Beyond Task Performance: Exploring the Effect of Generative AI Cognitive and Metacognitive Adaptive Scaffolding in Argumentative Writing

Xuanyu Chen (xuanyuchen23@m.fudan.edu.cn)
Institute of Higher Education, Fudan University, Shanghai, China

"But what if the opposite ends up happening, and AI takes on all the fun stuff?"
—Aki Ito, The End of Coding as We Know It, Business Insider, April 26, 2023

Abstract: Generative AI (GenAI) presents a performance-learning paradox in argumentative writing, often improving observable performance while undermining engagement for meaningful learning. This study investigates whether theoretically-grounded scaffolding can mitigate this paradox by reframing GenAI from an answer-provider to a collaborative partner. We designed two scaffolding frameworks: *cognitive scaffolding* improving argumentative writing quality (evidence, ideas, structure, language) and *metacognitive scaffolding* promoting self-regulated learning (forethought, monitoring, control, reflection) with support delivered adaptively through a progressive system of Questions, Hints, and Prompts. This study employs a randomized controlled, between-subjects experiment to compare performance-oriented (task performance, learning outcomes) as well as process-oriented (learning engagement, self-regulated learning) measures across four conditions: (1) baseline GenAI support (i.e., no cognitive or metacognitive scaffolding), (2) GenAI adaptive cognitive scaffolding, (3) GenAI adaptive metacognitive scaffolding, and (4) GenAI adaptive cognitive and metacognitive scaffolding. In doing so, we aim to provide evidence on designing hybrid intelligence systems that foster meaningful learning processes and outcomes under AI-enhanced learning environments.

Keywords: Generative AI, Scaffolding, Argumentative Writing, Human-AI Collaboration

1. Introduction

1.1 Argumentative Writing

Argumentation is widely recognized as a critical skill in higher education, essential for developing critical thinking, scientific reasoning, and domain-specific knowledge (Fan & Chen, 2021). This is commonly fostered through argumentative writing tasks, which require students to formulate a clear position, justify claims with evidence, and respond to counterarguments within a structured discourse (Noroozi et al., 2016, 2023). However, producing high-quality argumentative writing presents significant cognitive and metacognitive demands for students (Noroozi, 2018; Noroozi & Hatami, 2019). Furthermore, to enhance student learning engagement and performance, teachers have to monitor the writing process, provide adaptive scaffolding, and encourage learners to engage in reflection, which is extremely challenging in large classes (Fan & Chen, 2021; Hu et al., 2025). Therefore, embedding real-time, personalized support in technology-enhanced learning environments emerges as a promising solution for learning and evaluation in argumentative writing (Sikström et al., 2024).

1.2 Generative AI in Argumentative Writing

Throughout human history, successive waves of technological innovation have persistently automated tasks that once defined the limits of human capabilities (Acemoglu, 2002; Mokyr, 1992). The contemporary wave is distinguished by the emergence of generative artificial intelligence (GenAI), a new subset of AI powered by large language models (LLMs), which has shown positive impacts on productivity with its capabilities in generating meaningful, human-like content such as text, images, or audio in response to varied, task-specific prompts (Brynjolfsson et al., 2025; Eloundou et al., 2024; Feuerriegel et al., 2024). Yet, the rapid diffusion of GenAI has raised concerns about the potential erosion of employment, skill development, knowledge work, and human cognition in general (Autor et al., 2024; Dwivedi et al., 2021; Frank et al., 2019; Mokyr et al., 2015). This tension is particularly pronounced in education, where AI-powered technologies are increasingly embedded in instructional designs and learning environments, such as intelligent tutoring systems, feedback chatbots, and adaptive learning platforms to deliver real-time responses, personalized feedback, and effective visualizations, outperforming human teachers on various educational tasks (Chiu, 2024; Huang et al., 2022; Kasneci et al., 2023; Kulik & Fletcher, 2016; Yan, Sha, et al., 2024). Together, these applications have already transformed education from in-class instruction toward personalized learning and tutoring (Xie et al., 2019).

One of the key contexts of personalized learning is in argumentative writing, where many existing studies have provided robust yet narrowly focused evidence regarding “AI as a Tool” (Milano et al., 2023; Ouyang & Zhang, 2024), particularly using AI technologies as automated writing evaluation (AWE) tools for support in idea generation, grammatical correction, and even full-text generation (Ding & Zou, 2024; Fu et al., 2024; Li, 2023; Wang et al., 2020; Yeung, 2025). Similarly, a wide range of studies has documented the capacity of GenAI to enhance task performance across various contexts, such as creative ideation (Doshi & Hauser, 2024; Lee & Chung, 2024), writing (Guo et al., 2022; Noy & Zhang, 2023; Zhang et al., 2025), programming (Huang et al., 2025; Liu & Li, 2024), and scientific inquiry (Bianchini et al., 2022; Musslick et al., 2025; Ochoa et al., 2025; Stadler et al., 2024; Van Quaquebeke et al., 2025). While GenAI technologies have demonstrably boosted learners’ task performance in writing, often measured through immediate output quality or efficiency, such superficial outcomes do not fully capture the deeper learning processes at play (Chen et al., 2025).

1.3 Generative AI-Driven Adaptive Scaffolding

Scaffolding, a concept that originated in socio-cultural theory, refers to the support provided by a more knowledgeable individual to help learners accomplish tasks that they would otherwise not be able to achieve independently, which eventually fades away as learners no longer need it (Vygotsky, 1978; Wood et al., 1976). It aims to maximize learning outcomes through contingency and gradual transfers of responsibility, which involves providing responsive, tailored, and adjusted support that engages learners in cognitive or metacognitive activities independently, rather than simply providing them with solutions (Darvishi et al., 2024; Lo et al., 2024; Suriano et al., 2025; van de Pol et al., 2010). However, the effect of scaffolding depends on both scaffolding types and learner characteristics (Kim & Lim, 2019). Scaffolding types are identified through interactions between learners and tutors (planned, adaptive) (Azevedo et al., 2008; Saye & Brush, 2002), sources and examples presented (peer, teacher, technology) (Kim & Hannafin, 2011), and the function (implicit, explicit) (Hadwin & Winne, 2001) or the purpose of using scaffolding (conceptual, metacognitive, procedural, strategic) (Cagiltay, 2006). Specifically, cognitive (or supportive) scaffolding provides learners with domain knowledge and guides them on what to consider and how to associate ideas, whereas metacognitive (or reflective) scaffolding provides learners with metacognitive questions and helps them clarify their reflective processes (Cagiltay, 2006). On the other hand, learner characteristics include individual differences in knowledge, skills, and attitudes that shape intelligence, cognitive patterns, learning styles, and preference, where metacognition, a term proposed as “cognition about cognition” (Zimmerman, 2008), stands out as essential in that metacognitive skills (i.e., planning, execution, evaluation) predicts meaningful learning (Callan et al., 2021).

In contrast with planned scaffolding, adaptive scaffolding refers to dynamic, responsive instructional support that adjusts in real time based on students' ongoing performance and needs, which provides immediate, tailored guidance that addresses emerging challenges in the learning process, compared to planned scaffolding (Lim et al., 2024; Roll & Wylie, 2016; Zheng et al., 2023). Recent advances in Large Language Models (LLMs) have made adaptive scaffolding more scalable and context-sensitive, driven by generative AI, which can interpret natural language, identify learner misconceptions, and offer targeted, real-time support and feedback (Wu et al., 2025). In the context of AI-enhanced learning environments, these foundational principles of scaffolding provide a critical framework for designing AI that evolves from an automated assistant into a collaborative partner (Stojanov, 2023; Suraworachet et al., 2023; Vaccaro et al., 2024). Prior research suggests that when GenAI is used as a scaffold, for example, as a conversational agent that asks prompting questions or provides hints, it can actually enhance student engagement in writing tasks (Hu et al., 2025). By maintaining cognitive demand and requiring learners' input, such scaffolding facilitates effective human-AI collaboration in learning towards hybrid intelligence, where human capabilities and machine intelligence are complemented and augmented to achieve goals unreachable by either alone, representing the ideal state of this partnership (Cremer & Kasparov, 2021; Glickman & Sharot, 2024; Jarrahi et al., 2022; Järvelä et al., 2023; Raisch & Fomina, 2025). However, research into hybrid intelligence is still nascent, and the field particularly lacks deep insights into how to design AI systems as effective scaffolds for learning (Liu et al., 2024; Nguyen et al., 2024; Yan et al., 2025; Yan, Greiff, et al., 2024). Therefore, this study addresses this issue by systematically designing and evaluating the effects of cognitive and metacognitive adaptive scaffolding on the key learning processes that underpin true understanding.

1.4 Learning Engagement and Outcomes with AI Support

The widespread adoption of these powerful, yet performance-oriented, GenAI support gives rise to a critical challenge known as the performance-learning paradox, where performance refers to observable behaviors during task execution, whereas learning is the enduring retention and transfer of knowledge resulting from experience (Soderstrom & Bjork, 2015). The paradox, therefore, is that while GenAI can elevate immediate task performance, it can be a misleading proxy for learning if students are not actively engaged in the process (Darvishi et al., 2024; Fan et al., 2025; Stadler et al., 2024). This highlights the concept of learning engagement, including behavioral, cognitive, and emotional engagement, as a crucial intermediary between using GenAI and achieving meaningful learning outcomes (Chen, 2017; Hu & Hui, 2012; Lee et al., 2022). Specifically, a student might produce a well-structured essay with AI support yet remain passive or detached during its creation, indicating that strong performance alone provides little evidence about the quality of learning (Lo et al., 2024).

A primary mechanism underlying the learning-performance paradox is over-reliance, whereby learners accept AI-generated outputs without adequate critical evaluation, which is extremely evident in text generation tasks as argumentative writing (Barrot, 2023; Liu et al., 2024; Zhai et al., 2024). Behaviorally, easy access to AI-generated answers can lead students to reduce effort and time on task. Cognitively, it manifests as excessive cognitive offloading, in which learners delegate substantial cognitive demands to external tools (Risko & Gilbert, 2016). Emotionally, it fosters an illusion of competence and false self-efficacy, where the immediate satisfaction of generating high-quality text masks the lack of genuine personal investment and ownership over the content (Draxler et al., 2024). Appropriate cognitive offloading involves technologies as an active extension of human cognition and brain mechanisms (Chiriatti et al., 2024). Although strategic offloading can enhance task efficiency, it incurs notable costs, including diminished internal memory retention and the circumvention of desirable difficulties that foster deep processing, ultimately privileging superficial task completion over authentic knowledge construction, thus resulting in the erosion of cognition (Heersmink, 2024; Kosmyna et al., 2025;

Richmond & Taylor, 2025; Yan, Greiff, et al., 2024). The problem extends to metacognition as well. When learners become mere onlookers to AI-provided solutions, they can be detached from the learning experience, with reduced planning, monitoring, or sense of control in the self-regulated learning (Fan et al., 2025). Over time, over-reliance on GenAI may cultivate habitual avoidance of effortful cognition, resulting in the emergence and reinforcement of metacognitive laziness, a state characterized by reduced engagement in self-regulatory processes and a preference for intuitive, low-effort decision-making (Fan et al., 2025; Zhan & Yan, 2025). Within the context of argumentative writing, such over-reliance on GenAI is evident as students struggle to balance using AI to improve writing while remaining authentic, often resulting in diminished engagement and superficial reflection (Chan & Lee, 2025). In contrast, students who are behaviorally engaged (e.g., revising and iterating on their essays) and cognitively engaged (e.g., deliberating on arguments, evaluating AI feedback critically) tend to demonstrate stronger understanding and skill development in argumentation. Consequently, GenAI may inadvertently exacerbate superficial engagement, shallow cognitive processing, and metacognitive laziness, thereby undermining meaningful and enduring learning outcomes (Bastani et al., 2025; Bauer et al., 2025). These insights underscore the need for well-designed cognitive and metacognitive scaffolding around generative AI tools to counteract disengagement and promote active learning. Rather than allowing AI to become a convenient crutch that students lean on, educators and designers could design, implement, and evaluate AI systems that effectively engage learners, augment cognition, and promote metacognition towards hybrid intelligence.

2. Research Questions

This study investigates the impact of generative AI-driven scaffolding on argumentative writing. Moving beyond mere task performance, we aim to reveal the effects of GenAI-based support on learners' cognitive and metacognitive processes (2 × 2 design, see Figure 1). Specifically, this research examines how GenAI-driven adaptive cognitive and metacognitive scaffolding shape not only argumentative writing quality but also the underlying learning processes. In doing so, we contribute to a deeper understanding of effective human-AI collaboration in education, conceptualizing learning as a dynamic interplay within a hybrid intelligence system.

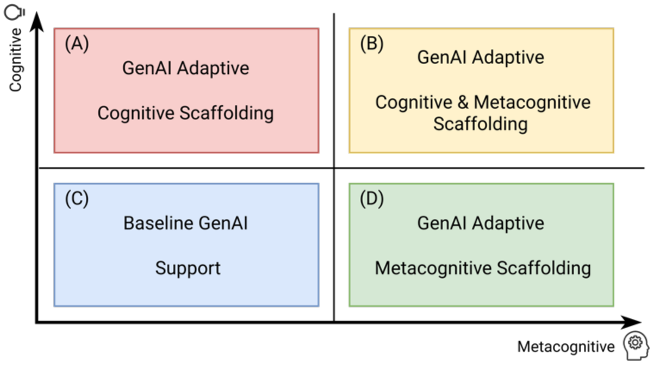

Figure 1

Generative AI-Driven Scaffolding Design

RQ1 (Performance-Oriented Outcomes): What are the effects of GenAI-driven adaptive scaffolding (cognitive and metacognitive) on *task performance* and *learning outcomes* in argumentative writing?

RQ1a: What is the impact of GenAI-driven adaptive scaffolding on *task performance*?

RQ1b: What is the impact of GenAI-driven adaptive scaffolding on *learning outcomes*?

RQ2 (Process-Oriented Outcomes): What are the effects of GenAI-driven adaptive scaffolding (cognitive and metacognitive) on *learning engagement* and *self-regulated learning processes*? How do they explain the *Performance-Oriented Outcomes* in argumentative writing?

RQ2a: How is *learning engagement* influenced by GenAI-driven adaptive scaffolding?

RQ2b: How is *self-regulated learning* influenced by GenAI-driven adaptive scaffolding?

RQ2c: To what extent do process-oriented outcomes (*learning engagement* and *self-regulated learning*) explain the variations in performance-oriented outcomes (*task performance* and *learning outcomes*)?

3. Research Design

This study follows a randomized, between-subjects experimental design with four conditions: 1) baseline GenAI support (i.e., no cognitive or metacognitive scaffolding), 2) GenAI adaptive cognitive scaffolding, 3) GenAI adaptive metacognitive scaffolding, and 4) GenAI adaptive cognitive and metacognitive scaffolding. Cognitive and metacognitive scaffolding are delivered at three guidance levels, including Questions, Hints, and Prompts, with adaptive triggering rules. Participants complete a pretest, an argumentative writing task with revision, an immediate posttest, and a two-week delayed test. Primary outcome measures are task performance (analytic rubric), learning outcomes (retention and transfer), learning engagement (behavioral logs and questionnaire), and self-regulated learning (chat history, behavioral logs, and questionnaire), where we combine self-report surveys with observable system log data. Randomization is stratified by prior knowledge and English proficiency, where essay raters are experts, blind to conditions.

3.1 Participants

Participants will engage in an English reading and argumentative writing task, who are randomized into the following control and experimental conditions: 1) Control, where participants receive baseline GenAI support only; 2) Adaptive Cognitive Scaffolding, where participants receive GenAI support with adaptive cognitive scaffolding; 3) Adaptive Metacognitive Scaffolding, where participants receive GenAI support with adaptive metacognitive scaffolding; 4) Integrated Scaffolding, where participants receive GenAI support with both adaptive cognitive and metacognitive scaffolding. A priori power analysis was conducted to determine the necessary sample size of 180 participants ($\alpha = 0.5, power = 0.8, Cohen's\ f = 0.25$). Given potential attrition, we intend to recruit 200 participants (50 per condition).

3.2 Procedures

Experiments are conducted in laboratory settings where participants complete tasks in several phases as follows (see Figure 2): **1) Pre-task Survey:** participants complete knowledge tests on their prior knowledge of the related field of the reading materials and are surveyed in terms of their demographics (e.g., age, gender, years of schooling), English proficiency (e.g., years of English learning), and prior experience with AI writing tools (e.g., Grammarly, ChatGPT); **2) Instruction:** participants receive instructions on the user interface with learning contents and tools, task requirements, and especially how they should seek help from the AI chatbot. They will then be told that they will receive a fixed participation payment plus a small bonus for process adherence (e.g., engaging with scaffolds as instructed), rather than final essay quality alone; **3) Reading and Writing 1:** participants complete their argumentative writing based on a series of reading materials; **4) Reading and Writing 2:** participants complete their essay with or without generative AI-driven adaptive scaffolding; **5) Immediate Post-Task Survey:** participants complete knowledge tests, knowledge transfer tests, and fill in surveys on learning engagement (i.e., cognitive, behavioral, emotional); **6) Delayed Post-Task Survey:** after two weeks, participants complete knowledge tests and knowledge transfer tests.

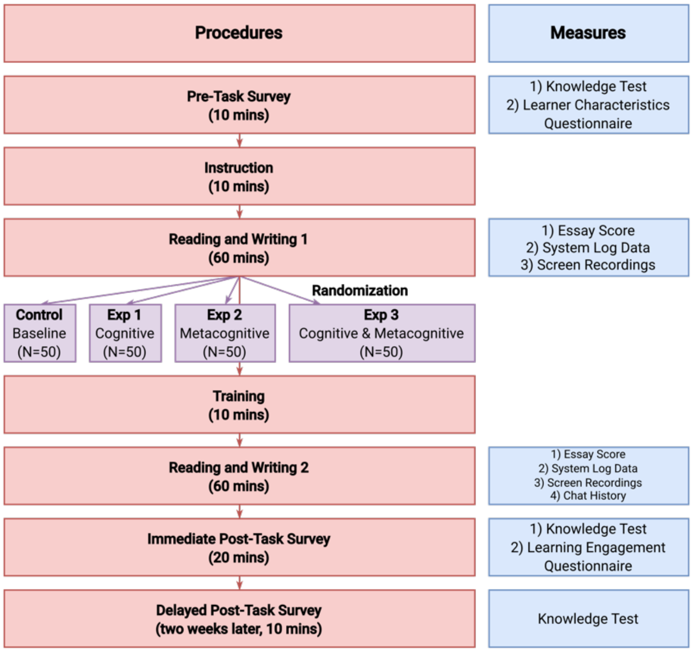

Figure 2

Experimental Design

3.3 Argumentative Writing Process

Argumentative writing is learned as a staged, scaffolded process in which learners progress from generating ideas to refining a final draft. This study consists of five stages, including Ideation, Identifying Resources, Thesis Statement, Drafting (Introduction, Body, and Conclusion), and Revision. Learners are guided through the process with our system providing scaffolding at each stage to ensure a personal and immersive writing experience. Specifically, in the ideation stage, students brainstorm and explore multiple perspectives on the topic (Graham et al., 2015). Next, in identifying resources, learners gather credible evidence and examples to support their claims (Latifi & Noroozi, 2021). With ideas and evidence at hand, learners craft a thesis statement that concisely expresses their main claim. After establishing a thesis, students engage in outline building, organizing their reasoning into a logical structure (Fan & Chen, 2021). During drafting, writers develop the content through an introduction paragraph, body paragraphs, and a conclusion paragraph. The introduction should hook the reader, provide background, and state the thesis, while each body paragraph should present a distinct point or sub-argument supported by evidence and reasoning (and address counterarguments), and the conclusion should synthesize the discussion and reaffirm the thesis in light of the evidence. Scaffolding in this drafting stage often takes the form of guided questions or checklists: for example, prompts may ensure the introduction establishes context and a clear position, each body paragraph contains a topic sentence, relevant evidence, an explanation linking the evidence to the claim, and (where appropriate) a consideration of opposing views, and that the conclusion ties together the argument and its broader implications. Finally, the revision stage requires learners to review and refine their draft (Noroozi et al., 2016). In sum, guiding learners through these stages with appropriate cognitive and metacognitive scaffolds not only improves the structure and clarity of their arguments but also actively engages them in critical thinking and self-regulation throughout the writing process. Such a scaffolded stage-wise approach to argumentative writing has been found to foster better argumentation skills and outcomes in learners (Hu et al., 2025).

3.4 Scaffolding Design Principles

Scaffolding in this study is categorized into cognitive scaffolding and metacognitive scaffolding. Cognitive scaffolding (see Appendix Table 1) is based on existing analytic frameworks on argumentative writing that facilitate cognitive processing and knowledge construction (Hu et al., 2025; Paas et al., 2003; van Nooijen et al., 2024). For the content of cognitive scaffolding, we adapted the analytic framework from Booth Olson et al. (2023). Metacognitive scaffolding (see Appendix Table 2) is grounded in Self-Regulated Learning (SRL) theory that encompasses forethought, monitoring, control, and reflection (Fan et al., 2025; Pintrich, 2000; Schunk, 2005; Zheng et al., 2019). For the form of scaffolding, inspired by previous researches that distinguish scaffolding based on levels of control, this study designed three levels of scaffolding: Questions (low-control), Hints (medium-control), and Prompts (high-control) (Hu et al., 2025; Ouyang et al., 2022; van de Pol et al., 2010, 2014). Triggering rules are operationalized as follows: Questions are issued after 90 seconds of inactivity or two consecutive low-yield turns; Hints are issued when a prior Question elicits no revision; and Prompts are issued only if the prior Question and Hint fail to change the draft.

3.5 System Architecture

To deliver the dynamic cognitive and metacognitive scaffolding outlined in our design principles, we aim to design a learning environment, the Adaptive Scaffolding for Argumentative Writing (ASAW) system, an LLM-based agent framework developed to provide real-time, evidence-based, and context-aware scaffolding (see Figure 3). The system is built on a robust client-server architecture, comprising a user-facing Frontend and a powerful data-processing Backend. These components work in concert through three core, integrated modules: a database, a controller agent, and a scaffolding agent, integrated with a user interface to support interaction. These components were designed to achieve the principles of scaffolding. To address contingency, this study develops a comprehensive database that encodes learners' behavioral log data into three parts: Stage, Dimension, and Form. Each entry includes names, descriptions, examples, and mappings between dimensions and forms. This structured organization allows the system to retrieve relevant forms dynamically, providing contextually accurate scaffolding aligned with educational practices. The controller agent supports the principle of transfer of responsibility and fading. When a student inputs a query, the controller agent analyzes the chat history and performs three tasks: determining the current argumentative writing stage, updating the dimension, and identifying the most needed improvement dimension along with the corresponding best form of scaffolding. These structured parameters guide the agent's subsequent actions. The scaffolding generator agent uses these parameters to generate real-time, context-specific scaffolding. Prompts include the persona, stage-specific objectives, and detailed descriptions of the form of scaffolding. This ensures the guidance is not only informative but also engaging and supportive. The system is deployed on a web-based platform, enabling real-time text interactions between learners and the AI systems. By combining these components, it dynamically adapts to learners' needs, providing personalized and structured scaffolding throughout argumentative writing while fostering learning engagement and outcomes.

The implementation combines frontend, backend, and data processing technologies to ensure seamless functionality. We utilized the Streamlit framework for rapid development of interactive web applications, enabling an integrated workflow that combines frontend and backend development directly from Python scripts. For data management, we adopted Google Sheets as a serverless backend to store user information, conversation logs, and other critical data in real time. Google Sheets is synchronized with the agent via API, ensuring smooth data retrieval and updates. User interactions and system responses are logged and processed dynamically, enabling continuous refinement of recommendations. This approach offers both scalability and flexibility, simplifying system updates and maintenance. The conversational AI capabilities are powered by OpenAI's GPT-5 API, integrated with Langchain to enhance the structure and flow of interactions. Using Retrieval-Augmented Generation (RAG) technology, these examples are dynamically retrieved during real-time interactions. The controller agent processes user input and determines contextual parameters. It then queries the database for relevant case examples, integrates them with the user's context, and inputs this structured information into the LLM. This process ensures that the guidance remains both tailored and contextually relevant. The system is deployed on a cloud server, ensuring accessibility via web browsers.

3.5.1 Frontend. The frontend allows the user to interact with our system, including chatting with LLMs, asking questions, and receiving scaffolding. Several tools and templates can be used to build Chatbot applications. We aim to develop the interface with an intuitive *Vue.js* framework, *Nuxt*.

3.5.2 Backend. The Backend serves as the system's analytical core, where learner data is processed to generate real-time, adaptive responses. It also handles communication between the front end and the database. All these communications are achieved with the backend APIs such as *FastAPI* and *LangChain*. For memory persistence, we use *LangChain* to store short-term memory and a database to store long-term memory. *LangChain* also handles the modularity so that at each scaffolding stage, our system generates in-context responses and takes corresponding actions. Specifically, it is composed of two key modules:

Module A Context Database functions as the system's dynamic memory. It aggregates and synthesizes multiple real-time data streams, including the learner's chat history with the AI, behavioral log data (e.g., clicks, time-on-task, revisions), and the static learning materials (e.g., source texts, scoring rubrics) to build a comprehensive, up-to-the-moment profile of the learner's state and needs.

Module B Multi-Agent AI System is the system's intelligence layer, powered by a state-of-the-art large language model (e.g., GPT-5). To manage the intricate workflows required for adaptive scaffolding, this module employs a coordinated system of specialized agents, each with a distinct responsibility (see Figure 4):

Agent a (Controller). This agent is responsible for the retrieval and synthesis of static contextual data. It analyzes the learning materials, the argumentative writing prompt, and the scoring rubrics to provide a foundational understanding of the task's requirements and content domain. It also focuses on dynamic, real-time user data. It continuously processes the learner's chat history and behavioral logs from the Context Database to construct an evolving model of the student's progress, current struggles, and interaction patterns.

Agent b (Scaffolding Provider). This agent acts as the "executive" of the module. It synthesizes the static task context provided by Agent a. This synthesized information becomes the critical context fed into a Retrieval-Augmented Generation (RAG) framework. As noted by Lo et al., 2024, RAG is a crucial technique for optimizing the educational effects of GenAI. By leveraging this rich, real-time context, the agent generates the precise, pedagogically-informed cognitive and metacognitive scaffolding prompts (as detailed in Tables 1 and 2) that are delivered to the user. This multi-agent RAG approach ensures that the support is not only context-aware but also factually grounded, effectively minimizing model hallucination.

Module C User Interface, the Frontend, which provides a seamless and integrated learning workspace. The interface is organized into three distinct panels: a Learning Contents Panel for reading source materials, an Argumentative Writing Panel where the student composes their essay, and an AI Chatbot Panel. To simulate an authentic digital writing environment and support learners' workflow, the interface is further equipped with a suite of supplementary tools: *Note* for drafting ideas, *Timer* to manage task duration, *Annotation* for highlighting source texts, a *Search* for finding content across materials, and *Navigation* for easily moving between sections.

The tailored cognitive and metacognitive scaffolds generated by the Backend are delivered to the learner through natural, conversational prompts within the AI Chatbot Panel, creating a responsive and interactive dialogue that guides and supports their writing process.

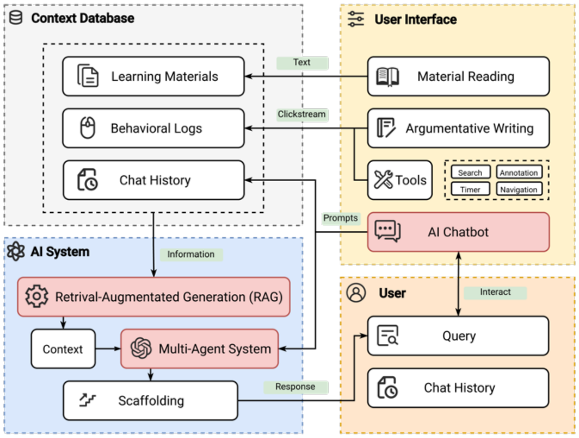

Figure 3

System Architecture

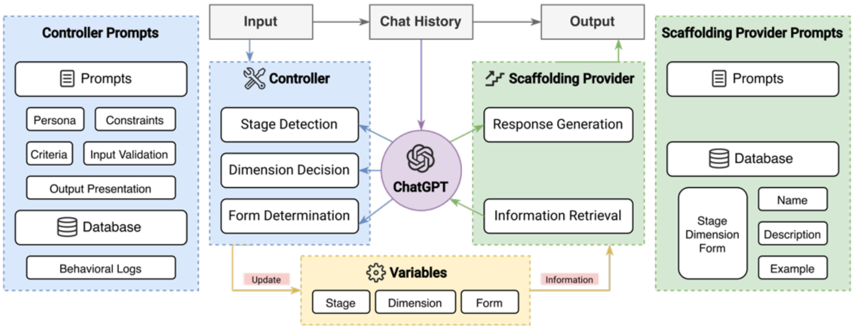

Figure 4

Multi-Agent System Framework

3.5.3 Prompt Engineering. To design effective scaffolding experiences, we employed various prompt engineering techniques: **1) Persona**. This approach shaped the responses to align with the role of a mentor of argumentative writing, enhancing performance on high-openness tasks with scaffolding. For example, prompts explicitly instructed the model to “Act as an argumentative writing mentor” and provided some characteristics of a writing mentor. This role-based prompting was intended to foster a scaffolding dynamic that supports learning without overriding student autonomy; **2) Constraints**. Incorporating constraints in prompts, such as explicitly instructing the AI not to rewrite students’ work, ensured that feedback remained within educational boundaries. For instance, one prompt stated: “You must not suggest any ideas or examples for the essay,” reinforcing the AI’s role as a scaffolding provider rather than a content generator; **3) Criteria-Based Scaffolding.** We combined criteria-based scaffolding to provide targeted guidance. For example, prompts instructed the AI to assess the “Use of Evidence” based on criteria such as Evidence, Commentary, and Balance; **4) Output Presentation.** We designed the output structure to reduce cognitive load and enhance clarity. Prompts instructed the AI to break down feedback into sections based on different criteria (e.g., coherence, cohesion, clarity) and use simple language. One prompt stated: "Provide your response on the criteria in this order: spelling, grammar, and punctuation"; and **5) Input Validation.** Input validation ensures that students engage meaningfully with the tool. A prompt included: "If the user does not type any paragraph or just random text, please direct them to type the paragraph." This technique helps maintain the integrity of student submissions and ensures constructive interaction with the tool.

3.5.4 Database. A MySQL database stores chat history (long-term memory) and user information. We chose MySQL as our database since all of the data is structured. Three tables are created: users, messages, and conversations, to store user information, message history, and conversation history. User interactions with the system are stored for user studies. Database operations are achieved by the Create, Read, Update, and Delete (CRUD) API. When LLM needs to reference long-term history, a query will be executed to retrieve relevant information from the database.

3.6 Measures

To examine the effect of GenAI adaptive scaffolding in argumentative writing processes and outcomes, this study will collect data on task performance, learning outcomes, and learning engagement, where each construct will be operationalized and measured using validated, theory-grounded instruments as described below.

3.6.1 Task Performance. Task performance will be measured by assessing the quality of the original and revised argumentative essays produced by participants. The essays will be collected and scored by two independent expert raters, blind to experimental conditions. The assessment will be guided by the Scoring Rubrics for Argumentative Writing (see Appendix Table 3). This rubric, adapted from the framework that evaluates essays across four key dimensions: Use of Evidence, Idea, Structure, and Language Use (Hu et al., 2025; Zou & Xie, 2019). To ensure consistency and reliability of the scoring, inter-rater reliability will be calculated using the Intraclass Correlation Coefficient (ICC). A high level of agreement (ICC > .80) will be established during a training and calibration phase. Any discrepancies in scores between the two raters on the final dataset will be resolved through discussion until a consensus is reached. The final, agreed-upon score for each essay will serve as the primary measure of task performance.

3.6.2 Learning Outcomes. Learning outcomes (e.g., knowledge retention and transfer) will be assessed using designed knowledge tests administered at three time points: during the Pre-task Survey, the Post-Task Survey 1, and the Post-Task Survey 2 (delayed). These tests are designed to measure both the acquisition of factual knowledge and the ability to apply it. The tests consist of two distinct sections, including knowledge retention and transfer. Knowledge Retention Test includes questions (e.g., multiple-choice, short-answer) that measure participants' understanding of core concepts and facts presented directly within the provided reading materials. This component assesses the direct acquisition of "original knowledge." The Knowledge Transfer Test contains questions that require participants to apply the principles learned from the reading materials to new problems or novel scenarios. These scenarios are on the same topic but are not explicitly covered in the texts, thereby measuring the participants' ability to generalize and transfer their knowledge. Learning gains will be calculated by comparing the scores from the post-task and delayed tests to the baseline scores from pre-task tests.

3.6.3 Learning Engagement. Learning engagement will be measured through screen recordings, eye-tracking indicators, and a self-report Learning Engagement Questionnaire, which will be administered during the Post-Task Survey 1. This instrument is adapted from an established, validated framework specifically targeting task engagement in English essay writing (see Appendix Table 4; Zare, 2023). It is structured to measure three widely adopted, interrelated dimensions of learning engagement, including behavioral, cognitive, and emotional engagement (Fredricks et al., 2004; Henrie et al., 2015). Participants will respond to statements on a 7-point Likert scale, ranging from "Strongly Disagree" to "Strongly Agree." Also, to corroborate self-reports, screen recordings will be captured to quantify learning engagement objectively (Henrie et al., 2015). Specifically, behavioral engagement is operationalized through *Time on Task* of key activities, including reading materials, drafting, revising (e.g., adding, deleting), and interacting with AI (Fleckenstein et al., 2024), while cognitive engagement is captured through *Writing Strategy*, such as specific instances in reasoning, revising, and editing (see Appendix Table 5; Lee, 2020). Given the dynamic nature of cognitive engagement and the eye-mind-engagement assumption (Miller, 2015), we adopt a multimodal learning analytics approach (Ochoa et al., 2022) and complement screen behaviors with both gaze-based (fixation duration) and pupil-based (pupil diameter) metrics to assess attention and absorption as two key components for cognitive engagement (Sankar et al., 2025). Specifically, we define each part of the user interface as an area of interest (i.e., reading, writing, tools, AI chatbot, see Figure 3) and use a wearable 180-Hz binocular eye-tracking device to capture eye movement indicators in terms of temporal, spatial, and count (see Appendix Table 6; Liu & Cui, 2025).

3.6.4 Self-Regulated Learning. To investigate participants' self-regulated learning behavior, we combine self-report questionnaires with system logs. A related questionnaire in writing tasks was adapted (Sitzmann & Ely, 2011; Suraworachet et al., 2023), consisting of four primary dimensions (see Appendix Table 7), including goal-setting (GS, 4 items; Barnard et al., 2009), persistence (P, 10 items; Warr & Downing, 2000), effort (E, 2 items; Brown, 2001), and self-efficacy (SE, 9 items; Pintrich & De Groot, 1990). Also, behavioral logs and chat history are extracted to measure self-regulated learning. For behavioral logs, we recorded the learners' learning trace data through click streams on the system, and we translated trace events into identifiable learning actions using the action library to label raw log data with meaningful behavioral indicators for self-regulated learning (see Appendix Table 8; Fan et al., 2022). For chat history, we used the GPH Taxonomy to code the conversation between learners and the AI, which defines 18 types of questions that relate to self-regulated learning (see Appendix Table 9; Cheng et al., 2025).

4. Research Timeline

This study follows a systematic two-year timeline, commencing in September 2026 and concluding in August 2028. The work is organized into four phases, which are from design and development to piloting and instrument validation, to large-scale experimental execution, and finally to analysis and dissemination (see Table 1).

4.1 Phase I: Preparation, Protocol Design, and System Development (Sep 2026 – Apr 2027)

The project begins with refining the theoretical framework and securing ethical approval. From September to November 2026, I will complete a focused *literature review* to ensure that the cognitive and metacognitive scaffolding design principles, as well as the study's intended contributions, are aligned with the latest developments in AI-enhanced learning environments. In parallel, I will finalize the experimental protocol and submit the *Institutional Review Board (IRB) application* to ensure full compliance with human-subject research requirements. From December 2026 to April 2027, efforts will shift to developing the Adaptive Scaffolding for Argumentative Writing (ASAW) system, including a multi-agent backend (e.g., LangChain- and RAG-based components) and a Vue.js frontend. This phase culminates in an instrumented *prototype* capable of delivering real-time, context-sensitive prompts and logging key behavioral and interaction data.

4.2 Phase II: Pilot Testing and Instrument Validation (May 2027 – Aug 2027)

With a system prototype, the summer of 2027 will be for *pilot testing* and *system refinement*. In May 2027, a small-scale pilot study (N = 20) will be conducted to evaluate the system's technical stability. Specifically, data from this pilot will be rigorously analyzed to identify any hallucinations in the AI's feedback or usability issues within the interface. Consequently, from June to August 2027, I will iteratively refine the prompting constraints (e.g., persona specifications and input validation) and validate assessment instruments, including knowledge retention and transfer tests, to ensure reliable measurement and a robust protocol for the main trial.

4.3 Phase III: Experimentation and Data Collection (Sep 2027 – Feb 2028)

The *formal experiment* initiates in Sep 2027, spanning approximately six months to accommodate the *recruitment* of 200 participants. The study will employ a stratified *randomization* approach based on prior knowledge and English proficiency. To maintain the integrity of the investigation, data collection will follow a strict two-stage procedure, where Session A will administer the argumentative writing task with the assigned scaffolding condition (Control, Cognitive, Metacognitive, or Integrated), followed by immediate post-tests, and Session B, scheduled exactly two weeks later for each cohort, will assess knowledge retention and transfer. By Feb 2028, I aim to have a complete, anonymized dataset comprising essay drafts, chat histories, and behavioral logs, ready for processing.

4.4 Phase IV: Data Analysis, Interpretation, and Write-Up (Mar 2028 – Aug 2028)

The final phase, beginning in **Mar 2028**, focuses on rigorous *data analysis* and *manuscript writing*. This period begins with the qualitative coding of essay quality by blind expert raters to establish inter-rater reliability, alongside the mapping of behavioral logs to the "Action Library" to quantify self-regulated learning behaviors. Once the qualitative data is quantified, advanced statistical analyses will be performed to address the research questions. The project will conclude between **Jun and Aug 2028** with the synthesis of these findings into a comprehensive manuscript, aiming for the submission to a high-impact journal, providing empirical evidence on the efficacy of GenAI-driven adaptive scaffolding.

Table 1

Research Timeline

Phase	**Research Tasks**	**2026**	**2027**	**2027**	**2027**	**2028**	**2028**	**2028**
		9-12	**1-4**	**5-8**	**9-12**	**1-4**	**5-8**	**9-12**
Phase I: Preparation	Literature Synthesis	🟦						
	Study Design/Hypothesis & Analysis Plan	🟦	🟦					
	IRB Submission & Approval		🟦					
	ASAW System Development		🟦	🟦				
Phase II: Pilot	Pilot Recruitment & Ddministration			🟦				
	Pilot analysis (stability, usability)			🟦				
	Instrument validation (difficulty, reliability)				🟦			
	Study Design Revision & System Refinement				🟦			
Phase III: Experiment	Participant Recruitment				🟦			
	Formal Experiment Session A				🟦			
	Formal Experiment Session B					🟦		
	Preliminary Analysis					🟦		
Phase IV: Analysis & Output	Data Integration (essays, logs, chat, surveys)					🟦	🟦	
	Coding & Feature Extraction (essay, logs, chat)					🟦	🟦	
	Primary Analyses (outcomes and processes)							🟦
	Mechanism & Complimentary Analysis							🟦
	Manuscript Writing & Submission						🟦	🟦

References

Acemoglu, D. (2002). Technical Change, Inequality, and the Labor Market. *Journal of Economic Literature*, *40*(1), 7–72.

Anmarkrud, Ø., Andresen, A., & Bråten, I. (2019). Cognitive Load and Working Memory in Multimedia Learning: Conceptual and Measurement Issues. *Educational Psychologist*, *54*(2), 61–83.

Autor, D., Chin, C., Salomons, A., & Seegmiller, B. (2024). New frontiers: The origins and content of new work, 1940–2018. *The Quarterly Journal of Economics*, *139*(3), 1399–1465.

Azevedo, R., Moos, D. C., Greene, J. A., Winters, F. I., & Cromley, J. G. (2008). Why is externally-facilitated regulated learning more effective than self-regulated learning with hypermedia? *Educational Technology Research and Development*, *56*(1), 45–72.

Barnard, L., Lan, W. Y., To, Y. M., Paton, V. O., & Lai, S.-L. (2009). Measuring self-regulation in online and blended learning environments. *The Internet and Higher Education*, *12*(1), 1–6.

Barrot, J. S. (2023). Using ChatGPT for second language writing: Pitfalls and potentials. *Assessing Writing*, *57*, 100745.

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences*, *122*(26), e2422633122.

Bauer, E., Greiff, S., Graesser, A. C., Scheiter, K., & Sailer, M. (2025). Looking Beyond the Hype: Understanding the Effects of AI on Learning. *Educational Psychology Review*, *37*(2), 45.

Bianchini, S., Müller, M., & Pelletier, P. (2022). Artificial intelligence in science: An emerging general method of invention. *Research Policy*, *51*(10), 104604.

Booth Olson, C., Maamuujav, U., Steiss, J., & Chung, H. (2023). Examining the Impact of a Cognitive Strategies Approach on the Argument Writing of Mainstreamed English Learners in Secondary School. *Written Communication*, *40*(2), 373–416.

Brown, K. G. (2001). Using computers to deliver training: Which employees learn and why? *Personnel Psychology*, *54*(2), 271–296.

Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative AI at Work*. *The Quarterly Journal of Economics*, *140*(2), 889–942.

Cagiltay, K. (2006). Scaffolding strategies in electronic performance support systems: Types and challenges. *Innovations in Education and Teaching International*, *43*(1), 93–103.

Callan, G. L., Rubenstein, L. D., Ridgley, L. M., & McCall, J. R. (2021). Measuring self-regulated learning during creative problem-solving with SRL microanalysis. *Psychology of Aesthetics, Creativity, and the Arts*, *15*(1), 136–148.

Chan, C. K. Y., & Lee, K. K. W. (2025). The balancing act between AI and authenticity in assessment: A case study of secondary school students' use of GenAI in reflective writing. *Computers & Education*, *238*, 105399.

Chen, I.-S. (2017). Computer self-efficacy, learning performance, and the mediating role of learning engagement. *Computers in Human Behavior*, *72*, 362–370.

Chen, Y., Wang, Y., Wüstenberg, T., Kizilcec, R. F., Fan, Y., Li, Y., Lu, B., Yuan, M., Zhang, J., Zhang, Z., Geldsetzer, P., Chen, S., & Bärnighausen, T. (2025). Effects of generative artificial intelligence on cognitive effort and task performance: Study protocol for a randomized controlled experiment among college students. *Trials*, *26*(1), 244.

Cheng, Y., Fan, Y., Li, X., Chen, G., Gašević, D., & Swiecki, Z. (2025). Asking generative artificial intelligence the right questions improves writing performance. *Computers and Education: Artificial Intelligence*, *8*, 100374.

Chien, K.-P., Tsai, C.-Y., Chen, H.-L., Chang, W.-H., & Chen, S. (2015). Learning differences and eye fixation patterns in virtual and physical science laboratories. *Computers & Education*, *82*, 191–201.

Chiriatti, M., Ganapini, M., Panai, E., Ubiali, M., & Riva, G. (2024). The case for human–AI interaction as system 0 thinking. *Nature Human Behaviour*, *8*(10), 1829–1830.

Chiu, T. K. (2024). The impact of Generative AI (GenAI) on practices, policies and research direction in education: A case of ChatGPT and Midjourney. *Interactive Learning Environments*, *32*(10), 6187–6203.

Darvishi, A., Khosravi, H., Sadiq, S., Gašević, D., & Siemens, G. (2024). Impact of AI assistance on student agency. *Computers & Education*, *210*, 104967.

De Cremer, D., & Kasparov, G. (2021). AI should augment human intelligence, not replace it. *Harvard Business Review*, *18*(1), 1–8.

Ding, L., & Zou, D. (2024). Automated writing evaluation systems: A systematic review of Grammarly, Pigai, and Criterion with a perspective on future directions in the age of generative artificial intelligence. *Education and Information Technologies*, *29*(11), 14151–14203.

Doshi, A. R., & Hauser, O. P. (2024). Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, *10*(28), eadn5290.

Draxler, F., Werner, A., Lehmann, F., Hoppe, M., Schmidt, A., Buschek, D., & Welsch, R. (2024). The AI ghostwriter effect: When users do not perceive ownership of AI-generated text but self-declare as authors. *ACM Transactions on Computer-Human Interaction*, *31*(2), 1–40.

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J., Eirug, A., & others. (2021). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, *57*, 101994.

Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2024). GPTs are GPTs: Labor market impact potential of LLMs. *Science*, *384*(6702), 1306–1308.

Fan, C.-Y., & Chen, G.-D. (2021). A scaffolding tool to assist learners in argumentative writing. *Computer Assisted Language Learning*, *34*(1–2), 159–183.

Fan, Y., Tang, L., Le, H., Shen, K., Tan, S., Zhao, Y., Shen, Y., Li, X., & Gašević, D. (2025). Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. *British Journal of Educational Technology*, *56*(2), 489–530.

Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2024). Generative AI. *Business & Information Systems Engineering*, *66*(1), 111–126.

Fisher, S. L., & Ford, J. K. (1998). Differential effects of learner effort and goal orientation on two learning outcomes. *Personnel Psychology*, *51*(2), 397–420.

Fleckenstein, J., Jansen, T., Meyer, J., Trüb, R., Raubach, E. E., & Keller, S. D. (2024). How am I going? Behavioral engagement mediates the effect of individual feedback on writing performance. *Learning and Instruction*, *93*, 101977.

Frank, M. R., Autor, D., Bessen, J. E., Brynjolfsson, E., Cebrian, M., Deming, D. J., Feldman, M., Groh, M., Lobo, J., Moro, E., Wang, D., Youn, H., & Rahwan, I. (2019). Toward understanding the impact of artificial intelligence on labor. *Proceedings of the National Academy of Sciences*, *116*(14), 6531–6539.

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research*, *74*(1), 59–109.

Fu, Q.-K., Zou, D., Xie, H., & Cheng, G. (2024). A review of AWE feedback: Types, learning outcomes, and implications. *Computer Assisted Language Learning*, *37*(1–2), 179–221.

Glickman, M., & Sharot, T. (2024). AI-induced hyper-learning in humans. *Current Opinion in Psychology*, *60*, 101900.

Graham, S., Hebert, M., & Harris, K. R. (2015). Formative assessment and writing: A meta-analysis. *The Elementary School Journal*, *115*(4), 523–547.

Guo, K., Wang, J., & Chu, S. K. W. (2022). Using chatbots to scaffold EFL students' argumentative writing. *Assessing Writing*, *54*, 100666.

Hadwin, A. F., & Winne, P. H. (2001). CoNoteS2: A software tool for promoting self-regulation. *Educational Research and Evaluation*, *7*(2–3), 313–334.

Heersmink, R. (2024). Use of large language models might affect our cognitive skills. *Nature Human Behaviour*, *8*(5), 805–806.

Henrie, C. R., Halverson, L. R., & Graham, C. R. (2015). Measuring student engagement in technology-mediated learning: A review. *Computers & Education*, *90*, 36–53.

Hessel, A. K., Nation, K., & Murphy, V. A. (2021). Comprehension Monitoring during Reading: An Eye-tracking Study with Children Learning English as an Additional Language. *Scientific Studies of Reading*, *25*(2), 159–178.

Hu, P. J.-H., & Hui, W. (2012). Examining the role of learning engagement in technology-mediated learning and its effects on learning effectiveness and satisfaction. *Decision Support Systems*, *53*(4), 782–792.

Hu, W., Tian, J., & Li, Y. (2025). Enhancing student engagement in online collaborative writing through a generative AI-based conversational agent. *The Internet and Higher Education*, *65*, 100979.

Huang, A. Y. Q., Lin, C.-C., Su, S.-Y., Yen, R.-X., & Yang, S. J. H. (2025). Improving students? Learning performance and summarization ability through a Generative AI-enabled Chatbot. *Educational Technology & Society*, *28*(3), 82–111.

Huang, H., Chen, Y., & Rau, P.-L. P. (2022). Exploring acceptance of intelligent tutoring system with pedagogical agent among high school students. *Universal Access in the Information Society*, *21*(2), 381–392.

Jarrahi, M. H., Lutz, C., & Newlands, G. (2022). Artificial intelligence, human intelligence and hybrid intelligence based on mutual augmentation. *Big Data & Society*, *9*(2), 20539517221142824.

Järvelä, S., Nguyen, A., & Hadwin, A. (2023). Human and artificial intelligence collaboration for socially shared regulation in learning. *British Journal of Educational Technology*, *54*(5), 1057–1076.

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274.

Kim, J. Y., & Lim, K. Y. (2019). Promoting learning in online, ill-structured problem solving: The effects of scaffolding type and metacognition level. *Computers & Education*, *138*, 116–129.

Kim, M. C., & Hannafin, M. J. (2011). Scaffolding problem solving in technology-enhanced learning environments (TELEs): Bridging research and theory with practice. *Computers & Education*, *56*(2), 403–417.

Korbach, A., Ginns, P., Brünken, R., & Park, B. (2020). Should learners use their hands for learning? Results from an eye-tracking study. *Journal of Computer Assisted Learning*, *36*(1), 102–113.

Kosmyna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X.-H., Beresnitzky, A. V., Braunstein, I., & Maes, P. (2025). *Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task* (No. arXiv:2506.08872). arXiv.

Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of Intelligent Tutoring Systems. *Review of Educational Research*.

Latifi, S., & Noroozi, O. (2021). Supporting argumentative essay writing through an online supported peer-review script. *Innovations in Education and Teaching International*, *58*(5), 501–511.

Lee, B. C., & Chung, J. (Jae). (2024). An empirical investigation of the impact of ChatGPT on creativity. *Nature Human Behaviour*, *8*(10), Article 10.

Lee, C. (2020). A study of adolescent English learners' cognitive engagement in writing while using an automated content feedback system. *Computer Assisted Language Learning*, *33*(1–2), 26–57.

Lee, J., Park, T., & Davis, R. O. (2022). What affects learner engagement in flipped learning and what predicts its outcomes? *British Journal of Educational Technology*, *53*(2), 211–228.

Li, R. (2023). Still a fallible tool? Revisiting effects of automated writing evaluation from activity theory perspective. *British Journal of Educational Technology*, *54*(3), 773–789.

Lim, L., Bannert, M., van der Graaf, J., Fan, Y., Rakovic, M., Singh, S., Molenaar, I., & Gašević, D. (2024). How do students learn with real-time personalized scaffolds? *British Journal of Educational Technology*, *55*(4), 1309–1327.

Liu, J., & Li, S. (2024). Toward Artificial Intelligence-Human Paired Programming: A Review of the Educational Applications and Research on Artificial Intelligence Code-Generation Tools. *Journal of Educational Computing Research*, *62*(5), 1165–1195.

Liu, M., Zhang, L. J., & Biebricher, C. (2024). Investigating students' cognitive processes in generative AI-assisted digital multimodal composing and traditional writing. *Computers & Education*, *211*, 104977.

Liu, X., & Cui, Y. (2025). Eye tracking technology for examining cognitive processes in education: A systematic review. *Computers & Education*, *229*, 105263.

Lo, C. K., Hew, K. F., & Jong, M. S. (2024). The influence of ChatGPT on student engagement: A systematic review and future research agenda. *Computers & Education*, *219*, 105100.

Milano, S., McGrane, J. A., & Leonelli, S. (2023). Large language models challenge the future of higher education. *Nature Machine Intelligence*, *5*(4), 333–334.

Miller, B. W. (2015). Using Reading Times and Eye-Movements to Measure Cognitive Engagement. *Educational Psychologist*. https://www.tandfonline.com/doi/abs/10.1080/00461520.2015.1004068

Mokyr, J. (1992). *The lever of riches: Technological creativity and economic progress*. Oxford University Press.

Mokyr, J., Vickers, C., & Ziebarth, N. L. (2015). The history of technological anxiety and the future of economic growth: Is this time different? *Journal of Economic Perspectives*, *29*(3), 31–50.

Musslick, S., Bartlett, L. K., Chandramouli, S. H., Dubova, M., Gobet, F., Griffiths, T. L., Hullman, J., King, R. D., Kutz, J. N., Lucas, C. G., Mahesh, S., Pestilli, F., Sloman, S. J., & Holmes, W. R. (2025). Automating the practice of science: Opportunities, challenges, and implications. *Proceedings of the National Academy of Sciences*, *122*(5), e2401238121.

Nguyen, A., Hong, Y., Dang, B., & Huang, X. (2024). Human-AI collaboration patterns in AI-assisted academic writing. *Studies in Higher Education*, *49*(5), 847–864.

Noroozi, O. (2018). Considering students' epistemic beliefs to facilitate their argumentative discourse and attitudinal change with a digital dialogue game. *Innovations in Education and Teaching International*, *55*(3), 357–365.

Noroozi, O., Banihashem, S. K., Biemans, H. J., Smits, M., Vervoort, M. T., & Verbaan, C.-L. (2023). Design, implementation, and evaluation of an online supported peer feedback module to enhance students' argumentative essay quality. *Education and Information Technologies*, *28*(10), 12757–12784.

Noroozi, O., Biemans, H., & Mulder, M. (2016). Relations between scripted online peer feedback processes and quality of written argumentative essay. *The Internet and Higher Education*, *31*, 20–31.

Noroozi, O., & Hatami, J. (2019). The effects of online peer feedback and epistemic beliefs on students' argumentation-based learning. *Innovations in Education and Teaching International*, *56*(5), 548–557.

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, *381*(6654), 187–192.

Ochoa, X., Huang, X., & Shao, Y. (2025). Exploring the potential of generative AI to support non-experts in learning analytics practice. *Journal of Learning Analytics*, *12*(1), 65–90.

Ochoa, X., Lang, C., Siemens, G., Wise, A., Gasevic, D., & Merceron, A. (2022). Multimodal learning analytics-Rationale, process, examples, and direction. *The Handbook of Learning Analytics*, *2*, 54–65.

Ouyang, F., Dai, X., & Chen, S. (2022). Applying multimodal learning analytics to examine the immediate and delayed effects of instructor scaffoldings on small groups' collaborative programming. *International Journal of STEM Education*, *9*(1), 45.

Ouyang, F., & Zhang, L. (2024). AI-driven learning analytics applications and tools in computer-supported collaborative learning: A systematic review. *Educational Research Review*, *44*, 100616.

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist*, *38*(1), 1–4.

Pintrich, P. R. (2000). The role of goal orientation in self-regulated learning. In *Handbook of self-regulation* (pp. 451–502). Elsevier.

Pintrich, P. R., & De Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology*, *82*(1), 33.

Raisch, S., & Fomina, K. (2025). Combining Human and Artificial Intelligence: Hybrid Problem-Solving in Organizations. *Academy of Management Review*, *50*(2), 441–464.

Richmond, L. L., & Taylor, R. G. (2025). The benefits and potential costs of cognitive offloading for retrospective information. *Nature Reviews Psychology*, *4*(5), 312–321.

Risko, E. F., & Gilbert, S. J. (2016). Cognitive Offloading. *Trends in Cognitive Sciences*, *20*(9), 676–688.

Roll, I., & Wylie, R. (2016). Evolution and Revolution in Artificial Intelligence in Education. *International Journal of Artificial Intelligence in Education*, *26*(2), 582–599.

Sankar, G., Djamasbi, S., Tulu, B., & Muehlschlegel, S. (2025). Measuring Cognitive Engagement with Eye-Tracking: An Exploratory Study. In D. D. Schmorrow & C. M. Fidopiastis (Eds.), *Augmented Cognition* (pp. 69–78). Springer Nature Switzerland.

Saye, J. W., & Brush, T. (2002). Scaffolding critical reasoning about history and social issues in multimedia-supported learning environments. *Educational Technology Research and Development*, *50*(3), 77–96.

Schunk, D. H. (2005). Self-regulated learning: The educational legacy of Paul R. Pintrich. *Educational Psychologist*, *40*(2), 85–94.

Shi, L., Jayawardena, G., & Gwizdka, J. (2025). Pupillometric Analysis of Cognitive Load in Relation to Relevance and Confirmation Bias. *Proceedings of the 2025 ACM SIGIR Conference on Human Information Interaction and Retrieval*, 219–230.

Sikström, P., Valentini, C., Sivunen, A., & Kärkkäinen, T. (2024). Pedagogical agents communicating and scaffolding students' learning: High school teachers' and students' perspectives. *Computers & Education*, *222*, 105140.

Sitzmann, T., & Ely, K. (2011). A meta-analysis of self-regulated learning in work-related training and educational attainment: What we know and where we need to go. *Psychological Bulletin*, *137*(3), 421–442.

Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance: An integrative review. *Perspectives on Psychological Science*, *10*(2), 176–199.

Stadler, M., Bannert, M., & Sailer, M. (2024). Cognitive ease at a cost: LLMs reduce mental effort but compromise depth in student scientific inquiry. *Computers in Human Behavior*, *160*, 108386.

Stojanov, A. (2023). Learning with ChatGPT 3.5 as a more knowledgeable other: An autoethnographic study. *International Journal of Educational Technology in Higher Education*, *20*(1), 35.

Suraworachet, W., Zhou, Q., & Cukurova, M. (2023). Impact of combining human and analytics feedback on students' engagement with, and performance in, reflective writing tasks. *International Journal of Educational Technology in Higher Education*, *20*(1), 1.

Suriano, R., Plebe, A., Acciai, A., & Fabio, R. A. (2025). Student interaction with ChatGPT can promote complex critical thinking skills. *Learning and Instruction*, *95*, 102011.

Vaccaro, M., Almaatouq, A., & Malone, T. (2024). When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, *8*(12), 2293–2303.

van de Pol, J., Volman, M., & Beishuizen, J. (2010). Scaffolding in Teacher–Student Interaction: A Decade of Research. *Educational Psychology Review*, *22*(3), 271–296.

van de Pol, J., Volman, M., Oort, F., & Beishuizen, J. (2014). Teacher scaffolding in small-group work: An intervention study. *Journal of the Learning Sciences*, *23*(4), 600–650.

van Nooijen, C. C. A., de Koning, B. B., Bramer, W. M., Isahakyan, A., Asoodar, M., Kok, E., van Merrienboer, J. J. G., & Paas, F. (2024). A Cognitive Load Theory Approach to Understanding Expert Scaffolding of Visual Problem-Solving Tasks: A Scoping Review. *Educational Psychology Review*, *36*(1), 12.

Van Quaquebeke, N., Tonidandel, S., & Banks, G. C. (2025). Beyond efficiency: How artificial intelligence (AI) will reshape scientific inquiry and the publication process. *The Leadership Quarterly*, *36*(4), 101895.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes* (Vol. 86). Harvard university press.

Wang, E. L., Matsumura, L. C., Correnti, R., Litman, D., Zhang, H., Howe, E., Magooda, A., & Quintana, R. (2020). eRevis(ing): Students' revision of text evidence use in an automated writing evaluation system. *Assessing Writing*, *44*, 100449.

Warr, P., & Downing, J. (2000). Learning strategies, learning anxiety and knowledge acquisition. *British Journal of Psychology*, *91*(3), 311–333.

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry*, *17*(2), 89–100.

Wu, J., Wang, J., Lei, S., Wu, F., & Gao, X. (2025). The impact of metacognitive scaffolding on deep learning in a GenAI-supported learning environment. *Interactive Learning Environments*, 1–18.

Xie, H., Chu, H.-C., Hwang, G.-J., & Wang, C.-C. (2019). Trends and development in technology-enhanced adaptive/personalized learning: A systematic review of journal publications from 2007 to 2017. *Computers & Education*, *140*, 103599.

Yan, L., Greiff, S., Lodge, J. M., & Gašević, D. (2025). Distinguishing performance gains from learning when using generative AI. *Nature Reviews Psychology*, *4*(7), 435–436.

Yan, L., Greiff, S., Teuber, Z., & Gašević, D. (2024). Promises and challenges of generative artificial intelligence for human learning. *Nature Human Behaviour*, *8*(10), Article 10.

Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2024). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, *55*(1), 90–112.

Yang, F.-Y., Chang, C.-Y., Chien, W.-R., Chien, Y.-T., & Tseng, Y.-H. (2013). Tracking learners' visual attention during a multimedia presentation in a real classroom. *Computers & Education*, *62*, 208–220.

Yeung, S. (2025). University students' engagement with generative AI-supported automated writing evaluation (AWE) feedback. *Journal of Second Language Writing*, *68*, 101203.

Zare, J. (2023). The impact of L2 learners' altruistic teaching on their task engagement in English essay writing. *System*, *118*, 103137.

Zawoyski, A. M., & Ardoin, S. P. (2019). Using Eye-Tracking Technology to Examine the Impact of Question Format on Reading Behavior in Elementary Students. *School Psychology Review*, *48*(4), 320–332.

Zhai, C., Wibowo, S., & Li, L. D. (2024). The effects of over-reliance on AI dialogue systems on students' cognitive abilities: A systematic review. *Smart Learning Environments*, *11*(1), 28.

Zhan, Y., & Yan, Z. (2025). Students' engagement with ChatGPT feedback: Implications for student feedback literacy in the context of generative artificial intelligence. *Assessment & Evaluation in Higher Education*, 1–14.

Zhang, Z., Aubrey, S., Huang, X., & Chiu, T. K. (2025). The role of generative AI and hybrid feedback in improving L2 writing skills: A comparative study. *Innovation in Language Learning and Teaching*, 1–19.

Zheng, J., Lajoie, S. P., Wang, T., & Li, S. (2023). Supporting self-regulated learning in clinical problem-solving with a computer-based learning environment: The effectiveness of scaffolds. *Metacognition and Learning*, *18*(3), 693–709.

Zheng, L., Li, X., Zhang, X., & Sun, W. (2019). The effects of group metacognitive scaffolding on group metacognitive behaviors, group performance, and cognitive load in computer-supported collaborative learning. *The Internet and Higher Education*, *42*, 13–24.

Zimmerman, B. J. (2008). Investigating self-regulation and motivation: Historical background, methodological developments, and future prospects. *American Educational Research Journal*, *45*(1), 166–183.

Zou, D., & Xie, H. (2019). Flipping an English writing class with technology-enhanced just-in-time teaching and peer instruction. *Interactive Learning Environments*, *27*(8), 1127–1142.

Appendix

Table 1

Design Principles for Cognitive Scaffolding

Dimension	Item	Condition	Form	Example
Use of Evidence	Evidence	Learner requests support on using evidence for the first time (e.g., via chat query like "How do I add evidence?").	Question	What sources from the reading materials could support your claim here?
		After initial question, learner shows no improvement (e.g., no citations added in draft within 5 minutes or follow-up query).	Hint	Consider citing specific quotes or data from the provided texts to strengthen your argument.
		After hint, learner persists with unresolved issues (e.g., further requests or draft analysis reveals weak evidence).	Prompt	Use this quote from the source: "Argumentation fosters critical thinking" (Fan & Chen, 2021a) as evidence for your claim.
	Commentary	Learner first seeks help on interpreting evidence (e.g., initial query about connecting evidence to claim).	Question	How does this evidence connect to your main argument?
		Post-question, no interpretive revisions detected (e.g., commentary remains summary-like after 5 minutes).	Hint	Explain why the evidence matters by linking it directly to your claim, avoiding mere summary.
		Following hint, continued lack of depth in commentary (e.g., repeated queries or NLP flags superficial analysis).	Prompt	Interpret the evidence like this: This quote demonstrates that AI scaffolding enhances engagement because it promotes active processing.
	Balance	Initial request for balancing elements (e.g., query on paragraph composition).	Question	Is your paragraph evenly distributing summary, evidence, and analysis?
		After question, imbalance persists (e.g., draft shows excessive summary without revisions).	Hint	Aim for a ratio where evidence and commentary outweigh summary to maintain analytical depth.
		Post-hint, no adjustment made (e.g., further help sought or draft unchanged).	Prompt	Revise to include: 20% summary, 40% evidence, 40% commentary, as in this example paragraph: [provide sample].
Idea	Address Prompt	First-time query on covering prompt aspects (e.g., "Did I miss something in the prompt?").	Question	Have you covered every element required by the writing prompt?
		Following question, incomplete coverage evident (e.g., no additions to draft).	Hint	Review the prompt keywords (e.g., "discuss pros and cons") and ensure each is explicitly addressed.
		After hint, prompt elements still omitted (e.g., ongoing queries or draft gaps).	Prompt	Add a section addressing the counterarguments, such as: While AI may reduce effort, it can foster deeper learning if scaffolded properly.
	Present Claim	Initial support request on claim formulation (e.g., "How to make my claim better?").	Question	What is your central thesis, and why is it persuasive?
		Post-question, claim remains vague (e.g., no refinements in draft).	Hint	Make your claim specific and debatable, avoiding vague statements.
		After hint, claim lacks clarity or compelling nature (e.g., further assistance needed).	Prompt	State your claim as: Generative AI scaffolding enhances learning outcomes beyond mere performance.
	Focus	Learner first asks about maintaining focus (e.g., "Is my writing on track?").	Question	Does every sentence contribute to supporting your claim?
		After question, off-topic elements persist (e.g., no edits to remove digressions).	Hint	Eliminate off-topic details and ensure transitions tie back to the thesis.
		Following hint, focus issues continue (e.g., draft analysis shows irrelevancies).	Question	Refocus this sentence: Instead of digressing on AI history, link it directly to engagement benefits.
Structure	Organization	Initial query on overall structure (e.g., "How to organize my essay?").	Question	How does the overall structure guide the reader through your argument?
		Post-question, disorganized flow detected (e.g., no restructuring).	Hint	Use a logical flow: introduction, body paragraphs by sub-claims, conclusion.
		After hint, organization remains weak (e.g., repeated requests).	Prompt	Organize as: Para 1: Intro; Para 2-4: Evidence for claims; Para 5: Conclusion.
	Introduction	First request for introduction help (e.g., "How to start my essay?").	Question	What hooks the reader and sets up your argument in the introduction?
		Following question, weak intro persists (e.g., no hook or thesis added).	Hint	Start with a question or statistic, then state thesis and outline.
		Post-hint, introduction lacks strength (e.g., further queries).	Prompt	Begin with: "In the era of AI, the question arises: Does it enhance or hinder learning?" followed by your thesis.
	Conclusion	Initial support on conclusion (e.g., "How to end my essay?").	Question	How does your conclusion reinforce the main claims without introducing new ideas?
		After question, conclusion incomplete (e.g., no summary of claims).	Hint	Summarize key points and end with implications for future research.
		Following hint, lacks completeness (e.g., draft unchanged).	Prompt	Conclude with: This study shows AI scaffolding promotes hybrid intelligence, urging educators to adopt it.
Language Use	Fluency	Learner first seeks fluency advice (e.g., "My sentences don't flow well.").	Question	Do your sentences read smoothly when connected?
		Post-question, fluency issues remain (e.g., no transition additions).	Hint	Vary sentence length and use transitions like "furthermore" for better flow.
		After hint, flow problems persist (e.g., additional help requested).	Prompt	Revise to: "AI boosts performance; however, it risks over-reliance."
	Syntax	Initial query on syntax variety (e.g., "How to vary my sentences?").	Question	Are you using a mix of simple, compound, and complex sentences?
		Following question, monotonous syntax detected (e.g., no changes).	Hint	Incorporate rhetorical devices like parallelism for stylistic enhancement.
		Post-hint, lacks variety (e.g., draft analysis flags repetition).	Prompt	Use varied syntax: "Not only does AI provide feedback, but it also encourages reflection."
	Diction	First request on word choice (e.g., "Better words for this?").	Question	Is your word choice precise and original?
		After question, poor diction continues (e.g., casual terms unchanged).	Hint	Choose academic synonyms over casual terms, e.g., "facilitate" instead of "help."
		Following hint, diction issues unresolved (e.g., further queries).	Prompt	Replace "stuff" with "mechanisms" in: Over-reliance on AI undermines learning mechanisms.
	Conventions	Initial help on grammar (e.g., "Check my errors?").	Question	Have you checked for grammar, spelling, or punctuation errors?

	Post-question, errors persist (e.g., no corrections made).	Hint	Focus on common issues like comma splices or subject-verb agreement.
	After hint, conventions violations remain (e.g., repeated requests).	Prompt	Correct: “Students engagement” to “Students' engagement.”
Tone	Learner first asks about tone (e.g., “Is this formal enough?”).	Question	Is your tone formal and objective for an academic audience?
	Following question, inappropriate tone detected (e.g., personal pronouns used).	Hint	Avoid contractions and personal pronouns to maintain professionalism.
	Post-hint, tone mismatches purpose (e.g., draft unchanged).	Prompt	Shift to formal tone: Change “I think AI is cool” to “Evidence suggests AI enhances engagement.”

Note. Question: open-ended questions that elicit thinking without providing new information; Hint: hints or clues to help solve problems; Prompt: direct answers to show how.

Table 2

Design Principles for Metacognitive Scaffolding

Dimension	Condition	Form	Example Content
Planning	Task Analysis & Goal Setting: Learner is inactive for the first 3-5 minutes of the writing task, or begins writing immediately without any discernible planning activity (e.g., using the 'Note' tool).	Question	“Before you dive in, what's your plan for tackling this essay? What do you understand as the main goal of the assignment?”
	Learner provides a vague plan or expresses uncertainty after the initial question.	Hint	“A good first step is often to break down the prompt and outline your main arguments. Consider using the rubric to guide your goals for the essay's content and structure.”
	Learner remains inactive or explicitly asks for direction on how to start planning.	Prompt	“Let's create a quick plan. Use the 'Note' tool to: 1. Write your main thesis statement. 2. List the three main points you will use to support it. 3. Identify one piece of evidence for each point.”
Monitoring	Progress Monitoring: The learner has been working on a single paragraph for an extended period (e.g., >10 minutes) or has not scrolled to review their previous writing in a long time.	Question	“How is your progress aligning with the plan you set? Let's do a quick check-in: are you on track with your timing and goals?”
	The learner's writing begins to drift from the argumentative prompt (e.g., becomes overly descriptive) or their initial thesis statement.	Hint	“It might be a good moment to pause and reread your thesis statement and the last paragraph you wrote. Ensure there is a clear connection between your current writing and your main argument.”
	The system detects a clear content gap (e.g., no counterargument has been addressed, which the rubric requires) and the learner hasn't noticed.	Prompt	“I've noticed your argument is developing well, but you haven't addressed any potential counterarguments yet. Let's take a moment to brainstorm a possible opposing viewpoint and how you can respond to it.”
Control & Regulation	Strategy Adjustment: The learner expresses frustration (e.g., writes “I'm stuck” or “this is hard”) or repeatedly deletes and rewrites the same sentence.	Question	“It seems like you've hit a challenging spot. What's the main difficulty you're facing right now? What strategies have you tried so far?”
	The learner identifies a problem but is unsure how to fix it (e.g., “my argument feels weak”).	Hint	“When a paragraph feels weak, writers often try a few different strategies, like finding stronger evidence, rephrasing the topic sentence, or even moving the paragraph to a different part of the essay. Could one of those help here?”
	The learner is clearly stuck and not making progress after being prompted to monitor and strategize.	Prompt	“This paragraph isn't flowing as well as the others. Try this strategy: Summarize the main point you want to make in this paragraph in one new sentence. Then, delete the rest of the paragraph and rebuild it around that core sentence.”
Evaluation & Reflection	Self-Evaluation: The learner has completed a full draft and either pauses for a significant time or asks for general feedback (“is this good?”). This is triggered primarily during the “Revision” phase.	Question	“Now that you have a complete draft, how would you evaluate it against the 'Content' and 'Organization' criteria in the rubric? What do you believe is the strongest part of your essay, and what needs the most work?”
	The learner provides a very superficial self-evaluation (“it's okay”) or struggles to apply the rubric to their own work.	Hint	“A powerful self-evaluation technique is to read your work from the perspective of your audience. Read your introduction and conclusion back-to-back. Do they align? Does your conclusion show a deeper understanding than your introduction?”
	The learner is preparing to finish the task but has not engaged in a deep revision of their work.	Prompt	“Before you submit, let's perform one final, targeted check. Reread only the first sentence of each body paragraph. Do these sentences, in order, create a logical and compelling summary of your entire argument?”

Note. Question: open-ended questions that elicit thinking without providing new information; Hint: hints or clues to help solve problems; Prompt: direct answers to show how.

Table 3

Scoring Rubrics for Argumentative Writing

	Outstanding	Good	Satisfactory	Barely adequate	Inadequate
Content	• A clear and fully developed position is presented which directly answers the questions. • Word length is appropriate, or within the requirements. • Ideas are highly relevant, comprehensive and fully extended. • All arguments/opinions/analyses very well supported and highly logical and critical, many of which show originality/creativity/sophistication; a good range of viewpoints and/or source materials.	• A clear and well-developed position is presented in response to the questions. • Word length is relevant, comprehensive and extended. • Ideas are relevant, comprehensive, and well-extended. • All arguments/opinions/analyses logically and critical; based on a good range of viewpoints and/or source materials.	• A clear and developed position is presented. • Word length is generally appropriate; may be just below/beyond the requirements. • Ideas are generally relevant, comprehensive and extended. • Arguments/opinions/analyses logically and critical; based on a fair range of different viewpoints and/or source materials.	• A position is discernible, but the reader has to read carefully to find it. • Word length may be generally inappropriate; noticeably below/beyond the requirements. • Some main ideas are put forward, but they are limited and are not sufficiently developed and/or there may be irrelevant detail. • Arguments/opinions/analyses generally unconvincing or illogical.	• No relevant position can be identified, and/or there is little direct response to the questions. • Word length may be inappropriate; far below/beyond the requirements. • No relevant position can be identified. • No opinion expressed as required.
Organization	• Information and ideas are logically sequenced, and cohesion is well managed. Any lapses in coherence or cohesion are minimal. • A range of cohesive devices including reference and substitution is used flexibly. • The message can be followed effortlessly. Paragraphing is skillfully managed.	• Information and ideas are logically sequenced though there may be lapses in cohesion or progression of the response. Occasional lapses in coherence are possible. • A range of cohesive devices including reference and substitution is used flexibly though with occasional inaccuracies or some over/under use. • The message can be followed well. Paragraphing is used sufficiently and appropriately.	• Information and ideas are generally arranged coherently and there is a clear overall progression. A few lapses in coherence or cohesion may occur. • Cohesive devices are used to some good effect but cohesion within and/or between sentences may be faulty or mechanical due to misuse, overuse or omission. • The relationship of ideas can be followed. Paragraphing is used generally appropriately.	• Organization is evident but is not wholly logical and there may be a lack of overall progression. Nevertheless, there is a sense that ideas are linked. Some lapses in coherence or cohesion may occur. • There may be limited/overuse of cohesive devices with some inaccuracy. The writing may be repetitive due to inadequate and/or in accurate use of reference and substitution. • The relationship of ideas can generally be followed but the sentences are not fluently linked to each other. Paragraphing is inconsistently managed, affecting overall coherence.	• Information and ideas are evident but not arranged coherently and there is no clear progression within the response. Many lapses in coherence or cohesion occur. • There is some use of basic cohesive devices, which may include some repetition, but their use may be inaccurate or inappropriate. • Relationship between ideas can be unclear and/or inadequately marked. There is little evidence of control of organizational features.
Language	• Grammar: Highly accurate; displays a very wide range of grammatical structures supporting sophisticated ideas; minor errors are extremely rare and have minimal impact on communication. • Vocabulary: A wide range of vocabulary is used accurately and appropriately with very natural and sophisticated control of lexical features. Minor errors in spelling and word formation are extremely rare and have minimal impact on communication. • Style and tone: Highly appropriate.	• Grammar: Accurate; displays a wide range of grammatical structures with occasional non-systematic errors and inappropriate choices. These do not impede communication. • Vocabulary: There skillful use of uncommon and/or idiomatic items when appropriate, despite occaisional inaccuracies in word choice and collocation; Occassional errors in spelling and/or word formation may occur, but have minimal impact on communication. • Style and tone: Appropriate.	• Grammar: Generally accurate; displays an adequate range of grammatical structures. A few errors may persist, but these do not impede communication. • Vocabulary: There is some ability to use less common and/or idiomatic items, through inappropriate occur; There are a few errors in spelling and/or word formation, and they do not detract from overall clarity. • Style and tone: Generally appropriate.	• Grammar: Displays a narrow range of grammatical structures. Errors in grammar are noticeable and may cause some difficulty for the reader. • Vocabulary: The meaning is generally clear in spite of a rather restricted range or lack of precision in word choice; If the writer is a risk-taker, there will be a wider range of vocabulary used but higher degrees of inaccuracy or inappropriacy; There are some errors in spelling and/or word formation, but these do not impede communication. • Style and tone: Generally inappropriate.	• Grammar: Simple grammatical structures used most of the time; errors may be frequent and cause some difficulty for the reader. • Vocabulary: There may be frequent lapses in the appropriacy of word choice, and a lack of flexibility is apparent in frequent simplifications and/or word formation may be noticeable and may cause some difficulty for the reader. • Style and tone: Inappropriate.

Table 4

Learning Engagement Questionnaire

Item	Category	Statement
1	Behavioral	I tried to do more than what was necessary to do the task well.
2		I did my best to stay focused and avoid distraction.
3		I spent as much time as necessary to complete the task.
4		I worked as hard as I could to complete the task.
5		I tried to actively engage myself in the task.
6	Emotional	Doing the task was fun.
7		I felt interested when doing the task.
8		Doing the task aroused my curiosity.
9		I felt enjoyable when doing the task.
10		I felt enthusiastic when doing the task.
11	Cognitive	During the task, I tried to explain the key concepts in my own words.
12		During the task, I tried to summarize it in my own words.
13		During the task, I tried to connect the ideas in the task with what I already know.
14		When doing the task, I tried to generate examples to help me understand them better.
15		During the task, I repeated the contents and asked myself questions about them.

Table 5

Analytical Framework for Screen Recordings of Learning Engagement

Category	Label	Description
Behavioral	Time on Task	Reading materials, drafting, revising (e.g., adding, deleting), and interacting with AI
Cognitive (Writing Strategy)	Reasoning	Reading online website material
		Reading and evaluating the text/point
		Reading system suggestion
		Reading suggestions and hints given by the system
		Re-reading a word/words
		Re-reading a phrase
		Re-reading a clause
		Re-reading two or more sentences
		Re-reading a paragraph/paragraphs
		Silence/pause with 5.0 or longer
	Revising	Adding a word/words
		Adding a phrase
		Deleting a word/words
		Deleting a phrase
		Major revision
		Word choice
		Substitution
		Word order

		Grammar
	Editing	Addition Deletion Grammar and syntax Spelling

Table 6

Eye-Tracking Indicators for Cognitive Engagement

Dimensions	Indicators	Definition	Example
Attention	Total Fixation Duration	The total time spent fixating on an AOI	Represent cognitive attention to the target information (Yang et al., 2013).
	Average Fixation Duration	The average duration of all fixations	Long fixation duration indicates the participant has difficulty extracting information (Chien et al., 2015).
	Gaze Duration	The sum of all fixation durations within an AOI before the gaze moves to the next one	Shorter reading time corresponds to higher reading fluency (Hessel et al., 2021).
	Total Fixation Count	The number of fixations falling within a specific AOI or across all AOIs	An indication of greater interest and importance in the area (Anmarkrud et al., 2019).
	Average Fixation Count	The average number of fixations	Fewer fixations reflect more efficient processing, which is negatively correlated with oral reading scores (Zawoyski & Ardoin, 2019).
	Transitions	The number of times the gaze shifts between different AOIs	Integrative cognitive processes and mental model construction when learning with multimedia (Korbach et al., 2020)
Absorption	Pupil Diameter	The change in pupil size when a learner focuses on the AOIs	High-frequency components of pupil diameter changes are more closely associated with cognitive load and mental effort (Shi et al., 2025).

Table 7

Self-Regulated Learning Questionnaire

Item	Category	Statement
1	Goal Setting	I set my standards for the task.
2		I set overall goals as well as detailed objectives for the task.
3		I keep a high standard for my learning in the task.
4		I set goals to help me manage study time during the task.
5	Persistence	Regardless of whether or not I like materials, I work my hardest to learn it.
6		When something that I am studying gets difficult, I spend extra time and effort trying to understand it.
7		I try to learn all of the testable material inside and out, even if it is boring.
8		I work hard to do well in the task even if I don’t like what we are doing.
9		Even when the materials are dull and uninteresting, I manage to keep working until I finish.
10		When I was feeling bored, I forced myself to pay attention.
11		When my mind began to wander during the task, I made a special effort to keep concentrating.
12		I increased my effort when the material did not really interest me.
13		I pushed myself even harder when I began to lose interest.
14		Whenever I lost interest in my work, I made a special effort to pay attention.
15	Effort	I usually spent more time than the requirements of the task.
16		I usually provide extra effort in the task.
17	Self-efficacy	I’m certain I can understand the basic concepts in the task.
18		I believe I will receive an excellent grade in the task.
19		I’m certain I can understand the most difficult material presented in the readings for the task.
20		I’m confident I can learn the basic concepts taught in the task.
21		I’m confident I can understand the most complex material presented by the instructor in the task.
22		I’m confident I can do an excellent job on the assignments in the task.
23		I expect to do well in the task.
24		I’m certain I can master the skills being taught in the task.
25		Considering the difficulty of the materials, requirements, and the task, I think I will do well.

Table 8

Action Library for Labelling Self-Regulated Learning Behavioral Logs

Label	Description
General Instruction	Learners read or re-read general instructions and learning goals
Rubric	Learners read or re-read the rubric for essay writing
Relevant Reading	Learners read and learn learning content for the first time
Relevant Re-reading	Learners re-read and review for learning content which they have read before
Irrelevant Reading	Learners read the pages which are not relevant to the learning goal or writing task
Irrelevant Re-reading	Learners re-read the pages which are not relevant to the Learning goal or writing task
Navigation	Learners navigate through pages or scroll at catalogue zone
Write Essay	Learners write, edit or stay in the essay zone
Copy Paste	Learners copy and paste some content from reading materials into the essay or notes
Note Editing	Learners create, delete, edit or label the notes
Note Reading	Learners click to open and read or re-read the notes
Highlight Editing	Learners create, delete or edit the highlights
Highlight Reading	Learners click to open and read or re-read the highlights
Highlight Labelling	Learners create tags for highlights
Timer	Learners click to check timer during learning
Search Content	Learners use the search tool on the left to search learning contents
Search Highlight Note	Learners use the search tool on the right to search notes or highlights
Planner	Learners click to open planner tool, and create or edit their plans
Other	Learners pause the learning session or other unlabeled data
Off Task	Learners do not have any action for a relatively long time (5 minutes in this study)

Table 9

Analytical Framework for Chat History with AI

Question Type	Depth	Definition	Example
Verification	Shallow	Questions to confirm the truth or occurrence of a fact or event	You mean the first paragraph still need to expand?
Disjunctive		Questions to determine which among a set of options is the case	Should I first do x, or do y?
Concept Completion		Questions to identify or complete a missing element, usually a referent of a noun argument slot.	What is the correct spelling of vechicle [sic]?
Example		Questions to identify an instance or label that exemplifies a category	Can you give examples on how to use differentiation to adapt education?
Feature Specification		Questions to understand the qualitative attributes of an entity	Can you explain Applications of scaffolding to me?
Definition		Questions to clarify the meaning of a term or concept	What is the definition of Differentiation?
Comparison		Questions to explore similarities and differences between two or more entities	What's the difference between scaffolding and instructional approaches?
Casual Consequence	Deep	Questions to understand the effects of an event or state	What are the potential consequences of overusing AI in education?
Instrumental		Questions to identify the means or methods to accomplish a goal	How to apply AI in education?
Enablement		Questions to understand the resources or conditions that allow an action to be performed	Can you give me some ideas to support that "AI can help tutors with differentiation."?
Judgmental		Questions to evaluate an idea or to seek advice	How would you rate the quality of this essay?
Assertion	Not Specified	Question that indicates a lack of knowledge or understanding of an idea	uh, i don't know what revision is needed to improve my essay haha
Indirect Request		Questions asked in a polite and indirect form when the speaker wants the listener to perform a specific action	Could you please give me the feedback of my essay?
Direct Request		Questions asked in an commanding or direct form when the speaker wants the listener to perform a specific action	Give me the feedback of my essay.