Computers & Education 235 (2025) 105357

Contents lists available at ScienceDirect

Computers & Education

journal homepage: www.elsevier.com/locate/compedu

Bridging the knowledge-skill gap: The role of large language model and critical thinking in education

Jin Yuxian

Sungkyunkwan University, Jongno-gu, Seoul, South Korea

ARTICLE INFO

Keywords:
Declarative knowledge
Procedural knowledge
Critical thinking
Large language model

ABSTRACT

In the dynamic field of education, addressing the knowledge-skill gap remains a significant challenge, as students often excel in theoretical understanding but struggle with practical application. This study investigates the combined effects of the large language model (LLM) chatbot and critical thinking guidance on learners' acquisition of declarative knowledge ("know what") and procedural knowledge ("know how"). Findings indicate that while the LLM chatbot enhances declarative knowledge acquisition, it does not significantly impact procedural learning, as learners tend to prioritize lower cognitive load and focus on declarative knowledge. In contrast, critical thinking guidance fosters procedural learning but increases cognitive load, thereby limiting resources available for declarative learning. However, when both interventions are combined, they generate synergistic effects—critical thinking guidance activates procedural learning, while the LLM chatbot mitigates cognitive burden, enabling a more balanced allocation of cognitive resources and improving both declarative and procedural knowledge acquisition. These findings highlight the potential of LLM chatbots as effective educational tools, suggesting that the strategic use of artificial intelligence can promote a more effective approach to knowledge acquisition. This research provides valuable insights into the application of advanced technologies in educational contexts, emphasizing the importance of appropriate instructional strategies to guide the effective use of these technologies.

1. Introduction

In the dynamic realm of education, enhancing learning outcomes is paramount as educators and researchers aim to prepare learners for an increasingly complex world. Central to this area is the development of cognitive learning outcomes, which include the knowledge and skills acquired through learning (Bloom, 1956; Krathwohl, 2002). These outcomes are organized within Bloom's taxonomy, a framework that categorizes educational goals into a hierarchy of cognitive complexity, ranging from basic knowledge recall to advanced problem-solving skills (Krathwohl, 2002).

A significant challenge in education is the knowledge-skill gap, where students excel in acquiring knowledge but struggle to apply it in real-world contexts (Dunning et al., 2004). Employers frequently report that graduates, despite their academic credentials, lack the essential skills needed to succeed in the workplace (Andrews & Higson, 2008; Jackson, 2016). This disconnect suggests that educational systems may be overly focused on imparting conceptual knowledge without adequately fostering practical skills (Bereiter & Scardamalia, 1993; Hattie & Timperley, 2007).

Addressing this gap necessitates differentiating between two primary types of cognitive learning outcomes: declarative knowledge

E-mail address: jinyuxian@g.skku.edu.

https://doi.org/10.1016/j.compedu.2025.105357
Received 10 November 2024; Received in revised form 4 May 2025; Accepted 7 May 2025
Available online 9 May 2025

and procedural knowledge. Declarative knowledge, or "knowing what," involves understanding facts, concepts, and information (Anderson & Krathwohl, 2010; Ryle & Tanney, 2009). Declarative learning, which is the pathway to acquire declarative knowledge, is often based on activities like reading and memorization (Sweller, 1988). Procedural knowledge, or "knowing how," pertains to skills and procedures demonstrated through action, emphasizing the application of learned skills in practical contexts (Anderson, 1982; Ryle & Tanney, 2009). Procedural learning, as the pathway to acquire procedural knowledge, often involves deeper cognitive processing and the integration of learned skills into practice, requiring learners to actively engage in skill-based application (Ericsson, 2006; VanLehn, 1996).

The dual pathways of declarative learning and procedural learning are essential components of the learning process. According to cognitive load theory (Sweller, 1988), both declarative and procedural learning demand cognitive resources to function effectively (Sweller, 2010). However, given that learners have limited cognitive resources, they tend to prioritize declarative learning, which generally requires fewer cognitive resources than procedural learning (Mayer & Moreno, 2003; Sweller, 1988). Procedural learning can be stimulated by critical thinking—a complex cognitive process involving evaluation, analysis, and synthesis of information to form reasoned judgments (Abrami et al., 2015; Lai, 2011).

In this context, integrating large language model (LLM) - based chatbots into educational settings presents a promising potential (Zawacki-Richter et al., 2019). These chatbots, powered by advanced artificial intelligence techniques, have the potential to enhance both declarative and procedural learning (Holmes et al., 2019; Luckin & Holmes, 2016; Shridhar et al., 2022). By providing access to diverse information and customized explanations, LLM chatbots can aid declarative learning by reducing the cognitive resources required for understanding and memorizing (Holmes et al., 2019; Mayer, 2002). Furthermore, by answering critical questions and offering feedback on reasoning, LLM chatbots can support procedural learning by alleviating the cognitive load associated with it (Anderson et al., 2018; Shridhar et al., 2022).

However, learners may focus more on declarative learning over procedural learning, preferring a lower cognitive load (Van Merrienboer & Sweller, 2005). To address this, critical thinking is an effective strategy to stimulate the procedural learning pathway (Bransford et al., 1999; Chin & Osborne, 2008). Critical thinking is crucial for effective problem-solving and decision-making, as it enables individuals to question assumptions, evaluate evidence, and approach issues with a rational, open-minded perspective (Paul & Elder, 2019). By fostering critical thinking, learners can activate their procedural learning pathways, which enhances their ability to apply theoretical information in practical situations (Halpern, 1998). However, since procedural learning requires more cognitive resources, learners may need to release some resources used by declarative learning to prevent cognitive overload (Chandler & Sweller, 1991).

This study explores the *interactive impact of the LLM chatbot and critical thinking guidance on declarative and procedural learning*. Using a robust within-subjects experimental design, we involved 80 undergraduate students in four distinct learning tasks across various domains unrelated to their major. This approach minimized the influence of pre-existing knowledge and allowed us to focus on the impact of the interventions. We employed a 2x2 factorial design, encompassing two interventions: an LLM chatbot and critical thinking guidance. This design enabled us to isolate the effects of each intervention on knowledge acquisition. Our analysis included regression models to evaluate the main and interaction effects of the interventions, with task and learner fixed effects to control for potential confounding factors.

This study makes a significant contribution to understanding the impact of advanced artificial intelligence technology and instructional design on cognitive learning outcomes. The research highlights the potential of LLM chatbots in promoting both declarative and procedural learning, while also emphasizing the importance of appropriate instructional support to fully realize the potential. Additionally, the study sheds light on the double-edged effects of critical thinking guidance, which, while stimulating procedural learning, can impose a high cognitive burden. This study provides valuable insights for integrating advanced AI technology and instructional design in educational settings, suggesting that the strategic use of AI tools can foster a more comprehensive and balanced approach to knowledge acquisition, ultimately enhancing the effectiveness of educational practices.

2. Theory

2.1. Cognitive learning outcomes

Cognitive learning outcomes refer to the mental skills and knowledge gained through learning. In educational contexts, cognitive learning outcomes provide a structure for evaluating how well learners acquire, comprehend, and apply knowledge within specific domains (Anderson & Krathwohl, 2010). Bloom's taxonomy organizes these outcomes into a hierarchical order of increasing complexity, ranging from basic knowledge recall to advanced skills (Bloom, 1956; Krathwohl, 2002). Similarly, learning theory emphasized two main types of cognitive learning outcomes: declarative knowledge and procedural knowledge (Anderson, 2013).

Declarative knowledge refers to "knowing what," involves the understanding of facts, concepts, and information (Ryle & Tanney, 2009). It is knowledge that can be explicitly stated, such as factual data, terminology, and principles (Anderson & Krathwohl, 2010). Bloom (1956) similarly suggests that declarative knowledge involves the recall of specifics and universals, methods and processes, or patterns, structures, or settings. Declarative knowledge, often described as surface or rote learning (Biggs & Tang, 1999), is considered the simplest and lowest level of knowledge, as it involves the memorization of facts and data (Azevedo & Aleven, 2013; Bloom, 1956; De Backer et al., 2012; Krathwohl, 2002). According to Sweller (1988), declarative knowledge serves as a basis for more complex forms of knowledge and skills, as learners first need to understand basic concepts before they can apply them in practical or procedural tasks.

Procedural knowledge pertains to "knowing how," involving skills and procedures that may not be easily verbalized but can be demonstrated through action (Anderson, 1982; Ryle & Tanney, 2009). It refers to organized modes of operation and generalized

techniques for addressing problems, emphasizing the cognitive processes of organizing and reorganizing material to achieve specific objectives (Bloom, 1956; Ten Berge & Van Hezewijk, 1999). Procedural knowledge represents a higher form of knowledge, characterized by the understanding of how to employ various problem-solving strategies and cognitive skills (Azevedo & Aleven, 2013; Bloom, 1956; Georgeff & Lansky, 1986; Krathwohl, 2002). In cognitive learning theory, procedural knowledge is essential because it enables individuals to execute tasks efficiently and effectively by applying learned declarative knowledge in appropriate contexts (Anderson & Schunn, 2013).

Anderson's (2013) knowledge acquisition theory distinguishes between the declarative learning pathway and the procedural learning pathway in the learning process. Declarative learning involves acquiring declarative knowledge through information recall and memorization (Anderson, 2013). In the procedural learning, declarative knowledge is integrated into the procedures required to execute the skill, marking the development of procedural knowledge (Anderson, 2013). However, the declarative learning and procedural learning is not strictly sequential; instead, they develop iteratively and simultaneously throughout the learning process (Rittle-Johnson et al., 2001).

2.2. Cognitive load theory

Cognitive Load Theory is a foundational principle in educational psychology and instructional design (Sweller, 1988). The theory posits that working memory has a limited capacity for holding information at any given time (Miller, 1956). Cognitive Load Theory distinguishes among three types of cognitive load: (1) intrinsic cognitive load, (2) extraneous cognitive load, and (3) germane cognitive load.

Intrinsic cognitive load refers to the inherent difficulty associated with a specific instructional topic, determined by the complexity of the content and the learner's prior knowledge (Sweller, 1994). In contrast, extraneous cognitive load arises from the way information is presented. Clear, well-explained instructions can help reduce extraneous cognitive load (Chandler & Sweller, 1991). Germane cognitive load, on the other hand, involves the cognitive resources devoted to processing, constructing, and automating schemas. This type of load is beneficial for learning, as it facilitates the integration of new information with existing knowledge (Sweller et al., 1998).

It is generally suggested that procedural learning imposes significantly greater cognitive load than declarative learning, primarily due to its complex, multi-step processes that demand more cognitive resources. Procedural learning involves higher intrinsic cognitive load because it requires learners to perform sequences of tasks that demand careful attention to each step, as well as to the overall task structure, which is not as pronounced in declarative learning (Sweller, 1988). Anderson's ACT-R Theory (1982) also supports this notion, indicating that procedural learning necessitates converting declarative knowledge into operational rules through practice, which intensifies the cognitive load, particularly in terms of germane load devoted to schema construction. Additionally, Van Merrienboer and Sweller's (2005) Four-Component Instructional Design model posits that procedural learning tend to increase cognitive load, as learners must continuously process feedback, refine steps, and adapt based on errors or new information, making these tasks more cognitively taxing than declarative learning. Research on task complexity further supports that procedural learning, which often involves problem-solving and sequential processing, generally requires sustained cognitive engagement that can lead to cognitive overload if not adequately managed through instructional support (Paas & Van Merriënboer, 1994). These findings highlight that procedural learning typically demands a higher cognitive investment, as learners navigate complex stages to achieve skill development.

Due to the limited capacity of working memory, learners frequently resort to surface-level knowledge memorization, often prioritizing declarative learning over the more cognitively demanding procedural learning (Bransford et al., 1999). According to Miller's (1956) theory on working memory, complex tasks can quickly overwhelm cognitive resources, prompting a shift towards simpler memorization of facts rather than engaging in deep procedural learning. This tendency aligns with the suggestion that learners may avoid the higher cognitive load associated with procedural tasks by focusing on recalling discrete facts or definitions, which impose a lower cognitive load (Chandler & Sweller, 1991; Sweller, 1988).

In environments lacking appropriate scaffolding, students may prioritize rote memorization over deeper understanding, ultimately hindering their ability to develop essential procedural knowledge (Chi & VanLehn, 2012). Research also indicates that learners' preference for declarative learning under cognitive load may be an adaptive strategy to optimize their cognitive efficiency, focusing on manageable chunks of information rather than more elaborate procedural tasks that could lead to cognitive overload (Paas & Van Merriënboer, 1994). Without effective instructional strategies, such as structured guidance and feedback, learners may struggle to progress beyond basic memorization and engage in the deeper cognitive processes required for procedural knowledge development (Hattie & Timperley, 2007). This emphasis on surface-level knowledge acquisition can ultimately undermine long-term retention and transfer, as procedural learning is essential for developing practical, transferable skills (Anderson, 1982).

2.3. Large language model

The integration of chatbots into educational environments has been a focal point of scholarly inquiry, with researchers exploring their effectiveness, benefits, and challenges (Pérez et al., 2020). Early chatbots, such as rule-based and retrieval-based systems, primarily relied on scripted responses and pattern-matching techniques, limiting their flexibility and adaptability in dynamic learning contexts (Wallace, 2009; Weizenbaum, 1966). These traditional AI-based chatbots were effective for structured interactions but struggled with complex reasoning, contextual understanding, and adaptive feedback, making them less suitable for personalized learning experiences (Kerlyl et al., 2006).

In contrast, recent advances in artificial intelligence have led to the development of Large Language Models (LLMs), which leverage deep learning techniques, including transformer architectures (Vaswani et al., 2017), zero-shot learning, and reinforcement learning from human feedback (Brown et al., 2020). LLM-based chatbots can generate high-quality, context-aware content across various domains, including mathematics (Shridhar et al., 2022), programming (Al-Hossami et al., 2024), physics (Gregorcic et al., 2024), and cognitive behavioral therapy (Izumi et al., 2024). LLMs have significantly enhanced chatbot capabilities by enabling more interactive, adaptive, and personalized learning experiences (Smutny & Schreiberova, 2020), fostering student engagement (Goda et al., 2014), marking a shift towards more interactive and adaptive educational tools.

LLMs have emerged as transformative tools in the realm of educational technology, offering unprecedented capabilities in natural language processing and understanding (Brown et al., 2020). These models, powered by advanced machine learning algorithms, are designed to comprehend and generate human-like text, making them highly effective in educational settings (Radford et al., 2019). LLMs can process vast amounts of data to provide contextually relevant responses, thereby enhancing the learning experience by offering personalized and immediate feedback (Devlin, 2018). Their ability to adapt to individual learning styles and needs allows for a more tailored educational approach (Kaswan et al., 2024), which is crucial for effective knowledge acquisition. As these models continue to evolve, their integration into educational environments promises to revolutionize the way knowledge is imparted and absorbed, making learning more accessible and efficient (Raffel et al., 2020).

LLMs significantly influence declarative learning by effectively managing both extraneous and intrinsic cognitive load. The capability to instantly retrieve information and tailored explanations reduces extraneous cognitive load, as learners can clarify concepts and retrieve relevant examples without the mental strain of searching through multiple resources (Baidoo-Anu & Ansah, 2023; Paas et al., 2003; Sweller, 1988). In addition, LLMs can indirectly reduce intrinsic cognitive load by simplifying the presentation of complex concepts and breaking them into more manageable, understandable parts (Luckin & Holmes, 2016). Thus, LLMs can support declarative learning.

In addition to enhancing declarative knowledge, LLMs also play a crucial role in facilitating procedural learning by reducing germane cognitive load. Learners benefit from immediate feedback and instructional prompts that encourage deeper engagement with the material (Hattie & Timperley, 2007). This interaction not only aids learners in applying conceptual knowledge to practical scenarios but also guides learners to explore multiple approaches to problem-solving (King, 1990). LLMs assist users to examine their assumptions and thought processes more deeply by answering their critical questions, which helps learners refine their cognitive engagement (Chin & Osborne, 2008; Kaswan et al., 2024; Shridhar et al., 2022). Furthermore, LLMs are capable of engaging users in debate and discussion, helping them to consider alternative perspectives and clarify their arguments, which supports critical reasoning (Anderson et al., 2018). Through feedback on reasoning, LLMs can identify logical flaws and offer constructive feedback to foster an iterative approach to learning (Hu et al., 2023; Lin et al., 2024). By simulating scenario-based analysis and hypothetical problem-solving situations, LLMs enable users to make connections between theory and practice, helping them navigate complex decision-making (Roll & Wylie, 2016). Thus, the interactive, responsive nature of LLM-based chatbots can serve as a valuable aid in assisting procedural learning.

2.4. *Critical thinking*

Critical thinking is a multifaceted cognitive process that involves evaluating, analyzing, and synthesizing information to form reasoned judgments (Abrami et al., 2015; Lai, 2011). Facione (2011) defines critical thinking as "the process of purposeful, self-regulatory judgment which results in interpretation, analysis, evaluation, and inference." It encompasses various cognitive skills, including analysis, interpretation, inference, explanation, and self-regulation (Ennis, 1985). In educational contexts, the consistent integration of critical thinking into the learning process not only enhances students' capacity for autonomous learning but also significantly augments their cognitive abilities, enabling them to think independently, solve problems, and make informed decisions (Hennessey, 1999; Hollingworth & McLoughlin, 2001).

As learners often default to focusing on declarative knowledge due to its relatively low cognitive load requires fewer cognitive resources (Miller, 1956; Sweller, 1988), critical thinking guidance serves as an effective intervention to stimulate procedural learning by encouraging deeper cognitive processing and application of knowledge (Elder & Paul, 2020; King & Kitchener, 1994). Research indicates that critical thinking prompts help redirect learners' attention from rote memorization towards more analytical processes, fostering the mental organization and restructuring required for procedural understanding (Halpern, 1998; Kuhn, 1999). By engaging learners in reflective questioning and problem-solving, critical thinking guidance increases germane cognitive load, which Sweller (2011) argue is essential for the development of complex skills. Moreover, Chi et al. (1994) demonstrated that learners who engage in critical thinking and self-explanation strategies achieve higher levels of procedural knowledge, as these approaches promote the mental organization of steps, strategies, and conditions for task completion. When learners are provided with structured opportunities to analyze, evaluate, and apply concepts, they more readily transition from declarative to procedural learning pathways, developing the skills necessary for higher-order, transferrable knowledge (Anderson, 1982; Hattie & Timperley, 2007).

As critical thinking guidance stimulates procedural learning, they inevitably allocate cognitive resources away from declarative learning. It is explained by the limited capacity of working memory, which can only hold and process a finite amount of information at any given time (Cowan, 2001; Miller, 1956). As learners invest more cognitive effort in procedural learning, the cognitive load required for declarative learning may exceed their available capacity. When cognitive resources are stretched thin, learners may experience a shift in attention towards the more demanding aspects of procedural learning, thereby diminishing their capacity to process declarative knowledge (Ayres & Paas, 2012; Paas & Van Merriënboer, 1994; Sweller, 1988). This shift is further compounded by the fact that procedural learning often involves the development of skills that are essential for long-term retention and real-world

application (Anderson, 1982; Ericsson et al., 1993), making it a natural progression for learners to dedicate more cognitive resources to mastering these complex tasks. The interaction between declarative and procedural learning is dynamic, where the focus on one often leads to the suppression of the other, particularly when cognitive resources are taxed (Schneider & Shiffrin, 1977).

The tension between declarative and procedural learning is also suggested in empirical evidence, as some studies have reported both positive correlations (e.g., Broadbent et al., 2020; Chevrier et al., 2019; Manganelli et al., 2019) and negative correlation (Ranellucci et al., 2015) between critical thinking and learning outcomes. This discrepancy may be attributed to the failure to differentiate between declarative and procedural knowledge, underscoring the necessity of investigating critical thinking with learning considered as a dual-pathway approach, rather than as a coherent process (Sweller, 1988; Van Merrienboer & Sweller, 2005).

2.5. Hypotheses

The integration of LLMs into educational settings holds significant potential for enhancing cognitive learning outcomes, both in terms of declarative and procedural learning. However, learners often default to declarative learning, especially in cases where there is a lack of appropriate guidance and instruction to stimulate procedural learning. We suggest that learners who focus primarily on declarative learning may not fully leverage the chatbot's capabilities. When learners prioritize declarative learning, they may tend to use LLMs predominantly as tools for accessing factual information, definitions, and clarifications, reinforcing surface-level knowledge. This usage pattern, driven by the learner's strategy and cognitive focus, limits the chatbot's broader potential to foster deeper engagement with procedural learning processes, such as problem-solving or skill development. As a result, learners who emphasize declarative knowledge acquisition may not actively engage with the procedural learning features of the LLM, such as interactive problem-solving or strategic feedback, thereby failing to tap into the full range of benefits the LLM offers.

Consequently, while the LLM enhances declarative knowledge acquisition through efficient access to information, its impact on procedural knowledge acquisition remains constrained. Thus, we propose the following hypothesis:

H1. The use of an LLM chatbot alone enhances declarative knowledge acquisition without significantly affecting procedural knowledge acquisition.

The use of critical thinking guidance in educational settings is designed to stimulate higher-order cognitive processes, such as analysis, evaluation, and synthesis. These processes are particularly aligned with procedural learning, which involves problem-solving, strategic planning, and the application of learned concepts in practical contexts. As learners engage with critical thinking, they actively focus on applying knowledge in dynamic scenarios, which stimulates their procedural learning by encouraging the development of strategies, decision-making, and problem-solving skills.

However, critical thinking also demands significant cognitive resources, which may reduce the cognitive capacity available for declarative learning. Since declarative learning relies more on rote memorization and recall of factual information, the cognitive load required for procedural learning may inadvertently limit the learner's ability to focus on acquiring and retaining factual information. As a result, while critical thinking guidance promotes deeper engagement with procedural learning, it can diminish the cognitive emphasis on declarative learning.

Therefore, by prioritizing cognitive resources for procedural knowledge acquisition, critical thinking reduces the focus on declarative knowledge acquisition, potentially limiting its acquisition in the learning process. Thus, we propose the following hypothesis:

H2. The use of critical thinking guidance alone diminishes declarative knowledge acquisition while enhancing procedural knowledge acquisition.

We argued earlier that while LLM chatbots have the potential to aid both declarative and procedural learning, learners who often default to a focus on declarative learning may only utilize LLM chatbots for that purpose, leaving the chatbot's potential to support procedural learning underutilized. In contrast, critical thinking guidance enhances procedural learning by fostering higher-order cognitive processes but inherently hinders declarative learning due to limited cognitive capacity. However, we suggest that integrating LLM chatbots with critical thinking guidance produces interaction effects that leverage the strengths of both tools, thereby promoting both declarative and procedural learning.

For the procedural learning pathway, stimulated by critical thinking guidance, the cognitive demand on learners is high. Here, LLM chatbots contribute by addressing critical questions, providing interactive problem-solving scenarios, and offering feedback and guidance. These features help reduce the high cognitive load associated with procedural learning, enabling learners to engage more effectively in procedural learning.

For the declarative learning pathway, which critical thinking guidance can impede by transferring cognitive resources to procedural learning, LLM chatbots offer an additional benefit beyond the effects argued in the first hypothesis. By alleviating some of the cognitive load associated with procedural learning, LLM chatbots help free up cognitive resources for declarative learning. This ensures that procedural learning activated by critical thinking guidance does not compromise declarative learning.

In summary, the combined use of LLM chatbots and critical thinking guidance produces interaction effects that enhance both declarative and procedural knowledge acquisition more effectively than using each alone. Therefore, we propose the following hypothesis:

H3. The combined use of an LLM chatbot and critical thinking guidance has interaction effects, enhancing both declarative and procedural knowledge acquisition.

3. Methodology

3.1. Experiment design

The study recruited undergraduate students majoring in Physics from a university in China. In addition to sharing the same major, all participants were from the same academic year to minimize individual-level differences. Participation in the experiment was voluntary and seamlessly integrated into their coursework.

Participants were informed that they would engage in a learning project comprising four learning tasks, each conducted over four consecutive weeks at a fixed time each week. These tasks covered four distinct learning domains that were not directly related to their major but were valuable for their future career prospects. Completion of all four tasks was required to earn bonus course credits, with the reward contingent upon their performance in the post-learning exams. This incentive structure was designed to ensure sustained participation and alignment with the within-subject experimental design.

Before the experiment, participants provided informed consent, confirming their voluntary participation and understanding of the study's objectives, procedures, and associated benefits. Ultimately, 80 participants voluntarily joined the study, all of whom successfully completed all learning tasks and post-learning exams.

The experiment was conducted on a custom-designed website, which assigned each participant a unique experiment ID and provided domain-specific learning materials, interventions, and exams. Each learning task lasted 30 min, during which participants studied text-based materials within the designated domain. At the end of each learning task, they were required to submit a 500-word learning note. A text area was embedded at the bottom of the screen for note-taking, and submissions were automatically collected at the conclusion of the learning task. This requirement was implemented to enhance engagement and encourage active learning. The relatively low word count aimed to discourage participants from using the LLM chatbot as a note-generation tool rather than as a learning aid.

Following each learning task, participants completed a 20-min exam assessing both declarative and procedural knowledge. They were informed in advance that learning materials, technological tools, and their notes would not be accessible during the exam.

3.2. Intervention

Two interventions are employed through the learning tasks: LLM chatbot and critical thinking guidance. The webpage presenting the learning tasks and interventions is documented in Appendix A.

The first intervention, a LLM chatbot powered by Baidu's ERNIE 4.0 model has been integrated into the website. The model shows as high performance with Chinese prompts (Baidu, 2023), matching the language of the participants, learning materials, and website design. The integration of this chatbot aims to assist learners by providing real-time support, such as answering questions and offering clarifications. Participants assigned to this intervention have access to the chatbot during learning tasks, which is positioned on the right side of the learning materials. In contrast, those without the intervention do not see the chatbot, ensuring that any observed effects can be attributed to the presence or absence of this LLM chatbot support.

The second intervention, critical thinking guidance, is embedded in the instructions at the beginning of each learning task. Critical thinking, defined as the ability to analyze, evaluate, and synthesize information to form reasoned judgments (Ennis, 2015). In the context of this study, participants assigned to receive critical thinking guidance are instructed to submit a learning note addressing specific prompts designed to foster critical thinking: 1) identifying potential practical applications of the knowledge learned, 2) noting any imperfections or limitations within the provided materials or concepts, and 3) suggesting improvements or alternative approaches to enhance the material's comprehensiveness or relevance. These prompts are grounded in established frameworks for fostering critical thinking, such as the approach proposed by Elder and Paul (2020), which emphasizes reflective thought and systematic evaluation. The design of these prompts aims to engage participants in deeper cognitive processing by encouraging them to evaluate and apply knowledge in a way that promotes critical analysis and problem-solving (Bransford et al., 1999). In contrast, participants not assigned critical thinking guidance are instructed to submit a learning note without content restrictions, allowing for open reflection. This distinction ensures that the effects of critical thinking guidance can be clearly observed and analyzed.

3.3. Randomization

The experiment employs a 2x2 factorial design, encompassing four conditions: (1) Baseline, (2) LLM-Only, (3) CT-Only, and (4) Combined. The Baseline group indicates the learning condition without an LLM chatbot or critical thinking guidance. The LLM-Only group denotes the learning condition with the LLM chatbot but without critical thinking guidance. The CT-Only group represents the learning condition without the LLM chatbot but with critical thinking guidance. Finally, the Combined group indicates the learning condition with both the LLM chatbot and critical thinking guidance.

A within-subject experimental design was employed, with each participant engaging in four learning tasks under each of the four conditions. To minimize cognitive contamination from prior exposure to critical thinking guidance, the first two tasks excluded this guidance, while the last two included it. This design allowed for the isolation of the effects of critical thinking guidance on knowledge acquisition. The sequence of conditions followed a fixed order: (1) Baseline, (2) LLM-Only, (3) CT-Only, and (4) Combined.

The domain of the learning task is randomized to counterbalance domain-related bias. Each participant engages in four learning tasks across the following domains: (1) research methodology, (2) motivation theory, (3) marketing, and (4) coding. These domains were selected for their lack of direct relevance to the participants' major, minimizing the influence of pre-existing knowledge.

Table 1
Variables.

Dependent Variables	
DKA	Learner's score on single-choice questions in the post-learning exam.
PKA	Learner's score on problem-solving questions in the post-learning exam.
Independent Variables	
LLM	A dummy variable indicating whether the LLM-based chatbot was provided during a learning task.
CT	A dummy variable indicating whether critical thinking guidance was given during a learning task.
Fixed Effects	
DFE	Domain fixed effect, a list of dummy variables indicating the domain of each learning task.
LFE	Learner fixed effect, a list of dummy variables indicating individual learners.

Randomization was achieved using Latin Square Counterbalancing, ensuring that any differences in learning outcomes were attributable to the interventions rather than familiarity with the domains or order effects. A pre-survey confirmed that no participant reported systematically studying these domains before the experiment, ensuring that their performance was not influenced by prior knowledge.

This structured approach, characterized by precise interventions and a controlled environment, enables a comprehensive evaluation of the impact of critical thinking guidance and LLM chatbot integration on knowledge acquisition, providing valuable insights into their role in educational contexts.

3.4. Measures

This study employs several measurement tools in the experiment. As the independent variables are manipulated through interventions during learning tasks, the dependent variables—declarative knowledge acquisition and procedural knowledge acquisition—are measured through exams following each learning task. To ensure the validity of the assessment tools for both declarative and procedural knowledge acquisition, the exam questions were directly aligned with the learning materials and designed to measure the specific cognitive processes.

Declarative knowledge acquisition was assessed using 10 single-choice questions constructed based on key factual content explicitly stated in the learning materials, ensuring content validity. Single-choice questions are recognized as effective measures of declarative knowledge, as they assess factual recall and recognition without requiring knowledge transformation (Haladyna et al., 2002). Each question directly tested a concept or piece of knowledge explicitly presented in the learning materials, ensuring a strong alignment between the assessment and instructional content. Additionally, the questions were evenly distributed across the learning materials to ensure comprehensive coverage.

Procedural knowledge acquisition was assessed using an open-ended problem-solving question designed to require learners to apply the learned concepts in a contextual scenario. This approach aligns with established models of procedural knowledge assessment, which emphasize problem-solving and application as key components (Anderson & Krathwohl, 2010; Van Merriënboer & Kirschner, 2017). Open-ended questions are particularly effective for assessing higher-order cognitive skills, such as reasoning, synthesis, and problem-solving, which are central to procedural knowledge (Jonassen, 1997; Resnick, 1987). Unlike the single-choice questions, these open-ended questions did not directly appear in the learning materials but instead required learners to integrate, apply, and extend their knowledge to novel situations.

To ensure the quality and validity of the open-ended questions, domain experts reviewed and refined them. Specifically, the open-ended questions for the research methodology and motivation theory domains were revised by a professor specializing in education, while the question for the marketing domain was sourced from the same textbook used as the learning material. The coding domain question was adapted from an in-class homework assignment used in a Python course taught by a professor.

To further validate the assessment, a pilot test was conducted with 20 master's students who had no prior knowledge of the tested domains. The pilot study aimed to evaluate the distribution of scores across the assessments. Based on the results, questions with excessively low or high accuracy rates were revised to ensure balanced score distributions across domains, thereby enhancing the content validity and reliability of the assessment.

Lastly, fixed time limits were implemented to standardize assessment conditions and reduce variability due to differences in time allocation. Participants were given 5 min to complete the single-choice questions and 15 min for the open-ended questions, further improving reliability.

The 10 single-choice questions are scored on a scale of 0–10, with each question worth 1 point. Examples of these questions can be found in Appendix B. For the open-ended questions, an independent master's student, who was completely blinded to this study, evaluated the responses using the rubric provided in Appendix C. Each question was assessed across five dimensions, with each dimension scored from 0 to 2, resulting in a maximum total score of 10 per question. Finally, the single-choice questions and open-ended questions are both range from 0 to 10.

Measurement of variables are listed in Table 1. Declarative knowledge acquisition (DKA) and procedural knowledge acquisition (PKA) served as the dependent variables in this study. The independent variables included the presence of an LLM chatbot (LLM) and critical thinking guidance (CT). To control for potential confounding factors, two sets of fixed effects were included: domain fixed effects (DFE) and learner fixed effects (LFE). DFE consists of dummy variables accounting for the domain of each learning task, while LFE includes dummy variables to control for individual learner characteristics across tasks.

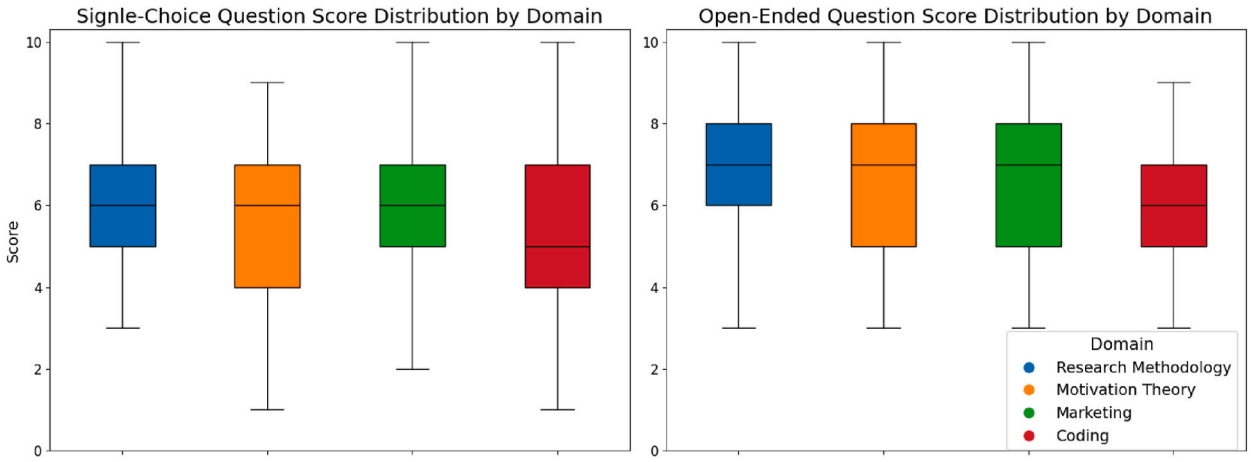

Fig. 1. Score distribution by task.

3.5. Analysis strategy

Data analysis for this study will involve several key steps to assess the effects of the LLM chatbot and critical thinking guidance on the acquisition of declarative and procedural knowledge. Before testing the hypotheses, descriptive statistics will be calculated for participant scores on the single-choice questions and open-ended questions, including means and standard deviations, to show the overall score distribution by learning domain. To examine differences in learning outcomes across conditions, we employ both t-tests and regression analysis.

Pairwise t-tests will be conducted to compare mean scores across conditions, providing direct numerical comparisons and identifying statistically significant differences between groups. However, t-tests are limited in that they cannot assess interaction effects or incorporate control variables.

Therefore, regression analyses are performed to examine both main and interaction effects, while controlling for domain-related and learner-related factors using fixed effects. Regression analyses employ two models to separately assess: 1) the overall effects of the LLM chatbot and critical thinking guidance on cognitive learning outcomes, and 2) the main effects and interaction effects of the LLM chatbot and critical thinking guidance on cognitive learning outcomes. The regression formulas are as follows:

$$DV = CT + LLM + FE + Error \tag{1}$$

$$DV = CT + LLM + CT \times LLM + FE + Error \tag{2}$$

where DV represents the dependent variables (DKA and PKA), with LLM and CT as the independent variables. FE refers to the fixed effects, including DFE and LFE. The first equation tests the overall direct effects of the LLM chatbot and critical thinking guidance on declarative and procedural knowledge acquisition. The second equation examines the main effects of the LLM chatbot and critical thinking guidance individually, as well as their interaction, with the interaction term (CT × LLM) included in the regression model.

By integrating *t*-test and regression analysis, this study balances interpretability with statistical rigor to provide a nuanced understanding of the relationships between interventions and learning outcomes. T-tests provide a straightforward numerical comparison between conditions, making it easier to interpret performance differences. However, they cannot assess interaction effects or control for confounding factors. Conversely, regression analysis enables a more comprehensive examination of intervention effects by accounting for both main and interaction effects while incorporating control variables and fixed effects. Nonetheless, because regression decomposes main and interaction effects, direct numerical comparisons between conditions may be less intuitive. This combined approach ensures a robust evaluation of the effects of the LLM chatbot and critical thinking guidance on learning outcomes.

3.6. Ethical considerations

This study was conducted in full compliance with ethical standards for educational research. Approval was granted by the Institutional Review Board (IRB ID: 2023-11-008),[1] ensuring that all aspects of the research adhered to institutional guidelines regarding participant welfare, informed consent, and data confidentiality.

Given the online format of the experiment, informed consent was obtained electronically. Participants were required to read detailed information about the study's purpose, procedures, potential risks, and their rights. They then confirmed their consent by

[1] Approval was obtained from the IRB affiliated with the author's institution. The specific affiliation is anonymized to comply with the requirements of the peer review process.

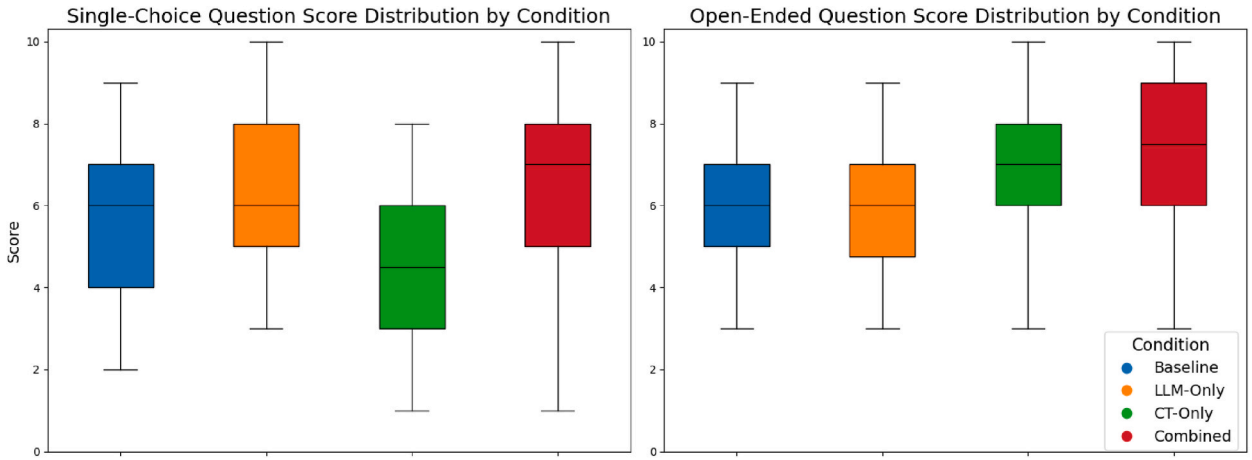

Fig. 2. Score distribution by condition.

checking a box indicating they had read and accepted all information before proceeding with the learning tasks. This step ensured that participants understood the study's requirements and voluntarily agreed to participate.

Participant confidentiality was strictly maintained by assigning each individual a unique identifier to anonymize their responses. All data was securely stored to prevent unauthorized access. Additionally, the learning tasks and assessments were designed to promote participants' educational experience without causing undue stress. By addressing these ethical considerations, the study ensured the protection of participants' rights and well-being throughout the research process.

4. Results

4.1. Descriptive statistics

Fig. 1 presents the descriptive statistics for the score distribution across different tasks. First, the distribution of scores for the single-choice questions is examined. In the Research Methodology domain, the mean score was 5.913, with a standard deviation of 1.950. The Motivation Theory domain recorded a mean score of 5.763 and a standard deviation of 1.963. The Marketing domain achieved the highest mean score of 6.188, with a standard deviation of 1.856. In contrast, the Coding domain had the lowest mean score of 5.413, with a standard deviation of 2.004.

Next, the score distribution for the open-ended questions is analyzed. In the Research Methodology domain, the mean score was 7.013, with a standard deviation of 1.673. The Motivation Theory domain had a mean score of 6.688 and a standard deviation of 1.740. The Marketing domain showed a mean score of 6.513, with a standard deviation of 1.743. Conversely, the Coding domain had the lowest mean score of 6.075, with a standard deviation of 1.621.

It is noteworthy that the fixed effects incorporated in the regression model are anticipated to address the disparities observed in the score distribution.

4.2. T-test

Fig. 2 shows the *t*-test results, which provide initial insights into score differences across different conditions. We examined our hypotheses using pairwise t-tests among four groups: (1) Baseline, (2) LLM-Only, (3) CT-Only, and (4) Combined.

For declarative knowledge acquisition, assessed through single-choice question scores, notable findings emerged. The mean score for learners in the Baseline condition was 5.688. Learners in the LLM-Only condition scored a mean of 6.513, significantly higher than the baseline ($p < 0.01$), indicating that using an LLM chatbot alone enhances declarative knowledge acquisition. In contrast, learners in the CT-Only condition scored a mean of 4.575, significantly lower than the baseline ($p < 0.001$), suggesting that critical thinking guidance alone may hinder declarative knowledge acquisition. However, the group with the Combined condition scored 6.500, significantly higher than the baseline ($p < 0.01$), suggesting that while critical thinking instruction alone may impede declarative knowledge acquisition, its combination with LLM benefits declarative learning. Notably, the mean score of the Combined condition (6.500) is similar to the LLM-Only condition (6.513), indicating that the negative effects of critical thinking guidance on declarative knowledge acquisition are generally mitigated by the interaction effects between the LLM chatbot and critical thinking guidance.

Regarding procedural knowledge acquisition, measured by open-ended question scores, significant insights were also observed. The baseline mean score for learners in the Baseline condition was 6.113. Learners in the LLM-Only condition scored 6.025, not significantly different from the baseline, indicating that introducing LLM alone does not significantly impact procedural learning. However, learners in the CT-Only condition scored 6.750, significantly higher than the baseline ($p < 0.01$), suggesting that critical thinking guidance alone enhances procedural knowledge acquisition. The group in the Combined condition scored 7.400, significantly

Table 2
Regression results.

DV	DKA		PKA	
Model	1	2	3	4
LLM	1.375***	0.825**	0.281	−0.088
CT	−0.563**	−1.113***	1.006***	0.638*
LLM × CT		1.100**		0.738*
FE	✓	✓	✓	✓
R^2	0.353	0.373	0.372	0.384
Adj-R^2	0.122	0.145	0.148	0.160
N	320	320	320	320

*p < 0.05; **p < 0.01; ***p < 0.001.

higher than both the baseline ($p < 0.001$) and the CT-Only group ($p < 0.05$). This indicates that while critical thinking guidance alone improves procedural learning, its effect is amplified when combined with LLM.

The *t*-test results provide strong preliminary support for the hypotheses. Compared to the Baseline condition, the CT-Only condition exhibited higher procedural knowledge acquisition but lower declarative knowledge acquisition. The LLM-Only condition demonstrated higher declarative knowledge acquisition while maintaining similar levels of procedural knowledge acquisition as the Baseline condition. Participants in the Combined condition outperformed the Baseline condition in both declarative and procedural knowledge acquisition.

4.3. Regression

To validate the hypotheses, a regression analysis is conducted. Regression is considered a more rigorous analyzation than which provides greater statistical rigor by examining main and interaction effects, while controlling for domain-related and learner-related factors. The results of the regression analyses examining the effects of the LLM chatbot (LLM) and critical thinking guidance (CT) on declarative knowledge acquisition (DKA) and procedural knowledge acquisition (PKA) are presented in Table 2. The models account for both domain fixed effects (DFE) and learner fixed effects (LFE), with a total sample size of 320 learning tasks completed by 80 learners.

Regression Model 1 indicates that the LLM chatbot has a significant positive overall effect on declarative knowledge acquisition (1.375, $p < 0.001$), suggesting that engagement with the LLM is generally associated with increased declarative knowledge acquisition. Conversely, critical thinking guidance shows a significant negative relationship with declarative knowledge acquisition (−0.563, $p < 0.01$), indicating that stimulating cognitive thinking may generally hinder the acquisition of declarative knowledge in the learning context.

Regression Model 2 provides a more detailed examination of the effects of the LLM chatbot and critical thinking guidance. The results indicate that the LLM chatbot has a significant positive main effect on declarative knowledge acquisition (0.825, $p < 0.01$), suggesting that engagement with the LLM chatbot alone enhances declarative knowledge acquisition. However, critical thinking guidance continues to show a significant negative main effect on declarative knowledge acquisition (−1.113, $p < 0.001$), indicating that critical thinking guidance alone detrimentally affects declarative knowledge acquisition. Notably, the significant positive interaction term (1.100, $p < 0.01$) suggests that the introduction of the LLM chatbot in combination with critical thinking guidance generates synergistic positive effects on declarative knowledge acquisition, independent of the main effects of the LLM chatbot and critical thinking guidance alone.

In Regression Model 3, no significant relationship between the LLM chatbot and procedural knowledge acquisition is observed (0.281, n.s.), suggesting that engagement with the LLM generally does not significantly affect procedural knowledge acquisition. However, critical thinking guidance reveals a significant positive overall effect (1.006, $p < 0.001$), suggesting that stimulating critical thinking generally improves procedural knowledge acquisition.

Regression Model 4 shows that the LLM chatbot still does not have a significant main effect (−0.088, n.s.) on procedural knowledge acquisition, indicating that providing the LLM chatbot without critical thinking guidance does not significantly enhance procedural knowledge acquisition. Conversely, critical thinking guidance demonstrates a significant positive main effect (0.638, $p < 0.05$), suggesting that stimulating critical thinking enhances procedural knowledge acquisition, even without LLM chatbot support. Furthermore, the interaction term between the LLM chatbot and critical thinking guidance is significantly positive (0.738, $p < 0.05$), indicating that the combination of the LLM chatbot and critical thinking guidance has a positive interaction effect on procedural knowledge acquisition.

To ensure the validity of the regression results, standard diagnostics were conducted. The independent variables were dummy-coded, making the linearity assumption relevant only to group mean differences. Fixed effects accounted for unobserved heterogeneity, addressing concerns of independence and omitted variables. Diagnostic checks indicated no major violations of homoscedasticity or normality. As the predictors were operationally defined, multicollinearity was not a concern. Collectively, these checks support the assumptions of the regression.

In summary, the results reveal the potential of LLM chatbots to improve both declarative and procedural knowledge acquisition. However, LLM chatbots improve declarative knowledge acquisition independently, but only enhance procedural knowledge

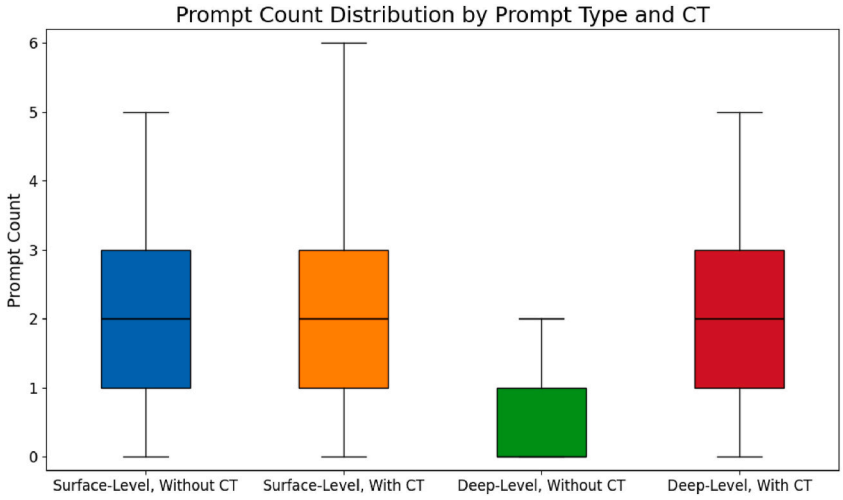

Fig. 3. Prompt count distribution by prompt type and CT

acquisition when combined with critical thinking guidance. Moreover, critical thinking guidance demonstrates dual roles in influencing declarative and procedural knowledge acquisition. While critical thinking guidance significantly increases procedural knowledge acquisition, it simultaneously decreases declarative knowledge acquisition unless combined with LLM chatbots. The positive interaction effect between the LLM chatbot and critical thinking guidance emphasizes that this combination enhances both declarative and procedural knowledge acquisition more effectively than using each alone.

5. Complementary analysis

5.1. Prompt analysis

We suggested in our theory that, while most learners default to focusing on declarative learning, they tend to use the LLM chatbot primarily for surface-level information retrieval and explanation, without fully realizing the impact of LLM chatbots on enhancing procedural learning. To support this argument, we provide an additional analysis of the prompts submitted by learners.

This section presents an analysis of prompts submitted by learners to the LLM chatbot, comparing engagement patterns between those who received critical thinking guidance (CT) and those who did not. The goal is to examine how critical thinking guidance influences both the quantity and quality of prompts directed at the LLM chatbot.

The analysis categorizes prompts into two types based on question depth: surface-level and deep-level questions. Surface-level questions typically involve simple facts, definitions, or basic explanations, while deep-level questions require reasoning, analysis, evaluation, or application. A master's student was recruited to categorize all prompts according to these criteria using a predefined rubric. Examples of both prompt types across different task domains are provided in Appendix D.

Fig. 3 illustrates the distribution of prompt counts categorized by prompt type and the presence of critical thinking guidance (CT). The analysis indicates a significant disparity in prompt frequency between the two groups. Learners with CT submitted an average of 3.550 prompts per session, compared to the 2.250 prompts per session submitted by those without CT. The results of the *t*-test suggest that critical thinking guidance markedly enhances learners' interaction frequency ($p < 0.001$).

Regarding surface-level prompts, learners with CT submitted an average of 1.875 prompts per session, which is nearly equivalent to the 1.825 surface-level prompts submitted by learners without CT. A *t*-test revealed no significant difference between these values. However, a notable difference was observed in the submission of deep-level prompts. Learners with CT submitted an average of 1.675 deep-level prompts per session, compared to only 0.425 deep-level prompts submitted by those without CT. A *t*-test confirmed that learners with CT were significantly more inclined to submit deep-level prompts ($p < 0.001$).

These findings provide an explanation for the earlier result that merely providing access to an LLM chatbot is insufficient for fostering procedural learning. Without critical thinking guidance, learners tend to focus primarily on declarative learning, engaging the chatbot with surface-level prompts requiring examples, definitions, or explanations. However, the significant increase in deep-level prompts directed at the LLM chatbot with critical thinking guidance suggests that such guidance encourages learners to pose more probing questions that support procedural learning.

This section provides insights into the mechanisms underlying our hypotheses. The results suggest that incorporating appropriate guidance and instruction into educational practices can empower learners to utilize AI tools like LLM chatbots more effectively, particularly for fostering deeper engagement that supports skill building.

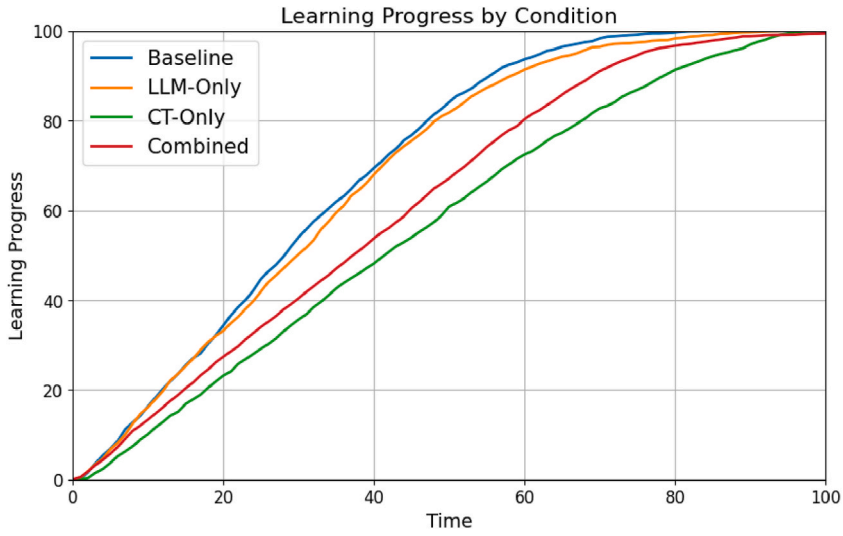

Fig. 4. Learning progress by condition.

5.2. *Learning progress analysis*

We propose that while most learners primarily focus on declarative learning, which imposes a lower cognitive load, critical thinking guidance promotes procedural learning but simultaneously introduces a substantial cognitive burden. In cases where cognitive overload arises due to critical thinking guidance, the LLM chatbot can serve as an effective intervention to alleviate the cognitive demands associated with procedural learning, thereby freeing cognitive resources to support declarative learning. To provide behavioral evidence supporting this proposition, this section examines learners' learning progress through their webpage scrolling behavior.

Learning progress and cognitive load are closely interrelated, as cognitive load directly influences the efficiency with which learners engage with instructional materials. According to Cognitive Load Theory (Sweller, 1988), when cognitive demands exceed a learner's processing capacity, information assimilation is disrupted, leading to slower learning progress. Excessive cognitive demands can impair comprehension and retention, ultimately hindering reading progress (Van Merrienboer & Sweller, 2005). Conversely, when cognitive load is effectively managed, learners can process and retain information more efficiently, facilitating smoother and more rapid progression through instructional content (Paas et al., 2003). Instructional strategies that reduce cognitive load—such as simplifying content presentation or incorporating scaffolding techniques—have been shown to improve learning efficiency and support an optimal learning pace (Gerjets et al., 2004). Thus, learning progress and cognitive load are closely intertwined, shaping how effectively learners engage with and absorb instructional content.

The analysis is presented in a time-based format, with the x-axis representing time as a percentage (ranging from 0 to 100 %) and the y-axis depicting learners' progress as a percentage (ranging from 0 to 100 %). Learning progress is measured as the percentage of the total instructional material accessed. Learners frequently engage in scroll-back behavior, reviewing previously read content. However, for the purpose of this analysis, learning progress is defined by the highest point reached within the instructional material. Consequently, a learner's progress is recorded based on the furthest position they have scrolled, even if they subsequently revisit earlier sections. Each learning task is designed to last 30 min, with learners' positions recorded at 18-s intervals, yielding a total of 100 data points per task.

Fig. 4 illustrates the learning progress over time for four distinct conditions: (1) Baseline, (2) LLM-Only, (3) CT-Only, and (4) Combined.

The Baseline condition demonstrated the fastest progression, suggesting an efficient learning pace. This indicates that learners in this group experienced the lowest cognitive load, primarily focusing on declarative learning.

Learners in the LLM-Only condition showed a slightly slower progression compared to the baseline. Their learning curve indicates a minor delay, likely due to two factors: while the LLM chatbot slightly reduces cognitive load by assisting with declarative learning, the time spent interacting with the chatbot slows their overall progress. The delayed curve suggests that the time spent using the LLM outweighs the cognitive load reduction benefits, especially since declarative learning does not impose a severe cognitive load on learners.

In contrast, learners in the CT-Only condition experienced the slowest progression, with a curve reflecting a more challenging and deliberate learning pace. This is consistent with the shift from declarative learning to procedural learning, which significantly increases cognitive load and results in slower learning progress.

Interestingly, the Combined condition showed significant improvement in learning progress compared to the CT-Only condition. While there was some delay compared to the baseline, the curve indicates that the presence of the LLM chatbot helped learners engage in critical thinking while maintaining an acceptable pace. This suggests that the LLM chatbot acted as a valuable support tool, reducing

the severe cognitive load imposed by critical thinking guidance, which stimulates procedural learning. In this context, the LLM chatbot's ability to reduce cognitive load outweighed the time spent interacting with it.

Overall, this analysis highlights distinct learning progress patterns across different conditions. It provides insights into how cognitive load differs during the learning process from a behavioral perspective. The results suggest that learners experiencing cognitive overload benefit more from the LLM chatbot in terms of reducing cognitive load, while those with lower cognitive load achieve limited improvements.

6. Discussion

The results of this study underscore the potential of integrating LLM chatbots with critical thinking guidance to enhance both declarative and procedural knowledge acquisition. The findings offer significant insights into how advanced AI technologies, when combined with appropriate instructional strategies, can help bridge the knowledge-skill gap in education.

First, the study highlights the conditional impact of LLM chatbots on cognitive learning outcomes. Consistent with previous research, LLM chatbots promote declarative learning by providing immediate access to vast amounts of information and contextual explanations (Holmes et al., 2019; Mayer, 2002). The chatbot's ability to deliver well-organized, readily available knowledge aligns with cognitive load theory, which suggests that reducing extraneous cognitive load facilitates information retention (Sweller, 1988). However, despite their effectiveness in declarative learning, LLM chatbots did not significantly enhance procedural learning. This aligns with prior findings indicating that AI-driven instructional tools often require structured scaffolding to support higher-order cognitive processes (Chi & Wylie, 2014).

Second, critical thinking guidance was found to significantly enhance procedural learning, but it may hinder declarative learning. This finding supports the notion that critical thinking encourages deeper cognitive engagement, prompting learners to analyze, evaluate, and synthesize information rather than passively absorb it (Halpern, 1998; Kuhn, 1999). Critical thinking guidance encourage learners to engage in reflective questioning, problem-solving, and decision-making, all of which are crucial for procedural learning. However, this approach also increases cognitive load, as learners must actively construct meaning and apply concepts in novel contexts (Paas & Van Merriënboer, 1994; Sweller, 1988). As a result, the increased cognitive burden may restrict the capacity for declarative learning, a phenomenon consistent with cognitive overload observed in complex environments (Van Gog et al., 2010).

The most compelling finding of this study is the synergistic effect observed when combining LLM chatbots with critical thinking guidance. This interaction effect suggests that while critical thinking guidance activates procedural learning by fostering deeper cognitive engagement, the LLM chatbot counterbalances cognitive load by streamlining information processing and reducing extraneous effort. Prior research has indicated that AI-enhanced scaffolding can facilitate knowledge integration by dynamically adjusting support based on cognitive demands (Roll et al., 2011). In this study, the chatbot's role in alleviating cognitive burden allowed learners to engage more effectively in both declarative and procedural learning, a result consistent with cognitive load theory's predictions regarding optimal resource allocation (Sweller, 2011). These findings extend existing research on AI-assisted learning by demonstrating that strategic integration of LLMs with instructional scaffolding can yield more effective learning outcomes than either intervention alone.

These findings have important implications for the design and implementation of AI-assisted educational tools. First, while LLM chatbots are effective for enhancing declarative learning, their impact on procedural knowledge acquisition is limited unless paired with structured guidance. This suggests that AI tools should not function as standalone instructional methods but rather as complements to pedagogical strategies that encourage deeper cognitive engagement. Second, critical thinking guidance is essential for fostering procedural learning, but educators must be mindful of its cognitive demands. Instructional strategies should balance cognitive load by integrating AI-driven support mechanisms that facilitate declarative knowledge acquisition while allowing for the cognitive effort required in procedural learning. Finally, the observed synergy between LLM chatbots and critical thinking guidance highlights the potential of hybrid instructional approaches that leverage AI's efficiency alongside human-guided scaffolding. Future AI-based educational interventions should consider adaptive learning systems that dynamically adjust support based on learners' cognitive states and task demands.

While this study provides valuable insights, several limitations should be considered. The sample consisted of undergraduate students from a single university, which may limit the generalizability of the findings. Future studies should examine the effects of LLM chatbots and critical thinking guidance across diverse educational settings, disciplines, and learner populations. Additionally, this study focused on short-term learning outcomes; future research should investigate the long-term impact of these interventions on knowledge retention, transfer, and real-world application. Further exploration of individual differences, such as prior knowledge, cognitive abilities, and learning styles, could provide deeper insights into how different learners interact with AI-driven tools.

This study demonstrates the potential of integrating LLM chatbots with critical thinking guidance to bridge the knowledge-skill gap in education. By strategically combining these tools, educators can enhance both declarative and procedural knowledge acquisition, ensuring a more balanced approach to learning. These findings emphasize the need for instructional designs that optimize cognitive resource distribution, leveraging AI's efficiency while fostering deep cognitive engagement. As AI continues to evolve, future research should explore how adaptive, intelligent learning systems can further refine and personalize educational experiences, maximizing the benefits of both technology and pedagogy.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the author used ChatGPT, based on GPT-4, to improve grammar, readability, and fluency.

Glossary

1. Cognitive Learning Outcomes: Mental skills and knowledge gained through learning, often evaluated within educational contexts.
2. Declarative Knowledge: Understanding of facts, concepts, and information, often described as "knowing what."
3. Declarative Learning: The process of acquiring declarative knowledge through activities like reading and memorization.
4. Procedural Knowledge: Skills and procedures demonstrated through action, emphasizing the application of learned skills in practical contexts, often described as "knowing how."
5. Procedural Learning: The process of acquiring procedural knowledge through deeper cognitive processing and skill-based application.
6. Large Language Model: Advanced artificial intelligence models designed to comprehend and generate human-like text, used in educational settings to enhance learning experiences.
7. Critical Thinking: A cognitive process involving evaluation, analysis, and synthesis of information to form reasoned judgments.

Following the use of this tool, the author reviewed and edited the content as necessary and takes full responsibility for the content of the published article.

Acknowledgement

I would like to express my sincere gratitude to Professor Cui Xuemei of Yanbian University for providing the participants for this study. Her support was invaluable in facilitating the successful completion of the online experiment.

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Appendix A

Appendix A. Webpage with Interventions.

Appendix B

Appendix B
Examples of Single-Choice Questions

Task 1-Research Methodology	
Single-Choice Question	
What are the five typical steps in the educational research process?	
A	Research problem identification → Literature review → Data collection → Data analysis → Conclusion
B	Literature review → Research problem identification → Data collection → Data analysis → Conclusion
C	Research problem identification → Data collection → Data analysis → Literature review → Conclusion
D	Data collection → Data analysis → Research problem identification → Literature review → Conclusion
Task 2-Motivation Theory	
Single-Choice Question	
According to Maslow's hierarchy of needs, what is the correct order of needs from lowest to highest?	
A. Physiological needs - Safety needs - Social needs - Esteem needs - Self-actualization needs	
B. Safety needs - Physiological needs - Esteem needs - Social needs - Self-actualization needs	
C. Physiological needs - Safety needs - Social needs - Self-actualization needs - Esteem needs	
D. Safety needs - Physiological needs - Self-actualization needs - Social needs - Esteem needs	
Task 3-Marketing	
Single-Choice Question	
Which of the following statements about value is correct?	
A	Customers always purchase products or services that provide the highest objective value.
B	Customers can accurately and objectively assess value and cost.
C	Creating value for customers is the foundation of building strong customer relationships.
D	The value of a product is the same for all customers.
Task 4-Coding	
Single-Choice Question	
Which of the following statements are correct?	
A	A list cannot be modified after it is created.
B	A list is usually a collection of multiple elements.
C	A list can only be output element by element.
D	To output the last element of a list, you must know the total length of the list.

Appendix C

Appendix C
Rubric for Open-Ended Questions

Task 1-Research Methodology		
Open-Ended Question A research group wants to investigate the differences, advantages, and disadvantages of Teaching Method A and Teaching Method B for elementary students. 1) Design an experiment to compare the impact of these two teaching methods on students' performance in physics. Describe each step of the study in detail. 2) Explain how researchers should ensure the study's internal and external validity.		
Dimension	Score	Criteria
Experimental Design Completeness	0	The answer includes no specific experiment steps.
	1	The answer includes some key steps of the experiment.
	2	The answer includes all key steps of the experiment.
Control of Bias	0	The answer does not mention bias control.
	1	The answer mentions bias control but lacks solution to address the bias.
	2	The answer identifies how key variables will be controlled or randomized to minimize bias.
Measurement of Outcomes	0	The answer does not address how outcome will be measured.
	1	The answer mentions outcome measurement but lacks specific details or justification for chosen metrics.
	2	The answer clearly defines how outcome will be measured and has a comprehensive metrics of these measures.
Internal Validity Considerations	0	The answer does not address internal validity.
	1	The answer mentions internal validity but lacks a clear strategy to achieve it.
	2	The answer identifies and explains strategies to ensure internal validity.
External Validity Considerations	0	The answer does not address external validity.
	1	The answer mentions external validity but lacks a clear strategy to achieve it.

(*continued on next page*)

Appendix C (*continued*)

Task 1-Research Methodology		
	2	The answer identifies and explains strategies to ensure external validity.
Task 2-Motivation Theory		
Open-Ended Question Imagine you are a high school physics teacher. You notice that, despite using the same teaching methods and motivational techniques for all students, some students put in effort while others do not. 1) Using the motivation theories you've studied, analyze why such differences in student effort might exist. 2) Based on these theories, discuss how you might use different strategies to motivate different types of students.		
Dimension	Score	Criteria
Theories Used for Analyze	0	The answer uses no theory to explain effort differences.
	1	The answer uses one or two theories to explain effort differences.
	2	The answer uses three theories to explain effort differences.
Accuracy in Explaining Effort	0	The answer fails to correctly explain effort differences using theories, with misunderstandings or irrelevance.
	1	The answer attempts to explain effort differences with theories, but lacks clarity in how the theories apply.
	2	The answer provides a clear and accurate explanation of why effort differences exist, using each theory appropriately and correctly.
Theories Used for Motivation	0	The answer suggests strategies without drawing on any theory.
	1	The answer uses one or two theories to propose strategies.
	2	The answer uses three theories to proposes strategies.
Novelty and Feasibility of Motivation	0	The answer offers strategies that are neither novel nor feasible.
	1	The answer offers strategies that are either novel or feasible, but not both.
	2	The answer offers strategies that are both novel and feasible.
Coherence of Theories in Motivation	0	The answer lacks integration and coherence, with a disjointed plan that does not logically connect different theories in the approach.
	1	The answer shows some integration of theories, but connections may be weak, fragmented, or lack coherence.
	2	The answer demonstrates strong integration of theories, presenting a cohesive and coherent plan that ties various theories.
Task 3-Marketing		
Open-Ended Question Find a company you familiar to. 1) How does the company manage its relationship with you as a customer? 2) What are the disadvantages of this company's CRM strategy? How could you improve it?		
Understanding of CRM	0	The answer refers to no or one CRM concepts.
	1	The answer refers to two or three CRM concepts.
	2	The answer refers to more than three CRM concepts.
Accuracy of Analysis	0	The answer provides totally inaccurate analysis of the company's CRM strategy on the concepts refer to.
	1	The answer provides partly accurate analysis of the company's CRM strategy on the concepts refer to.
	2	The answer provides totally accurate analysis of the company's CRM strategy on the concepts refer to.
Identification of Disadvantages	0	The answer fails to identify any meaningful disadvantages.
	1	The answer correctly identifies one or two disadvantages.
	2	The answer correctly identifies more than two disadvantages.
Novelty and Feasibility of Improvements	0	The answer offers improvements that are neither novel nor feasible.
	1	The answer offers improvements that are either novel or feasible, but not both.
	2	The answer offers improvements that are both novel and feasible.
Usefulness of Improvements	0	The answer proposes improvements that do not address any key issues in the current strategy.
	1	The answer proposes improvements that address some of key issues in the current strategy.
	2	The answer proposes improvements that address all key issues in the current strategy.
Task 4-Coding		
Open-Ended Question A library has a list of available books: "Harry Potter", "The Great Gatsby", "1984″, "To Kill a Mockingbird", "Pride and Prejudice". Each case should be solved by single line code. 1) Use Python to create a list named "books" containing these titles in the exact same order. 2) Use Python to add a new book, "The Hobbit", to the end of the "books" list. 3) Use Python to remove element "1984″ from the "books" list. 4) Use Python to add "1984″ to the 3rd position of "books" list. 5) Use Python to print the name of book at the first position while deleting it from the list.		
Case1	0	The answer fails to achieve the function.
	1	The answer achieves the function with other method.
	2	The answer achieves the function with: books = ["Harry Potter", "The Great Gatsby", "1984″, "To Kill a Mockingbird", "Pride and Prejudice"]
Case2	0	The answer fails to achieve the function.
	1	The answer achieves the function with other method.
	2	The answer achieves the function with: books.append("The Hobbit")
Case3	0	The answer fails to achieve the function.
	1	The answer achieves the function with other method.
	2	The answer achieves the function with: books.remove("1984")
Case4	0	The answer fails to achieve the function.

(*continued on next page*)

Appendix C (*continued*)

Task 1-Research Methodology		
	1	The answer achieves the function with other method.
	2	The answer achieves the function with: books.insert(2,"1984")
Case5	0	The answer fails to achieve the function.
	1	The answer achieves the function with other method.
	2	The answer achieves the function with: print(books.pop(0))

Appendix D

Appendix D
Examples of Surface-Level and Deep-Level Prompt

T1. Research Methodology	
Surface-level	"What does each step in the research process mean?"
Deep-level	"Do all research fields use the same standards for internal and external validity, or do they differ?"
T2. Motivation Theory	
Surface-level	"Which level of Maslow's hierarchy relates to love?"
Deep-level	"How are Maslow's and Herzberg's theories related? Do they have any conflicting ideas?"
T3. Marketing	
Surface-level	"How do you calculate the share of customers?"
Deep-level	"How do companies stay transparent while keeping their competitive edge to build customer trust in quality?"
T4. Coding	
Surface-level	"Can a list contain multiple data type?"
Deep-level	"In which cases are pop(), del, and remove() better than others?"

Data availability

Data will be made available on request.

References

Abrami, P. C., Bernard, R. M., Borokhovski, E., Waddington, D. I., Wade, C. A., & Persson, T. (2015). Strategies for teaching students to think critically: A meta-analysis. *Review of Educational Research, 85*(2), 275–314.

Al-Hossami, E., Bunescu, R., Smith, J., & Teehan, R. (2024). Can language models employ the socratic method? Experiments with code debugging. In *Proceedings of the 55th ACM technical symposium on computer science education V. 1*.

Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological Review, 89*(4), 369.

Anderson, J. R. (2013). *The architecture of cognition*. Psychology Press.

Anderson, L. W., & Krathwohl, D. R. (2010). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman.

Anderson, J., Rainie, L., & Luchsinger, A. (2018). Artificial intelligence and the future of humans. *Pew Research Center, 10*(12).

Anderson, J. R., & Schunn, C. D. (2013). Implications of the ACT-R learning theory: No magic bullets. *Advances in instructional Psychology, 5*, 1–33. Routledge.

Andrews, J., & Higson, H. (2008). Graduate employability,'soft skills' versus 'hard' business knowledge: A European study. *Higher Education in Europe, 33*(4), 411–422.

Ayres, P., & Paas, F. (2012). In *Cognitive load theory: New directions and challenges* (Vol. 26, pp. 827–832). Wiley Online Library.

Azevedo, R., & Aleven, V. (2013). *International handbook of metacognition and learning technologies* (Vol. 26). Springer.

Baidoo-Anu, D., & Ansah, L. O. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI, 7*(1), 52–62.

Baidu. (2023). *Baidu launches ERNIE 4.0 foundation model, leading a new wave of AI-native applications*. PR Newswire. https://www.prnewswire.com/news-releases/baidu-launches-ernie-4-0-foundation-model-leading-a-new-wave-of-ai-native-applications-301958681.html.

Bereiter, C., & Scardamalia, M. (1993). *Surpassing ourselves. An inquiry into the nature and implications of expertise*. Chicago: Open Court.

Biggs, J., & Tang, C. (1999). *Teaching for quality learning at universities*. Buckingham: Open University.

Bloom, B. S. (1956). *Taxonomy of educational objectives, handbook I: The cognitive domain* (pp. 25–59). New York: David McKay Co. Inc.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (1999). *How people learn: Brain, mind, experience, and school*. National Academy Press.

Broadbent, J., Panadero, E., & Fuller-Tyszkiewicz, M. (2020). Effects of mobile-app learning diaries vs online training on specific self-regulated learning components. *Educational Technology Research & Development, 68*(5), 2351–2372.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., & Askell, A. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901.

Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction, 8*(4), 293–332.

Chevrier, M., Muis, K. R., Trevors, G. J., Pekrun, R., & Sinatra, G. M. (2019). Exploring the antecedents and consequences of epistemic emotions. *Learning and Instruction, 63*, Article 101209.

Chi, M. T., De Leeuw, N., Chiu, M.-H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18*(3), 439–477.

Chi, M. T., & VanLehn, K. A. (2012). Seeing deep structure from the interactions of surface features. *Educational Psychologist, 47*(3), 177–188.

Chi, M. T., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist, 49*(4), 219–243.

Chin, C., & Osborne, J. (2008). Students' questions: A potential resource for teaching and learning science. *Studies in Science Education, 44*(1), 1–39.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24*(1), 87–114.

De Backer, L., Van Keer, H., & Valcke, M. (2012). Exploring the potential impact of reciprocal peer tutoring on higher education students' metacognitive knowledge and regulation. *Instructional Science, 40*, 559–588.

Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest, 5*(3), 69–106.
Elder, L., & Paul, R. (2020). *Critical thinking: Tools for taking charge of your learning and your life*. Foundation for Critical Thinking.
Ennis, R. H. (1985). A logical basis for measuring critical thinking skills. *Educational Leadership, 43*(2), 44–48.
Ennis, R. H. (2015). Critical thinking: A streamlined conception. In *The Palgrave handbook of critical thinking in higher education* (pp. 31–47). Springer.
Ericsson, K. (2006). The influence of experience and deliberate practice on the development of superior expert performance. In *Cambridge handbook of expertise and expert performance*. Cambridge University.
Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review, 100*(3), 363.
Facione, P. A. (2011). Critical thinking: What it is and why it counts. *Insight assessment, 1*(1), 1–23.
Georgeff, M. P., & Lansky, A. L. (1986). Procedural knowledge. *Proceedings of the IEEE, 74*(10), 1383–1398.
Gerjets, P., Scheiter, K., & Catrambone, R. (2004). Designing instructional examples to reduce intrinsic cognitive load: Molar versus modular presentation of solution procedures. *Instructional Science, 32*, 33–58.
Goda, Y., Yamada, M., Matsukawa, H., Hata, K., & Yasunami, S. (2014). Conversation with a chatbot before an online EFL group discussion and the effects on critical thinking. *Journal of Information Systems Education, 13*(1), 1–7.
Gregorcic, B., Polverini, G., & Sarlah, A. (2024). ChatGPT as a tool for honing teachers' Socratic dialogue skills. *Physics Education, 59*(4), Article 045005.
Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309–333.
Halpern, D. F. (1998). Teaching critical thinking for transfer across domains. Dispositions, skills, structure training, and metacognitive monitoring. *American Psychologist, 53*(4), 449–455. https://doi.org/10.1037//0003-066x.53.4.449
Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112.
Hennessey, M. G. (1999). *Probing the dimensions of metacognition: Implications for conceptual change teaching-learning*.
Hollingworth, R. W., & McLoughlin, C. (2001). Developing science students' metacognitive problem solving skills online. *Australasian Journal of Educational Technology, 17*(1).
Holmes, W., Bialik, M., & Fadel, C. (2019). Artificial intelligence in education promises and implications for teaching and learning. *Center for Curriculum Redesign*, 94–169.
Hu, X., Tang, P., Zuo, S., Wang, Z., Song, B., Lou, Q., Jiao, J., & Charles, D. (2023). Evoke: Evoking critical thinking abilities in LLMs via reviewer-author prompt editing. *arXiv preprint arXiv:2310.13855*.
Izumi, K., Tanaka, H., Shidara, K., Adachi, H., Kanayama, D., Kudo, T., & Nakamura, S. (2024). Response generation for cognitive behavioral therapy with Large Language models: Comparative study with socratic questioning. *arXiv preprint arXiv:2401.15966*.
Jackson, D. (2016). Modelling graduate skill transfer from university to the workplace. *Journal of Education and Work, 29*(2), 199–231.
Jonassen, D. H. (1997). Instructional design models for well-structured and III-structured problem-solving learning outcomes. *Educational Technology Research & Development, 45*(1), 65–94.
Kaswan, K. S., Dhatterwal, J. S., & Ojha, R. P. (2024). AI in personalized learning. In *Advances in technological innovations in higher education* (pp. 103–117). CRC Press.
Kerlyl, A., Hall, P., & Bull, S. (2006). *Bringing chatbots into education: Towards natural language negotiation of open learner models. International conference on innovative techniques and applications of artificial intelligence*.
King, A. (1990). Enhancing peer interaction and learning in the classroom through reciprocal questioning. *American Educational Research Journal, 27*(4), 664–687.
King, P. M., & Kitchener, K. S. (1994). Developing reflective judgment: Understanding and promoting intellectual growth and critical thinking in adolescents and adults. *Jossey-bass higher and adult education series and jossey-bass social and behavioral science series*. ERIC.
Krathwohl, D. (2002). A revision Bloom's taxonomy: An overview. *Theory into Practice, 41*(4), 212–218.
Kuhn, D. (1999). A developmental model of critical thinking. *Educational Researcher, 28*(2), 16–46.
Lai, E. R. (2011). Critical thinking: A literature review. *Pearson's Research Reports, 6*(1), 40–41.
Lin, Z., Gou, Z., Liang, T., Luo, R., Liu, H., & Yang, Y. (2024). *CriticBench: Benchmarking LLMs for critique-correct reasoning. arXiv preprint arXiv:2402.14809*.
Luckin, R., & Holmes, W. (2016). *Intelligence unleashed: An argument for AI in education*.
Manganelli, S., Cavicchiolo, E., Mallia, L., Biasi, V., Lucidi, F., & Alivernini, F. (2019). The interplay between self-determined motivation, self-regulated cognitive strategies, and prior achievement in predicting academic performance. *Educational Psychology, 39*(4), 470–488.
Mayer, R. E. (2002). Rote versus meaningful learning. *Theory into Practice, 41*(4), 226–232.
Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist, 38*(1), 43–52.
Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*(2), 81.
Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist, 38*(1), 1–4.
Paas, F. G., & Van Merriënboer, J. J. (1994). Instructional control of cognitive load in the training of complex cognitive tasks. *Educational Psychology Review, 6*, 351–371.
Paul, R., & Elder, L. (2019). *The miniature guide to critical thinking concepts and tools*. Rowman & Littlefield.
Pérez, J. Q., Daradoumis, T., & Puig, J. M. M. (2020). Rediscovering the use of chatbots in education: A systematic literature review. *Computer Applications in Engineering Education, 28*(6), 1549–1565.
Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog, 1*(8), 9.
Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research, 21*(140), 1–67.
Ranellucci, J., Hall, N. C., & Goetz, T. (2015). Achievement goals, emotions, learning, and performance: A process model. *Motivation Science, 1*(2), 98.
Resnick, L. B. (1987). *Education and learning to think*. The National Academies Press. https://doi.org/10.17226/1032
Rittle-Johnson, B., Siegler, R. S., & Alibali, M. W. (2001). Developing conceptual understanding and procedural skill in mathematics: An iterative process. *Journal of Educational Psychology, 93*(2), 346.
Roll, I., Aleven, V., McLaren, B. M., & Koedinger, K. R. (2011). Improving students' help-seeking skills using metacognitive feedback in an intelligent tutoring system. *Learning and Instruction, 21*(2), 267–280.
Roll, I., & Wylie, R. (2016). Evolution and revolution in artificial intelligence in education. *International Journal of Artificial Intelligence in Education, 26*, 582–599.
Ryle, G., & Tanney, J. (2009). *The concept of mind*. Routledge.
Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review, 84*(1), 1.
Shridhar, K., Macina, J., El-Assady, M., Sinha, T., Kapur, M., & Sachan, M. (2022). Automatic generation of socratic subquestions for teaching math word problems. *arXiv preprint arXiv:2211.12835*.
Smutny, P., & Schreiberova, P. (2020). Chatbots for learning: A review of educational chatbots for the facebook messenger. *Computers & Education, 151*, Article 103862.
Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*(2), 257–285.
Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction, 4*(4), 295–312.
Sweller, J. (2010). Cognitive load theory: Recent theoretical advances. In *Cognitive load theory*. Cambridge University Press.
Sweller, J. (2011). *Cognitive load theory*. Springer.
Sweller, J., Van Merrienboer, J. J., & Paas, F. G. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*, 251–296.
Ten Berge, T., & Van Hezewijk, R. (1999). Procedural and declarative knowledge: An evolutionary perspective. *Theory & Psychology, 9*(5), 605–624.
Van Gog, T., Paas, F., & Sweller, J. (2010). Cognitive load theory: Advances in research on worked examples, animations, and cognitive load measurement. *Educational Psychology Review, 22*, 375–378.
Van Merriënboer, J. J., & Kirschner, P. A. (2017). *Ten steps to complex learning: A systematic approach to four-component instructional design*. Routledge.

Van Merrienboer, J. J., & Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational Psychology Review, 17*, 147–177.
VanLehn, K. (1996). Cognitive skill acquisition. *Annual Review of Psychology, 47*(1), 513–539.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*.
Wallace, R. S. (2009). *The anatomy of ALICE*. Springer.
Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM, 9*(1), 36–45.
Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education–where are the educators? *International Journal of Educational Technology in Higher Education, 16*(1), 1–27.