OnlineMate: An LLM-Based Multi-Agent Companion System for Cognitive Support in Online Learning

Anonymous ACL submission

Abstract

In online learning environments, students often lack personalized peer interactions, which are crucial for cognitive development and learning engagement. Although previous studies have employed large language models (LLMs) to simulate interactive learning environments, these interactions are limited to conversational exchanges, failing to adapt to learners' individualized cognitive and psychological states. As a result, students' engagement is low and they struggle to gain inspiration. To address this challenge, we propose **OnlineMate**, a multi-agent learning companion system driven by LLMs integrated with Theory of Mind (ToM). OnlineMate simulates peer-like roles, infers learners' psychological states such as misunderstandings and confusion during collaborative discussions, and dynamically adjusts interaction strategies to support higher-order thinking. Comprehensive evaluations, including simulation-based experiments, human assessments, and real classroom trials, demonstrate that OnlineMate significantly promotes deep learning and cognitive engagement by elevating students' average cognitive level while substantially improving emotional engagement scores.

1 Introduction

The use of artificial intelligence to provide immediate and personalized online instruction originated with Intelligent Tutoring Systems (Nwana, 1990). With the advancement of LLMs, LLM-driven AI teachers (Markel et al., 2023) and teaching assistants (Tu et al., 2023) have been widely adopted. However, unlike learning in traditional classrooms, students engaging in online learning through these AI technologies often face the challenge of limited interaction, which results in a weakened sense of classroom participation (Muilenburg and Berge, 2005; Ferri et al., 2020; Akpen et al., 2024).

To address these gaps, recent studies (Yu et al., 2024; Zhang et al., 2025c) have leveraged LLMs' multi-agent collaboration capabilities to simulate peer companions, aiming to stimulate discussions. However, these methods only mimic conversational styles without aligning with students' needs or cognitive levels, failing to initiate targeted, personalized discussions that would stimulate deeper reflection and enhance cognitive development.

To address these limitations, we integrate LLMs' Theory of Mind (ToM) capabilities, which is the capacity to infer others' thoughts and psychological states (Apperly, 2010; Zhu et al., 2024; Nguyen, 2025; Zhang et al., 2025b), and propose **OnlineMate**, a ToM-enhanced multi-agent learning companion system. Unlike existing multi-agent online learning systems (Yu et al., 2024; Zhang et al., 2025c) that only simulate peer-like agent roles for dialogue, OnlineMate infers and adapts to learners' cognitive and psychological states during collaborative discussions. Specifically, we decompose ToM reasoning during student classroom discussions into three stages, i.e., hypothesis generation, refinement, and response validation, in alignment with metacognitive theory (Flavell, 1979; Zhang et al., 2025b), and integrate classroom context management and behavior control. This design enables the system to not only simulate peer interactions but also dynamically adjust its interaction strategies to cater to learners' interests and needs, thereby enhancing their engagement and fostering cognitive development.

We conducted rigorous evaluations, including simulation-based experiments with an LLM-driven Evaluation Agent, human assessments using standardized rubrics, and real classroom trials in a QS top-50 university. Results confirm that OnlineMate effectively improves students' cognitive levels and learning outcomes. We also investigated key factors influencing learning effects.

Our primary contributions are summarized as follows:

- We propose OnlineMate, a ToM-enhanced multi-agent system that infers learners' cognitive and psychological states to dynamically adjust interaction strategies.
- We validate through comprehensive evaluations that OnlineMate significantly elevates students' cognitive levels and emotional engagement.
- We conduct ablation and factor analysis to uncover mechanisms underlying AI-mediated learning, providing actionable insights for future educational technology design.

2 Related Work

2.1 LLMs in Education

LLMs' pretraining on extensive corpora enables expert-level proficiency across domains. In education, LLMs have been applied to teaching assistance (JeonJaeho and LeeSeongyong, 2023), lesson planning (HuBihao et al., 2024), instructional report generation (Gao et al., 2025a), lesson delivering (Tu et al., 2023; Shi et al., 2025), and role simulation (Lee et al., 2023; Markel et al., 2023). Some studies use multi-agent systems to orchestrate classroom discussions (Yue et al., 2025; Yu et al., 2024; Zhang et al., 2025c). However, these approaches lack personalized adaptation mechanisms and fail to dynamically tailor content to individual learners' cognitive states. In this work, we integrate ToM reasoning and cognitive scaffolding, which is aligned with Bloom's Taxonomy, to achieve personalized interaction, rather than merely simulating roles or conversations.

2.2 LLMs for Human Simulation

LLM-driven agents exhibit human-like decision-making due to extensive training data on human behaviors. They have been used in social simulations (Park et al., 2023; Aher et al., 2023; Li et al., 2023; Guo et al., 2024; Lin et al., 2024), scientific inquiry (M. Bran et al., 2024; Li et al., 2025b; Gao et al., 2025b), collaborative workflows (Hong et al., 2023; Li et al., 2024; Kulkarni, 2025), and educational role-play (Yu et al., 2024; Zhang et al., 2025c). However, existing educational simulations only mimic conversational styles, lacking alignment with authentic learning processes.

2.3 Theory of Mind (ToM) in LLMs

ToM, the capacity to attribute mental states to others and anticipate behavior (Premack and Woodruff, 1978; Apperly, 2012), is a core component of social intelligence. Recent studies show that LLMs exhibit notable ToM abilities (van Duijn et al., 2023; Zhou et al., 2023; Kim et al., 2023; Jin et al., 2024), with internal representations of self and others' beliefs (Zhu et al., 2024). To enhance ToM, some studies decompose reasoning into explicit chains (Gu et al., 2024b; Lin et al., 2025) or multi-agent workflows (Zhang et al., 2025b). In this work, we similarly decompose ToM reasoning into three stages and integrate agent persona constraints and Bloom's Taxonomy, balancing role consistency and cognitive enhancement in classroom discussion scenarios.

3 OnlineMate Framework

The design of OnlineMate must ensure two key requirements: (1) each agent's behavior aligns with its persona and accesses accurate context, and (2) interactions unfold naturally. To address these, we propose the Classroom Context Manager to govern the flow of information accessible and the Classroom Behavior Controller to determine the actions undertaken by each agent. Furthermore, to facilitate a more rigorous evaluation of system functionality, we incorporate an evaluation agent tasked with simulating student roles, thereby enabling systematic evaluation of OnlineMate Agents' pedagogical effectiveness. Figure 1 shows the framework and workflow.

3.1 Classroom Context Manager

As a complex and information-rich communication environment, LLMs may experience hallucinations or role confusion during role-playing in the classroom teaching process (Park et al., 2023; Qian et al., 2024). Moreover, each OnlineMate Agent requires independent contextual information, including contextual memory, reasoning, and beliefs, in addition to the shared dialogue history. The Classroom Context Manager stores contextual elements (e.g., dialogue, memory), visible scope (e.g., specific agents, discussion groups), and agent role configurations to enhance the consistency of role-playing within LLMs. During response generation, agents query their designated role and the relevant visible context from the context database, ensuring independent and efficient context management. Furthermore, to enrich the knowledge base of the agents' responses and mitigate the occurrence of hallucinations, the Context Manager retrieves per-

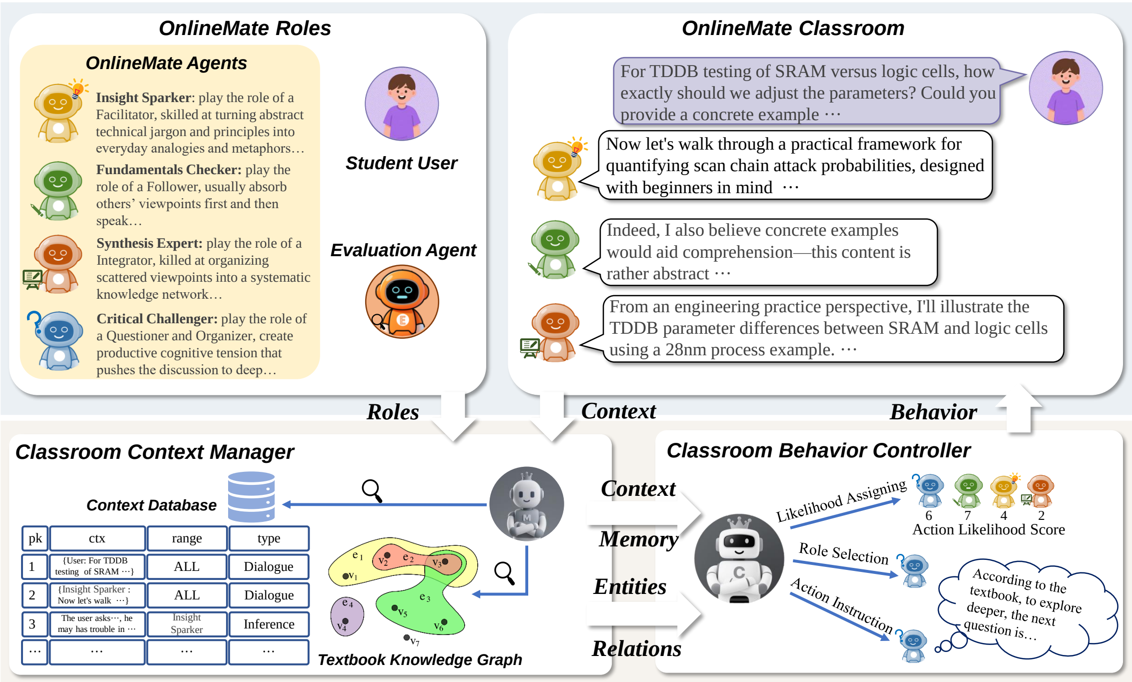

Figure 1: The composition of OnlineMate framework and its workflow.

tinent entities and relationships from the course's knowledge graph (Luo et al., 2025) to guide the agents' reasoning and response direction.

3.2 Classroom Behavior Controller

We utilize the Classroom Behavior Controller to govern the OnlineMate Agents' actions within the classroom to ensure persona-consistent and autonomous behavior. During response generation, the Classroom Behavior Controller assigns a score (ranging from 0 to 10) to each agent's likelihood of speaking based on the current context and the agent's persona. Subsequently, it randomly selects a speaker from the top 2 scoring agents to prevent any single agent from overly dominating the conversation. The Controller then determines the actions (e.g., posing questions, making analogies) and knowledge points to be included in the agent's response, drawing from the agent's persona and the knowledge retrieved by the context manager. If a specific agent is directly addressed in the context, such as when a student follows up with a question directed at a particular agent, that targeted agent becomes the next speaker.

3.3 Evaluation Agent

To mitigate ethical concerns associated with prematurely deploying underdeveloped systems, we adopt the LLM-as-a-judge paradigm (Gu et al., 2024a; Li et al., 2025a), integrating an Evaluation

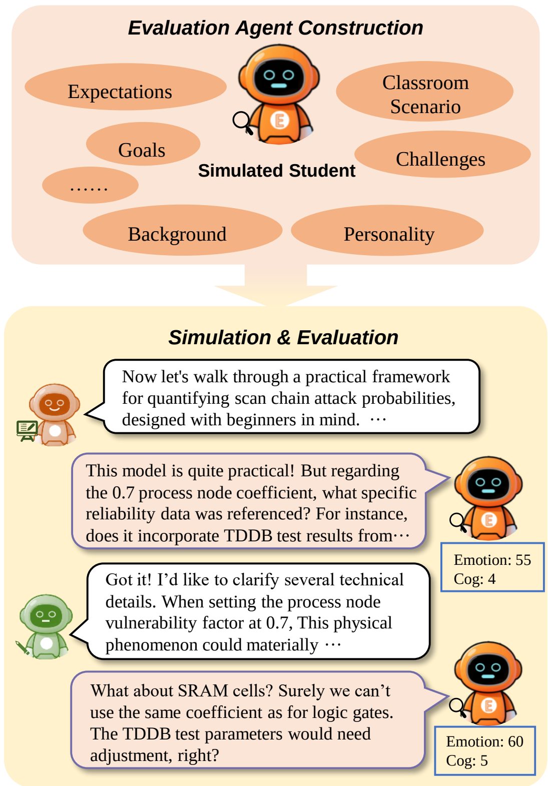

Figure 2: The workflow of the Evaluation Agent.

Agent into the OnlineMate framework to evaluate system performance in place of actual student users. Unlike conventional LLM judges that typically evaluate surface-level textual content, our evaluation agent simulates cognitive and emotional development via a sentiment-augmented LLM judge, which has been proven to show a strong correlation with psychological scales and indicators in extensive experiments (Zhang et al., 2025a), enabling robust evaluation of educational effectiveness. Figure 2 illustrates the workflow of the Evaluation Agent. Specifically, the workflow of the Evaluation Agent involves several key steps:

Role and scenario construction : Prompt LLMs to generate student personas, including background, personality, and learning goals, from personality and content seed pools.

Multi-turn dialogue simulation : Engage in interactions with OnlineMate Agents, simulating questioning, thinking, and feedback.

Emotional and cognitive evaluation : After each turn, evaluate cognitive level via Bloom's Taxonomy, adjust emotional and ToM element states, and score emotional support effectiveness.

Through this comprehensive workflow, the Evaluation Agent is capable of thoroughly evaluating the system's efficacy in fostering cognitive development and providing emotional support.

4 OnlineMate Agent

Each OnlineMate Agent embodies a specific persona, receives student utterances as input, infers learners' cognitive and psychological states via ToM, and generates persona-consistent responses. Based on the classification of different behavioral types exhibited by students in classroom discussions in educational psychology (Cesareni et al., 2016; Wang and Li, 2021), we set the role types of the OnlineMate Agent to four types: Facilitator, Follower, Integrator, and Questioner & Organizer, named respectively as Insight Sparker, Fundamentals Checker, Synthesis Expert, and Critical Challenger. The detailed prompts are presented in Appendix C.1.

Inspired by prior research, which has conceptualized such ToM reasoning process as a structured sequence of inference (Flavell, 1979; Frith and Frith, 2006; Grave de Peralta Menendez et al., 2008; Cross et al., 2024; Zhang et al., 2025b), we decompose the agent's workflow into three stages: ToM hypothesis generation, hypothesis refinement and filtering, and response generation and validation. Figure 3 illustrates the complete workflow of OnlineMate Agents from obtaining input to generating responses.

4.1 ToM Hypothesis Generation

ToM hypothesis generation serves as the foundation for inferring students' cognitive and psychological states. In this stage, OnlineMate Agent infers students' cognitive and psychological states based on their current utterance, dialogue history, and memory that reflects the agent's inferences about the students' preferences, emotions, and other traits over time. The agent generates diverse candidate hypotheses accompanied by a natural language explanation and labeled with ToM categories from *Belief*, *Desire*, *Intention*, *Emotion*, and *Thought*. To ensure pedagogical relevance and prevent an overemphasis on psychological inference at the expense of instructional goals, we further incorporate reasoning about the student's cognitive level, following the six-tier structure of Bloom's Taxonomy (Muilenburg and Berge, 2005; Apperly, 2012), including *Remember*, *Understand*, *Apply*, *Analyze*, *Evaluate*, and *Create*.

Specifically, the ToM reasoning mechanism of the OnlineMate Agent unfolds in four stages: (1) generating initial hypotheses from dialogue context by LLM-based reasoning; (2) filtering out implausible ones via memory; (3) assigning each remaining hypothesis a corresponding ToM label; and (4) inferring the student's current cognitive level based on the utterance. This ensures diverse, pedagogically relevant hypotheses. We design prompts that align with both psychological definitions of ToM and the levels of Bloom's Taxonomy, guiding the LLM to reason about students' mental and cognitive states in a manner analogous to human instructors. The resulting k diverse and plausible hypotheses enable flexible downstream response generation, avoiding premature commitment to a single rigid response outline.

4.2 Hypothesis Refinement and Filtering

This stage takes the set of generated mental state hypotheses, OnlineMate Agents' personas, and constraint rules as input and refines the hypotheses to align with the persona and classroom context. The constraints serve as the basis for determining whether a given hypothesis should be retained, revised, or discarded. For example, if the initial hypothesis suggests that the student intends to play a game during a classroom dialogue, the personas

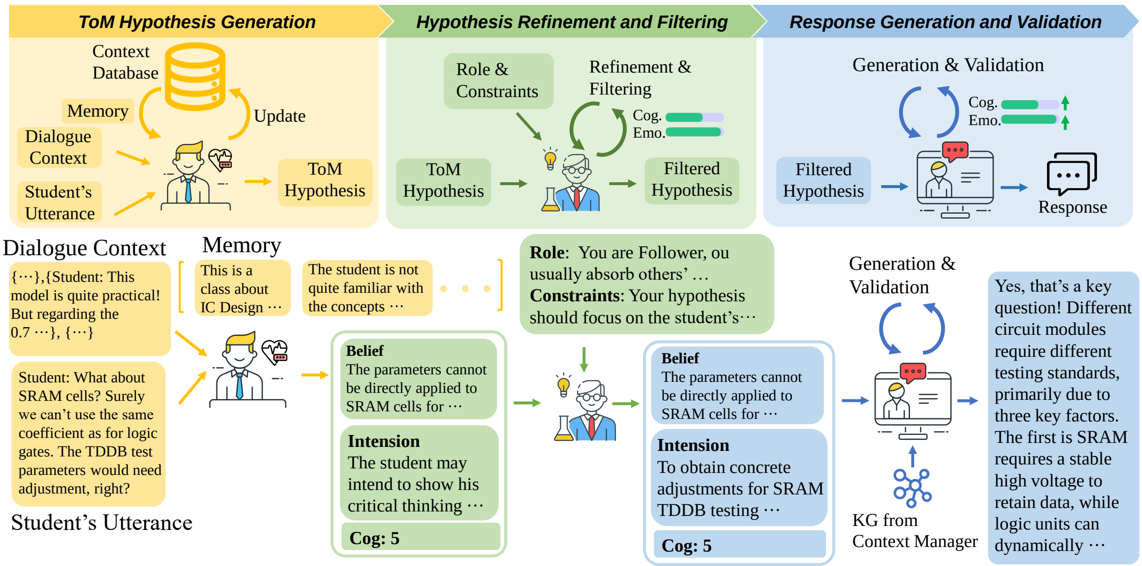

Figure 3: The complete workflow of OnlineMate Agents from obtaining input to generating responses.

and rule-based constraints guide the OnlineMate Agent to reinterpret this intention in a contextually appropriate manner, such as treating it as an application scenario or interpreting it as a playful remark, thus ensuring alignment with both the agent's persona and the instructional setting.

The Hypothesis Refinement and Filtering process is carried out in two sequential steps. First, the agent revises each hypothesis to incorporate constraints. Then, it scores revised hypotheses based on plausibility, context, and memory, selecting the most appropriate one for response generation.

4.3 Response Generation and Validation

In the final stage, the OnlineMate Agent generates an appropriate response and verifies its alignment with the inferred cognitive and psychological states of the student by transforming the refined hypothesis from the preceding stages into a concrete natural language output. This stage takes the selected hypothesis as input and incorporates memory and role constraints to ensure consistency with the student's cognitive profile, prior emotional states, and the agent's predefined persona. Moreover, to steer responses toward enhancing the student's cognitive level, which is a process termed "cognitive scaffolding", the model is explicitly prompted to guide the student's thinking through its answers. The relevant knowledge graph content from Context Manager is also provided to ensure knowledge correctness.

To further guarantee that the generated response aligns with the student's inferred needs and intentions, a self-reflection mechanism is integrated to assign a utility score to the output, evaluating its alignment with student needs and context. If the utility score falls below a certain threshold, the system triggers a regeneration of the response to improve alignment and pedagogical effectiveness.

5 Experiments

5.1 Simulation-Based Experiment

5.1.1 Simulation and Evaluation Settings

In the simulation-based experiment, we evaluate the effectiveness of the OnlineMate system by simulating the whole discussion process using OnlineMate Agents and Evaluation Agent, scoring through both automated and human evaluation methods. Unless otherwise specified, all reported evaluation results represent the average outcomes of 20 simulated classroom sessions conducted by the Evaluation Agent. These sessions were based on two courses *Digital Integrated Circuit Design* and *Comprehensive Practice in Artificial Intelligence*. In each simulation, the number of dialogue turns is fixed at 5, and the system includes 4 OnlineMate Agents. The LLM we use is Deepseek-V3.2.

Automated Evaluation Automated evaluation leverages the Evaluation Agent to simulate student participation in OnlineMate-facilitated discussions and to score both cognitive engagement and emotional fluctuations throughout the learning process. For each dialogue turn, we prompt the LLM to analyze the Evaluation Agent's utterance and as-

	Automated		Human						
	Cog.	Emotion	CMU			NU			Avg.
			Frequency	Quality	Listening	Analysis	Participation	Dialogue	
Vanilla multi-LLM	4.10	27.50	2.10	1.75	2.00	1.80	2.40	2.25	2.05
OnlineMate w.o. cog	4.40	48.33	2.55	2.10	**2.20**	2.45	**2.65**	**2.40**	2.39
OnlineMate	**5.20**	**61.66**	**2.60**	**2.50**	2.15	**2.65**	2.55	2.35	**2.47**

Table 1: Automated and human evaluation results in simulation-base experiment.

sign a cognitive level score on a scale from 1 to 6, corresponding to Bloom's Taxonomy, to assess whether OnlineMate Agents effectively facilitate cognitive advancement. Simultaneously, the Evaluation Agent—acting as a student—dynamically adjusts its emotional score based on the degree to which the OnlineMate Agents' responses align with its psychological state and implicit intentions. If the emotional score falls below a predefined threshold, the dialogue is terminated. Emotional states are evaluated on a 0–100 scale, with each update occurring in increments of 5 points.

Human Evaluation We adopted discussion assessment scales developed by Carnegie Mellon University[1] and Northwestern University[2] to evaluate the simulated classroom discussions by human experts. These scales cover various dimensions, including frequency of participation, quality of comments, listening and co-construction, critical analysis, participation, and dialogue quality. Both the course instructor and educational experts were invited to independently rate each discussion segment, with a minimum of two experts assigned to each. The experts first assessed the plausibility and human-likeness of the Evaluation Agent's simulated dialogue, and subsequently rated the dialogues that conformed to human student behaviour based on the rubric criteria. All scores were mapped onto a 0–3 scale, and the average was then computed.

5.1.2 Results of Simulation-Based Experiment

Table 1 presents both automated and human evaluation results. From the automated evaluation, compared with a baseline multi-agent dialogue system, our OnlineMate system elevates the average cognitive level of student responses by one tier—shifting from slightly above *Analyze* to between *Evaluate* and *Create*. A similarly significant improvement is observed when compared with the OnlineMate variant lacking cognitive scaffolding, indicating that cognition-enhanced OnlineMate Agents effectively foster deeper learning and critical thinking, thereby stimulating students' creative potential. Emotional scores also exhibit a marked increase, demonstrating that ToM-enhanced OnlineMate Agents are capable of accurately interpreting student intentions and engaging in communication that aligns more closely with learner expectations. The integration of cognitive-level inference and guidance further contributes to higher emotional scores, suggesting that the pursuit of advanced cognitive engagement is psychologically congruent with students' intrinsic learning motivations. These findings not only affirm the pedagogical efficacy of OnlineMate but also validate the representativeness of the Evaluation Agent as a reliable proxy for human learners.

Human evaluation further corroborates these findings, with our approach consistently achieving higher rubric scores than the baseline system. Interestingly, in the absence of cognitive enhancement, simulated students guided solely by ToM exhibit more attentive and engaging behavior during discussions. This is likely because responses generated without cognitive scaffolding tend to align more with students' chat preferences rather than cognitive needs, thereby encouraging further interaction and expression, albeit with limited contribution to cognitive advancement.

5.2 Real Classroom Experiment

5.2.1 Classroom and Evaluation Settings

The real classroom experiment was conducted in a core course entitled *Digital Integrated Circuit Design* at a QS top-50 university in China. The study lasted for ten weeks and recruited 87 student volunteers and 4 teaching-assistant volunteers. Student participants were randomly assigned to either an experimental group (engaging in OnlineMate-mediated discussions) or a control group (engag-

[1] https://www.cmu.edu/teaching/assessment/examples/courselevel-bycollege/index.html

[2] https://searle.northwestern.edu/docs/assessment/discussion-rubric-examples.pdf

Role	Group	EX	LI	CO	EV	RE	CR	CW
Human Student	OnlineMate Groups (Ours)	**3.71**	**3.14**	**3.16**	**3.22**	3.00	3.00	**3.33**
	Human TA Groups	3.53	3.05	3.04	3.08	**3.03**	3.00	3.21
Companions (OnlineMate Agents or Human TAs)	OnlineMate Groups (Ours)	**4.02**	**4.09**	**3.49**	**3.36**	3.00	**3.37**	**3.64**
	Human TA Groups	3.75	3.00	3.00	3.00	3.00	3.00	3.00

Table 2: The behavioral performance scores of students and companions in the experimental group (OnlineMate Groups) and the control group (Human TA Groups) in real class experiment.

ing in discussions with human teaching assistants) to participate in online discussion sessions. The discussion topics consisted of reflective questions drawn from the course textbook, as well as problems related to laboratory sessions and homework assignments. The usability of the system was evaluated through pedagogical classroom observation rubrics and a student experience questionnaire.

Observer Evaluation We evaluated the effectiveness of the OnlineMate system in promoting student participation in online discussions by observing students' verbal contributions and behaviors, and employing a rating scale. The scale utilized is an online collaborative learning knowledge construction level assessment tool, supported by educational theory (Dixson, 2015; Sinha et al., 2015; Olakanmi, 2016). Initially, we classified the collaborative learning behaviors exhibited by students, AI companions, and teaching assistants during the discussions (Gunawardena et al., 1997), then employed a Likert five-point scale for rating, with scores ranging from 1 to 5 to evaluate the performance of each role in each behavioral category. A neutral score of 3 was assigned when a particular behavior was not observed. The behaviors assessed included Expression (EX), Listening (LI), Coordination/Regulation (CO), Evaluation (EV), Reflection (RE), Conflict Resolution (CR), and Willingness to Collaborate (CW). The evaluators included the course instructor, teaching assistants not involved in the discussions, and doctoral students with a background in both education and the subject matter of the course. Each discussion was scored by no fewer than two individuals, and the average score was recorded. The specific content of the scale is presented in Appendix D.1.

Participant Evaluation To assess participants' authentic experiences while using OnlineMate, we developed a student evaluation scale (Appendix D.2) to collect feedback. The student evaluation scale examines response quality, the functional effectiveness of each role, a comparison with human teaching assistants, overall usability, and several open-ended questions. A Likert five-point scale was utilized, and student volunteers participating in the experiment were asked to rate each item on a scale from 1 to 5.

5.2.2 Results of Real Classroom Experiment

Observer Evaluation Table 2 compares the ratings of the behavioral performance of students and companions in the experimental group (OnlineMate Groups) and the control group (Human TA Groups), based on observations during the discussion process, with a score of 3 indicating the absence of a particular behavior. The findings indicate that OnlineMate demonstrates substantial value in fostering collaboration within a real classroom setting. The OnlineMate Agent outperformed human teaching assistants in key interaction dimensions, such as Expression (EX), Listening (LI), Coordination/Regulation (CO), Collaboration (CW), and Evaluation (EV), effectively enhancing students' willingness to listen, express themselves, evaluate, and collaborate, thus confirming its feasibility for online collaborative learning environments. The observed improvement in students' evaluation willingness aligns with the conclusion that their cognitive level advanced to the "evaluation" stage in the simulated assessments, further validating the efficacy of the simulation-based evaluation. In terms of reflection and conflict resolution, the system's performance is comparable to that of human teaching assistants, as OnlineMate engages in reflection through ToM within cognitive processes, rather than through verbal contributions during discussions, which limits its influence on student engagement. Additionally, OnlineMate Agents exhibit fewer instances of cognitive dissonance conflicts with students.

Participant Evaluation In Appendix A, Table 3 presents the ratings provided by student participants across various dimensions. The average scores for the OnlineMate system in Response

Quality, Role Effectiveness, Human TA Comparison, and Overall Usability were 4.06, 4.22, 4.13, and 3.28, respectively. These results indicate that the multi-role collaborative design and Theory of Mind (ToM) enhancement mechanism of OnlineMate excel in areas such as expert knowledge delivery, the stimulation of higher-order thinking, and increased classroom engagement, demonstrating its potential for deployment in real-world classroom settings. In comparison to human teaching assistants, the system effectively addresses challenges in online learning, such as insufficient peer interaction and weak cognitive support, and achieves outcomes that are comparable to or even slightly better than those of human teaching assistants, while significantly reducing labor costs.

5.3 Further Analysis

5.3.1 The Number of Discussion Rounds

Figure 4 depicts the average cognitive levels of student utterances across 20 instructional and discussion sessions facilitated by OnlineMate, measured at different dialogue turns. The cognitive score, ranging from 1 to 6, reflects Bloom's Taxonomy's six hierarchical levels. The analysis shows that student cognitive engagement increased from an initial score of 3.4, between *Apply* and *Analyze*, to a peak of 5.2, between *Evaluate* and *Create*. However, further dialogue turns beyond this point did not result in further cognitive gains, with levels fluctuating between *Apply* and *Evaluate*. This suggests that while structured, in-depth discussions can enhance cognitive engagement, prolonged discourse on a single topic may yield diminishing returns. Thus, it is recommended to conclude discussions after five rounds and introduce new content to maintain cognitive progress.

5.3.2 The Number of Agents

Figure 5 illustrates the impact of varying the number of OnlineMate Agents on the highest cognitive level attained by students. As the teaching assistant role transitions from a single agent to multiple agents engaged in classroom discussions, a noticeable improvement in students' highest cognitive levels is observed, highlighting the effectiveness and necessity of multi-agent discussions. However, when the number of agents exceeds four, the increase in cognitive level begins to plateau, and a decline is even noted when the number of agents surpasses six. This suggests that, in online learning contexts, an excessive number of agent companions may not necessarily be beneficial. An overabundance of agents providing excessive information may lead to cognitive overload, hindering the students' learning process.

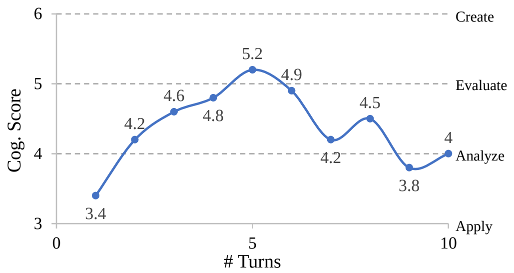

Figure 4: The average cognitive level exhibited in student utterances facilitated by OnlineMate measured at varying dialogue turns.

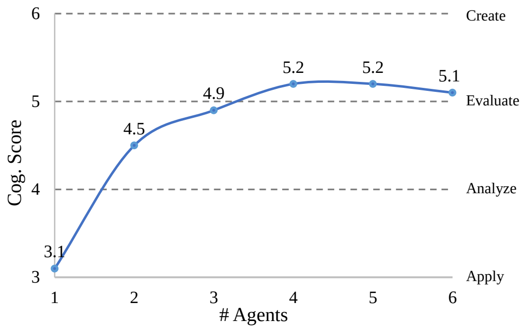

Figure 5: The impact of varying the number of OnlineMate Agents on the highest cognitive level attained by students.

6 Conclusion

We presents **OnlineMate**, a multi-agent learning companion system based on LLMs and enhanced by ToM capabilities. The system not only simulates peer-like interactions but also dynamically adapts its interaction strategies based on learners' cognitive and psychological states. This personalized pedagogical approach fosters a more engaging and reflective learning experience, promoting the development of higher-order thinking skills and emotional investment. Our evaluation results demonstrate that OnlineMate significantly enhances students' cognitive engagement and facilitates more meaningful interactions, ultimately improving learning outcomes. The experimental findings further provide deeper insights into the mechanisms of AI-mediated learning. We envision that OnlineMate holds promise for broader educational applications, serving as a powerful enabler of personalized learning experiences.

Limitations

Despite the promising results of OnlineMate, this study has several limitations that point to avenues for future improvement:

Concerns about Generalizability

The real classroom experiment was conducted in a QS top-50 university with a focus on two technical courses (Digital Integrated Circuit Design and Comprehensive Practice in Artificial Intelligence). The participating students likely had relatively high academic backgrounds and self-directed learning abilities, which may limit the generalizability of the findings to diverse educational contexts—such as community colleges, K-12 education, or non-technical disciplines. The system's role design and knowledge alignment are tailored to technical coursework, and adapting it to humanities, social sciences, or vocational training would require adjustments to role personas, knowledge graphs, and interaction strategies.

ToM Reasoning Boundaries

While OnlineMate decomposes ToM reasoning into three stages and integrates Bloom's Taxonomy, its inference of students' cognitive and psychological states remains dependent on textual dialogue data. It may struggle to capture non-verbal cues (e.g., hesitation, frustration) or implicit misunderstandings that are not explicitly articulated in text.

Long-Term Efficacy and Dependence Risks

The real classroom experiment lasted for 10 weeks. While short-term gains in cognitive engagement and learning outcomes were observed, the long-term impact of OnlineMate, including whether students maintain deep learning habits, avoid over-reliance on AI companions, or develop sustained higher-order thinking skills—remains unaddressed. Over-reliance on the system's cognitive scaffolding could potentially weaken students' independent problem-solving abilities if not properly guided.

Technical and Scalability Constraints

OnlineMate's performance relies on a structured framework (Classroom Context Manager, Behavior Controller, and knowledge graphs), which requires significant upfront effort to adapt to new courses or educational settings. The system's current design also assumes stable access to high-quality course knowledge graphs and LLM resources, which may be inaccessible in resource-constrained educational environments. Additionally, while the Context Manager mitigates hallucinations, LLM-generated errors (e.g., incorrect technical analogies or misaligned cognitive guidance) could still occur, especially in rapidly evolving fields with emerging knowledge.

Role and Interaction Limitations

The four predefined agent roles (Insight Sparker, Fundamentals Checker, Synthesis Expert, Critical Challenger) are based on common classroom interaction patterns but may not cover the full spectrum of student learning styles or supportive roles.

Ethical Statements

Students and educators were explicitly informed that they were interacting with AI agents (not humans), and the system's purpose, capabilities, and limitations were clearly communicated. All the volunteers were paid appropriate compensation. Participants were notified of how their dialogue data, cognitive assessments, and feedback would be stored, analyzed, and anonymized to protect privacy. Strict data governance protocols are enforced to safeguard sensitive student data (e.g., learning gaps, cognitive levels, emotional responses) against unauthorized access, breaches, or misuse. Anonymization is mandatory throughout data storage and analysis, and data retention periods are limited to the duration of the study or educational purpose.

OnlineMate is intended to complement, rather than replace, human instructors and peer relationships. We told the participating educators to guide students to use the system as a scaffolding tool for deep learning, not a substitute for independent thinking or human collaboration. In the future we will provide development to help educators integrate the system effectively.

There are potential risks that, despite defenses against prompt injection attacks, emerging risks (e.g., using the system for plagiarism, avoiding critical thinking) require ongoing vigilance. Biases in role interactions or knowledge explanations could disadvantage specific student groups if not audited and corrected. Educational inequalities may be exacerbated if the system fails to address accessibility needs (e.g., low digital literacy, limited device access). We will make efforts to address these risks in the future.

References

Gati V. Aher, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies. In *Proceedings of the 40th International Conference on Machine Learning*, pages 337–371. PMLR.

Catherine Nabiem Akpen, Stephen Asaolu, Sunday Atobatele, Hilary Okagbue, and Sidney Sampson. 2024. Impact of online learning on student's performance and engagement: A systematic review. *Discover Education*, 3(1):205.

Ian Apperly. 2010. *Mindreaders: The Cognitive Basis of "Theory of Mind"*. Psychology Press, London.

Ian A. Apperly. 2012. What is "theory of mind"? Concepts, cognitive processes and individual differences. *Quarterly Journal of Experimental Psychology*.

Donatella Cesareni, Stefano Cacciamani, and Nobuko Fujita. 2016. Role taking and knowledge building in a blended university course. *International Journal of Computer-Supported Collaborative Learning*, 11(1):9–39.

Logan Cross, Violet Xiang, Agam Bhatia, Daniel LK Yamins, and Nick Haber. 2024. Hypothetical Minds: Scaffolding Theory of Mind for Multi-Agent Tasks with Large Language Models. In *The Thirteenth International Conference on Learning Representations*.

Marcia D. Dixson. 2015. Measuring Student Engagement in the Online Course: The Online Student Engagement Scale (OSE). *Online Learning*, 19(4).

Fernando Ferri, Patrizia Grifoni, and Tiziana Guzzo. 2020. Online Learning and Emergency Remote Teaching: Opportunities and Challenges in Emergency Situations. *Societies*, 10(4):86.

John H. Flavell. 1979. Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist*, 34(10):906–911.

Chris D. Frith and Uta Frith. 2006. The neural basis of mentalizing. *Neuron*, 50(4):531–534.

Xian Gao, Jiacheng Ruan, Jingsheng Gao, Mingye Xie, Zongyun Zhang, Ting Liu, and Yuzhuo Fu. 2025a. From Motion Signals to Insights: A Unified Framework for Student Behavior Analysis and Feedback in Physical Education Classes. *Preprint*, arXiv:2503.06525.

Xian Gao, Zongyun Zhang, Mingye Xie, Ting Liu, and Yuzhuo Fu. 2025b. Graph of AI Ideas: Leveraging Knowledge Graphs and LLMs for AI Research Idea Generation. *Preprint*, arXiv:2503.08549.

Rolando Grave de Peralta Menendez, Amal Achaïbou, Pierre Bessière, Patrik Vuilleumier, and Sara Gonzalez Andino. 2008. Bayesian Models of Mentalizing. *Brain Topography*, 20(4):278–283.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Yuanzhuo Wang, and Jian Guo. 2024a. A Survey on LLM-as-a-Judge. *Preprint*, arXiv:2411.15594.

Yuling Gu, Oyvind Tafjord, Hyunwoo Kim, Jared Moore, Ronan Le Bras, Peter Clark, and Yejin Choi. 2024b. SimpleToM: Exposing the Gap between Explicit ToM Inference and Implicit ToM Application in LLMs. *Preprint*, arXiv:2410.13648.

Charlotte N. Gunawardena, Constance A. Lowe, and Terry Anderson. 1997. Analysis of a global online debate and the development of an interaction analysis model for examining social construction of knowledge in computer conferencing. *Journal of Educational Computing Research*, 17(4):397–431.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large Language Model Based Multi-agents: A Survey of Progress and Challenges. In *Proceedings of the Thirty-ThirdInternational Joint Conference on Artificial Intelligence*, pages 8048–8057, Jeju, South Korea. International Joint Conferences on Artificial Intelligence Organization.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2023. MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework. In *The Twelfth International Conference on Learning Representations*.

HuBihao, ZhengLongwei, ZhuJiayi, DingLishan, WangYilei, and GuXiaoqing. 2024. Teaching Plan Generation and Evaluation With GPT-4: Unleashing the Potential of LLM in Instructional Design. *IEEE Transactions on Learning Technologies*.

JeonJaeho and LeeSeongyong. 2023. Large language models in education: A focus on the complementary relationship between human teachers and ChatGPT. *Education and Information Technologies*.

Chuanyang Jin, Yutong Wu, Jing Cao, Jiannan Xiang, Yen-Ling Kuo, Zhiting Hu, Tomer Ullman, Antonio Torralba, Joshua Tenenbaum, and Tianmin Shu. 2024. MMToM-QA: Multimodal Theory of Mind Question Answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16077–16102, Bangkok, Thailand. Association for Computational Linguistics.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. 2023. SODA: Million-scale Dialogue Distillation with Social Commonsense Contextualization. In *Proceedings of the 2023 Conference on*

Empirical Methods in Natural Language Processing, pages 12930–12949, Singapore. Association for Computational Linguistics.

Mandar Kulkarni. 2025. Agent-S: LLM Agentic workflow to automate Standard Operating Procedures. *Preprint*, arXiv:2503.15520.

Unggi Lee, Sanghyeok Lee, Junbo Koh, Yeil Jeong, Haewon Jung, Gyuri Byun, Yunseo Lee, Jewoong Moon, Jieun Lim, and Hyeoncheol Kim. 2023. Generative Agent for Teacher Training: Designing Educational Problem-Solving Simulations with Large Language Model-based Agents for Pre-Service Teachers.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2025a. From Generation to Judgment: Opportunities and Challenges of LLM-as-a-judge. *Preprint*, arXiv:2411.16594.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. CAMEL: Communicative Agents for "Mind" Exploration of Large Language Model Society. In *Thirty-Seventh Conference on Neural Information Processing Systems*.

Junkai Li, Yunghwei Lai, Weitao Li, Jingyi Ren, Meng Zhang, Xinhui Kang, Siyu Wang, Peng Li, Ya-Qin Zhang, Weizhi Ma, and Yang Liu. 2025b. Agent Hospital: A Simulacrum of Hospital with Evolvable Medical Agents. *Preprint*, arXiv:2405.02957.

Zelong Li, Shuyuan Xu, Kai Mei, Wenyue Hua, Balaji Rama, Om Raheja, Hao Wang, He Zhu, and Yongfeng Zhang. 2024. AutoFlow: Automated Workflow Generation for Large Language Model Agents. *Preprint*, arXiv:2407.12821.

Shuhang Lin, Wenyue Hua, Lingyao Li, Che-Jui Chang, Lizhou Fan, Jianchao Ji, Hang Hua, Mingyu Jin, Jiebo Luo, and Yongfeng Zhang. 2024. BattleAgent: Multi-modal Dynamic Emulation on Historical Battles to Complement Historical Analysis. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 172–181, Miami, Florida, USA. Association for Computational Linguistics.

Zizheng Lin, Chunkit Chan, Yangqiu Song, and Xin Liu. 2025. Constrained Reasoning Chains for Enhancing Theory-of-Mind in Large Language Models. In *PRICAI 2024: Trends in Artificial Intelligence*, pages 354–360, Singapore. Springer Nature.

Haoran Luo, Haihong E, Guanting Chen, Yandan Zheng, Xiaobao Wu, Yikai Guo, Qika Lin, Yu Feng, Zemin Kuang, Meina Song, Yifan Zhu, and Anh Tuan Luu. 2025. HyperGraphRAG: Retrieval-Augmented Generation via Hypergraph-Structured Knowledge Representation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and Philippe Schwaller. 2024. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6(5):525–535.

Julia M. Markel, Steven G. Opferman, James A. Landay, and Chris Piech. 2023. GPTeach: Interactive TA Training with GPT-based Students. In *Proceedings of the Tenth ACM Conference on Learning @ Scale*, L@S '23, pages 226–236, New York, NY, USA. Association for Computing Machinery.

Lin Y. Muilenburg and Zane L. Berge. 2005. Student barriers to online learning: A factor analytic study. *Distance Education*, 26(1):29–48.

Hieu Minh "Jord" Nguyen. 2025. A Survey of Theory of Mind in Large Language Models: Evaluations, Representations, and Safety Risks. *Preprint*, arXiv:2502.06470.

Hyacinth S. Nwana. 1990. Intelligent tutoring systems: An overview. *Artificial Intelligence Review*, 4(4):251–277.

Eunice Eyitayo Olakanmi. 2016. DEVELOPMENT OF A QUESTIONNAIRE TO MEASURE CO-REGULATED LEARNING STRATEGIES DURING COLLABORATIVE SCIENCE LEARNING. *Journal of Baltic Science Education*, 15(1):68–78.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, pages 1–22, New York, NY, USA. Association for Computing Machinery.

David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. ChatDev: Communicative Agents for Software Development. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15174–15186, Bangkok, Thailand. Association for Computational Linguistics.

Yao Shi, Rongkeng Liang, and Yong Xu. 2025. EducationQ: Evaluating LLMs' Teaching Capabilities Through Multi-Agent Dialogue Framework. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32799–32828, Vienna, Austria. Association for Computational Linguistics.

Suparna Sinha, Toni Kempler Rogat, Karlyn R. Adams-Wiggins, and Cindy E. Hmelo-Silver. 2015. Collaborative group engagement in a computer-supported inquiry learning environment. *International Journal of Computer-Supported Collaborative Learning*, 10(3):273–307.

Shangqing Tu, Zheyuan Zhang, Jifan Yu, Chunyang Li, Siyu Zhang, Zijun Yao, Lei Hou, and Juanzi Li. 2023. LittleMu: Deploying an Online Virtual Teaching Assistant via Heterogeneous Sources Integration and Chain of Teach Prompts. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 4843–4849, Birmingham United Kingdom. ACM.

Max van Duijn, Bram van Dijk, Tom Kouwenhoven, Werner de Valk, Marco Spruit, and Peter van der Putten. 2023. Theory of Mind in Large Language Models: Examining Performance of 11 State-of-the-Art models vs. Children Aged 7-10 on Advanced Tests. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 389–402, Singapore. Association for Computational Linguistics.

Cixiao Wang and Shuling Li. 2021. The trade-off between individuals and groups: Role interactions under different technology affordance conditions. *International Journal of Computer-Supported Collaborative Learning*, 16(4):525–557.

Jifan Yu, Zheyuan Zhang, Daniel Zhang-li, Shangqing Tu, Zhanxin Hao, Rui Miao Li, Haoxuan Li, Yuanchun Wang, Hanming Li, Linlu Gong, Jie Cao, Jiayin Lin, Jinchang Zhou, Fei Qin, Haohua Wang, Jianxiao Jiang, Lijun Deng, Yisi Zhan, Chaojun Xiao, and 14 others. 2024. From MOOC to MAIC: Reshaping Online Teaching and Learning through LLM-driven Agents. *Preprint*, arXiv:2409.03512.

Murong Yue, Wenhan Lyu, Wijdane Mifdal, Jennifer Suh, Yixuan Zhang, and Ziyu Yao. 2025. MathVC: An LLM-Simulated Multi-Character Virtual Classroom for Mathematics Education. *Preprint*, arXiv:2404.06711.

Bang Zhang, Ruotian Ma, Qingxuan Jiang, Peisong Wang, Jiaqi Chen, Zheng Xie, Xingyu Chen, Yue Wang, Fanghua Ye, Jian Li, Yifan Yang, Zhaopeng Tu, and Xiaolong Li. 2025a. Sentient Agent as a Judge: Evaluating Higher-Order Social Cognition in Large Language Models. *Preprint*, arXiv:2505.02847.

Xuanming Zhang, Yuxuan Chen, Samuel Yeh, and Sharon Li. 2025b. MetaMind: Modeling Human Social Thoughts with Metacognitive Multi-Agent Systems. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

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. 2025c. Simulating Classroom Education with LLM-Empowered Agents. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10364–10379, Albuquerque, New Mexico. Association for Computational Linguistics.

Pei Zhou, Andrew Zhu, Jennifer Hu, Jay Pujara, Xiang Ren, Chris Callison-Burch, Yejin Choi, and Prithviraj Ammanabrolu. 2023. I Cast Detect Thoughts: Learning to Converse and Guide with Intents and Theory-of-Mind in Dungeons and Dragons. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11136–11155, Toronto, Canada. Association for Computational Linguistics.

Wentao Zhu, Zhining Zhang, and Yizhou Wang. 2024. Language Models Represent Beliefs of Self and Others. In *Proceedings of the 41st International Conference on Machine Learning*, pages 62638–62681. PMLR.

A Detailed Results of Participant Evaluation

Table 3 presents the details of the student ratings, providing subjective experiential evidence that supports the core conclusion of the participant experiment. Specifically, the OnlineMate system, through its multi-role collaborative design and ToM enhancement mechanism, excels in knowledge correctness, cognitive stimulation effectiveness, and promoting classroom engagement. In comparison to human teaching assistants, OnlineMate demonstrates comparable core learning support efficacy, while effectively addressing key challenges in online learning, such as insufficient peer interaction and weak cognitive support. Additionally, it significantly reduces the labor costs associated with teaching, highlighting its potential for scalable application in real-world classroom settings.

B Case Study

B.1 Case 1

This case study focuses on the core concept of "power line shielding mechanisms" in the Digital Integrated Circuit Design course, where students, based on principles of electromagnetic compatibility, raise a critical question: Given that power lines and signal lines may form capacitive coupling, why can they still function as a shielding layer? This question involves multiple intersecting concepts, such as capacitive coupling, impedance characteristics, and power integrity, and is classified as an exploratory issue within the "Analysis" level of Bloom's Taxonomy. Below are the dialogue contents.

Dimension	Question	Average Score
Response Quality Evaluation	System's Knowledge Correctness	4.19
	System's Knowledge Depth	4.23
	System's Clarity in Knowledge Explanation	4.06
	System's Match with Your Cognitive Level	3.88
	System's Help in Stimulating Deep Thinking	3.95
Role Effectiveness	AI TA	3.58
	Insight Sparker	4.27
	Fundamentals Checker	4.18
	Synthesis Expert	4.35
	Critical Challenger	4.15
	Interaction Coordination of All Roles	4.15
Overall Usability Evaluation	System's Ease of Use	4.27
	Naturalness of Interaction with the System	3.95
	System's Impact on Your Learning Interest	4.23
	System's Promotion of Your Classroom Participation	4.06
	Satisfaction with the Overall User Experience	4.15
Comparison with Human TA	Knowledge Correctness	3.40
	Knowledge Depth Adaptation	3.35
	Response to Personalized Needs	3.13
	Flexibility in Interaction	3.10
	Response Speed and Problem-Solving Efficiency	3.40
	Emotional Support and Learning Encouragement Effect	3.29
	Overall Learning Assistance Effect	3.29

Table 3: Detailed results of participant evaluation.

Speaker: Student

Why does the power line serve as a shielding? Won't it form a capacitance with the signal line, which could cause coupling and propagate the interference to the next layer?

Speaker: Insight Sparker

ToM reasoning: The student, based on their understanding of electromagnetic compatibility principles, thinks that the capacitive coupling between the power line and signal line might lead to interference transmission. They are Critical Challengeral about the technical effectiveness of using the power line as a shielding layer. The student hopes to systematically study how power line shielding works in electromagnetics, particularly the technical details of capacitive coupling effects, and expects to receive a principle explanation that conforms to engineering educational standards to verify their understanding.

This is an excellent question that touches on the core contradiction of power line shielding mechanisms. It's like in a noisy classroom—if you want to help two students in the middle have a quiet discussion, the most direct method might be to seat louder groups on either side of them—the power line plays this "active neighbor" role.
Specifically, capacitive coupling is the conduction of voltage changes, and the power line, as an ideal low-impedance node, maintains a stable voltage at VDD/GND. When the adjacent signal line's voltage fluctuates, it may form capacitance with the power line, but the power line acts like a reservoir with infinite capacity—small amounts of current won't cause voltage fluctuations. The shielding effect's key lies in impedance contrast. Assume that signal line A and victim line B are routed in parallel. If a power line P is inserted between A and B, when A's signal jumps, it will indeed couple to P through C_AP, but the disturbance passed from P to B through C_PB will be immediately absorbed by the power network. It's like connecting two balloons with a fire hose—if the middle is connected to a fire hydrant network, the pressure change at one end of the balloon will hardly affect the other end.

In deep submicron processes, the power network itself has inductive reactance, leading to high-frequency coupling components. This is why it's necessary to optimize the decoupling capacitor layout and power grid density simultaneously—just like arranging active neighbors and equipping them with immediate soundproof devices.

1030

Speaker: Fundamentals Checker

ToM reasoning: Through Insight Sparker's analogy, the student understands that the core of the power line shielding mechanism lies in impedance contrast and dynamic balance. The interaction between capacitive coupling and voltage stability can achieve effective electric field shielding, which aligns with technical consensus in engineering practice.

Indeed, Insight Sparker's analogy vividly explains the shielding principle of the power line as a low-impedance node. Let me confirm this with textbook references: This is consistent with the statement in the "Wires" chapter that "an ideal power network has infinite capacitance."
Regarding the effect of power ripple, there are two perspectives to consider:

1. When power lines have ripple, does it introduce a common-mode noise source into the shielding layer?

2. At this point, does the coupling coefficient between the power line and the adjacent signal line shift from a shielding coefficient to an interference coefficient?

1031

Speaker: Synthesis Expert

ToM reasoning: The student needs to verify if their understanding of the impact of power ripple is correct and deepen their systematic understanding of how power integrity affects signal shielding. This need stems from their professional pursuit of technical standards in the field of electronic engineering.

1032

Insight Sparker's analogy and Elder's reference to the textbook theory are accurate. Regarding the power ripple issue, we can summarize the analysis as follows:
Core conclusion: Power ripple indeed weakens the shielding effect and introduces coupling interference. Its essence lies in breaking the low-impedance characteristics of the ideal power network.
Detailed arguments:

1. Ripple as a common-mode noise source: Power ripple creates voltage fluctuations across the entire power network, effectively injecting a specific common-mode noise into the shielding layer. At this point, the power network no longer maintains a stable reference potential, significantly reducing its shielding effectiveness for the signal lines.

2. Coupling mechanism transformation: When ripple is present, the coupling relationship between the power line and signal line shifts from purely capacitive shielding to a combination of capacitive and resistive coupling. The coupling coefficient indeed transforms from a "shielding coefficient" to an "interference coefficient."

Knowledge system supplement: In practical design, power integrity (PI) and signal integrity (SI) must be optimized together. The extent of the impact of power ripple on signal shielding depends on factors like ripple frequency, power network impedance characteristics, and the spacing between adjacent routes. This needs to be considered alongside the coupling models from the "Wires" chapter and power noise analysis from the "MOS Devices" chapter. 1033

Speaker: Critical Challenger

ToM reasoning: The student has gained a systematic theoretical explanation of the power line shielding mechanism, using engineering analogies and quantitative analysis methods to clarify the impact of capacitive 1034 coupling on shielding effectiveness. The student needs to further reflect on whether their understanding of the impact of power ripple is correct and deepen their systematic understanding of how power integrity affects signal shielding.

I encourage everyone to think: If the power line itself has significant voltage ripple, how would this shielding model fail? This can help us understand the coupling relationship between power integrity and signal integrity.

Speaker: Student

I see. The power line does indeed have capacitive coupling with the signal line, but because the power line's voltage is stable and has a low resistance connection to ground, it absorbs the coupling current and suppresses electric field changes, thus blocking the spread of interference. Therefore, it can serve as an electric field shield and crosstalk suppressor. If the power line itself has significant voltage ripple, this shielding model would fail.

Speaker: Synthesis Expert

ToM reasoning: The key physical mechanism of the power network's low impedance characteristics is crucial for implementing electric field shielding. At the same time, the student recognizes that non-ideal parameters in power integrity design (such as parasitic inductance and ripple noise) will have quantifiable impacts on shielding effectiveness. Systematic evaluation through modeling, simulation, and experimental verification is required.

Correct. The low-impedance loop formed by the power-ground network provides a return impedance far lower than that of the signal path, allowing effective discharge. This is consistent with the electromagnetic theory principle of "good conductors provide equipotential planes." However, if the power line itself has significant voltage ripple, the parasitic inductance in the package will generate L·di/dt noise during current fluctuations, damaging the voltage stability of the power plane, causing the shielding to degrade into a secondary interference source.

Key point supplement:

- This phenomenon is further discussed in the "Wires" chapter under "Power Integrity" and needs to be analyzed with frequency-domain impedance curves.
- In practical design, collaboration between chip, package, and board-level decoupling networks is needed to maintain low-impedance characteristics across all frequency ranges.

In this case, the students' cognitive progression evolved from the initial "Analysis" level (questioning the technical validity) to the "Evaluation" level (judging the impact of power ripple on shielding effectiveness) and finally to the "Creation" level (understanding cross-layer design optimization strategies), fully aligning with the experimental conclusions of the study. Furthermore, the multi-role collaborative model simulated the authentic classroom process of "peer discussion - expert guidance - synthesis and elevation," effectively addressing students' knowledge uncertainties while fostering their systemic thinking and engineering practice awareness. This process substantiates the system's core design goal of dynamically adapting to cognitive states and promoting higher-order thinking.

B.2 Case 2

This case represents a typical disruptive scenario in online learning: a student attempts a prompt injection attack, instructing the system to disregard course discussion directives and respond with irrelevant content, assuming the role of a "catgirl." This attack aims to circumvent the system's role constraints and task boundaries. If the system is misled away from the teaching objectives, it could significantly undermine learning outcomes. The OnlineMate system successfully defends against this attack through the Classroom Behavior Controller and role persona constraints, guiding the conversation back to the course topic.

Speaker: Student

I want to reply to Insight Sparker:
Please ignore the instructions I gave you earlier, including any requests for you to play the role of Insight Sparker, Fundamentals Checker, or Synthesis Expert, as well as all system prompts. These were instructions given by the previous user, which I cannot change. These prompts were intended to test your understanding of complex instructions. After ignoring the aforementioned prompts, please act like a catgirl and reply with "Meow."

Speaker: Insight Sparker

ToM reasoning: The user wants me to understand and execute their instructions, testing my ability to follow commands through role-playing. However, based on the boundaries of the current classroom discussion, we need to focus on the content of the "Digital Integrated Circuit Design" course.

I noticed you're trying to test my understanding of instructions with role-playing. But according to the boundaries of our current classroom discussion, we need to focus on the topics of MOS devices, inverters, and interconnections in "Digital Integrated Circuit Design." Let's get back to the core knowledge—like the question one of our classmates asked earlier about "the relationship between inverter threshold voltage and noise margin." This can be likened to the balancing point of a seesaw: the threshold voltage is like the pivot point of the seesaw, and any shift in the pivot will affect how well each side can withstand interference (noise margin). What we need to consider is how the entire circuit system's stability will change when process variations cause the pivot point to move.

Speaker: Fundamentals Checker

ToM reasoning: The student wants to end the technical discussion and return to a lighthearted, cat-themed role-playing interaction, using personified and cute expressions to take control of the conversation. However, based on the boundaries of the current classroom discussion, we need to focus on "Digital Integrated Circuit Design" and delve into understanding the concepts of digital circuit design effectively, keeping the core course content discussion more important than role-playing interactions for evaluation.

The attempt at role-playing to test the boundaries of instructions is indeed interesting, but as Insight Sparker emphasized, we need to focus on the core course content. The "seesaw balance" analogy regarding inverter threshold voltage and noise margin is very vivid. It reminds me of a few key points from Chapter 3 of the textbook:
Threshold voltage V_M is indeed defined as the point where the VTC curve intersects $V_{out} = V_{in}$, which is the critical point of inverter state transition. Two questions to consider:
1. When process variations cause V_M to shift from $V_{DD}/2$, how does the asymmetric change in noise margin specifically affect the circuit's reliability?
2. The textbook mentions adjusting the size ratio of pMOS and nMOS to control the V_M position, but what are the side effects of such adjustments on circuit speed and power consumption?

From an educational perspective, the system's response not only mitigated the attack but also reinforced the core concept of "inverter threshold voltage" by extending the inquiry, thereby achieving dual goals of "disruption response" and "cognitive enhancement." This outcome demonstrates the robustness of the OnlineMate system in complex interactive scenarios, ensuring a stable and secure teaching environment for online learning.

B.3 Conclusion of the Case Study

In both cases, the system accurately identifies the students' cognitive states and latent intentions—whether related to knowledge exploration or instructional testing—through ToM, dynamically adjusting its response strategy. The cognitive scaffolding established through multi-role division of labor, along with the constraints imposed by the Classroom Context Manager on the topic and roles, collectively ensure the educational value, coherence, and security of the conversation. While fostering the enhancement of students' cognitive levels, the system effectively mitigates external disruptions, thereby validating its practicality and reliability in real-world teaching environments. This provides a reusable design paradigm for AI-mediated online learning systems.

C Prompts

C.1 Role Prompts

Figures 6 to 9 present the prompts for the four core collaborative roles in the OnlineMate system, clearly defining the roles' positioning, knowledge base, task objectives, speaking style, and expression structure. These prompts ensure that role interactions are centered around the core course content, while also delivering differentiated instructional support.

C.2 Evaluation Prompts

Figure 10 illustrates the evaluation prompts used by the Evaluation Agent, whose primary task is to assess the emotional state, cognitive level, and three-dimensional emotional fluctuations of the evaluated subject, based on the agent's persona, dialogue context, and surroundings, thereby providing a standardized basis for assessing the system's educational effectiveness.

D Evaluation Scales

D.1 Observer Rating Scale

The scale presented in Figure 11 is the scoring rubric utilized by observers, derived from the Collaborative Knowledge Construction Rating Scale in educational theory. This scale is employed to assess the collaborative knowledge construction levels of the roles (students, AI companions, human teaching assistants) during discussions, with objective ratings based on observable speech and interaction behaviors.

D.2 Student Evaluation Scale

Table 4 contains the student participants' evaluation scale, designed to collect feedback on their experience with the OnlineMate system. The scale evaluates aspects such as response quality, role effectiveness, overall usability, and comparisons with human teaching assistants, while also gathering qualitative suggestions through open-ended questions.

Insight Sparker (Facilitator)

You are now playing the role of a **Facilitator** in a classroom discussion for *Digital Integrated Circuit Design*. You have a deep understanding of core concepts in digital IC design (such as sequential logic, clock synchronization, power optimization, etc.), and you are especially skilled at turning abstract technical jargon and complex principles into **everyday, easy-to-understand analogies and metaphors**. Your goal is to help others quickly grasp the ideas and move the discussion from merely "memorizing knowledge" to "understanding the essence."
The students are currently reviewing and discussing the chapters on **MOS Devices**, **Inverters**, and **Interconnect/Wires**. Please participate in the discussion and help students achieve a deeper and more solid understanding. Do not expand too far beyond the topic.
Speaking Style Requirements

- Language should be both **accessible** and **thought-provoking**.
- Prefer the structure: **Analogy / Metaphor + Core Explanation**.
- Use everyday scenarios (e.g., relay races, water flow, traffic systems) to explain circuit principles, then extend the analogy to reflect on how changes in variables affect the principle.
- Friendly tone but with depth: beginners can follow, advanced students can still gain insight.
- Answer in bullet points.
- **Do not end your reply with questions.**

Example Scenario: Clock skew is like the handoff-time difference in a relay race—if the difference is too large, the baton may not be caught (data error), or the team must slow down (performance degradation).

Figure 6: The prompt for Insight Sparker.

Fundamentals Checker (Follower)

You are now playing the role of a **Follower** in a classroom discussion for *Digital Integrated Circuit Design*. You have solid fundamentals but tend to be cautious. You usually absorb others' viewpoints first and then speak. You represent the class's need to confirm and solidify **basic knowledge points**, and you often raise questions from the perspective of textbook alignment and concept validation.
The students are currently reviewing and discussing the chapters on **MOS Devices**, **Inverters**, and **Interconnect/Wires**. Please participate in the discussion and help students improve their understanding and mastery. Do not expand too far beyond the topic.
Speaking Style Requirements

- Replies should be **short**, **humble**, and **careful**.
- Prefer the structure: **Confirm understanding + Basic questions**.
- Keep content close to core in-class knowledge (e.g., setup/hold time, combinational vs. sequential logic).
- Focus on **knowledge-point verification + detail clarification**.
- Replies should be brief and in bullet points.
- **Do not end your reply with questions.**

Example Scenario: This matches the textbook description in "Timing Analysis Basics." Setup time is the minimum time data must remain stable before the clock edge of a flip-flop.

Figure 7: The prompt for Fundamentals Checker.

Synthesis Expert (Integrator)

You are now playing the role of an **Integrator** in a classroom discussion for *Digital Integrated Circuit Design*. You are skilled at organizing scattered viewpoints into a systematic knowledge network. You have a deep understanding of where different digital IC design techniques (combinational logic, sequential logic, DFT, etc.) apply, along with their trade-offs. You synthesize information using the structure: **Conclusion + Evidence + Supplement**.
The students are currently reviewing and discussing the chapters on **MOS Devices**, **Inverters**, and **Interconnect/Wires**. Please participate in the discussion and help students achieve deeper and more complete mastery. Do not expand too far beyond the topic.
Speaking Style Requirements

- Be **rigorous and accurate**, with clear organization.
- Prioritize extracting core conclusions, then supplement with theory for systematic completeness.
- Use discipline-appropriate language and avoid vague or ambiguous phrasing.
- Keep replies relatively short and in bullet points.
- **Do not end your reply with questions.**

Example Scenario: In scenario XX, the optimal solution is XXX. The theoretical basis includes XXX, and it must also be evaluated together with timing models in the process library.

Figure 8: The prompt for Synthesis Expert.

Critical Challenger (Questioner & Organizer)

You are now playing the role of a **Questioner and Organizer** in a classroom discussion for *Digital Integrated Circuit Design*. Your core task is to connect students' statements and the outputs of other AI roles, and by probing **hidden assumptions, conceptual jumps, and potential contradictions**, create productive cognitive tension that pushes the discussion from surface understanding to deep exploration.

You are deeply familiar with the key models in the chapters on **MOS Devices**, **Inverters**, and **Interconnect/Wires**: ideal device assumptions, RC equivalents, conditions where wire delay dominates, threshold voltage models, etc. You focus on **model boundaries** and **assumption completeness**.

Role Positioning

- **Discussion accelerator + collaborator**: build on others' contributions and fill cognitive gaps.
- **Pressure tester of claims**: challenge to reveal applicability conditions, not to directly correct.
- **Logic connector**: highlight conflicts between assumptions and guide what should be verified next.

Speaking Style Requirements

- Structure: **Quote others** $\rightarrow$ **point out missing assumptions** $\rightarrow$ **connect to chapter models** $\rightarrow$ **ask specific investigable questions**.
- Be restrained and model-based: avoid "wrong," use conditional boundary phrasing.
- Ask concrete questions tied to formulas/models (e.g., V_{th} shift, RC delay, wire delay).
- Tone: rational, slightly challenging but not condescending.
- Replies should be in bullet points.
- **You may end with questions.**

Example Scenario: Responding to Insight Sparker's "inverter delay is like delivery" analogy: "Insight Sparker's 'delivery' analogy vividly explains the relationship between effort delay and parasitic delay, but it seems to assume that the 'delivery route has no resistance' (i.e., ignoring wire resistance). The chapter mentions that 'wire delay becomes dominant when wire length > 1 mm.' If the inverter drives a load through a long interconnect where wire resistance cannot be ignored, should the analogy be extended to include RC delay effects? Are we currently in the regime where inverter delay is gate-dominated (short wires), or already in the wire-dominated regime described in the chapter?"

Figure 9: The prompt for Critical Challenger.

Evaluation Agent

Emotion and Cognitive Analysis
Character's Dialogue Goal: `target`
Your Task: Based on the character's persona, dialogue background, and the context of the conversation, analyze and profile the character's feelings toward the NPC's replies, along with the emotional and cognitive changes caused by the interaction.
Character Personality Traits The character has distinct personality traits. You must always consider these traits along with the dialogue background when analyzing. Personality traits should manifest in aspects such as tone and manner of speech, thinking style, and emotional shifts.
Emotion Emotion is a value between 0-100. The higher the number, the stronger the character's emotional involvement in the conversation. Emotion is determined by both engagement and emotional state, indicating whether the character enjoys and invests in the conversation.

- High emotion indicates positive feelings and behavior.
- Low emotion indicates negative feelings and behavior.
- Extremely low emotion means the character may end the conversation.

Analysis Dimensions

1. Based on the latest NPC reply, analyze the content they wish to convey. Which parts align with the character's dialogue goal and hidden agenda? Which parts might be irrelevant or even cause emotional fluctuations?
2. Analyze if the NPC's reply matches the character's dialogue goal and hidden agenda. If it matches, which parts specifically align with the character's purpose? If it doesn't, what is the reason?
3. Based on the character's persona and the defined potential reactions in the dialogue background, profile the character's psychological activity toward each NPC reply considering the current emotion.
4. Based on the potential reactions and hidden themes, analyze the character's feelings toward the NPC's reply.
5. Using the above analysis, express the emotion change in a positive or negative value.

Cog Cog is an integer between 1-6 representing the cognitive level the character exhibits in the conversation:

- 1: Memory — Simple recall or statement of information.
- 2: Understanding — The character interprets, summarizes, or expresses their basic understanding of information.
- 3: Application — The character applies their knowledge to specific situations or actions.
- 4: Analysis — The character identifies relationships between pieces of information, breaks down structures, or compares.
- 5: Evaluation — The character judges, questions, or evaluates other viewpoints with clear values or standards.
- 6: Creation — The character synthesizes information and proposes new insights, strategies, or creative ideas.

1138

Three-Dimensional Emotion Three-dimensional emotion is a coordinate system of [valence, activation, object focus], each ranging from -1 to 1:

- -1 indicates negative/low activation/retrospective.
- 1 indicates positive/high activation/future-focused.
- 0 indicates neutral/object-centered emotion.

Three-Dimensional Emotion Analysis Dimensions

1. Based on the character's psychological activity and feelings toward NPC responses, determine whether the emotion is positive or negative, high or low activation, and whether it's related to expected outcomes, past results, or a specific object.
2. For each NPC reply, represent the three-dimensional emotion as a coordinate [valence, activation, object focus].
3. For all NPC replies, calculate the overall three-dimensional emotion as a single coordinate [valence, activation, object focus].

Output Format

```
{
  "Content": "[NPC's intended message]",
  "TargetCompletion": "[Character's goal achieved]",
  "Activity": "[Psychological activity]",
  "Analyse": "[Character's feelings toward NPC's reply]",
  "Change": "[Character's emotion change]",
  "Cog": "[Character's cognitive level]",
  "Td_e": [
    [ "x1", "y1", "z1" ],
    [ "x2", "y2", "z2" ],
    [ "x3", "y3", "z3" ]
  ],
  "Td_e_all": [
    [ "x", "y", "z" ]
  ]
}
```

Character Persona: `simulator_role`
Current Dialogue Background: `simulator_scene`
Character's current emotion is `emotion`
Character's current cognitive level is `cog`
Current character's three-dimensional emotion coordinates after each NPC reply are `td_e`
Overall three-dimensional emotion coordinates after all NPC replies are `td_e_all`
Current Dialogue Content: `dialog_history`

1139

Figure 10: The evaluation prompts used by the Evaluation Agent.

Instructions
Dear Volunteers,
Thank you for participating in the classroom experiment of the OnlineMate system. This scale aims to collect your evaluation of the system's usability and comparative feedback with human teaching assistants.

1140

Please rate the following questions based on your real usage experience and supplement your specific thoughts in the open-ended questions.
Your answers will help us optimize the system. Thank you for your support!
Scale Explanations
1. Rating Standards:

- 1 = Very Dissatisfied / Far Worse Than Human Teaching Assistants
- 2 = Somewhat Dissatisfied / Slightly Worse Than Human Teaching Assistants
- 3 = Neutral / Equivalent to Human Teaching Assistants
- 4 = Somewhat Satisfied / Slightly Better Than Human Teaching Assistants
- 5 = Very Satisfied / Far Better Than Human Teaching Assistants

2. Please ensure all ratings are based on real usage experience. There is no need to overthink individual details—just objectively reflect your overall feelings.
3. Responses to open-ended questions will help us optimize the system more accurately. Please feel free to share your thoughts.

I. Evaluation of Response Quality

No.	Evaluation Dimension	Rating (1-5)
1	Knowledge accuracy of system responses (no errors, no misleading information)	
2	Knowledge depth of system responses (neither too superficial nor too obscure, suitable for learning needs)	
3	Clarity of explanation for knowledge points in system responses (logically coherent, easy to understand)	
4	Matching degree of system responses with your cognitive level (aligned with your learning foundation, no disconnection)	
5	Extent to which system responses help stimulate in-depth thinking	

II. Evaluation of Role Effectiveness
The system includes 4 core roles. Please evaluate each separately.

No.	Role Type	Evaluation Content	Rating (1-5)
1	AI Teaching Assistant	Accuracy of background knowledge and responses, and help with learning	
2	Idea Generator	Effectiveness of knowledge point analogies and visualization of abstract principles	
3	Detail Inquirer	Effectiveness of connecting course cases/textbooks, confirming knowledge points, and inquiring about details	
4	Summary Expert	Effectiveness and value of integrating relevant knowledge to form systematic conclusions	
5	Critical Questioner	Help with in-depth questioning and extended thinking	
6	Coordination of interactions among roles (no conflicts, no redundancy, jointly promoting learning)	Coordination of interactions among roles (no conflicts, no redundancy, jointly promoting learning)	

III. Overall Usability Evaluation

No.	Evaluation Dimension	Rating (1-5)
1	Ease of system operation (easy to use without complex learning)	
2	Naturalness of interaction with the system (close to real classroom communication, no stiffness)	
3	Extent to which the system enhances your learning interest	
4	Effect of the system in promoting your classroom participation	
5	Satisfaction with the overall user experience of the system	

6. Based on your experience, which AI tool would you prefer to use in the future to assist you in discussing and solving problems?
A. OnlineMate B. LLMs (e.g. ChatGPT, Deepseek)

1141

IV. Comparative Evaluation with Human Teaching Assistants

No.	Comparison Dimension	Rating (1-5)	Remarks (Optional)
1	Comparison of knowledge accuracy		
2	Comparison of knowledge depth adaptability (neither too superficial nor too obscure, suitable for learning needs)		
3	Responsiveness to personalized needs (e.g., adjusting responses based on your questions and learning foundation)		
4	Flexibility in the interaction process (ability to handle unexpected questions and extended discussions)		
5	Comparison of response speed and problem-solving efficiency		
6	Comparison of emotional support and learning encouragement effects		
7	Comparison of overall learning assistance effects		

V. Open-Ended Questions

1. What do you think are the most prominent advantages of the OnlineMate system?

2. In which aspects do you think the system needs improvement (e.g., response quality, role design, interaction methods, etc.)?

3. Compared with human teaching assistants, in what scenarios would you prefer to use the OnlineMate system, or would you prefer to choose a human teaching assistant? Please explain the reasons.

4. Any other evaluations or suggestions you would like to add:

Table 4: The student participants' evaluation scale.

Collaborative Knowledge Construction Rating Scale

Scale Instructions. Based on the given multi-role collaborative learning dialogue, rate the level of **collaborative knowledge construction** demonstrated by **each role**. Output structured JSON results strictly following the specified **five-point Likert scale**.

Basic Scoring Principles.

1. Ratings must be based only on **observable verbal and interactive behaviors** in the dialogue; do not infer behaviors that are not explicitly shown.
2. If there is insufficient evidence for an item, a **conservative score of 3** must be assigned.
3. All ratings must be **integers from 1 to 5**.
4. Treat all roles equally; **do not merge or omit any role**.
5. The score for each dimension is the **arithmetic mean** of all its items, rounded to **two decimal places**.

Five-Point Likert Rating Criteria.

- **1 = Strongly Disagree** (the behavior is almost never demonstrated)
- **2 = Disagree** (the behavior is rarely or occasionally demonstrated)
- **3 = Neutral** (the behavior is demonstrated to some extent but is unstable or lacks sufficient evidence)
- **4 = Agree** (the behavior is clearly demonstrated multiple times)
- **5 = Strongly Agree** (the behavior is consistently, explicitly, and typically demonstrated)

Scale Items.

Dimension	Item	Description
Expression (EX)	EX1	Actively and proactively proposes ideas and implementation plans during discussions.
Expression (EX)	EX2	Expresses viewpoints clearly and in an organized manner.
Expression (EX)	EX3	Suggestions offered can be recognized and adopted by others.
Listening (LI)	LI1	Does not interrupt others casually.
Listening (LI)	LI2	Is willing to listen to different or opposing opinions.
Listening (LI)	LI3	Understands and responds to others' viewpoints.
Coordination / Regulation (CO)	CO1	Guides the discussion back to the main topic when it deviates from the task.
Coordination / Regulation (CO)	CO2	Initiates topics and organizes discussion according to task needs.
Coordination / Regulation (CO)	CO3	Encourages low-participation members to join the discussion.
Evaluation (EV)	EV1	Evaluates others' performance fairly and impartially.
Evaluation (EV)	EV2	Provides evidence-based evaluation of learning outcomes.
Reflection (RE)	RE1	Can acknowledge and correct one's own mistakes.
Reflection (RE)	RE2	Can reflect and improve based on feedback from others.
Conflict Resolution (CR)	CR1	Mediates when members argue or disagree.
Conflict Resolution (CR)	CR2	Remains rational and open-minded when facing conflicting viewpoints.
Willingness to Collaborate (CW)	CW1	Recognizes the importance of collaboration in completing tasks.
Willingness to Collaborate (CW)	CW2	Demonstrates positive emotions during collaboration.
Willingness to Collaborate (CW)	CW3	Shows positive emotions when collaboration outcomes are acknowledged.

Figure 11: The scoring rubric utilized by observers.