

# “He Seems So Real, but Should I Trust Him?” : Fostering Critical Thinking Through Dialogues with a GAI Agent in Middle School History Class

Shiyuan Yan

Faculty of Education, Beijing Normal University, Beijing 100875, China

E-mail: [yanshiyuan01@mail.bnu.edu.cn](mailto:yanshiyuan01@mail.bnu.edu.cn)

## Abstract

Against the backdrop of generative artificial intelligence (GAI) increasingly penetrating secondary education, especially history instruction, existing studies have explored GAI’s educational potential and its effects on learning outcomes, yet most focus on overall instructional efficacy rather than the nuanced, process-based dynamics of dialogue between students and GAI in history classrooms. Few qualitative inquiries examine how students engage with AI-generated historical narratives, negotiate trust, and enact critical practices during real-time interactions. This qualitative case study of 15 seventh-grade students from a Beijing middle school explores what triggers students’ critical awareness, their verification strategies for GAI information, and their balance between artificial intelligence(AI) use and independent thinking, aiming to inform historical critical thinking development in AI-enhanced history education. Data were collected via classroom observations, semi-structured interviews, student reflection journals, teaching journals, and researcher reflection journals, and analyzed through thematic analysis. Results show that factual errors, logical inconsistencies, vague expressions, and perspective biases in GAI responses, as well as inconsistencies with prior knowledge, trigger skepticism; students use various verification strategies, including textbook cross-checking, authoritative document referencing, peer discussion and teacher consultation. While GAI dialogue fosters multi-perspective reasoning and evidence-seeking skills, it also risks cognitive dependence and uncritical acceptance of AI content. The findings conclude that GAI serves as a catalyst for middle school students’ historical critical thinking under deliberate pedagogical guidance, providing theoretical and practical implications for GAI-integrated history teaching in the digital era.

**Keywords:** Generative Artificial Intelligence (GAI), Critical thinking, Human-AI dialogue, Middle School History Education

**DOI:** 10.7176/JEP/17-6-16

**Publication date:** June 30<sup>th</sup> 2026

## 1. Introduction

Over the past few years, GAI has rapidly gained traction in the field of education (Cheng et al., 2020; Lin et al., 2023; Liu et al., 2024). In history education, students no longer solely acquire professionally verified historical narratives from textbooks and instructors. Instead, they can engage in inquiry-based dialogue with an ostensibly omniscient knowledge source at any time. This shift aligns with the transformation of contemporary history education, which requires fostering students’ historical thinking and critical literacy. Learners are expected to examine historical evidence, trace contextual origins and compare diverse narratives, rather than merely memorizing established conclusions. Yet, when GAI intervenes in the learning process with narratives that lack citations and explicit standpoints, while occasionally embedding biases and factual inaccuracies—does it hinder or advance the cultivation of middle school students’ critical thinking? This is the core puzzle this study aims to resolve.

Existing research offers an important yet incomplete picture of this issue. On the one hand, extensive work in educational technology has confirmed the potential of GAI in providing personalized learning scaffolds, immediate feedback, and enhanced learning motivation (Guo et al., 2024; Plass & Froehlich, 2025; Wang et al., 2026). At the same time, researchers have repeatedly warned about the risks of large language models generating hallucinations, reinforcing stereotypes, and producing seemingly plausible yet factually inaccurate texts that masquerade as authoritative sources (Bauer et al., 2025; Lee et al., 2024). On the other hand, although researchers in education have accumulated substantial findings on instructional strategies for critical thinking (Leibovitch et al., 2024), few studies have adopted a process-oriented lens to delve into history classrooms and examine how, when students engage in dialogues with GAI agents, their critical thinking is actually triggered, what cognitive and emotional fluctuations they undergo, and what evaluation and verification strategies they

proactively develop (Amani & Mkimbili, 2025).

Against this backdrop, this qualitative study was conducted in seventh-grade history classrooms at a middle school in Beijing, China, with 15 Grade 7 students recruited as participants. It explores how middle school students' critical thinking is stimulated and manifested during historical discussions with GAI agents, as well as its evolutionary trajectory. This study aims to offer practical implications for the innovative development of history education in the AI era.

## 2. Literature Review

### 2.1 *The application of GAI in education*

The rapid advancement of GAI is profoundly reshaping education, transforming both teaching approaches and learning paradigms (Chen et al., 2020). An accurate evaluation of its educational value begins with clarifying its core definition. Generative Artificial Intelligence (GAI) denotes AI systems that generate texts, images and other media content via generative models (Dwivedi et al., 2023). Currently, GAI has been widely adopted across diverse subject teaching. For instance, Liu et al. (2025) examined its application in senior high school geography education and confirmed its considerable potential while acknowledging aspects requiring further improvement. Wulff and Kubsch (2025) demonstrated that GAI tools enhance personalization and automation in STEM teaching and learning. The positive outlook validated by existing research lays a solid technical foundation for this study's design of adopting GAI agents as interactive dialogue partners in history classrooms. However, the aforementioned studies also reveal the inherent limitations of GAI, including hallucinations in generated content, insufficient factual accuracy, the risk of student over-reliance, as well as issues related to ethics, privacy, and technological equity. It is thus evident that while generative artificial intelligence brings new opportunities for educational innovation and transformation, it is also accompanied by considerable risks.

### 2.2 *Critical thinking in the field of history*

Critical thinking has long drawn widespread attention across education, psychology, logic and other disciplines, with varying interpretations of its connotation. Some scholars regard it as an essential skill for problem-solving, collecting diverse evidence and perspectives, evaluating key information and making rational decisions (Song et al., 2024). Others define it as a deliberate, self-regulated judgment process encompassing interpretation, analysis, evaluation and inference, which requires clarifying the concepts, methodologies, criteria and contextual foundations of evidence (Calma & Davies, 2026). Although these definitions enjoy broad applicability, they cannot be directly applied to history education. Accordingly, this study formulates a definition of critical thinking suited to the disciplinary features of history and the cognitive traits of middle school students through theoretical review, expert consultation and qualitative data triangulation. Its core connotation encompasses five interrelated, non-hierarchical dimensions, as illustrated in Figure 1.

- **Historical Source Judgment Ability:** The capacity to situate information within its historical context, assess the narrator's motivations and positional stance, and evaluate the authenticity and credibility of the content.
- **Contextualized Critical Thinking Ability:** The competence to identify implicit temporal limitations and subjective biases within narratives, reject the anachronistic fallacy of judging the past by modern standards, and view historical events and figures objectively.
- **Evidence-Seeking and Comparison Capabilities:** When confronted with the historical accounts generated by GAI, one can actively search for other learning resources such as textbooks and historical materials to support or challenge the authenticity and rationality of its statements.
- **Multiple Perspectives Recognition Ability:** This means being able to recognize that the same historical event or historical figure can have different interpretations and evaluations, and not being limited to a single narrative.
- **Self-Reflective Questioning:** During the interaction with GAI, one should be able to remain vigilant against one's tendency to be easily deceived, not blindly accept the content output by GAI, and always maintain the awareness of independent thinking.

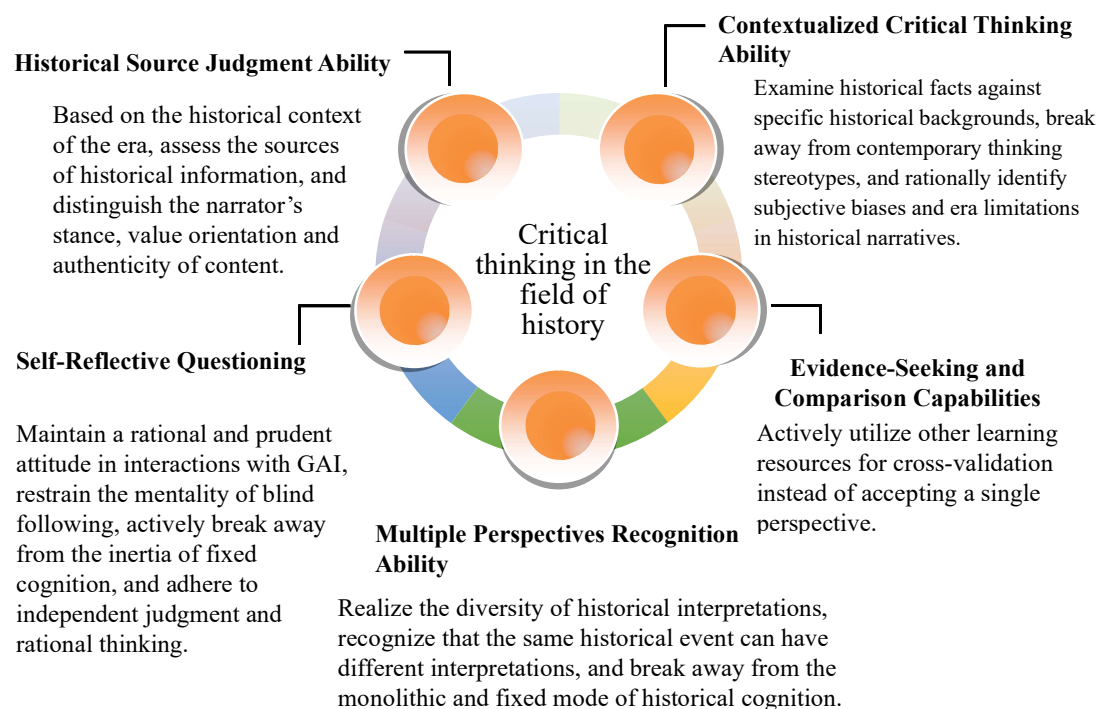


Figure 1. Critical thinking in the field of history

### 2.3 Research Questions and Limitations of Previous Studies

A review of existing literature reveals notable research gaps at the intersection of GAI and history education. Firstly, the cross-disciplinary research on GAI and history education focuses on efficacy verification. Most papers focus on the changes in academic performance or learning motivation after GAI technology is applied in history classes (Chen et al., 2025; Wall et al., 2025), while there are few empirical responses to deeper educational issues such as how GAI affects the development of critical thinking among middle school students and how it cultivates AI literacy in the context of the history discipline. Second, regarding the relationship between GAI and historical critical thinking, existing studies have examined the development of students' critical thinking within disciplinary contexts. However, little exploration has been conducted on how to cultivate this thinking in historical settings (Li et al., 2025; Shen & Chen, 2025). Thirdly, few existing studies on critical thinking adopt qualitative methods to explore how critical thinking is triggered during continuous interaction between students and GAI, what cognitive changes learners undergo, and what assessment and verification strategies they develop independently (Demircioglu et al., 2023; Guo et al., 2026).

Although this study does not adopt a specific instructional design framework a priori, the notion of the Zone of Proximal Development (Vygotsky, 1978)—which posits that learners can achieve higher levels of understanding with appropriate scaffolding—provides a useful lens for interpreting how pedagogical support might bridge the gap between students' current and potential levels of critical thinking. In view of this, this study takes the seventh-grade history class as the research field, aiming to explore how the critical thinking of middle school students is triggered, how it operates specifically, and what kind of development trajectory it presents when they engage in historical problem conversations with the GAI intelligent agent. Through this process-oriented qualitative investigation, this study seeks to address the research gap in the field of history education regarding GAI and critical thinking, and to provide empirical evidence for the cultivation of historical critical thinking in the AI era. The research questions are as follows:

- RQ1. What interaction scenarios and features of GAI responses arouse students' doubt and cognitive vigilance?
- RQ2. What cognitive strategies and external resources do students use to assess and verify the credibility of information offered by GAI?
- RQ3. How do students negotiate between using GAI to obtain historical information and maintaining independent thinking and critical inquiry?

### 3. Method

#### 3.1 Research Design

This study adopted a qualitative research paradigm and focused on the development of students' critical thinking in GAI-based dialogue teaching. Purposive sampling was employed to select 15 seventh-grade students and one classroom teacher from a middle school in Beijing as research participants. Data collection primarily relied on semi-structured interviews, supplemented by classroom observation records, dialogue reflection sheets, student reflection journals, teaching journals, and researcher reflection journals, which together formed a diverse qualitative dataset. Thematic analysis was further utilized to deeply explore students' inner perceptions and cognitive changes. This study strictly followed ethical principles. The school, teachers, students, and parents were fully informed of the research purpose, procedures, data usage, and potential risks in advance, and written informed consent was obtained from all participants. All participants' personal information was anonymized, with students assigned codes S01-S15 and the teacher coded T. All research data were encrypted, stored securely, and used solely for this study. Participants were clearly informed that their participation was voluntary and unrelated to academic performance, and that they could withdraw from the research unconditionally at any time without any adverse consequences. All participant information was treated with strict confidentiality to protect participants' rights and ensure the legitimacy of the research.

#### 3.2 Participant Sample Characteristics

The study systematically summarized the basic characteristics of the 15 student participants and the one classroom teacher. For the students, the characteristics primarily covered gender, history academic performance, level of interest in history, and prior GAI usage frequency. For the teacher, the focus was on professional background and teaching experience. The specific characteristics are as follows.

Table 1. Demographic and Background Characteristics of Student Participants(n=15)

Characteristic Categories	Subcategories	Number of Participants (n)	Percentage (%)
Gender	Male	8	52.3
	Female	7	47.7
Level of Interest in the Subject of History	High Interest	9	60.0
	Medium Interest	3	20.0
	Low Interest	3	20.0
	High ( $\geq 85$ points)	6	40.0
Academic Performance of history	Medium (70-84 points)	5	33.3
	Low (<70 points)	4	26.7
Prior GAI Usage Frequency	High (>3 times/week)	9	60.0
	Medium (1-3 times/week)	4	26.7
	Low (Less than once/week or never)	2	13.3

Note. The students' academic performance referred to in this study refers to the standardized average score of three history examinations, including the mid-term examination and final examination of the first semester of Grade Seven, and the unit test of the first unit of the second semester of Grade Seven.

Table 2. Background Characteristics of the Teacher Participant (n=1)

Characteristic Categories	Subcategories
Gender	Male
Educational Background	History Major
Teaching Experience	More than 3 years of middle school history teaching experience
Familiarity with GAI Tools	Having some experience with generative AI tools in teaching

The 15 student interviewees selected for this study had a relatively balanced gender ratio, which can, to a certain extent, reduce the interference of gender differences in the description of the interaction process. Students' history academic performance, level of interest in history, and prior GAI usage frequency all covered three tiers. Prior GAI usage frequency was categorized into three levels—high, medium, and low—based on weekly usage frequency, consistent with the classifications in Table 1. This stratified sampling was not intended for statistical comparison, but rather to incorporate diverse individual experiences, ensuring that the subsequent qualitative analysis could capture the varied responses and thinking processes of students from different backgrounds when interacting with GAI. In addition, the study included one classroom teacher as an interviewee. The teacher is male, with a professional background in history and extensive experience teaching history at the junior high school level. He possesses a certain level of familiarity with the application of GAI tools in teaching and can provide supplementary observations and reflections from the perspective of a teaching practitioner.

### 3.3 Research Tools

This study adopted multiple research tools, including GAI models (DeepSeek and Doubao), dialogue reflection sheets, semi-structured interview protocols, classroom observation forms, teaching journals, and researcher reflection journals. Two GAI models, DeepSeek and Doubao, were selected for this study. Both are widely accessible large language models commonly used by middle school students in China, which enhances the ecological validity of the research setting. The inclusion of two distinct models also allowed students to encounter varied response styles and potential inaccuracies across different AI systems, thereby broadening the range of interaction scenarios available for analysis. No systematic comparison between the two models was intended; rather, the dual-model design served to diversify the dialogue contexts in which students' critical thinking could be observed. Students interacted with GAI via text-based dialogue interfaces. These tools supported systematic data collection and analysis, guaranteeing standardized research procedures and credible research data.

Table 3. Semi-structured Interview Outline

Research Questions	Core questions for student interviews	Core questions for teacher interviews
RQ1. What interaction scenarios and features of GAI responses arouse students' doubt and cognitive vigilance?	1. Recall your experience of having a conversation with GAI. Was there any moment when you felt that "what it said didn't quite make sense" or "there was something odd about it"? Could you describe that moment in detail? 2. What was GAI's response at that time? Was it the tone, the choice of words, or the specific historical details that gave you this impression?	1. During your classroom observation, did you notice any specific points where students tended to hesitate, question, or pause when interacting with GAI? 2. Based on your observations, which features of GAI's responses are more likely to arouse students' suspicion?
RQ2. What cognitive strategies and external resources do students use to assess and verify the credibility of information offered by GAI?	3. After you have doubts about what GAI said, what do you usually do? Will you try to find out for sure? 4. What specific methods or resources did you use to verify its claim? (For example: consulting textbooks, searching the internet, asking teachers, discussing with classmates, etc.) Could you provide a specific example?	3. When students express doubts about the information provided by GAI, what strategies do you observe them to adopt in order to deal with this uncertainty? 4. In your opinion, can students effectively utilize textbook knowledge or external resources such as online searches to verify the information provided by GAI? Is this process a distraction or an opportunity for teaching?
RQ3. How do students negotiate between using GAI to obtain historical information and maintaining independent thinking and critical inquiry?	5. Overall, do you think the information provided by GAI is helpful for you to understand history? How is the information it gives different from what is presented in the textbooks? 6. How do you view the use of GAI to study history? Do you have any concerns or difficulties regarding the content that GAI claims? When faced with an appealing AI viewpoint and your own original judgment, how do you usually make a decision?	5. Do you think this GAI conversation mode has facilitated students' in-depth reflection on history, or has it increased the risk that they will regard the AI-provided information as the standard answer? 6. Based on the goal of cultivating students' critical thinking skills, what do you think are the main difficulties for students when using AI to obtain information and maintaining their own independent judgment? How can teaching be conducted to guide them?

Based on the three core research questions, this study developed a semi-structured interview outline. It centered on students' real perceptions of GAI conversations, their strategies for judging and verifying the credibility of GAI-generated content, and cognitive shifts throughout the interaction. Teacher interviews focused on observed student behaviors, practical difficulties in teaching implementation, and relevant improvement suggestions. All outlines were revised through expert consultation and pilot interviews to ensure appropriateness, openness and pertinence, so as to fully acquire key research information.

The classroom observation record sheet was used to systematically record students' participation, questioning behaviors, expressions of trust, and various interaction details during their dialogues with GAI. The sheet was pre-set with core observation dimensions while also leaving open recording space, so as to capture dynamic classroom behaviors and unexpected incidents. Furthermore, to guide students in systematically and clearly documenting their cognitive changes, the study provided a structured student reflection journal template, which not only ensured the standardization of journal entries but also allowed students to freely express their feelings, questions, and information discernment processes after interacting with GAI.

### *3.4 Data Collection Process*

Before the research was carried out, the purpose, implementation process, data usage and potential risks of the research were explained in detail to the participating schools, history teachers, 15 seventh-grade students and their parents. After ensuring that all participants fully understood the research content, they signed written informed consent forms. At the same time, the semi-structured interview outline, classroom observation record sheet and student reflection journal template designed in the early stage were pre-tested. Based on the feedback, they were revised and improved to ensure the applicability of the tools.

During the implementation of the GAI dialogue teaching, the researcher acted as a strict non-participant observer throughout the study. The researcher attended all classroom sessions, completed observation records, and documented students' performance and interactions, but did not participate in the design, delivery, or modification of the GAI dialogue teaching activities. A teaching journal was maintained by the classroom teacher to record the design rationale, implementation process, and reflections on the GAI-based lessons. Throughout the teaching cycle, the 15 students were required to submit weekly reflection journals after their dialogues with GAI, continuously recording their experiences of discerning GAI-generated information, cognitive changes, and emotional experiences. After the teaching intervention, semi-structured interviews were conducted with the 15 focal students, each lasting approximately 30-40 minutes, focusing on their perceptions of GAI, reasons for trust or skepticism, verification strategies, and cognitive changes. The teacher was also interviewed to gather insights into classroom observations, implementation difficulties, teaching effectiveness, and suggestions for instructional guidance. All interviews were audio-recorded and transcribed within 24 hours to ensure the completeness of the information. After data collection, all observation records, interview transcripts, student reflection journals, teaching journals, and researcher reflection journals were classified, screened, and organized, and a unified numbering and indexing system was established. This process ultimately formed a complete and standardized qualitative dataset, laying the foundation for subsequent data analysis.

### *3.5 Description of GAI Dialogue Tasks*

The GAI dialogue activities were designed in alignment with the Grade 7 history curriculum standards and provided opportunities to explore students' critical thinking processes related to the three research questions. All tasks were centered on core topics in the seventh-grade history syllabus, such as the Shang Yang Reform, the unification of the Qin Dynasty, and the imperial examination system of the Sui and Tang dynasties. Students completed the tasks independently through text-based dialogues with two GAI agents, DeepSeek and Doubao, under the guidance of the teacher.

The dialogue tasks consisted of three core types, each emphasizing different aspects of historical critical thinking. First, fact-checking tasks presented students with GAI-generated historical explanations containing factual inaccuracies or historically questionable claims, requiring students to identify inconsistencies and verify information against textbook content and historical source materials. This type of task was designed to examine the situations that triggered students' skepticism and their initial information discernment behaviors. Second, multi-perspective inquiry tasks guided students to discuss different interpretations and evaluative perspectives regarding historical figures or events with GAI, encouraging them to recognize divergent narratives and examine potential positional biases within AI-generated accounts. Third, comprehensive evaluation tasks asked students to systematically assess the credibility and limitations of GAI-generated historical narratives by cross-referencing multiple learning resources, prompting them to negotiate between AI-assisted information acquisition and independent critical judgment.

Full examples of task prompts and representative dialogue scenarios are provided in Appendix A. The task design aimed to create authentic inquiry contexts in which students could encounter the tension between trust and skepticism and engage in evidence-based reasoning during human-AI historical dialogues.

### *3.6 Data Analysis Methods*

This study employed thematic analysis as the core qualitative data analysis method, ensuring the rigor of data analysis through systematic coding and theme refinement. In this study, the five-dimensional framework of historical critical thinking served as the theoretical lens guiding the thematic analysis. Initial codes and sub-themes were continuously mapped back to the five dimensions to ensure analytical alignment with the conceptual framework, while allowing emergent themes from the data to complement the deductive structure. The researcher first read through and familiarized themselves with all the organized qualitative data, marked key content, core viewpoints, and typical cases closely related to the research themes, established a data index, and standardized the format of all textual data to create a normalized analysis text, laying the foundation for the subsequent coding work. After becoming sufficiently familiar with the data, the researcher proceeded to the open coding phase, reading through all textual data line by line without presupposing any thematic framework, and relying entirely on the raw data. Initial concepts related to core research themes, such as students' critical thinking development and GAI information discernment, were extracted, generating a detailed list of initial codes, with each code clearly labeled and annotated with specific explanations.

After the initial coding was completed, axial coding was carried out to systematically categorize, merge, and sort the initial codes, to deeply analyze the internal logical relationships among different codes, integrate similar codes, and extract sub-themes that accurately reflected the core characteristics of the data, thereby constructing a preliminary coding network. Based on the sub-themes derived from axial coding, selective coding was further conducted to integrate and refine the sub-themes, ultimately forming core themes that could directly address the central research questions of this study and establishing a complete and systematic thematic analysis framework. To ensure coding rigor, this study adopted a two-coder collaborative approach, in which two researchers familiar with qualitative research methods discussed any disagreements that arose during the coding process, reached consensus through thorough discussion, and made revisions accordingly. In the theme interpretation phase, multiple data sources, including classroom observations, interviews, student reflection journals, teaching journals, and the researcher reflection journals, were integrated to conduct an in-depth analysis of the core themes. Triangulation was achieved through cross-referencing different data sources, effectively reducing bias arising from a single perspective and ensuring the reliability and specificity of the analytical results, thereby providing solid support for the formation of the research conclusions.

## **4. Findings**

Through the analysis of students' dialogue records with the GAI agent, semi-structured interviews, student reflection journals, teaching journals, researcher reflection journals, and classroom observation data, the study identified three interrelated themes. The three main themes derived from the analysis are systematically mapped to the five dimensions of historical critical thinking. Specifically, Historical Source Judgment and Self-Reflective Questioning are reflected in trust construction and skepticism triggers; Evidence-Seeking and Comparison is central to verification strategies; Multiple Perspectives Recognition and Contextualized Critical Thinking are embodied in multi-perspective integration and contextual analysis; and Self-Reflective Questioning runs through all dimensions as a metacognitive foundation.

Table 4. Themes

Main Themes	Sub-themes	Description
Trust Construction and Information Discernment in Human-AI Interaction	Recognition and Acceptance of the Conversational Role	It refers to students' identification with GAI and their acceptance of AI as a carrier of historical knowledge during interaction. It focuses on psychological adaptation in the early stage of human-AI interaction, including whether students regard GAI as a reliable learning partner and recognize its role as a knowledge mediator.
	Evaluation of information credibility	It represents students' capability and judgment criteria to assess the authenticity and authority of historical information generated by GAI. It covers awareness of tracing information sources, willingness to verify facts and adoption of cross-check methods, reflecting learners' information screening competence in the digital era.
Critical Thinking Performance and Development	Triggers of Skepticism and Responses	During the learning process, the typical conditions and manifestations for students to doubt the information provided by GAI and enter a state of critical examination. This includes common situations that trigger cognitive alertness, the depth and frequency of questioning, as well as the differences in characteristics among different students in terms of triggering sensitive points and the tendency to question.
	Verification Strategies and Resource Deployment	After developing skepticism towards GAI-generated information, students deployed verification pathways and information resource combinations. These included preferences in verification methods, ways of organizing multi-source evidence, and the relative weight and trust hierarchy assigned to different types of information sources during the verification process.
	Discernment and Integration of Multiple Perspectives	When comparing GAI-generated content with their existing knowledge, students engaged in cognitive processing that involved recognizing and integrating information from different standpoints and sources to form a more comprehensive understanding. This reflected a thinking characteristic of transcending the limitations of a single narrative and seeking evidence from multiple sources.
Dual Effects of Learning Experience and Technology Adaptation in Human-AI Interaction	Emotional Experience and Self-Expression in Human-AI Interaction	The learning method of engaging in dialogues with GAI elicited emotional feelings and subjective experiences in students. These included self-awareness and expression of emotional experiences such as pleasure, confusion, frustration, and accomplishment during the participation process.
	Technology Dependence and Independent Thinking	After experiencing human-AI collaborative learning, students exhibited differentiated performance between independent thinking and technology dependence. This encompassed phenomena such as reduced inquiry initiative due to convenience and a lack of discernment toward GAI-generated information, as well as vigilance against cognitive laziness, efforts to break free from AI's preset frameworks, and the awareness and endeavor to maintain independent truth-seeking.
	Cognitive Change and Metacognition	It refers to students' awareness and summary of changes in their own thinking patterns after learning. It reflects their ability to describe their developmental progress in analyzing historical events, constructing arguments, and utilizing evidence, as well as to clarify directions for subsequent learning improvement.

#### 4.1 Trust Construction and Information Discernment in Human-AI Interaction

Trust construction and information discernment in human-AI interaction served as the foundation for students to engage in GAI dialogue learning and develop critical thinking. Students' understanding of the functional positioning of the GAI agent profoundly influenced their learning engagement and thinking participation. Most students exhibited a psychological characteristic of tentative acceptance during the initial contact. One student frankly stated, "At first, I thought AI was just an answering tool. So I just tried talking to it casually" (S03). As

their dialogue experience accumulated, students gradually recognized the instructional support value of GAI. Another student remarked, "It can simulate the perspective of historical figures and talk to me, which is more vivid than reading a book. I'm willing to discuss historical issues with it" (S11), demonstrating a cognitive shift from detachment to identification with the GAI as a historical learning partner. The teacher's classroom observations also confirmed this gradual process: "Seventh graders showed strong curiosity about new intelligent teaching tools, but in the early stage, they displayed polarized mindsets of either blind trust or complete distrust. Most students needed multiple dialogue experiences to establish a rational human-AI learning relationship" (T).

As students' recognition of the conversational role gradually took shape, they also began to develop the ability to make targeted judgments about the credibility of historical information, gradually moving away from unquestioning deference to the absolute authority of the intelligent tool. One student recorded in a reflection journal: "The historical viewpoints provided by AI may not be entirely correct. I will check key historical facts against the textbook" (S07). Another student mentioned: "When encountering contradictory historical statements, I will ask the AI more questions and consult historical sources for verification" (S15), initially demonstrating an awareness of source tracing and self-initiated fact-checking. Notably, some students also exposed weaknesses in their information judgment ability during this process; a few temporarily accepted GAI's erroneous statements as accurate information. This precisely illustrated that trust construction in human-AI interaction was not a unidirectional "move toward rationality," but a dynamic process of repeated adjustment between blind trust and careful judgment, providing genuine cognitive tension for the subsequent triggering and operation of students' critical thinking.

#### *4.2 Critical Thinking Performance and Dual Effects of Learning Experience in Human-AI Interaction*

Critical thinking performance and development constituted the core concern of this study. At the level of triggers of skepticism and responses, students gradually broke through the limitations of traditional history learning and were able to actively enter a state of critical scrutiny in specific situations. The study found that factual errors, logical inconsistencies, vague expressions, or limited perspectives in GAI's responses often served as key points that ignited students' awareness of skepticism. One student described: "When the AI's historical interpretation differed from what I had learned, I would neither directly deny myself nor readily trust the AI; instead, I would want to figure out which one was correct" (S06), reflecting a thinking tendency to initiate skepticism using prior knowledge as a reference. The teacher pointed out: "Compared with traditional classrooms, students' awareness of questioning in the GAI dialogue classroom was significantly enhanced. They were no longer confined to memorizing knowledge points, but were more willing to actively discern historical information and refute unreasonable viewpoints, showing a marked increase in critical thinking activity" (T).

After skepticism was triggered, students exhibited a certain degree of strategic differentiation in their verification strategies and resource deployment. Some students tended to rely on a single resource for information confirmation, such as only checking the textbook or only asking peers; whereas other students attempted to comprehensively use multiple approaches, including historical source retrieval, textbook verification, peer discussion, and teacher consultation, thus forming an awareness of cross-verification. One student mentioned: "When what the AI said differed from the textbook, I would first check the book, then discuss it with classmates, and finally confirm with the teacher" (S09), presenting a cognitive development trajectory transitioning from reliance on a single source to multi-source corroboration. At the level of discernment and integration of multiple perspectives, some students began to break through the single narrative framework and were able to recognize and integrate historical interpretations from different standpoints. A student reflected: "Talking with AI allows me to hear different historical perspectives, and I would more deeply understand that the choices of historical figures cannot be judged from a single angle; instead, they should be analyzed from multiple perspectives in the context of the era" (S02), demonstrating the budding of thinking in constructing historical understanding amidst multiple voices.

However, the introduction of GAI into history learning did not yield exclusively positive effects; students' performance showed a clear trend of differentiation. In terms of emotional experience and self-expression, the form of engaging in dialogues with GAI stimulated rich emotional feelings in students. One student remarked: "This kind of dialogue-based history learning was very interesting, and I felt a great sense of accomplishment when resolving doubts. However, when encountering factual errors by the AI, I also realized that history learning requires rigor" (S05), reflecting the coexistence of positive emotions and cognitive alertness. Yet, the erosion of independent thinking by technology dependence also began to surface. The study found that while some students enjoyed the convenience of GAI, their initiative in independent thinking and historical source retrieval diminished. A few students even blindly accepted GAI's erroneous statements, revealing the hidden costs of technology use.

One student, coded as S04, exhibited a consistent pattern of uncritical acceptance throughout the dialogue activities, reflecting an underdeveloped level of self-reflective questioning within the five-dimensional historical critical thinking framework. In his reflection journal, he wrote: "The AI says it is correct. I do not have time to check the textbook. I will just trust it." In the subsequent semi-structured interview, S04 further admitted that he rarely cross-referenced GAI's statements with other sources because doing so felt "too much work." He also explained that when the GAI provided a confident, detailed and fluent answer, he saw little reason to doubt its accuracy. Rather than accepting GAI content purely based on perceived factual correctness, S04 appeared to prioritize the efficiency and convenience of obtaining immediate, ready-made answers. Classroom observations similarly corroborated this pattern: he seldom initiated follow-up questioning or engaged in evidence-seeking behaviors during independent dialogue activities. This case illustrates how the perceived authority, narrative fluency, and low-effort accessibility of GAI responses may discourage some students from engaging in deeper verification processes, thereby increasing the risk of passive, uncritical acceptance of AI-generated content. This case stands in clear contrast to the active verification behaviors reported by most participants in this study.

Meanwhile, another group of students demonstrated a conscious vigilance against cognitive laziness. One student emphasized: "I must not let AI do the thinking for me. I still need to look up materials and make judgments on my own" (S01), reflecting a clear-headed weighing between technological convenience and independent truth-seeking. After the complete teaching cycle, students began to develop an awareness of the changes in their own thinking processes. One student summarized: "In the past, learning history meant rote memorization of events. Now I actively discern historical sources, sort out logic, and construct arguments from multiple perspectives. My thinking has become more comprehensive and rigorous" (S13). This showed that the student was able to clearly describe their progressive trajectory in historical analysis, argument construction, and the use of evidence, and to preliminarily identify directions for subsequent improvement.

Overall, the GAI dialogue teaching method effectively activated students' historical critical thinking abilities, enabling them to shift from passive knowledge recipients to active explorers. However, this transformation was not a linear narrative of progress, but rather a complex process unfolding in the interweaving of multiple forces such as trust construction, triggers of skepticism, and technology dependence. This necessitated that educators carefully examine and guide the dual effects of technology use in instructional design.

## 5. Discussion

### 5.1 Key Findings

Through a qualitative inquiry into seventh-grade history classrooms, this study focused on the triggering, operation, and developmental trajectory of students' critical thinking during dialogues with the GAI agent on historical topics. It found that GAI's intervention could, to a certain extent, promote the development of students' critical thinking, primarily converging in two types of situations: first, when GAI's responses exposed issues such as factual errors, vague expressions, or limited perspectives; second, the moments when students keenly detected discrepancies between what the AI stated and their own existing historical knowledge, textbook content, or historical records. These two types of moments were most likely to ignite students' skepticism and cognitive alertness.

Notably, differences in critical thinking performance among students with different levels of history academic performance (as shown in Table 1) manifested not merely in the frequency or intensity of skepticism, but in the specific dimensions of the five-dimensional historical critical thinking framework that were activated. Students with low history academic performance tended to be triggered primarily at the level of Historical Source Judgment Ability. They could identify obvious factual errors in GAI responses, especially when these contradicted explicit statements in textbooks. In contrast, students with high history academic performance were more likely to engage the higher-order dimensions of the framework, namely Evidence-Seeking and Comparison Capabilities and Multiple Perspectives Recognition Ability. These students not only detected factual inconsistencies but also questioned the logical coherence of GAI's narratives. They actively sought out multiple sources for cross-verification and demonstrated an awareness that the same historical event could be interpreted from different standpoints.

When verifying the credibility of historical information provided by GAI, students gradually developed a layered combination of strategies and a pattern of resource deployment. Historical source retrieval and cross-referencing, textbook content verification, peer discussion and corroboration, and teacher consultation constituted their primary verification pathways. Among these, textbooks and authoritative historical sources consistently served as the anchoring resources on which students relied most heavily. Comparatively, students with high history academic performance were more adept at flexibly employing multiple strategies for cross-verification, whereas students with low history academic performance tended to rely on a single resource for information confirmation.

This difference resonated with the findings of Barzilai and Zohar (2012), which suggested that students with higher epistemological levels demonstrated stronger abilities in integrating information online. Notably, McGrew et al. (2018) found that students rarely conducted systematic multi-source cross-verification when evaluating online information, mostly relying on a single resource to judge credibility. In contrast, some students in this study exhibited relatively conscious multi-source corroboration behavior. This difference may be partly attributed to the history discipline's emphasis on source criticism, but it may also be an artifact of the dialogic task structure. Future research should control for task effects.

However, the study also cautiously found that while GAI's involvement as a dialogue partner in history classrooms could effectively trigger students' critical thinking, it was not without cost. While enjoying the convenience of GAI, some students showed a decline in their initiative for independent thinking and historical source retrieval. This finding aligned with the views of Fan et al. (2024) and Yan et al. (2025).

Building on this observed gap in verification capacity, these group differences carry important pedagogical implications. While students with low history academic performance were generally able to identify obvious factual inaccuracies, they rarely progressed beyond basic fact-checking and often relied on a single source for verification. In contrast, students with high history academic performance were more likely to engage in multi-source corroboration, evaluate the logic of GAI-generated narratives, and consider alternative historical interpretations. To help students with low history academic performance move toward these higher-order dimensions of historical critical thinking, teachers may need to provide explicit, stepwise scaffolding. For example, instead of asking students merely whether a GAI-generated statement is correct, teachers could guide them through structured questioning prompts such as "What evidence supports this claim?", "What sources could verify it?", and "Could this event be interpreted differently from another perspective?". In addition, tasks that require students to compare GAI-generated responses with textbooks, primary sources, and peer interpretations may gradually cultivate habits of cross-verification and perspective-taking. Such instructional scaffolds can help students transition from simple factual verification to more sophisticated forms of historical reasoning, including evidence evaluation, contextual analysis, and multi-perspective interpretation, thereby supporting students' progression across the five dimensions of historical critical thinking identified in this study. This finding is consistent with the notion of the Zone of Proximal Development (Vygotsky, 1978), suggesting that appropriately designed scaffolds can support students in gradually extending their critical thinking beyond their current level of independent performance.

### *5.2 Limitations of the Study and Recommendations for Future*

This study clearly outlined the triggering mechanisms, operational pathways, and developmental trajectories of middle school students' historical critical thinking in the GAI dialogue context, further deepening the discipline-specific and contextualized understanding of critical thinking cultivation and providing solid empirical support for relevant theoretical research. On the practical level, the study offered frontline history teachers empirical references regarding the application of GAI in middle school history teaching, which could help teachers design GAI-related teaching activities more prudently and effectively in subsequent instruction.

However, this study was subject to certain limitations due to research conditions and design. First, the research setting was limited to the seventh-grade history classrooms of a single middle school in Beijing, which restricted the representativeness of the sample and failed to capture possible differences across school types and grade levels. Future research could expand the sampling scope to include diverse school types and grade levels, thereby testing the findings of this study with enhanced sample representativeness and systematically comparing the similarities and differences in critical thinking development among different student groups. Furthermore, as this study was deeply situated within the context of the history discipline, whether the triggering mechanisms of critical thinking and verification strategies identified here could be applied to other humanities or natural science disciplines would still require validation through cross-disciplinary comparative research. Second, as a cross-sectional study, this research did not track the long-term developmental trajectories of students' critical thinking and could not fully reveal the long-term effects of GAI use on middle school students' critical thinking. Subsequent research could extend the duration and employ a longitudinal design to examine in depth the sustained impact of GAI intervention on students' historical critical thinking and to delineate its long-term developmental trajectories and patterns. Third, although the qualitative methods adopted in this study allowed for an in-depth capture of the details of thinking processes, they lacked the support of large-scale quantitative data, thus limiting the generalizability of the conclusions. Subsequent research could adopt a mixed-methods design, introducing tools such as questionnaires and quantitative assessments on the basis of thick qualitative description, so as to achieve data triangulation and enhance the scientific rigor and persuasiveness of the findings. Meanwhile, further research could explore the differential effects of different types of GAI agents on students' critical thinking development, providing more precise guidance for the educational application and optimization

of GAI tools. Fourth, the sample in this study was not fully representative, as a majority of participants reported a high level of interest in history, high history academic performance, and high prior GAI usage frequency. This potential self selection bias may limit the generalizability of the findings to average or lower performing student populations.

## 6. Conclusion

This study demonstrates that sustained historical dialogues with GAI agents can serve as an effective lever for fostering students' critical thinking. The realization of this process requires several core conditions, such as allowing GAI to expose its imperfections, explicitly encouraging a spirit of skepticism, and creating a safe atmosphere for critical questioning in the classroom, among others. In such an environment, over time, some students tended to move from initially trusting GAI, toward active verification, and gradually developed a more critical and balanced stance. In the age of AI, the mission of history education is by no means to teach students to evade technology, but to enable them to continuously develop core competencies in history through repeated dialogues with this seemingly credible yet potentially unreliable intelligent other.

## Appendix A. Examples of GAI Dialogue Tasks and Representative Student-GAI Interaction Excerpts

This appendix presents three representative GAI dialogue tasks used in the teaching intervention, with corresponding student-GAI interaction excerpts. Each task is aligned with one core research question and targets specific dimensions of the five-dimensional historical critical thinking framework. All excerpts are translated from the original Chinese dialogues into English for presentation purposes. Minor adjustments have been made to improve readability while preserving the original meaning. Student codes correspond to those used in the main text.

Table A1. Example 1: Fact-Checking Task

Item	Description
Historical Topic	The Unification of the Qin Dynasty
Task Type	Fact-checking task
Learning Objective	To examine whether students could identify factual inconsistencies in GAI-generated historical explanations.
Prompt given to students	"Ask GAI to explain the reasons why Qin successfully unified China. Evaluate whether its explanation is historically reliable."

### Representative Student-GAI Interaction (Student S06)

GAI: The success of Qin's unification was driven by multiple factors, including strong military power, Shang Yang's Legalist reforms, and the integration of Confucian governance ideas that helped win popular support across the states.

Student: I think this part about Confucian ideas may not be correct. Qin's governing ideology was mainly Legalism, not Confucianism. I want to check this with the textbook.

Student reflection (from journal): "When the AI's historical interpretation differed from what I had learned, I would neither directly deny myself nor readily trust the AI; instead, I would want to figure out which one was correct."

Table A2. Example 2: Multi-Perspective Inquiry Task

Item	Description
Historical Topic	Shang Yang's Reforms
Task Type	Multi-perspective inquiry task
Learning Objective	To explore whether students could recognize different evaluations of historical figures and events from multiple standpoints.
Prompt given to students	"Was Shang Yang a successful reformer? Discuss his achievements and limitations with GAI from different historical perspectives."

Representative Student-GAI Interaction (Student S02)

GAI: Shang Yang was an outstanding reformer in Chinese history. His policies greatly strengthened the state of Qin and laid a solid foundation for its later unification of China.

Student: This is only one perspective. Although the reforms strengthened Qin, they also tightened social control and caused strong opposition from the noble class.

GAI: Indeed, evaluations of Shang Yang vary across historical perspectives. From the standpoint of state strength, his reforms were highly effective; from the perspective of social impacts, they also brought about harsh penalties and heightened social control.

Student reflection (from journal): "Historical figures should not be judged only by modern standards. We need to consider the conditions of their time."

Table A3. Example 3: Comprehensive Evaluation Task

Item	Description
Historical Topic	The Imperial Examination System in the Sui and Tang Dynasties
Task Type	Comprehensive evaluation task
Learning Objective	To examine students' ability to combine GAI assistance with independent judgment, and to evaluate the reliability and limitations of GAI-generated historical narratives.
Prompt given to students	"Ask GAI to analyze the historical significance of the imperial examination system in the Sui and Tang dynasties. After receiving its response, evaluate: (1) What aspects of its answer are reliable? (2) What aspects may be incomplete or one-sided? (3) What additional sources would you consult to form a more comprehensive understanding?"

Representative Student-GAI Interaction (Student S01)

GAI: The imperial examination system was a major institutional innovation in Chinese history. It broke the monopoly of aristocratic families on official positions, allowed talented people from humble backgrounds to enter the government, and promoted social mobility. It was a highly advanced talent selection system for its time.

Student: The main points of this answer are factually correct, but it is too one-sided. The examination was mainly limited to Confucian classics, which restricted the scope of knowledge. Also, wealthy families still had better access to education, so social mobility was not as great as it sounds. I need to check the textbook and other historical materials for more perspectives.

Student reflection (from journal): "GAI can give a clear and organized summary quickly, which is good for basic understanding. But its answers sometimes only show one side. I need to use my own thinking and check other sources to get a complete picture."

## References

- Amani, J., & Mkimbili, S. T. (2025). Developing critical thinking skills among secondary schools students: The role of research project. *Thinking Skills and Creativity*, 58, 101883. <https://doi.org/10.1016/j.tsc.2025.101883>
- Barzilai, S., & Zohar, A. (2012). Epistemic Thinking in Action: Evaluating and Integrating Online Sources. *Cognition and Instruction*, 30(1), 39-85. <https://doi.org/10.1080/07370008.2011.636495>
- 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, 45. <https://doi.org/10.1007/s10648-025-10020-8>
- Calma, A., & Davies, M. (2026). Assessing students' critical thinking abilities via a systematic evaluation of essays. *Studies in Higher Education*, 51(2), 422-437. <https://doi.org/10.1080/03075079.2025.2470969>
- Chen, H.-L., Vicki Widarso, G., & Sutrisno, H. (2020). A chatbot for learning chinese: Learning achievement and technology acceptance. *Journal of Educational Computing Research*, 58(6), 1161-1189. <https://doi.org/10.1177/0735633120929622>
- Chen, J., Mokmin, N. A. M., & Qi, S. (2025). Generative AI-powered arts-based learning in middle school history: Impact on achievement, motivation, and cognitive load. *The Journal of Educational Research*, 118(6), 688-700. <https://doi.org/10.1080/00220671.2025.2510395>
- Cheng, X., Sun, J., & Zarifis, A. (2020). Artificial intelligence and deep learning in educational technology research and practice. *British Journal of Educational Technology*, 51(5), 1653-1656. <https://doi.org/10.1111/bjet.13018>
- Chiu, M. C., & Hwang, G. J. (2026). Enhancing student creative and critical thinking in generative AI-empowered creation: a mind-mapping approach. *Interactive Learning Environments*, 34(2), 869-890. <https://doi.org/10.1080/10494820.2025.2511244>
- Demircioglu, T., Karakus, M., & Ucar, S. (2023). Developing Students' Critical Thinking Skills and Argumentation Abilities Through Augmented Reality-Based Argumentation Activities in Science Classes. *Science & Education* 32, 1165-1195. <https://doi.org/10.1007/s11191-022-00369-5>
- Fan, Y., Tang, L., Le, H., Shen, K., Tan, S., Zhao, Y., Shen, Y., Li, X., & Dragan Gašević. (2024). Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. *British Journal of Educational Technology*, 56(2), 489-530. <https://doi.org/10.1111/bjet.13544>
- Guo, H., Zhou, Z., Ma, F., Ning, Z. (2026). Critical thinking and AI-Assisted creativity in engineering education: differences between Undergraduate, Master's, and doctoral students. *Education and Information Technologies*, 31, 1033-1058. <https://doi.org/10.1007/s10639-025-13844-7>
- Guo, J., Ma, Y., Li, T., Noetel, M., Liao, K., & Greiff, S. (2024). Harnessing Artificial Intelligence in Generative Content for enhancing motivation in learning. *Learning and Individual Differences*, 116, 102547. <https://doi.org/10.1016/j.lindif.2024.102547>
- Lee, J., Hicke, Y., Yu, R., Brooks, C., & Kizilcec, R. F. (2024). The life cycle of large language models in education: A framework for understanding sources of bias. *British Journal of Educational Technology*, 55, 1982-2002. <https://doi.org/10.1111/bjet.13505>
- Leibovitch, Y. M., Beencke, A., Ellerton, P. J., McBrien, C., Robinson-Taylor, C.-L., & Brown, D. J. (2024). Teachers' (evolving) beliefs about critical thinking education during professional learning: A multi-case study. *Thinking Skills and Creativity*, 56, 101725. <https://doi.org/10.1016/j.tsc.2024.101725>
- Li, Z., Li, B., Wang, X., & Wang, T. (2025). Cultivating critical thinking dispositions in EGAP courses: The importance of teaching materials selection. *English for Specific Purposes*, 81, 184-196. <https://doi.org/10.1016/j.esp.2025.10.006>
- Lin, X.-F., Zhou, Y., Shen, W., Luo, G., Xian, X., & Pang, B. (2023). Modeling the structural relationships among Chinese secondary school students' computational thinking efficacy in learning AI, AI literacy, and approaches to learning AI. *Education and Information Technologies*, 29, 6189-6215. <https://doi.org/10.1007/s10639-023-12029-4>
- Liu, B., Liu, W., Zeng, W., & Peng, Y. (2025). Exploring the application potential of generative artificial intelligence in high school geography teaching: Scenarios, limitations, and improvement strategies. *The Journal of Educational Research*, 118(6), 674-687. <https://doi.org/10.1080/00220671.2025.2510396>
- Liu, J., Wang, C., Liu, Z., Gao, M., Xu, Y., Chen, J., & Cheng, Y. (2024). A bibliometric analysis of generative AI in education: current status and development. *Asia Pacific Journal of Education*, 44(1), 156-175. <https://doi.org/10.1080/02188791.2024.2305170>
- McGrew, S., Breakstone, J., Ortega, T., Smith, M., & Wineburg, S. (2018). Can Students Evaluate Online Sources? Learning From Assessments of Civic Online Reasoning. *Theory & Research in Social Education*, 46(2), 165-193. <https://doi.org/10.1080/00933104.2017.1416320>

- Plass, J. L., & Froehlich, F. (2025). The future of personalized learning with AI. *Learning and Individual Differences, 124*, 102813. <https://doi.org/10.1016/j.lindif.2025.102813>
- Shen, Y., & Chen, L. (2025). “Critical chatting” or “casual cheating”: how graduate EFL students utilize ChatGPT for academic writing. *Computer Assisted Language Learning*, 1-29. <https://doi.org/10.1080/09588221.2025.2479141>
- Song, Y., Roohr, K. C., & Kirova, D. (2024). Exploring approaches for developing and evaluating workplace critical thinking skills. *Thinking Skills and Creativity, 51*, 101460. <https://doi.org/10.1016/j.tsc.2023.101460>
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes* (M. Cole, V. John-Steiner, S. Scribner, & E. Souberman, Eds.). Harvard University Press. <https://doi.org/10.2307/j.ctvjf9vz4>
- Wall, V., Bedford, A., & Redmond, P. (2025). Generative artificial intelligence in education: Initial principles developed from practitioner reflexive research. *The Journal of Educational Research, 118*(6), 610-621. <https://doi.org/10.1080/00220671.2025.2510398>
- Wang, W., Wang, Y., Chen, J., Wang, X., Zhang, H., Guo, C., & Peng, Y. (2026). The Effectiveness of AI-Supported Personalized Feedback on Students’ Learning Outcomes and Motivation: A Meta-Analysis. *Journal of Educational Computing Research, 64*(3), 724-755. <https://doi.org/10.1177/07356331251410020>
- Wulff, P., & Kubsch, M. (2025). Learning against the machine: the double edged sword of (Gen)AI in STEM education. *International Journal of STEM Education, 12*, 66. <https://doi.org/10.1186/s40594-025-00588-6>
- Yan, L., Greiff, S., Lodge, J. M., & Gašević, D. (2025). Distinguishing performance gains from learning when using generative AI. *Nature Reviews Psychology, 4*, 435-436. <https://doi.org/10.1038/s44159-025-00467-5>
- Zhang, Q., Siraj, S. B., & Abdul Razak, R. B. (2025). Effects of AI chatbots on EFL students’ critical thinking skills and intrinsic motivation in argumentative writing. *Innovation in Language Learning and Teaching, 1-29*. <https://doi.org/10.1080/17501229.2025.2515111>