**Evaluator:** _____
**Date :** _____
**Session ID:** _____

**General Instructions:**

Good day! Thank you for participating in our evaluation. Please evaluate the dialogue based on the following criteria. Your responses will help measure the effectiveness of the roles in the discussion.

Your task is to carefully read the provided dialogue script and identify the speaker for each line. The possible speakers include:

- **A - Teacher**
- **B - Below Average Student (Student_BA)**
- **C - Average Student (Student_A)**
- **D - Above Average Student (Student_AA)**

For each line in the script, select the speaker that best represents who is delivering the dialogue. Please base your judgment on the **content, complexity, and clarity** of the response.

Your evaluation will help us analyze how well different participants contribute to the discussion. There are no right or wrong answers; we are interested in your perceptions.

---

Students are categorized based on their knowledge level, engagement, and questioning behavior

**Below Average Students (Student_BA)** have limited understanding, engage minimally, and rarely ask questions, often expressing confusion.

**Average Students (Student_A)** have a moderate grasp of the topic, participate in discussions without dominating, and occasionally ask clarifying questions.

**Above Average Students (Student_AA)** demonstrate strong understanding, actively engage in discussions, and frequently ask insightful questions that deepen the conversation.

---

**Consent and Data Usage Statement for Annotators**
Thank you for your participation in this study. Before we proceed, we would like to inform you about the nature of your involvement and how your data will be used.
Your evaluations will be used solely for research purposes and will be documented as part of the study's findings. We want to assure you that no personal data will be collected, stored, or analyzed at any stage of the research. All responses you provide will remain anonymous and will only be used as part of the study's dataset.
By participating, you acknowledge that you understand these terms and consent to the use of your assessments in this research while ensuring full compliance with data privacy and confidentiality guidelines. If you have any concerns or questions, please feel free to ask before proceeding.

Figure 5: **Evaluation Form for Classroom Dialogue** This is the first page where general instruction and consent were discussed with the administrators before they answered the questionnaire.