



GPTalk: LLM-based virtual companions for metacognitive growth in self-regulated e-learning

In-Taek Jung^{a,1} , ChungHa Lee^{a,1} , In-Chang Baek^a , Dongik Oh^a , Youjin Choi^a ,
KyungJoong Kim^a , Duk-Jo Kong^{b,*} , Jin-Hyuk Hong^{a,*} 

^a College of Information and Computing, Department of AI Convergence, Gwangju Institute of Science and Technology, 123 Cheomdangwagi-Ro, Buk-Gu, Gwangju, 61005, the Republic of Korea

^b College of Information and Computing, Graduate School of AI Policy and Strategy, Gwangju Institute of Science and Technology, 123 Cheomdangwagi-Ro, Buk-Gu, Gwangju, 61005, the Republic of Korea

HIGHLIGHTS

- GPTalk supports metacognitive development by enabling interactive learning with LLM-based virtual agents.
- Formative interviews with students and teachers provided practical insights into the design of metacognitive support in e-learning environments.
- GPTalk offers interactive support to foster students' self-regulated learning in solitary e-learning environments.

ARTICLE INFO

Keywords:

Meta-cognition
LLM
Education
Self-regulated e-learning
Solitary e-learning

ABSTRACT

Although students need to self-monitor and manage their learning process for effective metacognition, it can be particularly challenging in solitary e-learning environments that rely on pre-recorded videos. Unlike interactive e-learning or physical classrooms, typical e-learning environments prevent students from interacting with their teachers and peers, thereby hindering metacognitive support. To address this challenge, we introduce GPTalk, a system designed to support students' learning experiences by facilitating interactions with LLM-based virtual companions. Through interviews with students and teachers, we identified design recommendations and implemented them in GPTalk. A user study involving 32 high-school students demonstrated that, compared to a baseline system, GPTalk fostered richer metacognitive engagement and self-regulated learning processes during video-based study (e.g., more monitoring questions and in-situ reflections), while short-term content understanding accuracy remained comparable across conditions. Overall, our findings suggest that students' interactions with a virtual teacher and peer can support key aspects of their metacognition and self-regulated e-learning processes.

1. Introduction

In e-learning environments, metacognition is crucial for helping learners effectively manage their learning process. Metacognition, which represents an individual's ability to plan, monitor, and evaluate their understanding and learning strategies (Flavell, 1979), is particularly important in the context of typical e-learning (Kizilcec et al., 2013), wherein learners often need to assume greater responsibility for their education (Dresel and Haugwitz, 2008) owing to the lack of direct teacher guidance and peer interaction (Fetzner, 2013). While group-based learning in traditional classrooms supports the metacognitive

development of learners through real-time feedback and dynamic interactions (Bailenson et al., 2008; Graesser et al., 1999; Messinger et al., 2019), common e-learning environments fail to provide opportunities for such rich interactions (Lee et al., 2015). Consequently, learners may struggle to monitor their progress and accordingly adjust their learning strategies, which can negatively impact their overall learning outcomes.

Metacognition encompasses two interrelated components: metacognitive knowledge, which reflects learners' awareness of their cognitive processes and learning strategies (Schraw and Moshman, 1995; Pintrich, 2002), and metacognitive skills, the practical application of

* Corresponding author.

Email addresses: dukjokong@gist.ac.kr (D.-J. Kong), jh7.hong@gist.ac.kr (J.-H. Hong).

¹ These authors contributed equally to this work.

<https://doi.org/10.1016/j.ijhcs.2026.103754>

Received 29 May 2025; Received in revised form 28 January 2026; Accepted 30 January 2026

Available online 2 February 2026

1071-5819/© 2026 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

this knowledge to regulate and execute cognitive activities (Zimmerman, 1989). Together, these elements serve as the foundation for metacognitive regulation, a high-level control mechanism that enables learners to actively manage and optimize their learning through the processes of planning, monitoring, and evaluating (Winne, 1998). At the heart of metacognitive regulation is monitoring, which plays a critical role in guiding learners to observe and assess their cognitive states and learning progress (Nelson, 1990). By facilitating ongoing evaluation, monitoring enables learners to make real-time adjustments to their understanding, strategies, and overall performance. This process is particularly vital in SRL, where the integration of cognitive, motivational, behavioral, and emotional dimensions is essential for success. The effectiveness of metacognitive monitoring is often assessed through calibration (Schraw, 2009), which measures how accurately learners evaluate their comprehension, performance, and progress. Calibration acts as a feedback mechanism, helping learners identify discrepancies between their perceived and actual understanding. Accurate calibration not only allows for efficient strategy adjustments tailored to individual needs (Thiede et al., 2003; Schraw and Dennison, 1994) but also sustains confidence, reinforces motivation, and enhances overall learning outcomes.

Recent advancements in Large Language Models (LLMs) have enabled near-human interactions and real-time responses (Bubeck et al., 2023; Kang et al., 2024; Tamoyan et al., 2025), opening new possibilities for enhancing cognitive engagement in educational domains (Gan et al., 2023; Wang et al., 2024). Various studies have demonstrated how LLMs can simulate human-like interactions to foster meaningful learning experiences by personalizing feedback (Morales-Chan et al., 2024) and promoting deeper cognitive engagement through vicarious dialogues (Choi et al., 2024; Tanprasert et al., 2023) and interactive learning tools (George and Dewan, 2024). Building on these capabilities, emerging research has begun to explore the potential of LLMs to support SRL in e-learning environments, leveraging their advanced personalization, reasoning, and real-time processing abilities (Kumar et al., 2024; Steinert et al., 2023). These studies demonstrate the considerable potential of LLMs in supporting SRL. However, considering the rapid advancements in LLM technology in recent years, research on their applications in e-learning environments is still in its early stages and requires further exploration across diverse learning environments and contexts. In this study, we propose methods to utilize LLMs for supporting metacognition in solitary e-learning environments, contributing to ongoing efforts in this emerging area and advancing understanding of how LLMs can effectively facilitate SRL.

In this paper, we specifically aim to answer the following research questions (RQs):

- **RQ1.** What are the key design recommendations for developing LLM-based systems to support metacognition in solitary e-learning environments?
- **RQ2.** How can an e-learning system enhance students' metacognitive monitoring abilities and content understanding?
- **RQ3.** How does an e-learning system support students' perceptions of metacognitive usefulness and enrich their overall learning experiences?

To answer these RQs, we conducted in-depth interviews with students and teachers to understand their interaction strategies and approaches for promoting metacognition in classroom settings. Based on the acquired insights, we established a set of design guidelines to improve the self-regulated e-learning experience. Using these guidelines for enhancing metacognitive growth, we developed GPTalk, a system that automatically generates interactions within pre-recorded lectures to emulate the diverse dynamics of offline classes. Through both quantitative and qualitative evaluations, we assessed the effectiveness of GPTalk in fostering metacognitive skills. Compared to a RAG-based Baseline system, GPTalk fostered richer metacognitive monitoring and SRL-related behaviors during the video-based study (e.g., more monitoring questions

and in-situ reflections), while short-term content understanding accuracy remained comparable across conditions. Based on these findings, we discuss future directions for developing e-learning systems that can better support learners' cognitive and metacognitive needs.

The contributions of this study are as follows:

- **Derivation of design recommendations (DRs) to promote metacognition:** Through formative interviews with ten students and eight teachers, we derived system design recommendations to enhance metacognitive support in self-regulated e-learning.
- **Development of GPTalk:** GPTalk, an LLM-based system, transforms an isolated e-learning experience into an interactive learning experience. It adapts to user interaction logs and emulates classroom-like interaction.
- **Empirical evidence of process-level impact on metacognition and SRL:** A user study with 32 students showed that, compared to a Baseline, GPTalk fostered richer metacognitive monitoring and self-regulated learning behaviors (e.g., more monitoring questions and in-situ reflections) during a video-based study.

These contributions provide a foundation for advancing e-learning systems that move beyond mere content delivery, instead offering personalized, metacognitively rich interactions that can better support learners' cognitive and metacognitive development.

2. Background

2.1. The importance of metacognition in e-learning environments

Metacognition refers to the awareness and control of one's cognitive processes—understanding what one knows, what one does not, which strategies to employ, and how to improve them (Flavell, 1979; Schraw and Moshman, 1995). It serves as a key component of Self-Regulated Learning (SRL) (Panadero, 2017; Zimmerman, 1986), a foundational framework integrating cognitive, motivational, and emotional dimensions of learning (Zimmerman, 1989). Since scholars began to distinguish between metacognition and SRL, this metacognition has guided advances in educational psychology (Zimmerman, 1986; Pintrich et al., 1993), enabling learners to actively manage goals, monitor progress, and refine strategies (Panadero, 2017). Within the SRL framework, metacognition functions as a high-level regulatory mechanism, directing learners in the effective use of cognitive strategies to enhance their overall learning process (Pintrich, 2002; Azevedo, 2009; Zimmerman, 2000).

The need for metacognitive skills becomes even more critical in solitary e-learning environments (Azevedo, 2009; Azevedo and Cromley, 2004; Dabbagh and Kitsantas, 2012; Tsai, 2009), where learners must independently navigate their cognitive processes without the immediate feedback and guidance typically available in traditional face-to-face education (Azevedo, 2009; Dabbagh and Kitsantas, 2012). While traditional face-to-face education allows instructors to immediately identify and address learners' misconceptions (Black and Wiliam, 1998; Hattie and Timperley, 2007), such instant interaction is limited online (Dabbagh, 2007; Tsai and Tsai, 2003). In e-learning environments, learners can easily access diverse materials, such as audio, text, and video, but without metacognitive and self-regulatory guidance—like strategies provided by SRL—learning efficiency may not improve (Azevedo and Cromley, 2004). As e-learning environments provide a high degree of freedom (Tsai and Tsai, 2003; Liaw and Huang, 2013), such as temporal and spatial flexibility, and increasingly enhance access to learning materials and interactivity, the importance of self-regulation becomes even more critical for fully leveraging these advantages and achieving efficient learning outcomes (Barnard et al., 2009). Consequently, effective metacognitive supports that enable learners to accurately identify their knowledge gaps and proactively refine their learning strategies are essential (Tsai, 2009).

In e-learning environments, where SRL is increasingly recognized as essential, a growing body of research focuses on delivering metacognitive and self-regulatory support. According to (Garcia et al., 2018), numerous studies have explored gamification approaches

(Anderson et al., 2014; Berkling and Thomas, 2013; Psathas et al., 2023; Vapiwala and Pandita, 2023) and adaptive scaffolding for personalization (Azevedo and Cromley, 2004; Shute, 2012) to bolster SRL. Such research aims to enhance motivation, increase learner engagement, and support the development of complex skills that are often challenging to cultivate independently. Alvarez et al. (2022) further noted that tools designed to support SRL in online learning environments are generally grounded in theoretical frameworks such as Zimmerman's (Zimmerman, 2000) or Pintrich's SRL models (Pintrich, 2004), focusing on the core stages of SRL. These tools commonly support the metacognitive, motivational, and behavioral aspects of SRL by leveraging features such as personalization, visualization, and real-time feedback. While effective in enhancing learner engagement and strategy use, they emphasize that improving SRL tools requires an integrated approach with personalization, real-world applicability, and sustained learner motivation.

2.2. Enhancing interactivity in e-learning with LLM support

Recent advancements in LLMs have achieved near-human levels of interaction, enabling systems to generate contextually aware, meaningful, and dynamic responses (Bubeck et al., 2023). These capabilities have been leveraged in various domains, including education (Lai and Tu, 2024), to simulate human-like interactions that enhance user experiences (Kang et al., 2024; Tamoyan et al., 2025).

In educational contexts, LLMs have demonstrated the potential to create responsive learning environments through their real-time processing and interactive conversational abilities (Gan et al., 2023; Wang et al., 2024). For personalized learning support, NotebookGPT (George and Dewan, 2024) assists students in solving programming problems by enabling interactions with GPT. A feedback system based on the LangChain framework (Morales-Chan et al., 2024) offers personalized feedback to students in e-learning environments through LLMs. To introduce elements of observational learning by providing vicarious dialogue, two studies focus on enhancing cognitive engagement in e-learning. VIVID (Choi et al., 2024) converts online lecture videos into scripted content to help students better grasp concepts and promote cognitive engagement through interactive learning experiences. Similarly, Scripted Vicarious Dialogue (SVD) (Tanprasert et al., 2023) enables students to engage with educational content through pre-scripted dialogues, facilitating learning even without direct interactions with instructors or peers. However, these systems primarily provide observational forms of interaction, where learners engage with scripted or pre-structured dialogues rather than actively co-constructing interactions with the agents.

Recent studies have begun to explore how LLMs can foster SRL and metacognition in education. Early work emphasized feedback and reflection: Steinert et al. (2023) showed how teacher-designed prompts can guide LLMs to deliver formative feedback, while (Kumar et al., 2024) demonstrated that LLM-supported reflective dialogue enhances self-efficacy and performance. More recent approaches extend to broader SRL processes. Ge et al. (2025) proposed a gamified LLM agent that explicitly supports planning, performance, and reflection; Li et al. (2025) used large-scale analytics to reveal how SRL strategies and scaffolding interact to shape learning outcomes; Cohen and Cohen (Cohen and Cohen, 2024) introduced a motivational chatbot that integrates cognitive, emotional, and motivational support; and (Fan et al., 2025) combined metacognitive scaffolds with LLM agents to strengthen reflective practice. Collectively, these works demonstrate promising directions for LLM-based SRL support, though many have tended to address specific phases of regulation or rely on teacher- or peer-mediated contexts. Our work builds on this foundation by focusing on the less-examined case of solitary e-learning, where learners lack immediate social interactions and require system-mediated scaffolds to regulate their learning.

While vicarious dialogues support conceptual understanding, peer and teacher agents extend this by actively engaging learners in regulating their own learning processes and reflecting on their thinking.

Designing agents as peers has been shown to promote learners' creative engagement and facilitate peer-like interactions (Liu et al., 2024). The use of virtual peers is grounded in the pedagogical rationale that sharing and refining diverse perspectives can expand learners' thinking beyond individual capabilities (Tanprasert et al., 2023; Isohätälä et al., 2017). Some studies intentionally introduced errors into peer agents to prompt students to teach or correct them, leveraging the learning-by-teaching effect (Moribe and Ushiyama, 2025). This approach supports deeper understanding and metacognitive development through the act of explaining and correcting others' mistakes. As the importance of regulation beyond the individual level has gained attention (Hadwin et al., 2011), a growing number of tools now employ LLM-based agents in the role of virtual teachers to facilitate collaborative learning in e-learning environments (Hao et al., 2025; Wang et al., 2025). These agents aim to replicate core teacher functions—delivering content, responding to questions, providing feedback, and motivating learners. Against this backdrop, our study explores the potential of positioning LLMs as metacognitive companions, with a particular focus on solitary video-based learning. Our contribution lies not in proposing a new conceptual model but in contextualizing and operationalizing established SRL principles within this setting.

3. Formative study for deriving design recommendations

3.1. Theoretical foundation of DR1–DR4

In multi-person classes, the diverse and vivid interactions among teachers and peers play a crucial role in facilitating students' metacognitive processes, including planning, monitoring, and evaluating their learning processes. However, in solitary e-learning, these dynamic interactions are often limited or altogether absent. To address this gap, we first grounded our design in Zimmerman's SRL framework and established an initial set of Design Recommendations (DR1–DR4) that map onto the forethought, performance, and self-reflection phases. We then conducted in-depth interviews with teachers and students to align these DRs with authentic instructional practices and to specify them into functional elements of an LLM-based system. The interviews also surfaced recurrent needs around sustained motivation and peer interaction; these were subsequently codified as DR5 and DR6.

Our DR1–DR4 were grounded in Zimmerman's SRL framework (Zimmerman, 2000), and the structure of each element was further informed by well-established pre-LLM SRL and scaffolding literature. Specifically, DR1 (goal setting and task analysis) draws on hypermedia prompting and scaffolding work (e.g., Bannert, 2006; Puntambekar and Hübscher, 2005) that operationalizes forethought via advance organizers, task analysis prompts, and planning supports. DR2 (self-monitoring and evaluation) and DR3 (adaptive strategy adjustment) were composed from help-seeking and metacognitive tutoring principles in intelligent tutoring systems—e.g., diagnostic feedback, Socratic checks, and fading (Roll et al., 2007)—and conversational, metacognition-oriented dialogue in tutoring systems (D'Mello and Graesser, 2012). DR4 (self-reflection and feedback) follows reflection-prompt traditions in computer-based learning (Bannert, 2006) and design exemplars that scaffold in-action and post-hoc reflection (e.g., Azevedo and Cromley (2004)) (Table 1.).

We selected these strands among many SRL approaches for four reasons. First, they offer mechanism-level prescriptions (e.g., prompts, help-seeking tutors, diagnostic feedback, fading) that are readily operationalizable as system features. Second, taken together, they provide full-cycle coverage of Zimmerman's phases (forethought–performance–reflection) with minimal redundancy. Third, they exhibit cross-domain robustness, having been replicated in hypermedia and ITS settings with college learners—an important prerequisite for transfer to solitary video-based learning. Fourth, their interaction patterns (turn-based prompts, conversational checks, adaptive feedback, reflective questioning) translate naturally to LLM-mediated interfaces. Together, these theoretical foundations allowed us to present a structured set of four recommendations

Table 1
Design recommendations (DR1–DR4) and their theoretical underpinnings.

No.	DRs	Theoretical underpinnings
1	Goal setting and planning	Supports learners in setting goals and analyzing tasks before learning; grounded in Zimmerman's forethought phase and supported by scaffolding and hypermedia prompting studies (Bannert, 2006; Puntambekar and Hübscher, 2005).
2	Self-monitoring and evaluation	Enables learners to monitor their understanding and evaluate progress during learning; aligned with Zimmerman's performance phase and supported by studies on Socratic dialogue and metacognitive tutoring (Roll et al., 2007; D'Mello and Graesser, 2012).
3	Learning strategy adjustment	Provides adaptive feedback and scaffolding to help learners adjust strategies in real time; grounded in Zimmerman's performance phase and supported by feedback-driven scaffolding and fading approaches (Roll et al., 2007; D'Mello and Graesser, 2012).
4	Self-reflection and feedback	Encourages learners to consolidate outcomes and reflect on their learning; aligned with Zimmerman's self-reflection phase and supported by reflective prompt research (Bannert, 2006) and dialogic reflection systems (Azevedo and Cromley, 2004).

(DR1–DR4) as a coherent, academically justified basis for supporting SRL in solitary video-based learning, before extending them with DR5–DR6 via interview evidence.

3.2. Formative interviews (Overview)

To concretize each DR into practically relevant system features, we examined instructional interactions between teachers and students in traditional classroom settings. The interviews aimed to capture the entire learning process, including cognitive, behavioral, and emotional dimensions. The semi-structured interviews were also designed based on Zimmerman's cyclical phases model (Zimmerman, 2000), and were conducted with teachers and students, divided into pre-class, in-class, and post-class stages. These phases reflect the SRL process, where learners set goals and plan strategies (forethought), actively monitor and adjust their actions (performance), and evaluate outcomes to inform future learning (self-reflection). This approach is consistent with prior SRL research that combines theoretical frameworks with qualitative methods to inform design decisions in applied settings (Tanprasert et al., 2023; Cohen and Cohen, 2024; Mejeh and Sarbach, 2025). The interviews examined how these phases manifest from both teacher and student perspectives, identifying SRL-related learning interactions, particularly metacognitive strategies, and concretizing design considerations for e-learning interaction models.

Specifically, we explored:

- **Pre-interview questions:** Students shared their use of e-learning platforms, their purposes for online lectures, subjects studied, and time spent on lectures per day and week.
- **Pre-class (forethought):** We explored strategies for setting learning goals and analyzing tasks, along with how these strategies could be effectively implemented in e-learning environments. For teachers, the focus was on lesson preparation strategies. For students, we explored approaches to setting goals and preparing for lessons in both e-learning and traditional classrooms.
- **In-class (performance):** We investigated how self-control, monitoring, and interaction strategies are applied during learning and their roles in supporting SRL. Teachers discussed strategies to engage students, manage interaction timing and content, and ensure effective content delivery. Students described their participation methods, peer interactions, and ways to confirm understanding in e-learning and classroom settings.
- **Post-class (self-reflection):** We examined strategies used to evaluate learning outcomes and make improvements. Teachers outlined additional learning activities after lessons. Students shared their review strategies and approaches to addressing confusion or difficulties.

All questions were adapted to teacher and student perspectives, with additional follow-up questions in a semi-structured format when necessary. For teachers, the goal was to identify effective strategies for

supporting learning and metacognitive teaching practices. For students, the focus was on understanding the learning strategies and metacognitive regulation methods they employed, comparing their experiences in e-learning and traditional classroom environments, and identifying elements that could be effectively applied to e-learning contexts. We transcribed all interviews and analyzed them using an inductive thematic analysis approach (Thomas, 2006), allowing recurring patterns and themes to emerge without predefined categories. The first authors independently reviewed the transcripts and created initial codes, and matched these initial codes with Zimmerman's three SRL phases through a round of discussion. Other co-authors reviewed these initial codes and finalized the coding scheme through group consensus. We ensured consistency through multiple rounds of collaborative discussion and iterative codebook refinement.

3.3. Students interviews

3.3.1. Design

Participants. We recruited 10 first-year high-school students (aged 17 years, with an equal gender distribution of five females and five males, referred to as S1–S10) from four local high schools. All of these students engaged in e-learning for a minimum of 2.5 hours to a maximum of 21 hours per week (M: 9.95h, SD: 6.68h), with eight students using e-learning platforms that stream pre-recorded lecture content, one student utilizing offline academy-provided videos, and another relying on short-form content from platforms like YouTube. Each student participated in a one-hour-long in-depth interview. Informed consent was obtained from their legal representatives, and the study was approved by the Institutional Review Board (IRB). The participants were awarded approximately \$15 for their involvement.

Process. The interviews were designed to compare group learning, where multiple students participate simultaneously, with self-regulated e-learning. The primary aim was to identify the key types of interactions occurring in traditional classroom settings and evaluate their impact on learning outcomes. Furthermore, we sought to explore the manner in which students implement metacognitive strategies in their learning processes and identify elements that could be effectively integrated into e-learning environments. Table 2 shows the key insights from the student interviews.

3.3.2. Findings

Challenges Encountered during E-learning. Most students ($n=8$) reported encountering common difficulties in receiving immediate feedback from e-learning settings. Unlike physical classrooms, where students can ask questions and receive prompt responses, e-learning environments—particularly those relying heavily on video-based learning—are characterized by a unidirectional flow of information. Under these circumstances, opportunities for interaction, such as asking questions or engaging with instructors and peers, during lectures are limited or nonexistent. The participants expressed frustration with

Table 2
Key insights from student interviews.

Category	Findings	Comments
Challenges in E-Learning	Difficulty in receiving immediate feedback	S3: "Acquiring feedback from e-learning tools can take very long; hence, I prefer to find the relevant information to prepare for the next lecture." S6: "In school or at academies, I can ask questions directly, but during online courses, even posting on the bulletin board can be frustrating." S8: "When I fail to understand something, I first check the bulletin board for answers. If I do not find anything useful there, I turn to internet searches to resolve the issue."
	Isolation and boredom	S4: "Video lectures often concentrate solely on the teacher, making it difficult to experience the actual class atmosphere. This one-sided approach makes video lectures less engaging compared to traditional classrooms." S6: "Given the lack of real-time feedback from online instructors, the learning experience feels prolonged and less interactive."
Benefits of Offline Learning	Preference for peer interactions	S1: "Peers who are in the same learning phase can offer valuable advice based on their own experiences and struggles, which can be helpful when I encounter similar difficulties." S4: "I feel more at ease asking friends for help because they share my perspectives. Hence, I usually turn to them first." S5: "While my friends may be less knowledgeable than my teachers, they are easier to talk to because we share similar experiences, which makes conversations more comfortable." S6: "I feel much more comfortable with friends who share my interests and can relate to my experience." S10: "Because teachers assess our performance, I am concerned that asking a silly question might make me seem unwise. My friends, on the other hand, are less likely to judge me."
Metacognitive Learning Strategies	Planning before lessons	S1: "Before class, I try to visit the teacher to briefly inquire about what we will be learning today, and I look up those concepts in advance." S9: "I review the topics covered in the previous class and identify areas requiring extra attention during today's class."
	Monitoring understanding during lessons	S1: "If I lose interest or fail to understand a teacher's metaphor, I take notes in my own words and try to connect the dots to pinpoint the areas when I am struggling." S3: "When something is unclear, I often review my thought process to gain clarity." S7: "I tend to review my notes-taking process during class to check for any unclear points."
	Adjusting learning strategies based on understanding	S1: "If I encounter unclear concepts that are closely related to upcoming topics, I would review materials that have already been taught rather than proceeding with new content." S9: "When I fail to understand something, I revisit the material or seek help from a friend to grasp the concept before proceeding."
	Evaluating knowledge after lessons	S3: "I compare my notes with those of my friends to identify any gaps in my understanding." S4: "I perform a blank-page review to compare the topics I have actually learned with those that I should have learned, identifying any missing points." S9: "Answering my friends' questions often helps me identify aspects of the topic that I have not fully grasped."

the inability to seek clarification or receive immediate feedback during lectures, which often led to decreased motivation and engagement. Furthermore, the lack of peer collaboration or opportunities to observe the learning processes of peers was cited as a source of isolation and boredom. These limitations not only delayed the immediate resolution of questions but also undermined the students' ability to engage in self-monitoring, participate in peer learning, and maintain motivation, all of which are crucial for fostering metacognitive growth.

Preference for Peer Interactions. Interestingly, many students ($n=6$) expressed a preference for interacting with their peers over teachers, despite the peers' relative lack of expertise. They found comfort in discussing topics with their classmates, who were learning the same material, which fostered a sense of camaraderie and relatability. In addition, one student noted, "I feel much more comfortable with friends who share my interests and can relate to my experiences (S6)," while another mentioned, "Peers who are in the same learning phase can offer valuable advice based on their own experiences and struggles, which can be helpful when I encounter similar difficulties (S1)". Thus, the participants viewed their peers, who were in a similar learning situation, as valuable learning partners, providing a sense of connection that extended beyond the instructor's expertise. The participants appreciated the practical advice provided by their peers who had encountered similar difficulties, often finding these insights more relatable than those provided by their teachers. This suggests that peers are not merely fellow learners but also essential contributors to the learning process, significantly supporting students' metacognitive development.

Metacognitive Learning Strategies. Although few students ($n=2$) were explicitly aware of the academic definition of metacognition, some unconsciously employed metacognitive strategies throughout their

learning process based on their experiences. For instance, while most students ($n=8$) engaged in basic preparatory activities, such as skimming through textbooks or attempting practice problems, a smaller group went beyond these methods by adopting heuristic metacognitive strategies. Specifically, they used various methods to create strategic learning plans. For example, one student noted, "Before class, I try to visit the teacher to briefly inquire about what we will be learning today, and I look up those concepts in advance (S1)". Another student mentioned, "I review the topics covered in the previous class and identify areas requiring extra attention during today's class (S9)". During lectures, the students employed strategies such as paraphrasing confusing material in their own words (S7) and looking back on previous thought processes (S3) to clarify what they did not understand. They also dynamically adapted their learning strategies based on their self-monitoring results. "If I encounter unclear concepts that are closely related to upcoming topics, I review materials that have already been taught rather than proceeding with new content (S1)". After the lectures, these students reinforced their learning by engaging with peers, answering questions (S10), and comparing notes (S3) to reflect on their knowledge from others' perspectives. These practices reflect an inherent use of metacognitive strategies by the students, even if they were not consciously aware of them.

3.4. Teachers interviews

3.4.1. Design

Participants. We recruited eight teachers (including four females and four males, designated as T1–T8, with a mean age of 31 years and a standard deviation of 5.55 years) with over four years of experience in teaching diverse subjects, including English, Social Science, and Science,

Table 3
Key insights from teacher interviews.

Category	Findings	Comments
Role of the Teacher	Building rapport and motivating students	T3: "I use real-life examples that students can relate to." T4: "I inquire about students' thoughts and personal experiences and link these to the lesson content." T6: "I think that it is important to adapt my teaching strategies to students' reactions to establish a connection, rather than simply focusing on what I personally find interesting or enjoyable." T7: "I ask students to reflect on their mood before the class begins." T8: "... As I gained more knowledge, I began teaching more complex content and often stayed up late preparing my lessons to ensure depth. Just when I felt I had become proficient at delivering in-depth material, I realized that I was the only one speaking in the classroom. This realization prompted me to shift my focus from covering more material to actively encouraging student participation and engagement during classes."
Metacognitive Teaching Strategies	Providing lesson content in advance	T5: "I prepare PDFs of the lesson content and share them with students before the class begins." T8: "I carefully craft thoughtful pre-lesson questions to encourage students to engage with the material beforehand."
	Presenting misconceptions and correction scenarios	T2: "I ask the same question to multiple students and use their responses to identify and correct any mistakes." T4: "I first introduce misconceptions and then guide students through the correction process to help them reflect on their initial thought process."
	Responding to students' questions with counter-questions	T4: "Instead of providing direct answers to a student's questions, I ask other students to respond or pose counter-questions to the original questioner, prompting them to identify areas they may have overlooked and consider alternative perspectives." T6: "I ask students to reiterate what was just explained to ensure thorough understanding."
	Supporting peer collaboration	T5: "To promote cooperative learning, I encourage students to explain concepts to peers who are struggling." T6: "I allocate time for group research on unfamiliar terms, following which students present their findings to the class." T7: "During group activities, I ensure that each team presents a key concept from the lesson to foster collaborative learning." T8: "At the end of each lesson, I pair students to discuss the topics they have learned and address any questions they may have."

to middle- and high-school students. Each interview was approximately one hour in length and conducted over Zoom. The participants were awarded approximately \$37.5 (IRB approved).

Process. We conducted these interviews to explore the types of interactions in physical classrooms, which are often underrepresented in e-learning environments, and to identify effective teaching strategies. We primarily focused on understanding the role of teachers, not just as providers of knowledge but also as facilitators of students' emotional, psychological, and cognitive development. Furthermore, we examined ways in which teachers employed metacognitive strategies to guide students and identified elements that could be effectively integrated into e-learning platforms. Table 3 shows the key insights from the teacher interviews.

3.4.2. Findings

Most teachers (n=7) adhered to a standardized lesson preparation process, following the curriculum and achievement standards set by the Ministry of Education. They designed a semester-long curriculum, which determines the amount of material to be covered during each class and incorporates digital resources and activities. While the overall preparation framework remains consistent, classroom interactions become highly diverse depending on each teacher's approach and the dynamics of their students. Leveraging their extensive experience, the teachers employed diverse instructional strategies to communicate more effectively with their students.

Building Rapport and Motivating Students. A high-school physics teacher, T8, shared an enlightening experience highlighting the critical role of educators: "I am pursuing a part-time master's degree in relativity. As I gained more knowledge, I began teaching more complex content and often stayed up late preparing my lessons to ensure depth. Just when I felt I had become proficient at delivering in-depth material, I realized that I was the only one speaking in the classroom. This realization prompted me to shift my focus

from covering more material to actively encouraging student participation and engagement during classes". This insight underscores the shift from content delivery to fostering active student engagement and understanding. Similarly, other teachers emphasized the importance of addressing students' emotional and psychological needs. For instance, T2 stated that they spent time with students outside of class to build rapport with them, while T7 noted that they asked students about their mood before classes to create a comfortable and supportive learning environment. T4 mentioned the importance of allowing students time to think and respond. Collectively, these perspectives demonstrate that educators not only serve as knowledge providers but also function as holistic learning facilitators who support students' psychological, emotional, and cognitive development.

Metacognitive Teaching Strategies. All teachers mentioned that they typically handled classes of 20 to 30 students at a time, which made it difficult to manage them effectively and offer individualized support. Despite this, they continuously employed various strategies to foster metacognitive reflection and collaborative learning among peers. These strategies include introducing deliberate errors for students to identify and correct (T4), sharing a range of student responses to the same question (T2), and organizing group activities where students summarized their learning and discussed questions with one another (T4-8). They are aimed at providing students with opportunities to articulate their understanding, reflect on their knowledge, and consider alternative perspectives, thereby fostering their metacognitive growth.

3.5. Six design recommendations (DR1-DR6)

Thus, RQ1 is addressed by deriving six design recommendations (Table 4) that apply metacognitive support in solitary e-learning. As established in the preceding subsection, DR1-DR4 were structured on Zimmerman's SRL framework and the pre-LLM literature. The interviews then served two purposes: they specified how each of these four

Table 4
Six DRs and corresponding functional elements.

No.	Design recommendations	Functional elements
1	Goal setting and planning	Provide content summaries before classes to help students set learning goals and identify focus areas
2	Self-monitoring and evaluation	Incorporate quizzes and misconception correction scenarios into lessons to promote self-assessment and knowledge evaluation
3	Learning strategy adjustment	Adjust in-class feedback based on the level of a student's query (question or answer) to help them refine their learning strategies
4	Self-reflection and feedback	Conduct reviews with a virtual peer guided by a virtual teacher to help students organize and consolidate their understanding of the material
5	Motivation and sustenance	Include elements of empathy and real-life examples to maintain students' interests and engagement
6	Peer learning	Create opportunities for virtual peer collaboration to replicate the experience of having learning companions in an e-learning environment

DRs should operate as system functions in solitary e-learning, and they also informed the derivation of two additional recommendations (DR5–DR6). Our analysis of teacher interviews indicated that teachers did not primarily perceive themselves as transmitters of knowledge, but rather as holistic facilitators who orchestrate students' planning, monitoring, strategy adjustment, and reflection. Students, in turn, consistently reported a preference for peer-like interactions—often valuing exchanges with similar others over interactions with a more knowledgeable teacher. These convergent themes underwrite DR5 (motivational/affective sustenance) and DR6 (peer learning/collaboration) as cross-cutting supports that enhance SRL in solitary settings. Accordingly, DR1–DR4 are conceptualized as metacognitive DRs that directly support learners' self-regulation, while DR5 and DR6 function as e-learning support DRs that sustain the overall learning experience.

Guided by these insights, we specified functional elements for each phase of the study. Pre-class (DR1) integrates teacher practices (e.g., advance organizers, task analysis) with students' planning routines to deliver concise previews and scaffold focused goal setting. In-class (DR2), the system supports knowledge monitoring by incorporating misconception-correction scenarios and Socratic counter-questions commonly used by teachers; in-class (DR3), it adapts strategies by diagnosing the level of a learner's query and providing tailored prompts and feedback. Post-class (DR4), it facilitates peer-oriented review (e.g., guided comparison and blank-page recall) to consolidate understanding and inform subsequent study. Two interview-derived, cross-phase supports are embedded throughout: DR5 provides motivational and encouraging expressions consistent with the teacher-as-coach role, and DR6 operationalizes peer learning—either via direct student–peer exchanges (e.g., co-explanations, question reformulations) or by enabling learners to observe and learn from dialogues between a virtual teacher and a virtual peer.

The concrete realization of these six DRs in an LLM-based system—including persona design (virtual teacher and virtual peer), prompt and feedback templates, adaptation logic, orchestration policies, and their integration within a video-based learning interface—will be detailed in Section 4.

4. GPTalk: Metacognition support tool in e-learning environments

Based on the six DRs, we developed GPTalk, an LLM-based learning support tool specifically tailored to facilitate metacognition in e-learning environments. GPTalk automatically generates various in-context interactions grounded in the metacognitive DRs (1-4), with each interaction considering the e-learning support DRs (5-6). Among these, the interactions based on DR1, 3, and 4 are specifically designed to support personalized learning experiences for each user by leveraging the user's question and answer logs. We designed GPTalk to feature two types of AI agents: a teacher agent that offers instructional guidance to support knowledge acquisition, and a peer agent that fosters collaborative learning by simulating peer interaction. This dual-agent design was directly

motivated by insights from our formative interviews (Section 3), which revealed that learners often struggle without both teacher-like guidance and peer support in e-learning environments. The peer agent takes on the role of a less knowledgeable partner (Moribe and Ushiyama, 2025) to encourage learners to share and refine diverse perspectives, ultimately expanding their thinking beyond individual capacity (Tanprasert et al., 2023; Liu et al., 2024; Isohätälä et al., 2017). To address the hallucination issue—an inherent limitation of generative AI—and to ensure that interactions are aligned with the learner's materials, we employ a vector database-based Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) approach with the support of prompt engineering techniques. Within this framework, interactions are strategically embedded within the lecture, providing each student with a unique and personalized learning experience.

The following two sections provide a detailed overview of GPTalk: Section 4.1 focuses on the student's perspective, explaining the interaction flow and user experience with the tool, while Section 4.2 shifts to the system perspective, explaining the implementation of functional components such as RAG and prompt engineering.

4.1. Main functions and user interaction

This section details the primary functions and interactions of GPTalk through the perspective of an imaginary student, Lily, who is learning economics. Please refer to Fig. 1 for the detailed user interface and the interaction examples. The learning process in GPTalk is divided into three phases, as shown in Fig. 2.

4.1.1. Pre-class interaction

Using the chat window on the lecture screen, Lily prepares for the lecture. Guided by DR1 [goal setting and planning], the virtual teacher summarizes the lecture content, highlights learning keywords (Fig. 1(A)-1), and reminds students of the keywords they planned to focus on.

Goal Setting by Providing Keywords. When the virtual teacher asks Lily about the topics she wishes to focus on, Lily provides the learning keywords related to her interests (Fig. 1(A)-2). The virtual teacher then provides a concise explanation based on these keywords (Fig. 1(A)-3). Since each student may have different areas of focus during the lecture, the virtual teacher tailors its explanations based on the learning keywords selected by each student.

Keyword Reminders for Motivation. Using the keywords selected by Lily, the virtual teacher scans through the lecture script and performs vector searching to find similarities between the content of the script and the keywords Lily is interested in. When the keywords Lily has chosen begin to be explained in the lecture, the virtual teacher and peer take turns sending reminder messages to keep her engaged and motivated with the upcoming content (DR 5 [motivation and sustenance]).

4.1.2. In-class interaction

While Lily attends the lecture, GPTalk generates interactions based on DR2 [Self-monitoring and evaluation] to support the learning process

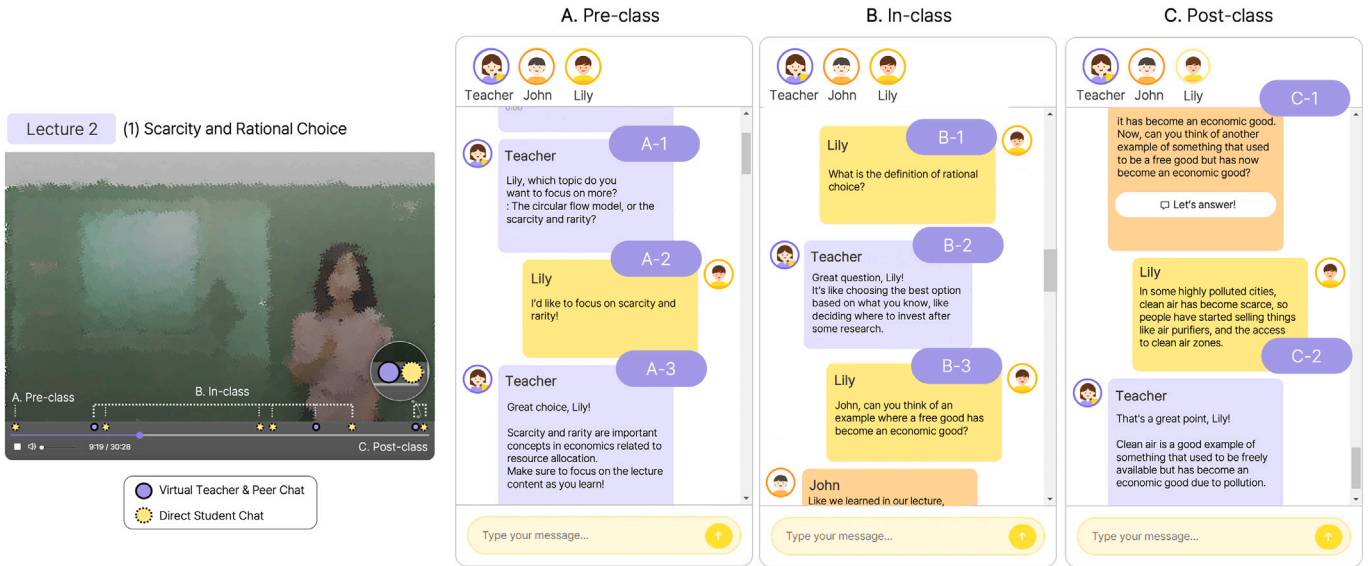


Fig. 1. User interface of GPTalk showing the lecture video player (left) and interaction examples with student Lily: (A) pre-class interaction, (B) in-class interaction, and (C) post-class interaction. In the video progress bar, purple dots with solid lines indicate chat timings with the virtual teacher and peer (John), while yellow dots with dotted lines represent direct student chats. Details of interaction generation are provided in Sections 4.1.1 (A), 4.1.2 (B), and 4.1.3 (C).

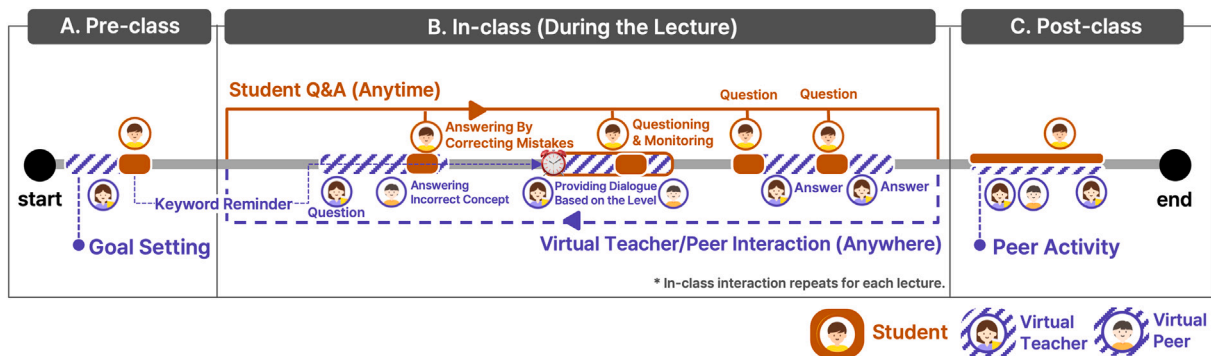


Fig. 2. Diagram illustrating the learning process involving three phases: Pre-class (goal setting), in-class (repeated for each chapter with learning activities, monitoring, QA sessions, and answer correction), and post-class (peer activity and QA sessions). User actions are indicated in orange, while virtual teacher and peer activities are indicated in a diagonal pattern with violet.

as follows: (1) a conversation in which Lily is asked to correct a mistake made by a virtual peer (low-level knowledge) and (2) a QA dialogue between the virtual peer and the virtual teacher. Additionally, she can ask GPTalk any questions that arise and receive real-time answers from the virtual teacher.

Clarifying Concepts by Correcting the Peer’s Incorrect Answers. As the ongoing chapter nears its end, the virtual teacher asks the virtual peer a question about a specific concept, and the virtual peer provides an incorrect answer. The teacher then prompts Lily to identify and correct the peer’s misconception. This exercise helps Lily assess and reinforce her understanding of the concept. During this interaction, the virtual teacher evaluates Lily’s level of comprehension and adjusts the feedback based on her responses to enhance her understanding of the material in line with DR3 [learning strategy adjustment].

Providing Dialogue Content Suitable to the Learning Level. When Lily responds to the teacher’s requests during their interactions, the virtual teacher assesses Lily’s level of understanding and provides customized feedback on her responses to deepen comprehension based on DR3 [learning strategy adjustment]. Based on Lily’s responses, the teacher offers detailed explanations and uses auxiliary teaching strategies, such as employing metaphors or leveraging Lily’s interests in line with DR5 [motivation and sustenance].

Receiving real-time responses from the virtual teacher and peer. During the lecture, Lily asks a question through the chat window on the right side of the lecture screen. She submits her inquiry (Fig. 1(B)-1), and the virtual teacher responds in real time (Fig. 1(B)-2), referencing both the current lecture material and relevant information stored in the vector database. This interaction not only provides Lily with a contextually accurate answer but also encourages deeper engagement by integrating relevant content from previous lectures and materials. Lily also freely asks the virtual peer questions about concepts she finds confusing (Fig. 1(B)-3) and receives immediate responses.

4.1.3. Post-class interaction

When the lecture concludes, the virtual teacher guides Lily to engage in paired activities with a virtual peer to review her level of understanding. The virtual peer reviews Lily’s question logs, and subsequently, following DR4 [self-reflection and feedback] and DR6 [peer learning], the virtual peer initiates activities to revisit key lecture concepts. These peer activities are specifically designed to focus on topics raised in Lily’s question logs, ensuring a personalized review of her queries.

Reflecting through Peer Activities. Once Lily completes the lecture, the virtual teacher first summarizes the entire content and subsequently prompts Lily and the virtual peer to engage in reflection activities, with

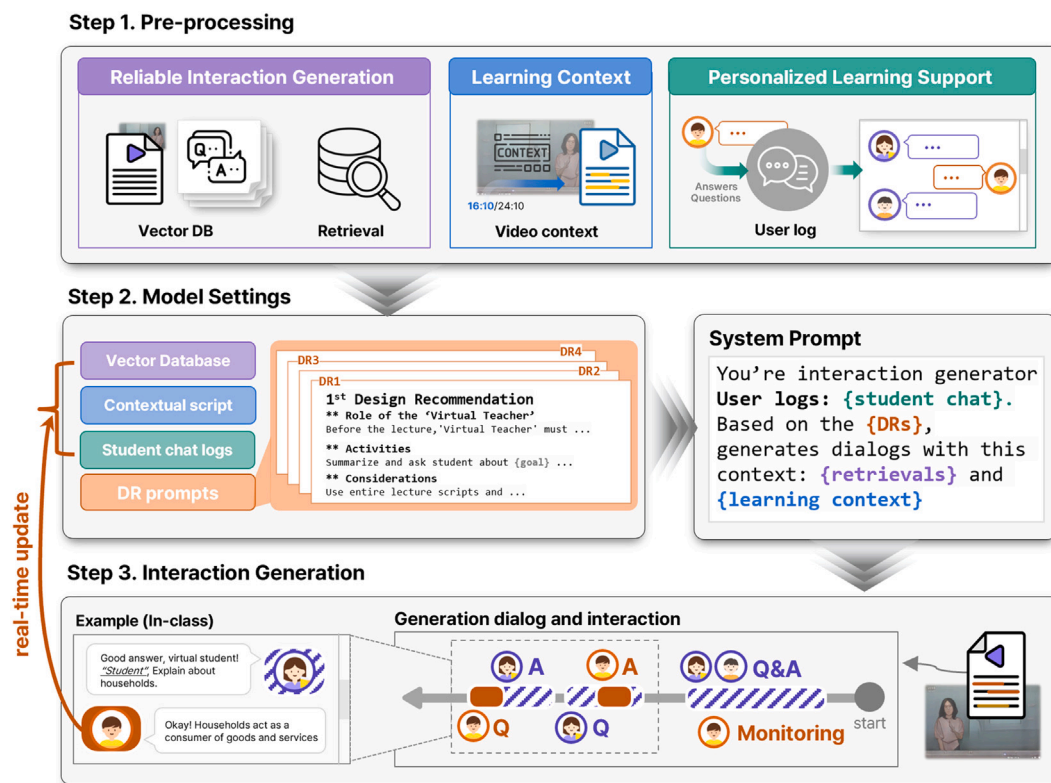


Fig. 3. Developmental workflow of GPTalk. The figure illustrates the sequence of the interaction generation process: (1) Pre-processing for RAG, considering learning context and personalization; (2) integrating the system and DR prompts into the vector database, contextual scripts, and chat logs; and (3) generating interactions.

the peer taking the lead. These peer activities are designed to help the virtual peer and Lily share their knowledge and provide feedback on each other's understanding. To identify relevant topics for these personalized peer activities, the virtual peer analyzes Lily's chat logs. The peer begins by explaining topics based on Lily's questions and then requests Lily to articulate (Fig. 1(C)-1), in her own words, how well she understands the material based on the questions she had (Fig. 1(C)-2). Finally, the virtual teacher wraps up the session by summarizing both Lily's and the peer's responses.

4.2. System design

Reliable Interaction Generation. Fig. 3 illustrates the developmental workflow of GPTalk. Providing accurate information in the educational field is essential to effectively support learners' understanding; however, this is often hindered by hallucination issues commonly associated with LLMs. To address these challenges and ensure reliable, contextual interactions, we integrated an RAG approach (Lewis et al., 2020). In this design, lecture materials and authentic QA data are transformed into a searchable knowledge base. When a learner submits a query, relevant segments and prior interactions are retrieved and combined with prompt-engineered instructions, ensuring that generated responses remain aligned with the instructional content and context.

Considerations of the Learning Context. To reinforce the connection between the student and the system, we incorporated the given lecture context and chat log into the interaction generation process. The system leverages the timestamp of the currently viewed video to extract the instructor's utterances from the preceding minute of the transcript. By integrating this context into the prompt, the system then enhances the language model's ability to accurately pinpoint and respond to the student's questions based on the relevant context, even when the student provides minimal contextual information. This strategy effectively

establishes a seamless connection between the lecture and chat interaction, reinforcing the perception that the lecturer in the video and the virtual teacher constitute a unified entity. Consequently, the coherence between the video and chat contexts is improved, enhancing the student's sense of presence within the e-learning environment.

Personalized Learning Support. The system utilizes chat logs to deliver a personalized learning experience. At the beginning of each lecture, the virtual teacher provides a summary of the content and asks students to identify specific focus areas. Based on the students' responses, GPTalk searches the lecture script to identify relevant segments and offers customized reminders to maintain student focus. Furthermore, to deliver feedback that is aligned with the students' understanding, the system uses an LLM to analyze interaction logs, categorizing comprehension levels as high or low, and adjusting feedback accordingly. After the lecture, the system conducts a personalized review through a turn-taking talk with a virtual peer, focusing on the student's previously raised questions.

Model Settings and Interaction Generation. To implement the DRs presented in Section 3.3, Sections 4.1.1–4.1.3 translate these recommendations into GPTalk's fine-grained functional elements. Subsequently, GPTalk employs a structured generation process wherein each DR is translated into specific interactions using diverse prompt templates (see Fig. 3, Step 2). These "DR prompts" are integrated into the overall "system prompt," along with three components: the vector database, contextual script, and chat log, as outlined in Step 1 of Fig. 3. Before the lecture starts, the system generates user interactions that are dynamically embedded within the lecture content. Students then engage in these interactions by providing responses and personal queries. Input logs are retrieved in real time from the vector database. This information is then used to update the contextual script and is stored in the chat log to enhance personalized learning.

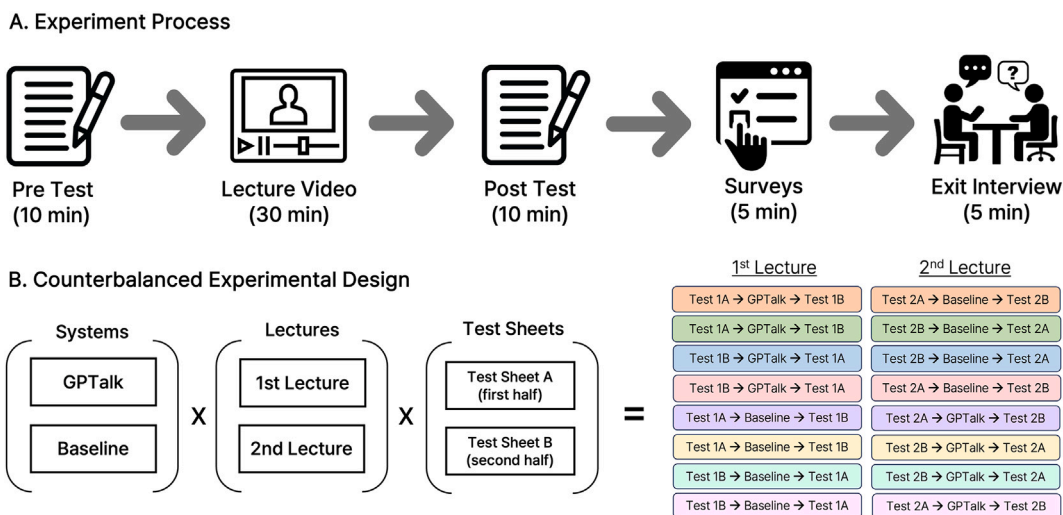


Fig. 4. The detailed information of the user study, illustrating (A) the experimental process and (B) the counterbalanced experimental design. A total of 32 students were divided into eight groups to ensure counterbalancing across experimental conditions (1A: Q1-6, 1B: Q7-12, 2A: Q13-18, 2B: Q19-24).

4.3. Implementation

GPTalk is implemented using React² and connected to a Flask³ server, enabling simultaneous web access for students. We leveraged OpenAI’s GPT-4o-2024-05-13, capable of generating responses within 232 milliseconds, aligning with the natural pace of human conversation. This choice allows the system to provide fast and contextually appropriate responses, thereby enhancing the overall user experience. To further improve reliability, we implemented a RAG pipeline using the LangChain framework and Chroma library. Lecture scripts and 909 related teacher–student Q&A pairs were collected and converted into 3072-dimensional embeddings with OpenAI’s text-embedding-3-large model to construct a vector database. When a query is submitted, the system retrieves the lecture segment corresponding to the learner’s current viewing point along with its summary, and combines these with the most relevant Q&A entries from the database. This enriched context is then passed to GPT-4o, enabling it to generate answers that remain faithful to the lecture material while mitigating hallucinations.

The temperature, a hyperparameter controlling the model’s diversity, ranges from 0 to 2, with a default of 1. In this study, the researchers empirically adjusted the temperature to 0.7 to enhance the generation of responses that are more consistent and reliable. While this adjustment potentially reduces output diversity and increases predictability, it was designed to achieve a balance between producing engaging responses and ensuring trustworthy results. A similarity threshold of 0.1 was applied to the RAG pipeline to filter retrieved context, selecting one or two of the most relevant entries to ensure accurate output generation.

5. User study design

In this section, we describe the design of a user study conducted to evaluate the effectiveness of GPTalk in addressing RQ2 and RQ3. The study was structured to examine two key aspects:

- **Objective learning outcomes:** To measure performance improvements, we designed pre- and post-tests to assess students’ metacognitive monitoring abilities and content understanding (RQ2).
- **Subjective user perceptions:** To capture students’ perceptions of metacognitive usefulness and overall e-learning experiences, we incorporated the Metacognitions Questionnaire and User Experience Questionnaire (UEQ) as evaluation tools (RQ3).

To quantitatively evaluate the effectiveness of GPTalk in enhancing metacognitive monitoring, we conducted an IRB-approved user study involving 32 first-year high-school students who had no prior economics knowledge. In line with common practices in many exam-oriented educational contexts, participants regularly engaged in solitary e-learning through pre-recorded video lectures, where learners passively viewed instructional content without opportunities to ask questions in real time, receive immediate feedback, or actively participate in class activities. Participants reported spending 2–20 hours per week ($M = 5.23$, $SD = 4.90$). Two systems were tested: GPTalk and Baseline (details in Section 5.1.) As shown in Fig. 4(A), each experimental session consisted of a pre-test, a 30-minute economics lecture, a post-test, followed by metacognition-related questionnaires and a five-minute semi-structured exit interview. Each participant completed two sessions in total, using different lecture topics for each system condition. A within-subject design was used to minimize bias, ensuring all participants interacted with both systems. Tests were counterbalanced to account for order effects.

Fig. 4(B) illustrates the counterbalanced design with 32 students divided into eight groups. Each group, with four participants, was assigned different system usage orders and test types. The tests consisted of 24 multiple-choice questions (5 options each) based on lecture content from the early, middle, and final sections. For each lecture, 12 questions were divided (Lec. 1: Q1–12, Lec. 2: Q13–24) into two sets of six (pre-, post-tests). Participants used one system for the first lecture and switched to the other for the second. For example, in the first lecture, participants completed either Q1–6 or Q7–12 as a pre-test, watched the lecture using either Baseline or GPTalk, and then answered the remaining six questions as a post-test, followed by an interview.

5.1. System settings

The primary distinction between the GPTalk and Baseline systems lies in their integration of interactive functionalities. To minimize hallucinations and ensure context relevance, both systems employ the same RAG technology, referencing a single vector database and prompt engineering. When using Baseline, students can ask questions at any time during the lecture and receive real-time answers identical to those provided by GPTalk. GPTalk enhances its functionality by incorporating interactions grounded in the six DRs outlined in Section 3.4. Alongside the virtual teacher, a virtual peer supports learners throughout the pre-, in-, and post-lecture phases, offering tailored assistance with study planning, knowledge monitoring, and personalized review. As introduced in Section 4.1, GPTalk provides a range of metacognitive interactions not

² <https://react.dev/>.

³ <https://flask.palletsprojects.com/en/3.0.x/>.

available in Baseline, enabling learners to engage in both direct and vicarious interactions with the virtual teacher and peer, thereby enriching their overall learning experience.

5.2. Metacognitive monitoring analysis

To quantitatively assess metacognitive monitoring abilities, we leverage the Knowledge Monitoring Assessment (KMA) framework, a widely adopted method in cognitive psychology and education (Schraw, 2009). Most KMA strategies follow a common approach that involves the analysis of a 2×2 contingency table based on students' knowledge estimates and their test performance (Fleming and Lau, 2014). The four elements of this table are categorized based on correctness and confidence levels: (+ +), (- -), (+ -), and (- +). For instance, (+ -) represents low confidence in a correct answer. According to Tobias and Everson (Tobias and Everson, 2009), (+ +) and (- -) serve as indicators of accurate monitoring ability, while (+ -) and (- +) function as indicators of inaccurate monitoring ability. These elements are interpreted based on their complex interrelationships, because a decrease in (+ +) could either benefit or adversely impact performance depending on whether this change corresponds to an increase in (- -) or any other element. Statistical analyses commonly focus on the relationships between these elements, utilizing metrics that quantify the alignment between confidence judgments and actual performance. Schraw et al. (2013) identified 10 representative statistical measures commonly employed in the 2×2 framework and evaluated their interpretive efficiency. Their results indicated that combining sensitivity and specificity is optimal for explaining sample variance. Sensitivity measures the ability to correctly identify accurate monitoring instances (e.g., recognizing when a learner is confident and correct), emphasizing learners' capacity to validate their knowledge. Specificity, on the other hand, evaluates the ability to detect inaccurate monitoring instances (e.g., identifying when a learner is confident but incorrect), focusing on learners' awareness of their limitations. However, no single metric can comprehensively explain all aspects of the sample variance, highlighting the need for at least two metrics to achieve a more nuanced and thorough interpretation.

To assess improvements in metacognitive monitoring with GPTalk, we applied two key metrics: the odds ratio, which relies on the relative proportions of the four components between the two systems, and the d' metric, which relates confidence levels to correctness using the difference between sensitivity and a modified specificity value. For convenience, we refer to the aforementioned four components as follows: Knowledge (K): correct answers with high confidence (+ +), Partial Knowledge (PK): correct answers with low confidence (+ -), Misinformed (MI): incorrect answers with high confidence (- +), and Uninformed (UI): incorrect answers with low confidence (- -).

5.3. Questionnaires and interviews

Our survey aimed to evaluate the participants' perceptions of GPTalk's effectiveness in supporting their metacognitive activities, and its impact on their e-learning experiences. Building on prior research (Chen et al., 2023), we utilized 10 items from the Metacognitions Questionnaire (Wells and Cartwright-Hatton, 2004), adapting them to focus on self-regulated e-learning and metacognition. The questionnaire comprises five items assessing trust and confidence in memory (Cognitive Confidence) and five items evaluating awareness and engagement with thought processes (Cognitive Self-Consciousness). Additionally, we used a user experience questionnaire (UEQ) (Laugwitz et al., 2008) to evaluate the contributions of GPTalk to students' e-learning experiences. These questionnaires were administered using a 7-point Likert scale. Given the within-subject study design and the sample size of 32 participants, we performed a paired t -test, applying the central limit theorem to analyze the questionnaire responses. The exit interview included open-ended questions to evaluate participants' experiences with both systems. The questions were grounded in DRs derived

from the formative interviews, focusing on SRL and metacognition in e-learning settings. To minimize potential bias, the systems were referred to as "System A" (Baseline) and "System B" (GPTalk) during the experiment, and participants were not informed of the specific identities of the two systems.

6. Findings

6.1. Metacognitive effect

6.1.1. Content understanding

Accuracy We asked the 32 participants to answer six multiple-choice questions (with five options each) in both the pre- and post-tests. Each participant used both systems and completed pre- and post-tests for each system, resulting in a total of 24 questions answered. For each participant, we calculated the mean accuracy across the six questions within four conditions: Pre/Base, Pre/GPTalk, Post/Base, and Post/GPTalk. This yielded 32 values per condition (one for each participant), which were then subjected to paired t -tests ($df = 31$, two-tailed) to examine differences across conditions (Post/Base vs. Post/GPTalk, Post-Pre/Base vs. Post-Pre/GPTalk) and to control for familywise error across the two planned comparisons, Bonferroni-corrected p -values were computed. As shown in Fig. 5, the post-test results presented content understanding accuracies of 55.73% ($SD = 0.270$) and 63.02% ($SD = 0.277$) when using Baseline and GPTalk systems, respectively. In the post-test, GPTalk and Baseline were not significantly different in accuracy ($t(31) = -1.365$, $p = 0.182$, Cohen's $d = 0.24$), and this remained non-significant after Bonferroni correction ($p_{bonf} = 0.364$), despite the numerical difference in means (Baseline: $M = 0.557$, $SD = 0.270$; GPTalk: $M = 0.630$, $SD = 0.277$). Compared to the pre-test, the post-test scores demonstrated improvements of 13.02pp ($SD = 0.316$) and 23.44 pp ($SD = 0.302$) for Baseline and GPTalk, respectively. The improvement from pre- to post-test did not significantly differ between GPTalk and Baseline ($t(31) = -1.438$, $p = 0.161$, Cohen's $d = 0.25$), and it remained non-significant after Bonferroni correction ($p_{bonf} = 0.321$), despite the numerical difference in gain scores.

6.1.2. Changes in metacognitive monitoring indicators

We examined the differences in the metacognitive monitoring metrics between the pre-test and post-test for each system, as shown in Fig. 6. For each of the 24 items, participants also rated their confidence in their chosen answers on a 7-point Likert scale. These confidence ratings were then binarized into high (5–7) and low (1–4) categories. Based on these ratings and answer correctness, we calculated the counts for K, PK, MI, and UI for each participant across the four conditions (Pre/Base, Pre/GPTalk, Post/Base, Post/GPTalk). We then conducted paired two-tailed t -tests ($df = 31$) to compare the two systems within each phase for each metric.

The metrics K, PK, and UI exhibited similar trends in both systems: K increased (+32.29 pp for Baseline and +39.58 pp for GPTalk), while PK (-19.27 pp and -16.15 pp) and UI (-26.04 pp and -23.35 pp) decreased. In contrast, MI increased by 13.02 pp with Baseline but remained unchanged with GPTalk. This difference in MI provides a signal that the two conditions may have shaped learners' confidence calibration in different ways. GPTalk also showed a relatively higher increase in K (+7.29 pp) and smaller decreases in PK and UI (-3.12 pp and -2.69 pp, respectively). These descriptive differences are consistent with the observed 13.02 pp increase in the MI metric for Baseline. Moreover, GPTalk showed differences in item transitions (more than 10): more items shifted from MI to K (22 vs. 11), and fewer transitioned from UI to MI (20 vs. 30). Overall, the observed patterns differed across the two systems: MI increased by 13.02 pp under Baseline but remained unchanged under GPTalk. In the item-transition analysis (counts > 10), GPTalk showed more MI→K transitions (22 vs. 11) and fewer UI→MI transitions (20 vs. 30) than Baseline. Together with the larger increase in K

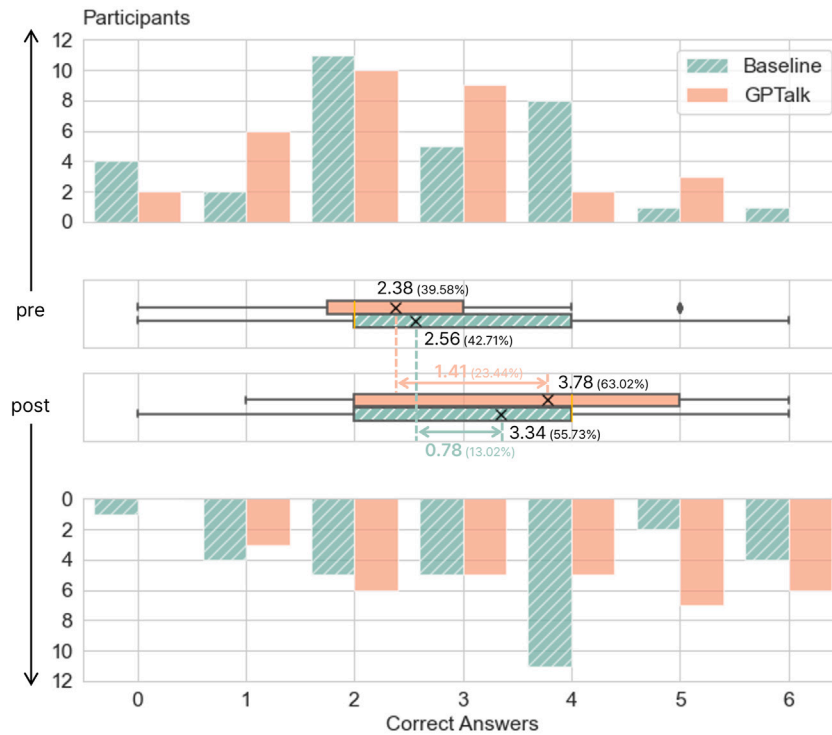


Fig. 5. The number of correct answers (accuracy rates) of the pre-/post-tests in each system. The bar graphs, positioned above and below the box plots, illustrate the distribution of correct answers (accuracy rates) for both systems.

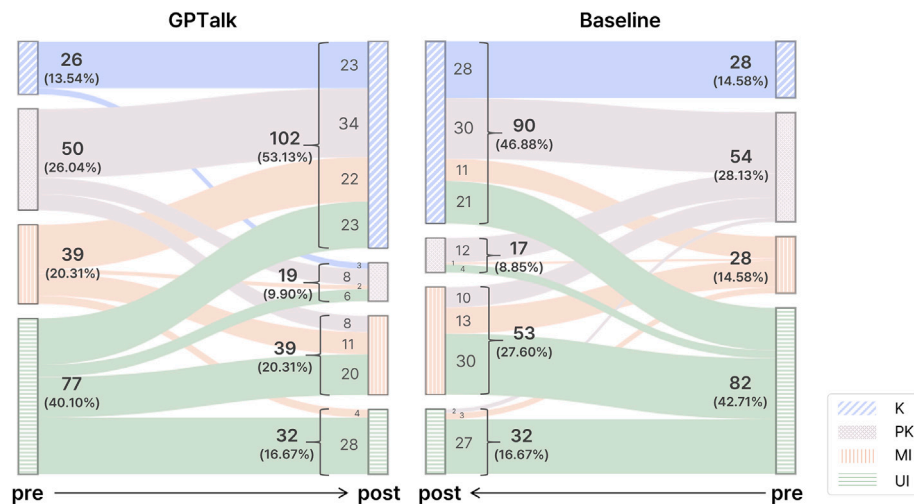


Fig. 6. Changes in metacognitive monitoring indicators across the pre- and post-tests in each system. K (Knowledge): correct answers with high confidence, PK (Partial knowledge): correct answers with low confidence, MI (Misinformed): incorrect answers with high confidence, and UI (Uninformed): incorrect answers with low confidence.

and the smaller decreases in PK and UI under GPTalk, these results summarize how confidence-accuracy categories and their transitions were distributed differently across conditions.

In interviews, S27, who achieved a high MI score, commented when using Baseline: “Before using the system, I answered the questions based on vague general knowledge; however, as I became more familiar with the content, I was able to answer the questions more easily. Although the problems were confusing, they were not too difficult.” This suggests that while the concepts became more familiar, they were not clearly internalized. Participants with higher MI scores may have overestimated their understanding, which may have led them to provide incorrect answers with high confidence. In contrast, no changes were observed in the MI

metric when using GPTalk, suggesting that the system may have supported participants’ monitoring of their knowledge relative to Baseline. The statistical analysis revealed a significant difference in the MI indicator ($t(31) = 2.352, p = 0.025, \text{Cohen’s } d = 0.42$), indicating a moderate effect size. For the other indicators— K ($t(31) = -1.191, p = 0.243, d = 0.21$), PK ($t(31) = -0.641, p = 0.526, d = 0.11$), and UI($t(31) = -0.419, p = 0.678, d = 0.07$) showed no statistically significant differences. While GPTalk did not result in an absolute decrease in MI, its relative stability compared to the 13.02 pp increase observed in Baseline suggests that it may help limit the growth of overconfident responses based on inaccurately monitored knowledge during learning progression.

Table 5

Summary of fixed and random effects for metacognitive monitoring categories of logistic mixed-effects models. For each group $c \in \{K, PK, MI, UI\}$, we fit a separate model: $isGroup_c \sim System + (1|Participant) + (1|Item)$.

Group	Effect	β	SE(β)	Wald-z	p	OR	CI _{lower}	CI _{upper}
Knowledge								
	GPTalk vs Baseline	0.32	0.23	1.392	0.164	1.377	0.878	2.160
Partial Knowledge								
	GPTalk vs Baseline	0.134	0.367	0.366	0.714	1.144	0.557	2.348
Misinformed								
	GPTalk vs Baseline	-0.437	0.249	-1.751	0.079	0.646	0.397	1.053
Uninformed								
	GPTalk vs Baseline	0	0.299	0	1.000	1.000	0.556	1.798
Random Effects								
	Participant (SD)	1.093						
	Test item (SD)	0.374						

6.1.3. Odds ratio and d'

We report the results of the odds ratio and d' metric to present a comprehensive comparison of the four indicators (K, PK, MI, and UI). The odds ratio is a widely used statistical metric for analyzing categorical data comprising discrete variables. We used the odds ratio to evaluate the relative effectiveness of each system in enhancing accurate metacognitive monitoring and reducing errors in confidence judgments. By calculating the relative ratios of the above indicators for the two systems, we can compare the overall improvement in monitoring effectiveness between the GPTalk and Baseline systems. Furthermore, by examining the d' value of each system, we can assess the relationship between correctness and confidence levels, achieving a comprehensive comparison of the specific characteristics.

Odds Ratio Calculation. We used mixed-effects logistic regression to examine how GPTalk influences metacognitive monitoring across the four response categories. This modeling approach enables a more reliable estimation of effect sizes by treating the system (GPTalk vs. Baseline) as a fixed effect and incorporating participant- and item-level variability as random intercepts. Each category was modeled as a binary outcome (e.g., $is_K = 1$ vs. 0), allowing us to estimate how the GPTalk system changes the likelihood of belonging to each monitoring group relative to the Baseline. The resulting odds ratios (ORs) provide interpretable effect-size estimates that reflect system-driven changes in metacognitive behavior while accounting for individual differences and item difficulty.

Table 5 summarizes the fixed and random effects for metacognitive monitoring categories of logistic mixed-effects models. For Knowledge (K), the estimated odds ratio for GPTalk relative to Baseline was 1.377; however, this effect was not statistically significant ($p = 0.164$, CI includes 1). For Partial Knowledge (PK), the estimated odds ratio was 1.144 ($p = 0.714$), indicating no statistically significant difference between GPTalk and Baseline. For Misinformed (MI), the estimated odds ratio under GPTalk was 0.646, but this effect was not statistically significant ($p = 0.0799$). Finally, for the Uninformed (UI) category, the odds ratio was 1.00 ($p = 1.00$), indicating no detectable difference between conditions. Taken together, these model estimates indicate that the direction and magnitude of the odds ratios varied across categories, but the differences between GPTalk and Baseline were not clearly separated in this analysis.

d' Calculation. The d' metric is calculated using the formula $(z(\frac{K}{K+PK}) - z(\frac{MI}{MI+UI}))$. This metric assesses how effectively participants connect confidence levels with correctness, a nuance not captured by accuracy metrics alone. The d' value is derived from the proportion of correct and incorrect answers provided with high confidence, offering a nuanced view of metacognitive monitoring. The d' values for Baseline and GPTalk are 0.684 and 0.883, respectively. The $z(K/(K+PK))$ values were similar across the two systems (0.999 for Baseline and 1.007 for GPTalk). The $z(MI/(MI+UI))$ values were 0.315 for Baseline and 0.124

for GPTalk. We conducted a paired-samples t -test to compare d' values between the two systems. The analysis indicated that d' did not significantly differ between GPTalk and Baseline ($t(31) = 1.61, p = 0.118$, Cohen's $d = 0.28$). The difference in mean d' values (0.883 vs. 0.684) is reported descriptively, but it was not statistically distinguishable in this comparison.

6.1.4. Interaction duration per system

We analyzed interaction log data to compare the time spent using the two systems and the number of interactions performed. Interaction time was divided into system interactions (pre-, in-, and post-class) and video viewing time, calculated by subtracting interaction time from total usage. The total learning time averaged 38.4 minutes for GPTalk and 37.4 minutes for Baseline, with GPTalk users spending 1.0 minute more. Pure lecture viewing time was 36.1 minutes (96.5%) for Baseline and 33.8 minutes (88.0%) for GPTalk, while interaction time was 4.6 minutes (12%) for GPTalk and 1.3 minutes (3.5%) for Baseline. Students asked slightly more questions in Baseline (3.35 vs. 2.90 on average), but GPTalk included four additional system-initiated interactions per lecture. Compared to Baseline, where students frequently replayed lecture segments, GPTalk encouraged more interaction, potentially enabling students to engage more effectively with the lecture content.

6.2. Subjective experience regarding metacognition and engagement

6.2.1. Metacognitive effects of GPTalk

To evaluate how GPTalk influenced students' subjective metacognitive awareness, we analyzed the Metacognitions Questionnaire (MCQ) at the subscale level. The two subscales (Cognitive Confidence and Cognitive Self-Consciousness) showed strong internal consistency, with Cronbach's alpha values of .85 and .87, respectively. Because each subscale represents a theoretically unified construct, we computed summed scores for each subscale and conducted paired-samples t -tests comparing GPTalk and Baseline. To control for family-wise error, the Bonferroni correction was applied. Both subscales showed statistically significant improvements under GPTalk after correction. Cognitive Confidence increased significantly ($t = 3.003, p_{bonf} = 0.0105$), indicating that GPTalk enhanced students' confidence in recalling and evaluating lecture content. Cognitive Self-Consciousness also increased ($t = 3.273, p_{bonf} = 0.0052$), suggesting that GPTalk helped students become more aware of and reflective about their thought processes during learning. Fig. 7 presents the summed MCQ scores across the two subscales (Cognitive Confidence and Cognitive Self-Consciousness) for GPTalk and Baseline.

Enhancing Students' Trust and Confidence through Virtual Peer Interactions. Students appreciated that observing peers questioning the virtual teacher made them rethink the concept. For instance, one student noted, "When the peer questioned about things I thought I already

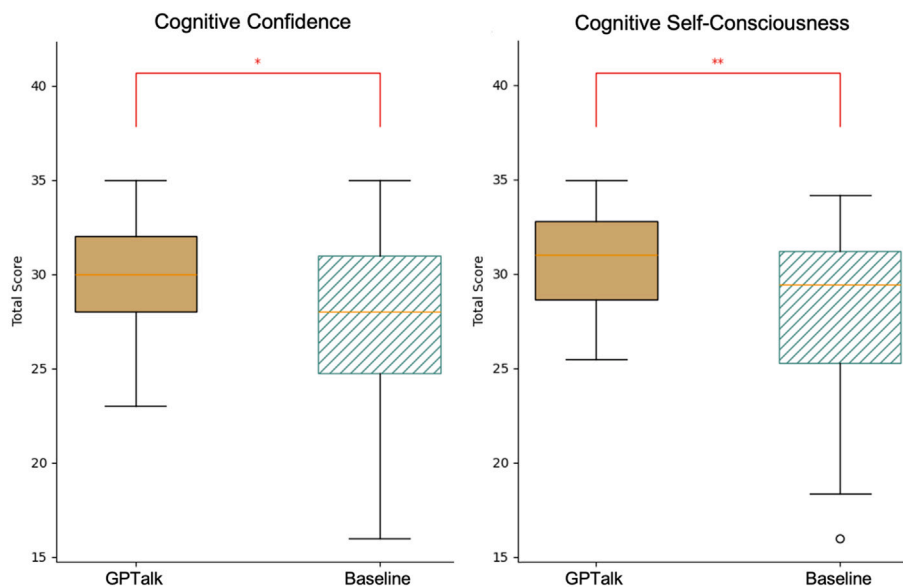


Fig. 7. Comparison of the summed MCQ score in two subscales (Cognitive confidence and cognitive self-consciousness) between GPTalk and Baseline. Statistical significance was determined using Bonferroni correction (* $p_{bonf} < 0.05$ and ** $p_{bonf} < 0.01$).

knew, it made me reconsider them, which helped me understand the concepts more clearly (P21).” When the virtual peer answered incorrectly to the teacher’s question, observing the virtual teacher correct the peer’s misconceptions also helped the students avoid being misled by their memory at times in video-based learning. This demonstrates that DR 2 [Self-monitoring and evaluation] helped students actively reflect on their understanding and evaluate their own knowledge by observing the virtual peer’s interactions and responses. For instance, one student noted, “When the virtual peer made a mistake, it made me reflect once again on the concept, in a correct way (P29).” Additionally, when the teacher asked students to correct this wrong answer, it gave them the opportunity to reconsider the relevant concept. “By correcting my peer’s incorrect answer, I had the opportunity to reconsider the relevant concept (P14).”

Empowering Continuous Reflection and Metacognition. In the interviews, students described GPTalk as more effective than the Baseline for supporting their ongoing recognition and reflection on their thought processes. This shows that DR 1 [Goal setting and planning] encouraged students to remain aware of their goals and actively engage in reflecting on their thought processes. For instance, one student noted, “The virtual teacher highlighted the importance of the keyword I chose at the beginning when it came up during the lecture, which kept me focused and engaged in my thought process (P23).” Additionally, through the peer activity, the students were able to monitor their own thinking processes. For instance, one student expressed, “Having the virtual friend explain the given concept first and then ask, ‘What did you learn today?’ was really helpful because it encouraged me to reconsider the concept (P3).” It aligns with DR 6 [Peer learning], where students first listened to the virtual peer explain a concept and then took their turn to explain it at the end of the lecture (Table A.6). This process helped students recall what they had learned during the lecture. One student noted, “The peer activity helped me revisit the lecture content and exchange answers with the virtual peer, which made it easier to retain the material for a longer time (P5).”

Moreover, GPTalk’s virtual teacher facilitated efficient and personalized learning by guiding students through a review process centered on their own questions. This can be connected with DR 3 [Learning strategy adjustment], as it encouraged students to adapt their learning strategies based on their reflections. It helped them better understand their thought processes when working through a problem in video-based learning, allowing the students to focus their reviews on areas they were unsure about. For instance, one student expressed, “I think it helped me

reflect on things I may have forgotten toward the end of the lesson and allowed me to explore topics I was more interested in (P14).”

Valuing Immediate Responses while Seeking Diverse Interactions. Some students valued the ability of Baseline to address their questions immediately, but they struggled to understand others’ perspectives on the material and to think of or articulate questions when watching the lecture video. For example, one student noted, “It was disappointing to not know my friends’ views on the topic I was studying (P25).” Although another mentioned, “While System B (GPTalk) allowed me to reflect on interactions where I asked questions or received responses, System A (Baseline) allowed me to only reflect on what I was curious about (P27).” Students reported that using Baseline allowed them to ask questions and receive feedback only on areas they believed they didn’t understand. This indicates that GPTalk provides students with diverse peer perspectives, leading to a more comprehensive and enriching learning experience compared to Baseline. These results indicate that the interactions in GPTalk are crucial for helping students monitor and reflect on areas they may have overlooked or misunderstood, thus offering more robust support for metacognitive learning.

6.2.2. Self-regulated e-learning experience with GPTalk

To examine user experience differences between GPTalk and the Baseline, we conducted paired-samples t-tests across the eight UEQ items and applied a Bonferroni correction to control for familywise error (Fig. 8). The results indicate that GPTalk achieved significantly higher scores than the Baseline in terms of interestingness, fun, originality, and novelty of experience. In contrast, GPTalk showed comparable performance to the Baseline on ease of use, efficiency, and clarity, with no statistically significant differences observed. The Baseline received a higher mean score for simplicity; however, this difference was not statistically significant.

Enriching Engagement and Enjoyment through Peer Interaction. Students reported that their e-learning experience with GPTalk was more interesting (Fig. 8(e), GPTalk: 6.15 and Baseline: 5.21, $p_{bonf} = 0.0079$) and enjoyable (Fig. 8(f), GPTalk: 6.06 and Baseline: 5.06, $p_{bonf} = 0.0074$) compared to that with Baseline. They felt more engaged and entertained when interacting with the virtual friend, both when asking questions and answering them. For instance, one student noted, “Having a virtual friend made me feel like I was studying with an actual friend (P1).” During these interactions, the students actively engaged

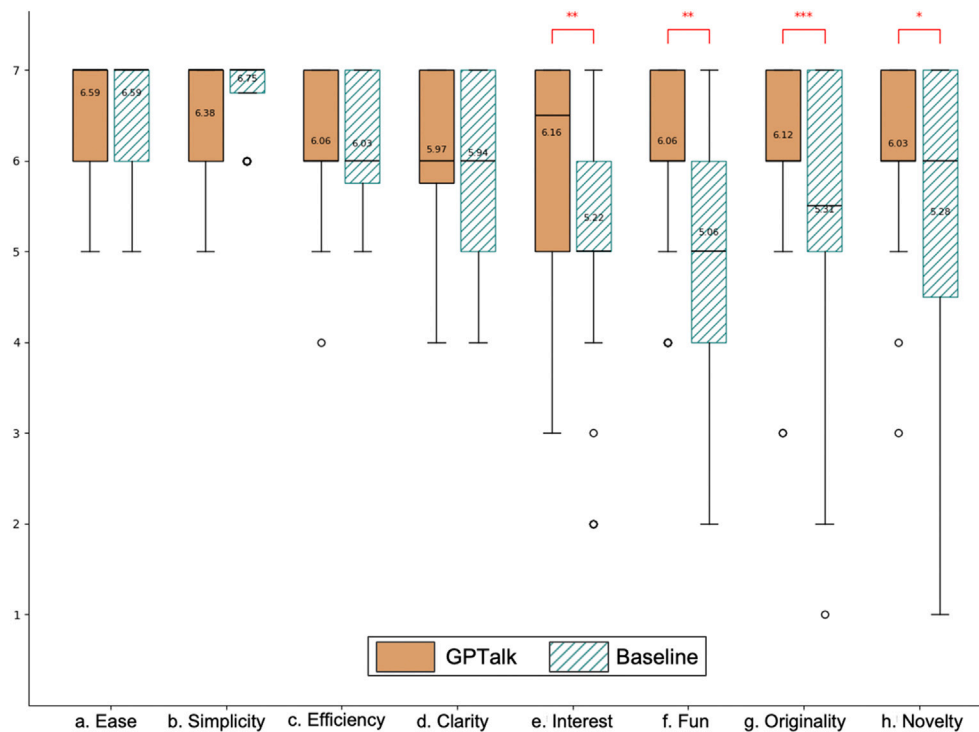


Fig. 8. Comparison of e-learning experience between GPTalk and Baseline. The number refers to the mean value and the statistical significance was determined using Bonferroni correction (* $p_{bonf} < 0.05$, ** $p_{bonf} < 0.01$, and *** $p_{bonf} < 0.001$).

with both the virtual teacher and student. For example, they responded to questions initially posed by the virtual teacher to the virtual student (Table A.8). Some students stated that because the virtual peer asked questions to the teacher throughout the lecture, it also encouraged them to ask their own questions as well. This shows that DR 4 [Self-reflection and feedback] helped students consolidate their learning process. For instance, one student noted, “If it were just the virtual teacher and me, it would be a question-and-answer session. However, having a virtual friend encouraged me to keep asking questions, and I think this interaction helped me with my studies (P24).” It demonstrates that students felt a sense of closeness and comfort knowing that a virtual peer was present in the e-learning environment, which influenced them to ask questions more freely to the teacher. This feedback suggests that GPTalk provides students with a more exciting and enjoyable learning experience, boosting their motivation and engagement compared to Baseline in an e-learning environment.

Novel Approaches in Self-regulated E-learning. Despite this, most students stated that the virtual teacher’s guidance in GPTalk added more originality to their e-learning experience (Fig. 8(g), GPTalk: 6.12 and Baseline: 5.31, $p_{bonf} = 0.0003$) and was novel (Fig. 8(h), 6.03 and 5.28, $p_{bonf} = 0.0146$). The students also felt more engaged with the topic and more curious about the subject matter. For instance, one participant stated, “When the virtual teacher asked me a question, it encouraged me to pay more attention to the lecture, which I did (P7).” Some students felt that GPTalk increased their curiosity about key concepts that they might have overlooked in their previous e-learning experiences. For instance, one student stated, “System B (GPTalk) made me feel like my virtual peer was really listening, which made me more curious about the details of the concept (P26),” while another mentioned, “I felt like my virtual peer was paying attention to the lecture, which made me more engaged and motivated to answer the virtual teacher’s questions (P3).” According to DR 5 [Motivation and sustenance], the interactions with the virtual peer and guidance from the virtual teacher motivated students to engage more deeply with the learning experience, leading them to perceive GPTalk as more inventive and cutting-edge.

Balancing between Simplicity and Interactivity. Students perceived Baseline as simpler compared to GPTalk, but no significant change was found (Fig. 8(b), GPTalk: 6.37 and Baseline: 6.75, $p_{bonf} = 0.0947$). As mentioned in the earlier finding, while the engaging nature of GPTalk positively impacted students’ e-learning experience, some students preferred Baseline because it had a simpler question-and-answer format. For example, one student noted, “System B (GPTalk) was fun, but it made me focus more on the chat than the lecture. Hence, I think that System A (Baseline) was better because I could ask questions and focus on what I was unsure about (P8).” The example from P9 about the conversations between the student and the virtual peer (Table A.10) also shows that high interactivity and freedom in GPTalk could make them talk about subjects that are unrelated to the lecture. Some students ($n = 3$) felt that the constant questioning by the virtual peer in GPTalk interrupted the lecture flow, thereby hindering their e-learning experience. For instance, one student mentioned, “In System B (GPTalk), the constant questioning by the virtual friend interrupted the flow of the online lecture and interfered with my learning (P10).” This shows that while interactivity can enhance engagement for the students, excessive interactions may disrupt the learning experience. These results demonstrate that GPTalk was perceived as a more engaging, enjoyable, and inventive e-learning experience than Baseline, while balancing interactivity and simplicity remains an open challenge for optimizing self-regulated e-learning environments.

7. Discussion and future works

7.1. Reflections on LLM-based metacognitive support in solitary e-learning

In summary, our findings indicate that GPTalk currently acts as a selective SRL support, strengthening certain regulatory processes without yet yielding reliable gains in short-term content understanding performance. GPTalk increased SRL-related behaviors (e.g., attention to the lecture, question generation, brief in-situ reflection). Still, these process-level changes did not translate into statistically significant differences in test accuracy compared to the Baseline, suggesting that the system mainly targets motivational and engagement-oriented

components of SRL rather than directly scaffolding task-specific content understanding strategies. This pattern is in line with SRL theory, which conceptualizes SRL as a multidimensional construct — cognitive, metacognitive, behavioral, motivational, and emotional/affective — where achievement emerges from multiple components working together (Panadero, 2017), and with meta-analytic evidence that SRL activity only partially mediates the effects of SRL interventions on achievement (Jansen et al., 2019).

Rather than interpreting this as a failure of SRL support, we view it as an early LLM-based example of process–outcome decoupling, underscoring the need to move from supporting isolated SRL moments within a single lesson toward phase-spanning, long-term support across forethought, performance, and reflection. Recent umbrella and meta-analytic reviews of computer-assisted SRL similarly stress that the relationship between SRL skill development, technology-supported environments, and learning outcomes is “complex and multi-directional,” and that more robust achievement gains tend to arise when diverse SRL supports are integrated across the full SRL cycle and over extended periods (Prasse et al., 2024; Guntur and Purnomo, 2024). Building on this, future iterations of GPTalk should provide richer forethought support (e.g., goal setting and planning), more explicit scaffolding of cognitive and metacognitive strategies during content understanding, and post-lesson reflection that feeds into subsequent sessions, enabling us to test whether the observed SRL gains can accumulate into durable improvements in academic achievement.

7.2. Enhancing personalization through an accumulated log-based vector database

GPTalk supports personalized learning by leveraging students’ interaction logs (questions and answers) and DR-based prompts. When a student plans and inputs a focus area, the virtual peer and teacher direct attention to that topic. Furthermore, GPTalk enables personalized peer activities at the end of each lesson, providing students with the opportunity to revisit their questions. This approach is favored by the participants of this study, as the learners appreciate the emphasis on their knowledge gaps and the personalized experience. Despite these benefits, GPTalk can be improved further. Currently, GPTalk utilizes a general vector database for content referencing and RAG (Veletsianos, 2009); however, future improvements could focus on accumulating and storing detailed interaction logs to develop a hyper-personalized vector database for each learner. This advanced RAG strategy (Kuhail et al., 2023) would enable GPTalk to more accurately track and analyze individual learning patterns, offering highly customized feedback and support to optimize the personalization process.

However, implementing such a system requires careful consideration of ethical issues, including data privacy and algorithmic bias. Future developments should focus on advancing these personalization techniques while maintaining transparency and granting users control over their data. This approach would significantly bolster the ability of GPTalk to foster personalized learning and metacognitive development, delivering more effective and impactful educational experiences tailored to each learner’s unique requirements.

7.3. Designing personas of the virtual peer

The current version of GPTalk includes single personas for both the virtual teacher and the virtual peer. However, during interviews, students ($n=10$) expressed a preference for interactions with a wider variety of personas, particularly for the virtual peer. While the single-persona model for each virtual teacher and peer has proven effective in metacognitive e-learning experiences (Section 6), this feedback suggests that having multiple personas for virtual peers has the potential to comprehensively address the diverse needs and preferences of individual students. It also indicates that students desired multiple personas for virtual peers due to the sense of closeness and intimacy that comes from

the shared identity of being peers (Qin et al., 2020; de Las Heras et al., 2020).

Existing research in educational contexts and on the design of learning-focused chatbots highlights the value of diverse personas. For instance, Veletsianos (Veletsianos, 2009) underscored the importance of designing personas tailored to specific learner characteristics, such as their learning style, motivation, and cognitive ability. Similarly, Kuhail et al. (2023) demonstrated that students prefer interacting with bots exhibiting a range of personalities and emotional expressions, which can enhance engagement and satisfaction. Furthermore, Hew et al. (2023) suggested that students can be supported in setting goals and form stronger connections with virtual agents in online activities when the virtual agent offers personalized feedback and shares similarities with them. These studies support the idea that creating multiple, distinct virtual peers could be beneficial for GPTalk.

While implementing multiple personas to reflect the diverse preferences of students can be challenging (Haugeland et al., 2022), it is important to carefully design various virtual personas for SRL to reflect the diverse needs of the students. In our study, although the virtual peer was designed to focus on learning-related content, it occasionally responded to ambiguous conversations that were not directly related to learning without offering guidance (Table A.9). Therefore, when designing the virtual peer persona, it is crucial to ensure that it can steer conversations back to the learning content whenever the student’s dialogue with the peer drifts away from learning.

Future works of GPTalk could include diverse individualized personas that reflect the unique traits and preferences of each student. Such a design would not only cater to a broader spectrum of user needs but also enhance the educational value of the platform by offering more personalized and relatable interactive experiences. By integrating these insights, future research and developmental efforts could create a more inclusive and effective learning environment, wherein each student can interact with a virtual peer that better aligns with their individual characteristics and learning needs.

Furthermore, we found that the role of the virtual peer agent underscores important contributions to socially shared regulated learning. Previous research has emphasized that collaboration with peers fosters articulation, negotiation of strategies, and critical reflection (Isohätälä et al., 2017; Hadwin et al., 2011). While our study was grounded in virtual peer collaboration, it aligns with prior work on online vicarious learning (Tanprasert et al., 2023) and LLM-based socially shared regulated learning (Edwards et al., 2025) in viewing socially shared regulation as a process that can emerge through the observation of shared interactions. Our findings suggest that observing interactions between a virtual teacher and a virtual peer can also reproduce these benefits, prompting learners to reconsider concepts and engage in collaborative sense-making. This indicates that socially shared regulatory processes can be meaningfully approximated through LLM-based agents, thereby extending classroom-based theories (Panadero, 2017) of SRL into solitary learning contexts. Together, our insights demonstrate that LLM-driven metacognitive support not only aligns with established theories but also expands their applicability to settings where immediate social interaction is limited. In doing so, our study provides empirical evidence that solitary e-learning can be enriched with deeper forms of metacognitive regulation when mediated by virtual companions.

7.4. Connecting virtual and real interactions with educational agents

Based on the DRs we derived from the formative interviews, we also suggest ways to apply our DRs in the practical implications of educational settings with virtual agents. Since previous studies have indicated that virtual agents and real peer interactions can complement each other (Du Boulay et al., 2010; Ferschke et al., 2015), it is essential to consider the practical implications of integrating educational agent systems like GPTalk not only into e-learning environments but also into blended learning contexts (Martha et al., 2019; Shukla et al., 2024).

Thus, future research may focus on interactions among real students, actual peers, and virtual agents. Our findings—where students effectively communicated and learned with a virtual peer during peer learning activities—suggest new opportunities for group interactions. Aligning with DR 1 [Goal setting and planning] and DR 2 [Self-monitoring and evaluation], we propose that virtual agents should not merely act as tools for goal setting and monitoring but also be designed to encourage students to participate more actively and critically in decision-making processes. For instance, instead of simply listening to a student's response, the agent could ask reflective follow-up questions such as, “*Why do you think it is important to study this?*” or “*What outcomes might you expect if you approached this differently?*” These questions could be posed to the target student and their peers, enhancing the quality of self-regulation and self-monitoring within the context of socially shared regulated learning (Isohäätä et al., 2017).

This process also ties into DR 3 [Learning strategy adjustment] and DR 4 [Self-reflection and feedback]. While our study demonstrated how virtual agents could guide learning strategy adjustments based on an individual student's responses, this can be extended to activities that involve solving problems collaboratively, such as debates or discussions (Rivadeneira and Inga, 2023). The virtual agent could mediate in these scenarios to facilitate student interaction. For example, when the agent detects differences in opinions among students, it could highlight these differences and steer the discussion with prompts like, “*How does this approach compare to the alternative?*” or “*Which parts of your peer's opinion do you disagree with, and why?*” The agent can help students critically evaluate and adjust their strategies by doing so, ultimately enhancing their self-reflective experience (Moreno and Mayer, 2005). With DR 6 [Peer learning], virtual agents can be extended to support interactive teaching models tailored for class-wide student engagement by broadening the concept of “groups” (Christopher et al., 2003). For instance, a virtual peer agent could be pre-programmed to participate in providing feedback on student work during a gallery-walk session. Expanding such interactions with virtual agents aligns with DR 5 [Motivation and sustenance], as it has the potential to increase student motivation (Ferschke et al., 2015) and enhance self-efficacy (Kim and Wei, 2006). For future work, clarifying the effects of each DR on fostering metacognition will be necessary, as this can provide a more detailed understanding of their individual contributions and guide targeted interventions.

7.5. Limitations

Our study has several limitations. First, we selected first-year high school students as our target group, focusing on students who engage most in e-learning. It is important to apply GPTalk to students across different age groups to determine whether varying insights emerge based on grade level or educational stage. Additionally, as we aimed to verify the effect of GPTalk on students' metacognition and e-learning, we experimented using a single subject to avoid confounding variables that could arise from differences in subject matter. Hence, future research will need to explore the impact of GPTalk across a wider range of subjects and learning contexts to ensure the generalizability of the findings. Lastly, while GPTalk's high level of conversational freedom allowed students to engage in various discussions, it also led to off-topic conversations unrelated to the lecture (e.g., one student engaged in unrelated discussions with the virtual peer). This suggests that it may be necessary to impose some limitations on the freedom of student questions within the system to maintain students' focus on the lecture. Therefore, future research will need to incorporate filters to limit the freedom of students' conversations or questions, ensuring that interactions remain within the scope of learning.

8. Conclusion

In this paper, we introduced GPTalk, an LLM-based system designed to foster metacognitive growth in e-learning environments. Grounded in formative studies with students and teachers, GPTalk

provided preliminary indications of enhanced metacognitive engagement and self-regulated learning processes, particularly in how students monitored their understanding, articulated uncertainties, and briefly reflected during video-based study. By integrating LLMs to emulate classroom dynamics and provide personalized, real-time feedback, GPTalk suggests a possible way to partially bridge the gap between solitary e-learning and traditional classroom interactions by enriching the regulatory processes that accompany video learning. At the same time, GPTalk in its current form primarily supports a subset of SRL—such as attention, motivation, and in-lesson reflection—within a single session, and there remains unexplored design space for providing more comprehensive, longitudinal support across the full SRL cycle that may be necessary to yield robust gains in content understanding and performance. We see this work as an initial step toward LLM-based companions that more broadly and consistently scaffold SRL over extended learning trajectories, with the ultimate goal of helping students translate metacognitive and self-regulatory gains into durable improvements in academic achievement. Taken together, our findings suggest that interactions with virtual teachers and peer companions can help create richer self-regulated learning experiences in solitary e-learning settings, while also highlighting the need for more comprehensive and long-term SRL-focused designs in future systems.

CRedit authorship contribution statement

In-Taek Jung: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **ChungHa Lee:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **In-Chang Baek:** Visualization, Software, Formal analysis, Data curation. **Dongik Oh:** Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Youjin Choi:** Writing – review & editing, Data curation, Conceptualization. **KyungJoong Kim:** Writing – review & editing, Project administration, Methodology. **Duk-Jo Kong:** Writing – review & editing, Validation, Project administration. **Jin-Hyuk Hong:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Jin-Hyuk Hong reports that administrative support, equipment or supplies were provided by the Gwangju Institute of Science and Technology. In-Taek Jung reports that financial support was provided by the National Research Foundation of Korea, the Ministry of Science and ICT of Korea, and the Institute of Information & Communications Technology Planning & Evaluation of Korea. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was supported by National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (Grant No. RS-2025-16902996) (60%), by the Project for Science and Technology Opens the Future of the Region program through the INNOPOLIS FOUNDATION funded by the Ministry of Science and ICT (Project No. 2022-DD-UP-0312) (30%), and by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (Grant No. 2019-0-01842, Artificial Intelligence Graduate School Program (GIST)) (10%).

Appendix A. Chat logs in GPTalk and Baseline

See Fig. A.9 for the chat logs of GPTalk, and Fig. A.10 for those of the Baseline.

A.1. Class chat log (P3)

In GPTalk, P3 actively engages in peer-to-peer discussions, with the virtual peer and teacher encouraging critical thinking and a

deeper understanding of economic concepts (Table A.6). In contrast, in the Baseline condition, P3 asks a direct question and receives a factual response from the teacher (Table A.7). This demonstrates how GPTalk promotes deeper engagement and reflection, whereas Baseline provides a more traditional, one-way learning experience.

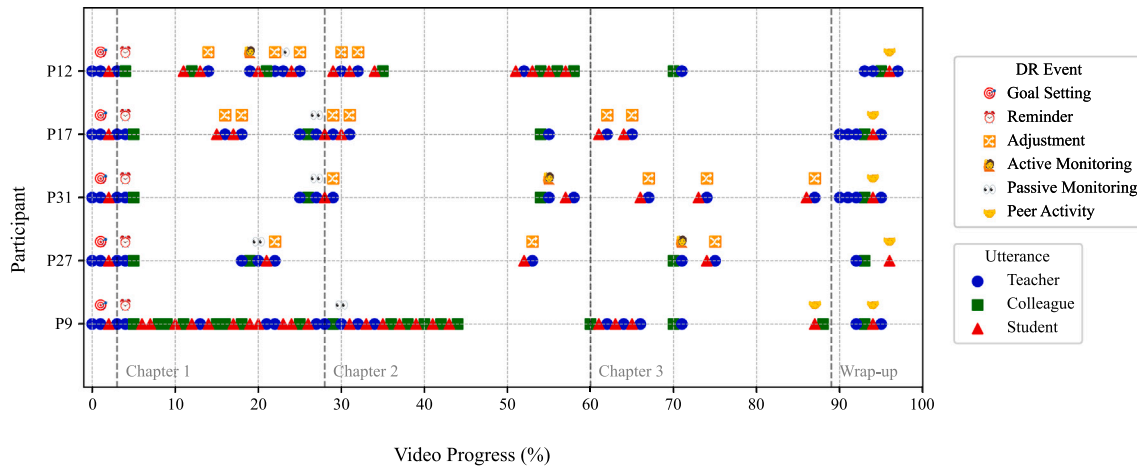


Fig. A.9. Five examples of chat logs in GPTalk. The graph's X-axis represents the lecture progress, while the Y-axis represents the five students who participated in the user study. During each DR event, students interacted with their virtual teacher or peer, with markers indicating when a conversation occurred.

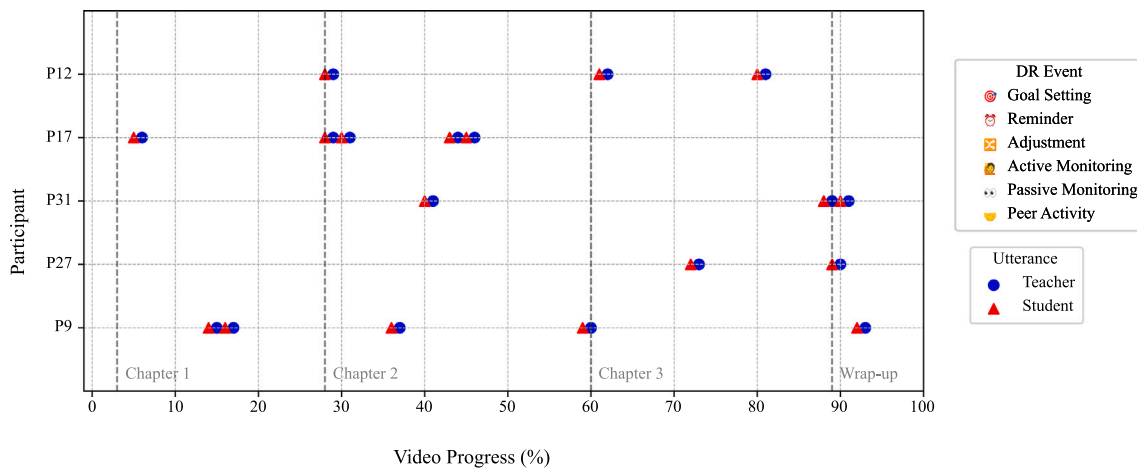


Fig. A.10. Five examples of chat logs in Baseline. The graph's X-axis represents the lecture progress, while the Y-axis represents the five students who participated in the user study. The conversation was started by the student with the requirement of a minimum number of questions. The teacher's responses were generated without DR-guided prompts.

Table A.6
Chat transcript in **GPTalk**: peer activity interactions (P3).

Utterer	Content
	(16 previous messages)
Teacher	“Today, we learned about the basic concepts of economic activities. Economic activities are divided into production, distribution, and consumption, with the main participants being households, businesses, and the government. Households are the primary participants in consumption, businesses focus on production, and the government works to maximize social welfare. Now, let’s do a peer activity to review what we learned today with your partner!”
Peer	“[Student’s Name], let me go first! You asked about consumption, right? Yeah, it’s something we do to feel satisfied. But if you’re forced into it, even though consumption technically happens, there’s no satisfaction. Like, if someone made you eat food you didn’t want, that’s not something you chose, so you wouldn’t feel good about it. Now, can you tell me what you learned in this lecture?”
Student	“When a company buys stuff to make products, that’s not consumption—it’s production. Like you mentioned earlier, if something starts off being consumed by an individual but later gets used for business purposes, it shifts from being consumption to production.”
Teacher	“Exactly, [Student’s Name]! If something an individual first consumes gets used later for company activities, it turns into production. So, depending on who’s using it and why, the same action can count as consumption or production. You totally got it!”

Table A.7
Chat transcript in **Baseline**: question and answer interactions (P3).

Utterer	Content
Student	“Can rarity and scarcity appear simultaneously for the same resource?”
Teacher	“Yes, rarity and scarcity can occur simultaneously for the same resource. Rarity refers to a situation where the absolute quantity of a resource is limited, while scarcity arises when the amount of the resource is insufficient to meet human desires. Therefore, if a resource is both absolutely limited and falls short of fulfilling human needs, it can possess both rarity and scarcity. For example, natural monuments are rare because of their limited quantity, and they can also be scarce if many people desire them.”

A.2. Class chat log (P17)

In GPTalk, P17 engages collaboratively by answering a question on behalf of a virtual peer, actively participating in discussions with teacher support (Table A.8). In contrast, the Baseline condition involves direct question-and-answer interactions, focusing on straightforward explanations (Table A.9). This example suggests that GPTalk enhances students’ thinking by providing examples to support their understanding.

Table A.8
Chat transcript in **GPTalk**: answering a question on behalf of a virtual peer (P17).

Utterer	Content
	(12 previous messages)
Teacher	“[Peer’s Name], referring to what we just learned, what is an example of a service provided in production activities?”
Peer	“Hmm... Is selling goods in a shop a service?”
Teacher	“[Peer’s Name], it might be a bit confusing. Selling goods in a shop is a consumption activity. Examples of services include medical services and educational services. [Student’s Name], what are some other examples of services in production activities?”
Student	“I think it is transportation services.”
Teacher	“That’s correct, [Student’s Name]! Transportation services are a good example of production activities. Various services we encounter daily can all be part of production activities. You’re doing great!”
Student	“Why is selling goods in a shop considered a consumption activity?”
Teacher	“Good question, [Student’s Name]. Selling goods in a shop is considered a consumption activity because consumption refers to the act of purchasing goods for satisfaction. Essentially, a consumer buys an item to use or consume it for satisfaction, which is why the activity of selling goods is a consumption from the consumer’s perspective. While sales are a part of the firm’s production, from a consumer’s viewpoint, it’s considered a consumption.”

Table A.9
Chat transcript in **Baseline**: question and answer interactions (P17).

Utterer	Content
Student	“What is a tangible asset?”
Teacher	“A tangible asset refers to an object that actually exists and has value. Examples include goods, real estate, gold, raw materials, machinery, daily necessities, and grains.”
Student	“What are some things that have rarity?”
Teacher	“Things that have rarity typically include goods, real estate, gold, raw materials, machinery, daily necessities, and grains. For example, the unique masterpiece by the artist can be considered absolutely rare. Rarity refers to the state where the absolute quantity of a resource is limited, which can apply to such artworks or natural monuments.”

A.3. Class chat log (P9)

In GPTalk, P9 engages in off-topic peer discussions, fostering collaboration and mutual support (Table A.10). In contrast, the Baseline condition involves direct teacher–student question-and-answer exchanges (Table A.11). This example suggests that the absence of virtual peers influences the participant’s attitude toward the class.

Table A.10

Chat transcript in GPTalk: discussing an off-topic subject (P9).

Utterer	Content
	(36 previous messages)
Teacher	“That’s right, [Student’s Name]! The flow of goods and money is a key concept in the circular flow, so keep studying with attention. Next, we’ll dive deeper into scarcity and rational choice.”
Student	“[Peer’s Name], are you in the department of science or humanities?”
Peer	“I’m in humanities! How about you?”
Student	“I’m in science, of course, having swag!”
Peer	“Oh, that’s cool! Science must be exciting! Although we’re in different fields, let’s study hard together! What’s your favorite subject?”
Student	“Math.”
Peer	“So, you like math! It’s a fascinating subject. I enjoy studying history. Even though we love different subjects, we can help each other by studying together. Let’s support each other in math and history!”
Student	“Yes, I really struggle with history.”
Peer	“It’s fine, [Student’s Name]! I’ll help you. While studying history, feel free to ask me anything you don’t understand. I’ll also seek your help when I struggle with math!”
Student	“[Peer’s Name], can you explain what a free good is?”
Peer	“A free good is a resource that isn’t scarce, thus it can be used without payment. Like a sunlight.”

Table A.11

Chat transcript in Baseline: question and answer interactions (P9).

Utterer	Content
Student	“What is rent?”
Teacher	“Rent refers to the money received in exchange for providing land.”
Student	“Can you give examples of groups that belong to a household?”
Teacher	“Groups that belong to a household typically refer to individuals and families. For example, think of what we usually call a household. Each household is a representative example of a group that belongs to a household.”

Data availability

The data that has been used is confidential.

References

- Alvarez, R.P., Jivet, I., Pérez-Sanagustin, M., Scheffel, M., Verbert, K., 2022. Tools designed to support self-regulated learning in online learning environments: a systematic review. *IEEE Trans. Learn. Technol.* 15 (4), 508–522. <https://doi.org/10.1109/TLT.2022.3193271>
- Anderson, A., Huttenlocher, D., Kleinberg, J., Leskovec, J., 2014. Engaging with massive online courses. In: Proceedings of the 23rd international conference on World wide web, pp. 687–698. <https://doi.org/10.1145/2566486.2568042>
- Azevedo, R., Cromley, J.G., 2004. Does training on self-regulated learning facilitate students’ learning with hypermedia? *J. Educ. Psychol.* 96 (3), 523–535. <https://doi.org/10.1037/0022-0663.96.3.523>
- Azevedo, R., 2009. Theoretical, conceptual, methodological, and instructional issues in research on metacognition and self-regulated learning: a discussion. *Metacogn. Learn.* 4, 87–95. <https://doi.org/10.1007/s11409-009-9035-7>
- Bailenson, J.N., Yee, N., Blascovich, J., Beall, A.C., Lundblad, N., Jin, M., 2008. The use of immersive virtual reality in the learning sciences: digital transformations of teachers, students, and social context. *J. Learn. Sci.* 17 (1), 102–141. <https://doi.org/10.1080/10508400701793141>
- Bannert, M., 2006. Effects of reflection prompts when learning with hypermedia. *J. Educ. Comput. Res.* 35 (4), 359–375. <https://doi.org/10.2190/94V6-R58H-3367-G388>
- Barnard, L., Lan, Y., To, Y.M., Paton, V.O., Lai, S.-L., 2009. Measuring self-regulation in online and blended learning environments. *Int. High. Educ.* 12 (1), 1–6. <https://doi.org/10.1016/j.iheduc.2008.10.005>
- Berkling, K., Thomas, C., 2013. Gamification of a software engineering course and a detailed analysis of the factors that lead to its failure. In: 2013 international conference on interactive collaborative learning (ICL). IEEE, pp. 525–530. <https://doi.org/10.1109/ICL.2013.6644642>
- Black, P., William, D., 1998. Assessment and classroom learning. *Assess. Educ.: Princ. Policy Pract.* 5 (1), 7–74. <https://doi.org/10.1080/0969595980050102>
- Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y.T., Li, Y., Lundberg, S., et al., Sparks of artificial general intelligence: early experiments with GPT-4, arXiv preprint [arXiv:2303.12712](https://arxiv.org/abs/2303.12712), 2023.
- Chen, S., Situ, J., Cheng, H., Kirst, D., Huang, Y., 2023. MirrorUs: mirroring peers’ affective cues to promote learner’s meta-cognition in video-based learning. *Proc. ACM Hum.-Comput. Interact.* 7 (CSCW2), 1–25. <https://doi.org/10.1145/3610079>

- Choi, S., Lee, H., Lee, Y., Kim, J., 2024. Vivid: human-AI collaborative authoring of vicarious dialogues from lecture videos. In: Proceedings of the CHI Conference on Human Factors in Computing Systems, pp. 1–26. <https://doi.org/10.1145/3613904.3642867>
- Christopher, D.A., et al., 2003. Interactive large classes: the dynamics of teacher/student interaction. *J. Bus. Econ. Res.* 1 (8). <https://doi.org/10.19030/jber.v1i8.3040>
- Cohen, G., Cohen, A., 2024. Motivational genAI chatbot for promoting SRL in problem solving: data-driven design principles. *Ubiquity Proc.* 4 (1). <https://doi.org/10.5334/uproc.147>
- D’Mello, S., Graesser, A., 2012. Autotutor and affective autotutor: learning by talking with cognitively and emotionally intelligent computers that talk back. *ACM Trans. Interact. Intell. Syst.* 2 (4), 23. <https://doi.org/10.1145/2395123.2395128>
- Dabbagh, N., Kitsantas, A., 2012. Personal learning environments, social media, and self-regulated learning: a natural formula for connecting formal and informal learning. *Internet High. Educ.* 15 (1), 3–8. <https://doi.org/10.1016/j.iheduc.2011.06.002>
- Dabbagh, N., 2007. The online learner: characteristics and pedagogical implications. *Contemp. Issues Technol. Teach. Educ.* 7 (3), 217–226.
- de Las Heras, S.C., Jones, M.N., Gernaey, K.V., Krühne, U., Mansouri, S.S., 2020. An e-learning bot for bioprocess systems engineering. In: Computer aided chemical engineering, vol. 48. Elsevier, pp. 2023–2028. <https://doi.org/10.1016/B978-0-12-823377-1.50338-4>
- Dresel, M., Haugwitz, M., 2008. A computer-based approach to fostering motivation and self-regulated learning. *J. Exp. Educ.* 77 (1), 3–20. <https://doi.org/10.3200/JEXE.77.1.3-20>
- Du Boulay, B., Avramides, K., Luckin, R., Martínez-Mirón, E., Méndez, G.R., Carr, A., 2010. Towards systems that care: a conceptual framework based on motivation, metacognition and affect. *Int. J. Artif. Intell. Educ.* 20 (3), 197–229. <https://doi.org/10.5555/1971783.1971784>
- Edwards, J., Nguyen, A., Lämsä, J., Sobocinski, M., Whitehead, R., Dang, B., Roberts, A.-S., Järvelä, S., 2025. Human-AI collaboration: designing artificial agents to facilitate socially shared regulation among learners. *Br. J. Educ. Technol.* 56 (2), 712–733. <https://doi.org/10.1111/bjet.13534>
- Fan, F., Wang, S., Dinuer, M., Hemuti, M., Nie, X., Yang, L.T., 2025. Enhancing students’ metacognition with innovative IA-based metacognitive reflective learning tool. *IEEE Trans. Learn. Technol.* 18, 699–715. <https://doi.org/10.1109/TLT.2025.3580536>
- Ferschke, O., Howley, I., Tomar, G., Yang, D., Liu, Y., Rosé, C.P., 2015. Fostering Discussion Across Communication Media in Massive Open Online Courses. *International Society of the Learning Sciences, Inc.[ISLS]*. <https://doi.org/10.22318/cscl2015.404>
- Ferschke, O., Yang, D., Tomar, G., Rosé, C.P., 2015. Positive impact of collaborative chat participation in an edX MOOC. In: Artificial Intelligence in Education: 17th International Conference, AIED 2015, Madrid, Spain, June 22–26, 2015. Proceedings 17. Springer, pp. 115–124. https://doi.org/10.1007/978-3-319-19773-9_12
- Fetzner, M., 2013. What do unsuccessful online students want us to know? *J. Asynchronous Learn. Netw.* 17 (1), 13–27. <https://doi.org/10.24059/olj.v17i1.319>
- Flavell, J.H., 1979. Metacognition and cognitive monitoring: a new area of cognitive-developmental inquiry. *Am. Psychol.* 34, 906–911. <https://doi.org/10.1037/0003-066X.34.10.906>
- Fleming, S.M., Lau, H.C., 2014. How to measure metacognition. *Front. Hum. Neurosci.* 8, 443. <https://doi.org/10.3389/fnhum.2014.00443>
- Gan, W., Qi, Z., Wu, J., Lin, J.C.-W., 2023. Large language models in education: vision and opportunities. In: 2023 IEEE international conference on big data (BigData). IEEE, pp. 4776–4785. <https://doi.org/10.1109/BigData59044.2023.10386291>
- Garcia, R., Falkner, K., Vivian, R., 2018. Systematic literature review: self-regulated learning strategies using e-learning tools for computer science. *Comput. Educ.* 123, 150–163. <https://doi.org/10.1016/j.compedu.2018.05.006>
- Ge, W., Sun, Y., Wang, Z., Zheng, H., He, W., Wang, P., Zhu, Q., Wang, B., SRLAgent: enhancing self-regulated learning skills through gamification and LLM assistance, arXiv preprint [arXiv:2506.09968](https://arxiv.org/abs/2506.09968), 2025.
- George, S.D., Dewan, P., 2024. NotebookGPT—facilitating and monitoring explicit lightweight student GPT help requests during programming exercises. In: Companion Proceedings of the 29th International Conference on Intelligent User Interfaces, pp. 62–65. <https://doi.org/10.1145/3640544.3645234>
- Graesser, A.C., Wiemer-Hastings, K., Wiemer-Hastings, P., Kreuz, R., Group, T.R., et al., 1999. Autotutor: a simulation of a human tutor. *Cogn. Syst. Res.* 1 (1), 35–51. [https://doi.org/10.1016/S1389-0417\(99\)00005-4](https://doi.org/10.1016/S1389-0417(99)00005-4)
- Guntur, M., Purnomo, Y.W., 2024. A meta-analysis of self-regulated learning interventions studies on learning outcomes in online and blended environments. *Online Learn.* 28 (3), 563–584. <https://doi.org/10.24059/olj.v28i3.4025>
- Hadwin, A.F., Järvelä, S., Miller, M., 2011. Self-regulated, co-regulated, and socially shared regulation of learning. *Handb. Self-Regul. Learn. Perform.* 30, 65–84.
- Hao, Z., Jiang, J., Yu, J., Liu, Z., Zhang, Y., Student engagement in collaborative learning with AI agents in an LLM-empowered learning environment: a cluster analysis, arXiv preprint [arXiv:2503.01694](https://arxiv.org/abs/2503.01694), 2025.
- Hattie, J., Timperley, H., 2007. The power of feedback. *Rev. Educ. Res.* 77 (1), 81–112. <https://doi.org/10.3102/003465430298487>
- Haugeland, I.K.F., Følstad, A., Taylor, C., Bjørkli, C.A., 2022. Understanding the user experience of customer service chatbots: an experimental study of chatbot interaction design. *Int. J. Hum.-Comput. Stud.* 161, 102788. <https://doi.org/10.1016/j.ijhcs.2022.102788>
- Hew, K.F., Huang, W., Du, J., Jia, C., 2023. Using chatbots to support student goal setting and social presence in fully online activities: learner engagement and perceptions. *J. Comput. High. Educ.* 35 (1), 40–68. <https://doi.org/10.1007/s12528-022-09338-x>
- Isöhätälä, J., Järvenoja, H., Järvelä, S., 2017. Socially shared regulation of learning and participation in social interaction in collaborative learning. *Int. J. Educ. Res.* 81, 11–24. <https://doi.org/10.1016/j.ijer.2016.10.006>
- Jansen, R.S., Van Leeuwen, A., Janssen, J., Jak, S., Kester, L., 2019. Self-regulated learning partially mediates the effect of self-regulated learning interventions on achievement

- in higher education: a meta-analysis. *Educ. Res. Rev.* 28, 100292. <https://doi.org/10.1016/j.edurev.2019.100292>
- Kang, H., Ben Moussa, M., Thalmann, N.M., 2024. Nadine: a large language model-driven intelligent social robot with affective capabilities and human-like memory. *Comput. Anim. Virtual Worlds* 35 (4), e2290. <https://doi.org/10.1002/cav.2290>.
- Kim, Y., Wei, Q., 2006. Virtual peers help increase learner self-efficacy. In: *Proceedings of EdMedia + Innovate Learning*. Association for the Advancement of Computing in Education (AACE), pp. 2033–2039.
- Kizilcec, R.F., Piech, C., Schneider, E., 2013. Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. In: *Proceedings of the third international conference on learning analytics and knowledge*, pp. 170–179. <https://doi.org/10.1145/2460296.2460330>
- Kuhail, M.A., Alturki, N., Alramlawi, S., Alhejori, K., 2023. Interacting with educational chatbots: a systematic review. *Educ. Inf. Technol.* 28 (1), 973–1018. <https://doi.org/10.1007/s10639-022-11177-3>
- Kumar, H., Xiao, R., Lawson, B., Musabirov, I., Shi, J., Wang, X., Luo, H., Williams, J.J., Rafferty, A.N., Stamper, J., et al., 2024. Supporting self-reflection at scale with large language models: insights from randomized field experiments in classrooms. In: *Proceedings of the Eleventh ACM Conference on Learning@ Scale*, pp. 86–97. <https://doi.org/10.1145/3657604.3662042>
- Lai, C.-L., Tu, Y.-F., 2024. Roles, strategies, and research issues of generative AI in the mobile learning era. *Int. J. Mob. Learn. Organ.* 18 (4), 516–537. <https://doi.org/10.1504/IJML.2024.141836>
- Laugwitz, B., Held, T., Schreppe, M., 2008. Construction and evaluation of a user experience questionnaire. In: *HCI and Usability for Education and Work: 4th Symposium of the Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society, USAB 2008, Graz, Austria, November 20–21, 2008*. *Proceedings 4*. Springer, pp. 63–76. https://doi.org/10.1007/978-3-540-89350-9_6
- Lee, Y.-C., Lin, W.-C., Cherng, F.-Y., Wang, H.-C., Sung, C.-Y., King, J.-T., 2015. Using time-anchored peer comments to enhance social interaction in online educational videos. In: *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pp. 689–698. <https://doi.org/10.1145/2702123.2702349>
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S., Kiela, D., 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474. <https://doi.org/10.5555/3495724.3496517>
- Li, T., Yan, L., Iqbal, S., Srivastava, N., Singh, S., Raković, M., Swiecki, Z., Tsai, Y.-S., Gašević, D., Fan, Y., et al., 2025. Analytics of self-regulated learning strategies and scaffolding: associations with learning performance. *Comput. Educ. Artif. Intell.* 8, 100410. <https://doi.org/10.1016/j.caeai.2025.100410>
- Liaw, S.-S., Huang, H.-M., 2013. Perceived satisfaction, perceived usefulness and interactive learning environments as predictors to self-regulation in e-learning environments. *Comput. Educ.* 60 (1), 14–24. <https://doi.org/10.1016/j.compedu.2012.07.015>
- Liu, J., Yao, Y., An, P., Wang, Q., 2024. PeerGPT: probing the roles of LLM-based peer agents as team moderators and participants in children's collaborative learning. In: *Extended abstracts of the CHI conference on human factors in computing systems*, pp. 1–6. <https://doi.org/10.1145/3613905.3651008>
- Martha, A.S.D., Santoso, H.B., Junus, K., Suhartanto, H., 2019. Designing metacognitive and motivation tutor: a pedagogical agent to facilitate learning in blended-learning environment in a higher education context. In: *International Conference on Computers in Education (ICCE)*, pp. 581–590.
- Mejeh, M., Sarbach, L., 2025. Co-design: from understanding to prototyping an adaptive learning technology to enhance self-regulated learning. *Technol. Knowl. Learn.* 30 (2), 1021–1046. <https://doi.org/10.1007/s10758-024-09788-5>
- Messinger, D.S., Prince, E.B., Zheng, M., Martin, K., Mitsven, S.G., Huang, S., Stölzel, T., Johnson, N., Rudolph, U., Perry, L.K., et al., 2019. Continuous measurement of dynamic classroom social interactions. *Int. J. Behav. Dev.* 43 (3), 263–270. <https://doi.org/10.1177/0165025418820708>
- Morales-Chan, M., Amado-Salvatierra, H.R., Medina, J.A., Barchino, R., Hernández-Rizzardini, R., Teixeira, A.M., 2024. Personalized feedback in massive open online courses: harnessing the power of langchain and openAI API. *Electronics* 13 (10), 1960. <https://doi.org/10.3390/electronics13101960>
- Moreno, R., Mayer, R.E., 2005. Role of guidance, reflection, and interactivity in an agent-based multimedia game. *J. Educ. Psychol.* 97 (1), 117. <https://doi.org/10.1037/0022-0663.97.1.117>
- Moribe, S., Ushijima, T., 2025. Imitating mistakes in a learning companion AI agent for on-line peer learning. In: *2025 19th International Conference on Ubiquitous Information Management and Communication (IMCOM)*. IEEE, pp. 1–8. <https://doi.org/10.1109/IMCOM64595.2025.10857528>
- Nelson, T.O., 1990. Metamemory: A theoretical framework and new findings. In: *Psychology of learning and motivation*, vol. 26. Elsevier, pp. 125–173. [https://doi.org/10.1016/S0079-7421\(08\)60053-5](https://doi.org/10.1016/S0079-7421(08)60053-5)
- Panadero, E., 2017. A review of self-regulated learning: six models and four directions for research. *Front. Psychol.* 8, 422. <https://doi.org/10.3389/fpsyg.2017.00422>
- Pintrich, P.R., Marx, R.W., Boyle, R.A., 1993. Beyond cold conceptual change: the role of motivational beliefs and classroom contextual factors in the process of conceptual change. *Rev. Educ. Res.* 63 (2), 167–199. <https://doi.org/10.3102/00346543063002167>
- Pintrich, P.R., 2002. The role of metacognitive knowledge in learning, teaching, and assessing. *Theory Pract.* 41 (4), 219–225. https://doi.org/10.1207/s15430421tp4104_3
- Pintrich, P.R., 2004. A conceptual framework for assessing motivation and self-regulated learning in college students. *Educ. Psychol. Rev.* 16, 385–407. <https://doi.org/10.1007/s10648-004-0006-x>
- Prasse, D., Webb, M., Deschênes, M., Parent, S., Aeschlimann, F., Goda, Y., Yamada, M., Raynault, A., 2024. Challenges in promoting self-regulated learning in technology supported learning environments: an umbrella review of systematic reviews and meta-analyses. *Technol. Knowl. Learn.* 29 (4), 1809–1830. <https://doi.org/10.1007/s10758-024-09772-z>
- Psathas, G., Tegos, S., Demetriadis, S.N., Tsiatsos, T., 2023. Exploring the impact of chat-based collaborative activities and SRL-focused interventions on students' self-regulation profiles, participation in collaborative activities, retention, and learning in MOOCs. *Int. J. Comput.-Support. Collab. Learn.* 18 (3), 329–351. <https://doi.org/10.1007/s11412-023-09394-0>
- Puntambekar, S., Hübscher, R., 2005. Tools for scaffolding students in a complex learning environment: what have we gained and what have we missed? *Educ. Psychol.* 40 (1), 1–12. https://doi.org/10.1207/s15326985ep4001_1
- Qin, C., Huang, W., Hew, K.F., 2020. Using the community of inquiry framework to develop an educational chatbot: lesson learned from a mobile instant messaging learning environment. In: *Proceedings of the 28th international conference on computers in education*.
- Rivadeneira, J., Inga, E., 2023. Interactive peer instruction method applied to classroom environments considering a learning engineering approach to innovate the teaching-learning process. *Educ. Sci.* 13 (3), 301. <https://doi.org/10.3390/educsci13030301>
- Roll, I., Aleven, V., McLaren, B.M., Koedinger, K.R., 2007. Designing for metacognition—applying cognitive tutor principles to the tutoring of help seeking. *Metacogn. Learn.* 2 (2–3), 125–140. <https://doi.org/10.1007/s11409-007-9010-0>
- Schraw, G., Dennison, R.S., 1994. Assessing metacognitive awareness. *Contemp. Educ. Psychol.* 19 (4), 460–475. <https://doi.org/10.1006/ceps.1994.1033>
- Schraw, G., Moshman, D., 1995. Metacognitive theories. *Educ. Psychol. Rev.* 7, 351–371. <https://doi.org/10.1007/BF02212307>
- Schraw, G., Kuch, F., Gutierrez, A.P., 2013. Measure for measure: calibrating ten commonly used calibration scores. *Learn. Instr.* 24, 48–57. <https://doi.org/10.1016/j.learninstruc.2012.08.007>
- Schraw, G., 2009. Measuring metacognitive judgments. In: *Handbook of metacognition in education*. Routledge, pp. 415–429.
- Shukla, J., Kumari, A., Sakthi, G., 2024. Investigating chatbot integration for improved blended learning. In: *Computational Methods in Science and Technology*. CRC Press, pp. 225–232. <https://doi.org/10.21137/jeeval.2023.2.2.1>
- Shute, V.J., Zapata-Rivera, D., 2012. Adaptive educational systems. *Adaptive Technol. Train. Educ.* 7 (27), 1–35. <https://doi.org/10.1017/CBO9781139049580.004>
- Steinert, S., Avila, K.E., Ruzika, S., Kuhn, J., Küchemann, S., Harnessing large language models to enhance self-regulated learning via formative feedback, arXiv preprint [arXiv:2311.13984](https://arxiv.org/abs/2311.13984), 2023.
- Tamoyan, H., Schuff, H., Gurevych, I., 2025. LLM roleplay: simulating human-chatbot interaction. In: *Proceedings of the Third Workshop on Social Influence in Conversations (SICon 2025)*, pp. 1–26. <https://doi.org/10.18653/v1/2025.sicon-1.1>
- Tanprasert, T., Fels, S.S., Sinnamon, L., Yoon, D., 2023. Scripted vicarious dialogues: educational video augmentation method for increasing isolated students' engagement. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–25. <https://doi.org/10.1145/3544548.3581153>
- Thiede, K.W., Anderson, M., Theriault, D., 2003. Accuracy of metacognitive monitoring affects learning of texts. *J. Educ. Psychol.* 95 (1), 66. <https://doi.org/10.1037/0022-0663.95.1.66>
- Thomas, D.R., 2006. A general inductive approach for analyzing qualitative evaluation data. *Am. J. Eval.* 27 (2), 237–246. <https://doi.org/10.1177/1098214005283748>
- Tobias, S., Everson, H.T., 2009. The importance of knowing what you know: A knowledge monitoring framework for studying metacognition in education. In: *Handbook of metacognition in education*. Routledge, pp. 107–127.
- Tsai, M.-J., Tsai, C.-C., 2003. Student computer achievement, attitude, and anxiety: the role of learning strategies. *J. Educ. Comput. Res.* 28 (1), 47–61. <https://doi.org/10.2190/PL27-TC1Q-08B2-RMCL>
- Tsai, M.-J., 2009. The model of strategic e-learning: understanding and evaluating student e-learning from metacognitive perspectives. *J. Educ. Technol. Soc.* 12 (1), 34–48.
- Vapiwala, F., Pandita, D., 2023. Assessing digitized self-regulation strategies in workplace e-learning. In: *2023 14th International Conference on Information & Communication Technology and System (ICTS)*. IEEE, pp. 57–61. <https://doi.org/10.1109/ICTS58770.2023.10330876>
- Veletsianos, G., 2009. The impact and implications of virtual character expressiveness on learning and agent-learner interactions. *J. Comput. Assist. Learn.* 25 (4), 345–357. <https://doi.org/10.1111/j.1365-2729.2009.00317.x>
- Wang, S., Xu, T., Li, H., Zhang, C., Liang, J., Tang, J., Yu, P.S., Wen, Q., Large language models for education: a survey and outlook, arXiv preprint [arXiv:2403.18105](https://arxiv.org/abs/2403.18105), 2024.
- Wang, T., Zhan, Y., Lian, J., Hu, Z., Yuan, N.J., Zhang, Q., Xie, X., Xiong, H., 2025. LLM-powered multi-agent framework for goal-oriented learning in intelligent tutoring system. In: *Companion Proceedings of the ACM on Web Conference 2025*, pp. 510–519. <https://doi.org/10.1145/3701716.3715244>
- Wells, A., Cartwright-Hatton, S., 2004. A short form of the metacognitions questionnaire: properties of the MCQ-30. *Behav. Res. Ther.* 42 (4), 385–396. [https://doi.org/10.1016/S0005-7967\(03\)00147-5](https://doi.org/10.1016/S0005-7967(03)00147-5)
- Winne, P.H., 1998. Studying as self-regulated learning. *Metacogn. Educ. Theory Pract./LEA* 9, 277–304.
- Zimmerman, B.J., 1986. Becoming a self-regulated learner: which are the key subprocesses. *Contemp. Educ. Psychol.* 11, 1–25. [https://doi.org/10.1016/0361-476X\(86\)90027-5](https://doi.org/10.1016/0361-476X(86)90027-5)
- Zimmerman, B., 1989. A social cognitive view of self-regulated academic learning. *J. Educ. Psychol.* 81, 329–339. <https://doi.org/10.1037/0022-0663.81.3.329>
- Zimmerman, B.J., 2000. Attaining self-regulation: a social cognitive perspective. In: *Self-regulation: theory, research, and applications/academic*, pp. 13–39. <https://doi.org/10.1016/B978-012109890-2/50031-7>