Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 09:2026
DOI: 10.5281/zenodo.22771708

A GENERATIVE AI-BASED INSTRUCTIONAL STRATEGY FOR CRITICAL THINKING AND EPISTEMIC AUTONOMY: A PRETEST–POSTTEST STUDY

IBRAHIM ENAD EIADA AL-MASAEED
PhD, Philosophy of Curriculum and Teaching Methods, Faculty of Educational Sciences, Al-Hussein Bin Talal University. Email:almasaediia@gmail.com

SARAH SALEH ALAWAD
PhD, Philosophy of Curriculum and Teaching Methods, Faculty of Educational Sciences, Al-Hussein Bin Talal University. Email:alawadsarah94@gmail.com

AHMAD AWAD TASHTOUSH*
PhD, Philosophy of Curriculum and Teaching Methods, Faculty of Educational Sciences, Al-Hussein Bin Talal University. Empirical Study | Undergraduate Higher Education.
*Corresponding Author Email: amadtashtosh1984@gmail.com

HEYAM ABDELKAREEM ALDIABAT
PhD, Philosophy of Curriculum and Teaching Methods, Faculty of Educational Sciences, Al-Hussein Bin Talal University. Email: sohelbtoush@gmail.com

HALA SALEM HAMAD AL HJOOJ
PhD, Philosophy of Curriculum and Teaching Methods, Faculty of Educational Sciences, Al-Hussein Bin Talal University. Email:uiyr46788@gmail.com

Abstract
This study examined a generative artificial intelligence (GenAI)-based instructional strategy using pretest and posttest responses from 96 undergraduate students, with 48 in each condition. The eight-week program emphasized independent verification, argument critique, and reflective justification. Outcomes were a 20-item critical-thinking score (0–40) and an 18-item epistemic-autonomy mean (1–5). Both pretests were controlled in the posttest models. The joint group contrast yielded Wilks lambda = .674, $F(2, 91) = 21.99$, $p < .001$. Adjusted differences were 6.82 critical-thinking points (95% CI [4.17, 9.47]; partial eta squared = .221) and 0.79 autonomy points (95% CI [0.53, 1.04]; partial eta squared = .295). Both outcome tests remained significant after Holm adjustment, and HC3 sensitivity intervals excluded zero. Students in the GenAI condition had higher adjusted posttest scores on both outcomes. The findings support further investigation of verification-centered GenAI instruction, with interpretation limited by the study design, measurement evidence, and the absence of reported delayed transfer outcomes.

Keywords: Generative Artificial Intelligence; Critical Thinking; Epistemic Autonomy; Higher Education; Pretest–Posttest Design; ANCOVA; AI Literacy.

1. INTRODUCTION

Generative artificial intelligence has entered the field of higher education at a very fast pace as large language models are able to generate explanations, examples, summaries, questions, and feedback in response to typical language prompts (Chan, 2023). This availability has brought about both opportunities of personalized support and quick generation of ideas, yet has led to issues of authorship, evaluation, factuality, and the

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 09:2026
DOI: 10.5281/zenodo.22771708

allocation of cognitive labour between students and AI (Chan and Hu, 2023; Kasneci et al., 2023).

The fluency of the output of GenAI does not translate into an educational value (Farrokhnia et al., 2024). Big language models produce statistically realistic sequences, not personally validated knowledge, and as such polished answers might contain a factual error, fake sources, underlying assumptions or culturally biased representations (Dwivedi et al., 2023). So that students think fluency is equivalent to truth, the tool can encourage superficial recognition instead of a judgment based on evidence (Miao and Holmes, 2023).

It is particularly problematic in undergraduate education, when students should be able to go beyond the process of reproduction of information to the analysis, evaluation, inference, and reasoning decision making (Abrami et al., 2015). Critical thinking is often seen as intentional and self-directed judgment, which entails the interpretation of information, the assessment of facts, the warranted conclusions and the articulation of the grounds underpinning such conclusions (Halpern, 1998).

Such thinking can be reinforced or undermined by GenAI in accordance with its incorporation within instructional activities. In case students demand final responses and send minimally edited output, the system can eliminate the constructive challenge in which a reasoning process occurs (Chi and Wylie, 2014). By asking questions, contrasting AI output with reliable sources, uncovering assumptions, developing counterarguments, and updating conclusions, GenAI can be the subject of criticism and a prompt to engage in higher-level thinking (Essien et al., 2024; Kasneci et al., 2023).

Another issue that is closely associated is the epistemic autonomy: the ability of the learner to have responsible control of knowing (Chinn et al., 2011). Students with epistemically autonomy do not just resist authority and demand to be opinionated; they decide what they need to know, gain a sense of the credibility of the information, weigh trust to the evidence they have, track the uncertainty, and take the final verdict. This concept links personal epistemology, epistemic cognition, reflective judgment and self-regulated learning (Muis, 2007).

GenAI and epistemic autonomy are thus paradoxical in relation to each other. On the one hand, the access to explanations and alternative perspective through conversations can broaden the options of students to explore questions without an expert (Barzilai and Chinn, 2018). Conversely, relying on a system which talks with undue confidence can shift epistemic power to the learner to a black box (Chinn et al., 2011). Conscientious instructional design should maintain the student ability to establish epistemic goals, evidence, and own the judgment (UNESCO, 2021).

The current research in higher education has charted student perceptions, potential benefits, academic-integrity issues and institutional policy requirements. The reviews always include the statements of promise and risk, as well as stating that the evidence base continues to be dominated by conceptual or descriptive studies or short-term studies

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 09:2026
DOI: 10.5281/zenodo.22771708

but not theory-driven interventions with quantifiable learning outcomes (Chan and Hu, 2023; Crompton and Burke, 2023; Lo, 2023; Tlili et al., 2023).

The current research covered this gap in the methodology by adopting the instructional approach where students operated GenAI by engaging in a repetitive process of question formulation, provisional generation, independent verification, argument analysis, revision, and reflective justification. The approach integrates explicit teaching in critical thinking with epistemic standards and constructive interactive learning exercises, thus defining AI literacy as disciplined judgment, as opposed to familiarity with tools.

1.1 Statement of the Problem

GenAI systems are becoming more accessible to undergraduate students prior to systematic training on how to perform verification of AI-created claims, evidence provenance, bias detection, or an understanding of when automated help should not be utilized (Miao and Holmes, 2023). Since these systems generate fluent and seemingly coherent answers, students can perceive linguistic confidence as the indicator of accuracy (Farrokhnia et al., 2024). Uncontrolled usage can also focus more on completing tasks quickly than thinking and promote students to delegate significant cognitive tasks to the system (Kasneci et al., 2023). This can lead to an illusion of achievement in academic performance without any development in the critical-thinking skills of students (Cotton et al., 2024).

A number of several studies have addressed this issue. The previous studies have found possible advantages of GenAI, such as personalized help, real-time feedback, generation of ideas, and academic writing support. Nevertheless, inaccuracy, privacy, academic integrity, overreliance, as well as the potential undermining of independent thought, were issues of concern in the past studies too. Essien et al. (2024) discovered that the most prominent gains through the use of AI text generators were in the lower levels of Bloom taxonomy, which casts doubt on the role of AI text generators in higher-order assessment and independent judgment (Chan and Hu, 2023; Crompton and Burke, 2023; Essien et al., 2024; Kasneci et al., 2023; Lo, 2023).

Accordingly, the research problem lies in the limited intervention-based evidence showing whether a structured GenAI instructional strategy can improve undergraduate students' critical thinking and epistemic autonomy. To overcome this issue, it is necessary to use AI-generated answers as drafts, which students have to analyze, check, criticize, and edit (Chi and Wylie, 2014). Independent source comparison, unsupported assumptions identification, counterargument, and uncertainty monitoring alongside staying accountable to final conclusions should also be included in the list of tasks that students should perform (Barzilai and Chinn, 2018).

1.2 Purpose of the Study

The aim of the work was to compare a structured GenAI-based teaching approach to the critical thinking and epistemic autonomy of undergraduate learners. The posttest results

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 09:2026
DOI: 10.5281/zenodo.22771708

of the experimental and control groups were compared and their results adjusted to both the baseline measures.

1.3 Objectives

- To design and implement an eight-week instructional strategy that uses GenAI for verification, argumentation, counterargument generation, source comparison, and reflective judgment.
- To estimate the adjusted effect of the strategy on students' posttest critical-thinking scores.
- To estimate the adjusted effect of the strategy on students' posttest epistemic-autonomy scores.
- To examine the combined multivariate effect of the strategy on both outcomes.
- To assess the robustness of the adjusted group differences using sensitivity analyses and confidence intervals.

1.4 Research Questions and Hypotheses

The study was guided by the following questions:

RQ1. What is the effect of the GenAI-based instructional strategy on undergraduate students' critical thinking after controlling for baseline critical thinking and epistemic autonomy?

RQ2. What is the effect of the GenAI-based instructional strategy on undergraduate students' epistemic autonomy after controlling for baseline epistemic autonomy and critical thinking?

RQ3. Does the strategy have a statistically significant combined effect on posttest critical thinking and epistemic autonomy after adjustment for both baseline outcomes?

The corresponding null hypotheses were:

H01. There is no statistically significant adjusted posttest difference between the experimental and control groups in critical thinking at alpha = .05.

H02. There is no statistically significant adjusted posttest difference between the experimental and control groups in epistemic autonomy at alpha = .05.

H03. There is no statistically significant multivariate group effect on the combined posttest outcomes after adjustment for both baseline measures at alpha = .05.

1.5 Significance of the Study

The study relates critical-thinking training to epistemic thinking and self-regulation by highlighting the role of learners in assessing evidence, keeping track of uncertainty, and giving ultimate judgments when using GenAI. It offers a one-stop solution to the analysis of changes in critical thinking and epistemic autonomy in terms of reliability, adjusting the

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 09:2026
DOI: 10.5281/zenodo.22771708

baseline, testing assumptions, confidence interval, and estimating the effect size. The study also presents a sequential instructional pattern which can be applied to other undergraduate courses with a constant cycle of questioning, verifying, comparing sources, assessing arguments, reflecting and justifying. This strategy can assist instructors to create GenAI-facilitated tasks that can encourage active learning and critical thinking without violating the accountability and integrity of academic work in the hands of students.

1.6 Conceptual and Operational Definitions

Generative Artificial Intelligence GenAI is, conceptually, machine-learning systems that can generate new text or other content based on user prompts (Kasneci et al., 2023; Miao and Holmes, 2023). Operationally, it is the conversational language model that served as an intentionally fallible collaborator in the generation of claims, counterarguments, questions and draft that students had to be able to scrutinize and amend.

Critical Thinking In its conceptual form, critical thinking is mindful, deliberate judgment that entails interpretation, analysis, assessment, inference, and explanation (Andreucci-Annunziata et al., 2023). In terms of operations it is the total score on the 20-item Critical Thinking Performance Test of the range 0-40, where a higher score indicates a better evidence-based reasoning.

Epistemic Autonomy Conceptually, epistemic autonomy is the self-governmental responsibility of knowledge seeking and assessment (Barzilai and Chinn, 2018; Chinn et al., 2011). It is operationally defined as the average rating on the 18-item Epistemic Autonomy Scale, which is scored between 1 and 5, including source independence, evidential justification, and metacognitive authorship.

Genai-Based Instructional Strategy An eight-week plan where students create questions, get AI to provide tentative answers, searching information in reliable sources, assessing the quality of arguments, making and testing counterarguments, revising the product, and recording why they believe AI ideas should be accepted or rejected.

2. LITERATURE REVIEW AND THEORETICAL FRAMEWORK

2.1 Generative AI as a Pedagogical Affordance and Epistemic Risk

GenAI differs from conventional search and tutoring tools because it can synthesize a context-sensitive response in real time and continue the exchange through follow-up prompts (Grassini, 2023; Kasneci et al., 2023). These features facilitate brainstorming, alternative explanation, formative feedback, and personalized examples, yet the identical conversational fluency can be misleading when it comes to the lack of source transparency, as well as solid factual grounding (Lim et al., 2023).

According to the affordance view, the relationship between the capabilities of the tools, the design of the task, the knowledge held by the learners, and the supervision by the

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 09:2026
DOI: 10.5281/zenodo.22771708

teachers brings about educational value (Chan, 2023). A prompt that asks a complete answer encourages a different mental process as does a prompt that asks conflicting explanations, evidence gaps, or a critique of the argument made by the learner. As a result, GenAI analyses must indicate how students actually use the output and not categorize a course as AI-supported (Tlili et al., 2023).

The epistemic dangers do not pertain to hallucination of fact only. Uncertainty can be flattened, dominant assumptions replicated, citations created, and arguments based on assumptions that do not justify their conclusions, by model responses (Dwivedi et al., 2023; Farrokhnia et al., 2024). These threats put the source triangulation, uncertainty calibration, and explicit reasoning criteria at the primary educational activities, rather than the technical safeguards that are not mandatory (UNESCO, 2021).

2.2 Critical Thinking

Critical thinking includes both cognitive skills and a disposition to use them. These are core skills interpreting information, relationships, assessing the reliability and relevance of evidence, warranted inferences and explanations (Ennis, 1985). The dispositional factor entails open-mindedness, persistence, readiness to rethink, and sensitivity to quality of reasons (Halpern, 1998).

Critical thinking is best taught explicitly, through practice on authentic problems, with a dialogue and metacognitive reflection (Abrami et al., 2008). The meta-analysis shows that interventions that are based on explicit instructions and subject-matter immersion tend to be more effective than those that presuppose the appearance of critical thinking as an incidental process (Abrami et al., 2015).Relations between claims, reasons, objections, and rebuttals can be externalized with the help of argument mapping and structured evaluation (Dwyer et al., 2012). It is especially applicable to GenAI since students have an opportunity to contrast a fluent response with an explicit map and find unsupported transitions that would otherwise be obscure (Dwyer et al., 2014).

The application of critical thinking is also not limited to academic achievement (Butler, 2012). The quality of real-world decision making and reduced exposure to adverse life experience has been related to measures of critical-thinking competence, which has led to the perception that evidence assessment and reflective judgment are transferable educational objectives (Butler et al., 2012).

2.3 Epistemic Autonomy

Personal epistemology research examines beliefs about the nature, source, certainty, and justification of knowledge. Lower-level jobs are more likely to interpret knowledge as definite, uncomplicated, and delivered through authority, but more advanced jobs acknowledge that assertions must be assessed relative to the facts, approach, setting, and alternative accounts (Barzilai and Chinn, 2018).

Being epistemically autonomous does not just involve having advanced beliefs. It involves effective performance in those instances where the learner has to decide on an epistemic purpose, design correct strategies, apply evidence criteria, and manage the knowing

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 09:2026
DOI: 10.5281/zenodo.22771708

process. Autonomy thus does not conflict with the valid dependence on experts; its characteristic is the measured dependence as opposed to a reflexive trust or distrust of authority (Chinn et al., 2011).

An explanation of how epistemic beliefs may influence task interpretation, goal setting, monitoring, strategy selection and evaluation can be explained through self-regulated learning (Muis, 2007). A student who believes that a plausible answer is adequate will cease to search sooner than a student who needs corroboration and is able to describe what he/she does not know. The latter trend is more epistemic authorship as the learner is still accountable to the criteria applied to arrive at a conclusion (Zimmerman, 2002).

Reflective-judgment research also explains how development involves a reliance on some authority to a use of reason in dealing with uncertainty. GenAI opens such opportunities regularly due to its answers being both useful and defeasible at the same time: students need to choose what to believe in, what to check, and how much a conclusion is justified (Barzilai and Chinn, 2024).

2.4 Integrated Theoretical Framework

The plan incorporates three perspectives that are complementary. To begin with, the AIR model of epistemic cognition focuses on epistemic goals, ideals or standards, and trustworthy procedures to embrace the goals (Barzilai and Chinn, 2018). Students hence start by defining what they should know, standards of a satisfactory response and the verification procedures prior to accepting an AI response (Chinn et al., 2011).

Second, ICAP framework anticipates more profound learning as the students leave the passive reception to active, constructive, and interactive learning. The intervention does not involve passive copying since it involves students needing to come up with critiques, rehearse arguments, and negotiate evaluations with classmates; these processes render the reasoning of the learner transparent and subject to debate (Chi and Wylie, 2014).

Third, self-guided learning emphasizes of planning, monitoring, control and reflection cycles. Students keep an AI-use journal, which contains prompts, sources of verification, identified weaknesses, revisions and the reasoning behind their final decisions. The aim of this routine is to increase metacognitive control, and avoid the development of convenience into epistemic dependence (Muis, 2007; Zimmerman, 2002).

Collectively, the frameworks forecast that only in the case of instructional constraints, evaluative authority is retained with the learner, and GenAI would be able to enhance critical thinking and epistemic autonomy (Abrami et al., 2015; Long and Magerko, 2020). The working mechanism is not exposure to AI but a habit of engaging in evidence-based, productive, and self-regulated epistemics (Ng et al., 2021).

2.5 Generative AI, Cognitive Offloading, and Productive Effort

The cognitive processes that students can still do after the implementation of GenAI are the determinants of the educational effects of GenAI. Providing a fully developed explanation, argument, or synthesis, students can save time and go around the process of comparison, structuring, and assessment in which the comprehension builds (Kasneci

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 09:2026
DOI: 10.5281/zenodo.22771708

et al., 2023). This potential is similar to cognitive offloading: a third party resource executes some portion of the task that the learner would. The concept of offloading is not necessarily good or bad, as the tools can liberate attention to work on more advanced tasks, and the value of saved effort will be determined by whether this energy will be spent on analysis and judgment or the tools will be eliminated as part of the learning experience (Lim et al., 2023).

Assistance and substitution are most applicable to critical thinking. GenAI can be used by a learner to generate candidate explanations, objections or examples and simultaneously assess their applicability and strength of evidence. In such a set-up, the model increases the content to be subjected to examination. Through comparison, when a fluent answer is accepted as a complete response, production and evaluation is transferred to the tool. The same technology can consequently assist or inhibit critical thinking as dictated by the rules of the task, standards of assessment and the level of accountability in the learner (Chi and Wylie, 2014).

The studies of critical-thinking teaching imply that effective work is more probable when the students tackle real-life issues, discuss them, have instructions, and train definite reasoning steps. These characteristics are more theoretically justified as compared to non-directed exposure to a computerized system (Abrami et al., 2008). GenAI strategy must be thought of as a guided pedagogy comprising of prompts, validation routines, peer challenge and revision demands and not as mere access to ChatGPT or an alternative model (Abrami et al., 2015).

Another concept that explains the role of feedback is this. The AI feedback can be used in real-time to assist students in finding an alternative or gaps, but speed can prompt learners to find the first plausible answer adequate (Zimmerman, 2002). Feedback that can be used in the educational process must result in additional thought: the students must recognize the assertion under dispute, find external supports, justify why a correction is required, and record the lack of clarity. These reactions make feedback more of a metacognitic process as opposed to an answer (Muis, 2007).

2.6 Source Evaluation, Epistemic Vigilance, and Calibrated Trust

GenAI makes the evaluation of the sources more challenging since a response can consist of both good statements, unsupported inferences, and fake details in a single fluent text. Traditional indicators like grammatical correctness, confident voice, and seeming cohesion are thus inaccurate proxies on truth (Chinn et al., 2011; Long and Magerko, 2020). Students require epistemic alert: they should enquire to where a statement was originated, what evidence would substantiate it, is the sources independent, and the strength of the conclusion. Such practices relate AI literacy to the existing descriptions of evaluations of evidence and epistemic cognition (Ng et al., 2021).

Measured trust gives a more justifiable educational objective than unconditioned acceptance or categorical rejection of AI. The right dependence is dependent on the degree of stakes of the task, surface of outside evidence, openness of the reaction, and the ability of the learner. Epistemic autonomy thus involves being aware when to consult

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 09:2026
DOI: 10.5281/zenodo.22771708

expertise and what kind of claim is beyond his/her comprehension to justify. An even expert or automatic support cannot make the learner not choose standards and justify the reasons to rely on them (Barzilai and Chinn, 2018).

Triangulation of the sources should not be superficial. Sometimes, online searches to confirm doubtful information can actually make it seem more accurate when users are exposed to low-quality information that restates or reinforces the same misleading information (Aslett et al., 2024). GenAI systems can also create false references or give the wrong bibliographic data, so the fact that multiple citations are made is not enough to say that it is accurate (Walters and Wilder, 2023). Each source should then be documented as to its origin, expertise, methodology, recency and independence instead of just the number of sources which seem to concur.

Another aspect of epistemic autonomy is uncertainty communication. Recent experimental results reveal that users can overrate the correctness of the answers provided by the LLM in case they base on the default explanations of the models (Steyvers et al., 2025). Prolonged explanations can also provide more confidence to the users without enhancing their discriminatory skills between the right and wrong answers (Steyvers et al., 2025). Students are thus advised to differentiate between known facts, possible interpretations and questions that cannot be answered as they evaluate the quality of their conclusions to the quality of available evidence. Requesting students to compose uncertainty statements can aid in understanding whether or not their confidence is suitably matched to the evidence and warrant of their conclusions (Barzilai and Chinn, 2024).

2.7 Instructional Conditions for Higher-Order Engagement

CAP framework can provide a valuable explanation of disparity in learning results of GenAI. The main difference between reading or slightly editing an AI response and creating a critique, creating a counterargument, and justifying a revision is between passive or active involvement and constructive involvement. There is an interactive element to peer comparison of judgments in that the students are asked to answer each other as opposed to splitting the work. The framework anticipates learning in a deeper level because the engagement shifts towards constructive and interactive levels (Chi and Wylie, 2014).

It can be aided by argument mapping, which makes the externalization of claims, evidence, warrants, objections, and rebuttals. An obvious structure diminishes the persuasive power of fluent prose and gives the students a chance to find ungrounded relationships (Dwyer et al., 2012). Empirical and conceptual studies on argument mapping indicate that it is capable of supporting the performance of critical-thinking when combined with teaching and practice, but the advantage should not be attributed to all disciplines and applications (Dwyer et al., 2014).

Group critique can also enhance argumentation in that students will have to state the evaluation criteria and answer opposing interpretations (Lo, 2023). Dialogue may reveal implicit assumptions, and demand justification which may not be forthcoming in a one-on-

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 09:2026
DOI: 10.5281/zenodo.22771708

one interaction with a model. Meta-analytic results distinguish dialogue, and real issues among fruitful characteristics of critical-thinking teaching, which justify their application in the instructional plan regardless of any assertion regarding GenAI itself (Abrami et al., 2015). Guidance by teachers is still paramount. Little-domain-knowledge students might fail to identify more advanced errors, and unbridled prompting will reward more superficial productivity (Crompton and Burke, 2023). Teachers have to choose activities where they can verify the results, exemplify the analysis of AI output, offer plausible sets of sources when necessary, and stepwise eliminate scaffolds. This position is aligned with the reviews that outline the potential of EFL as well as the issue of reliability, integrity, and equity of GenAI in higher education (Farrokhnia et al., 2024).

2.8 Measurement Issues and the Need for Direct Evidence

Evaluating the strategy requires measures aligned with the claimed outcomes. Product quality alone is ambiguous because a polished assignment can reflect model capability, editing skill, or teacher support rather than improved reasoning (Butler, 2012). A critical-thinking assessment should require students to analyze evidence, identify assumptions, draw warranted inferences, and explain uncertainty in unfamiliar tasks. Transfer tasks completed without immediate AI assistance are especially important for determining whether students acquired a capability they can exercise independently (Halpern, 1998).

Epistemic autonomy is also difficult to capture with self-report alone. Students may endorse statements about verification and independent judgment without enacting them under time pressure or when AI output appears convincing (Barzilai & Chinn, 2018; Muis, 2007). Stronger evidence would combine questionnaires with behavioral indicators, including source diversity, correction of fabricated citations, revision quality, appropriate rejection of unsupported claims, and justified help seeking. Process traces can show how a conclusion was reached, although they require clear privacy and consent procedures (Miao & Holmes, 2023). Reliability and validity provide different forms of evidence about measurement quality. Internal consistency indicates the degree to which item scores are related, but a high coefficient does not establish that the instrument measures the intended construct (DeVellis & Thorpe, 2021). Structural validity and measurement invariance should also be examined to determine whether scores represent the same construct across groups or measurement occasions (Józsa et al., 2023). When performance assessments depend on human judgment, researchers should report evidence of scorer agreement and use assessors who are unaware of participants' study conditions whenever feasible. These considerations are essential because a precise statistical estimate has limited meaning when the interpretation of the outcome scores is inadequately supported.

The current literature therefore supports a cautious research model: specify the instructional mechanism, measure reasoning directly, examine autonomous performance after scaffolds are removed, and document implementation. Existing studies and reviews establish relevance and plausible mechanisms, but they do not justify transferring a single effect estimate across populations or treating AI use as a uniform intervention (Chan & Hu, 2023; Essien et al., 2024; Tlili et al., 2023).

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 09:2026
DOI: 10.5281/zenodo.22771708

2.9 Structure of the GenAI-Based Instructional Strategy

Table 1: Eight-Week Instructional Strategy

Week	Instructional focus	Student product	Targeted process
1. Orientation	Capabilities, limitations, disclosure, privacy	Risk audit of an AI response	AI literacy; ethical awareness
2. Question design	Epistemic aims and high-quality prompts	Question-and-criteria sheet	Problem framing
3. Claim verification	Fact checking and source triangulation	Verification matrix	Evaluation; source independence
4. Argument mapping	Claims, evidence, warrants, assumptions	Argument map	Analysis; justification
5. Counterargument	Alternative explanations and rebuttals	Counterargument brief	Inference; open-mindedness
6. Bias and uncertainty	Bias, missing perspectives, confidence calibration	Uncertainty statement	Reflective judgment
7. Collaborative critique	Peer review of AI-supported reasoning	Revised position paper	Interactive engagement
8. Transfer and reflection	Independent application to a novel problem	Portfolio and AI-use audit trail	Metacognitive authorship

Note. The program comprised eight weeks of instruction.

2.10 Research Gap

Past studies on GenAI in higher education have mostly addressed the perceptions of students, usage patterns, academic writing, academic integrity, and overall opportunities and challenges of these technologies (Chan and Hu, 2023; Crompton and Burke, 2023; Lo, 2023). Even though a few studies have examined the correlation between GenAI and critical thinking, the evidence obtained is inconsistent and the studies are usually founded on descriptive or perception-based evidence, but not on structured intervention programs (Essien et al., 2024). Therefore, there is a dearth of evidence related to the ability of a verification-based GenAI instructional plan to generate meaningful changes in the performance of undergraduate students in terms of critical-thinking.

Another gap is the fact many people pay little or no attention to the possibility of having epistemic autonomy as an achievement of GenAI-supported teaching. The existing literature seldom addresses the question of whether students would maintain the responsibility of evidence selection, source evaluation, tracking of uncertainty, and final verdicts when using GenAI. There is a lack of research that explores critical thinking and epistemic autonomy in a combined study or considers the development of the two in a framework of a systematic instructional plan. Thus, the current research fills this gap with an investigation of the impact of a GenAI-based instructional approach on both outcomes and adjusting the baseline scores of the students and presenting adjusted differences, confidence intervals, and effect sizes.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 09:2026
DOI: 10.5281/zenodo.22771708

3. METHODOLOGY

3.1 Research Approach and Design

The study used a two-group pretest–posttest design to examine an eight-week GenAI-based instructional intervention. The data analyzed represented 96 undergraduate students and the results of the experimental and control conditions were compared, but considering both pretest results. Student records were considered as independent in the statistical models. The recruitment process, allocation process and the number of course sections is not given in the study documentation available; randomization and independence at the classroom are then not achievable based on this report.

3.2 Participants

The sample size was 96 undergraduate students, 48 students in each group. Participants ranged from 18 to 24 years of age (M = 20.49, SD = 1.33); 56 students (58.3%) were female and 40 (41.7%) were male, evenly distributed across groups. Table 2 presents the sample characteristics and baseline scores.

3.3 Variables

Instructional condition (0 = control, 1 = GenAI-based strategy) was the independent variable. Posttest critical-thinking and epistemic-autonomy scores were the dependent variables. Both matching pretest scores were covariated in the multivariate model and both follow-up ANCOVAs.

3.4 Intervention Procedures

The experimental plan was executed by two 75 minutes meetings a week during eight weeks. Every cycle entailed six steps: clarify the epistemic question; delineate requirements on an acceptable response; obtain a tentative GenAI response; check assertions by other sources; criticize and refine the argument; put in writing why certain recommendations were pursued, amended, or discarded. The control condition involved the same topics by use of readings, instructor explanation, discussion, and traditional written activities without the necessity to use GenAI. The two conditions were allotted equal time in instruction and similar final product. The structured verification-and-reflection process was thus the intended contrast, rather than more time of exposure.

3.5 Implementation Fidelity

The teaching progression included learning goals, directed thinking, self-checking, peer review, self-reflection, and similar amounts of time on task in the conditions. Quantitative ratings of fidelity and agreement of independent observers are not provided. As a result, the degree of compliance with each instructional element could not be measured independently of the analyses of outcomes.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 09:2026
DOI: 10.5281/zenodo.22771708

3.6 Instrumentation

3.6.1 Critical Thinking Performance Test

The Critical Thinking Performance Test consisted of 20 short scenario based questions that were spread over analysis (7 questions), evaluation of evidence (7 questions) and inference/explanation (6 questions). Each answer was rated as 0 (unsupported or irrelevant), 1 (partially justified), or 2 (well justified) to get a total of 0 to 40. It was informed by the already developed conceptions of critical thinking and dictated by the necessity of constructed or multi-response tasks that demonstrate reasoning, but not recognition alone (Halpern, 1998).

3.6.2 Epistemic Autonomy Scale

The Epistemic Autonomy Scale had 18 statements which were rated on the scales of 1 (strongly disagree) to 5 (strongly agree). There were three six-item domains that depicted the independence of the source, evidence justification, and metacognitive authorship. All the items were averaged to form the total score. The higher scores were the more responsibility was attributed to evidence search, calibrated trust, uncertainty monitoring, and final judgment (Barzilai and Chinn, 2018; Chinn et al., 2011; Muis, 2007).

3.7 Content validity and Scoring Quality

The blueprints of the instruments were aligned with the constructs presented in the theoretical framework. The response of critical-thinking was graded on the given 02 rubric, and the response of the epistemic-autonomy was graded on the given 15 scale. The available study documentation does not report expert-review procedures, content-validity indexes, cognitive-interview results, factor analyses, and interrater agreement. These are the limits to be taken into consideration when interpreting the scores; internal consistency is not enough to determine validity.

3.8 Reliability

Cronbachs alpha was used to estimate internal consistency with respect to the responses of students on the items. At pretest and posttest, coefficients were .931 and.916, respectively, and.921 and.932 for critical thinking and epistemic autonomy respectively. These coefficients characterize the consistency of scores across the study sample, and cannot, alone, determine construct validity.

3.9 Ethical Considerations

The participants of the study were undergraduates, and the findings are given in summary without references to students. The available study documentation does not specify the institutional ethics-review details, consent procedures and data-protection arrangements. Without that information, there is no allegation of ethics approval, exemption, or informed consent.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 09:2026
DOI: 10.5281/zenodo.22771708

3.10 Statistical Analysis

The student-response data and two main posttest results were analyzed. Missingness, item ranges, within-group/pooled alpha and composite scoring were verified. There were no claims of equivalence of the baseline levels and standardized mean differences. A multivariate linear model was tested that controlled a group coefficient vector that controlled both pretests; each outcome was then expressed as posttest = intercept + CT pretest + EA pretest + group + error. The analysis was performed by (1) removing a single term of the complete model and (2) to compute partial sums of squares. Outcome tests were conducted on the two main contrasts, Holm adjusted, nominal alpha.05. Adjusted means, differences in raw units, 95 percent model-based, and partial eta squared are presented. Individual paired significance tests, subgroup tests which were not necessary, and post hoc power calculations were not included. Shapiro–Wilk and median-centered Levene diagnostics were used to test residual normality and between-group residual spread. Interactions between both group bypretest were simultaneously tested in a larger model. Cook distance and leverage without the removal of cases were checked. HC3 sandwich standard errors are a sensitivity analysis with t(92) intervals. These diagnostics neither determine causal identification, nor independent measurement validity, nor extrapolation to data of clustered fields.

4. RESULTS

The findings are the summary of the analysis of the pretest and posttest answers of 96 undergraduate students after the implementation of the teaching program. Characteristics of the sample, estimations of reliability, descriptive contrasts, model diagnostics, and adjusted group contrasts are shown in table 2-10. The confidence intervals and p values are both conditional to the individual level statistical models and their assumptions.

4.1 Data Quality, Sample Characteristics, and Baseline Description

The 96 records were analyzed (48 in each condition); responses on items and composite scores were within the range of their allowances. The analyzed dataset did not have any missing records or values, and no cases were not included in the reported analyses. Table 2 explains the sample of study. Baseline differences are reported in a descriptive manner and nonsignificance is not interpreted as evidence of an equivalent or effective randomization.

Table 2: Sample Characteristics and Baseline Differences

Characteristic	Control (n=48)	GenAI (n=48)	Difference [95% CI]	SMD
Age, years	20.48 (1.22)	20.50 (1.44)	0.02 [-0.52, 0.56]	0.016
CT pretest	19.19 (11.13)	20.75 (10.90)	1.56 [-2.90, 6.03]	0.142
EA pretest	2.92 (0.90)	2.96 (1.00)	0.03 [-0.35, 0.42]	0.035
Female, n (%)	28 (58.3%)	28 (58.3%)	—	—
Male, n (%)	20 (41.7%)	20 (41.7%)	—	—

Note. Continuous values are M (SD). Difference = GenAI − control; confidence intervals use Welch standard errors. SMD = difference divided by pooled within-group SD. These are descriptive checks, not equivalence tests

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 09:2026
DOI: 10.5281/zenodo.22771708

4.2 Internal Consistency and Score Distributions

Table 3: Internal Consistency and Boundary Frequencies

Measure	Items	Alpha: all	Alpha: control	Alpha: GenAI	Floor / ceiling n
CT PRE	20	0.931	0.936	0.928	0 / 0
CT POST	20	0.916	0.906	0.897	1 / 6
EA PRE	18	0.921	0.913	0.929	0 / 0
EA POST	18	0.932	0.917	0.921	0 / 0

Note. Alpha is calculated on the raw item scores, at each time. The composite endpoints which are pooled N=96 are known as floor/ceiling counts.

Separations between groups in a posttest alpha may invalidate the pooled posttest alpha; within-group coefficients are provided as well. No rater reliability or construct validity is determined. High alpha means the internal consistency but does not prove that a test is assessing critical thinking, that a self-report is assessing autonomous behavior, or that the three mentioned subdomains are empirically separate. Hypothesis tests at the domain level and factor-model claims are thus not reported. Interpretation of these outcomes would be reinforced with independent validation and evidence of scorer agreement.

4.3 Unadjusted Pretest–Posttest Changes

Table 4: Observed Score Levels and Paired Changes

Outcome / group	Pretest M (SD)	Posttest M (SD)	Change M (SD)	Change 95% CI
CT / Control	19.19 (11.13)	20.79 (9.70)	1.60 (8.57)	[-0.88, 4.09]
CT / GenAI	20.75 (10.90)	28.50 (8.64)	7.75 (7.23)	[5.65, 9.85]
EA / Control	2.92 (0.90)	2.87 (0.93)	-0.05 (0.65)	[-0.24, 0.13]
EA / GenAI	2.96 (1.00)	3.68 (0.90)	0.72 (0.69)	[0.52, 0.92]

Note. CT range 0–40; EA range 1–5. Change = posttest − pretest. Intervals rely on paired-score variability, rather than independent pre/post samples. Within-group change is not the most important treatment comparison and it is descriptive.

Table 4 separates improvement within a condition and adjusted between condition contrast. The control EA mean shifts a bit downwards, which explains the reason the outcome cannot be defined as consistent across all groups. ANCOVA is used instead of individual paired significance tests to answer the above between-condition questions.

4.4 Model Diagnostics

Table 5: Diagnostic Checks for the Adjusted Models

Outcome	Check	Statistic	df	p
CT	Residual Shapiro–Wilk	0.995	—	0.971
CT	Residual variance (median Levene)	2.306	1, 94	0.132
CT	Both group × pretest terms jointly	0.540	2, 90	0.585
CT	Maximum Cook distance	0.071	—	—
EA	Residual Shapiro–Wilk	0.986	—	0.426
EA	Residual variance (median Levene)	0.018	1, 94	0.895
EA	Both group × pretest terms jointly	0.279	2, 90	0.757

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 09:2026
DOI: 10.5281/zenodo.22771708

EA	Maximum Cook distance	0.129	—	—

Note. The main-effects model is compared to a model which includes both group-by-pretest interactions in each slope test. Levene is used on fitted residues; it is not a guarantee but a diagnostic. Cases were retained irrespective of diagnostic results.

The two baseline covariates gave a r = 0.470 maximum leverage = 0.095. Nonsignificant diagnostic tests fail to establish distributional assumptions, linearity, and common slopes. Limited and discrete results are still a disadvantage. HC3 sensitivity estimates below are sensitive to heteroskedasticity, but not to unmeasured confounding, clustering or invalid measurement.

4.5 Joint Outcome Test: Research Question 3

Table 6: Multivariate Group Contrast Controlling Both Pretests

Test	Value	F	df	p
Wilks lambda	0.674	21.987	2, 91	< .001
Pillai trace	0.326	21.987	2, 91	< .001

Note. Group is a one-degree-of-freedom predictor and therefore, these two statistics give the same exact F transformation. They are not independent replication of a test, but alternative summaries.

The null hypothesis of equal zero adjusted group coefficients on the two outcomes was rejected: Wilks lambda = 0.674, F(2, 91) = 21.99, p < .001. The outcome after adjustment at baseline was correlated with instructional condition. This finding does not imply that 32.6 percent of student learning was mediated by GenAI, or that one of the outcomes mediated the other.

4.6 Critical Thinking: Research Question 1

Table 7: Full Adjusted Critical-Thinking Model

Source	Partial SS	df	MS	F	p	Partial η^2
CT pretest	2671.685	1	2671.685	62.772	< .001	0.406
EA pretest	66.294	1	66.294	1.558	0.215	0.017
Group	1109.586	1	1109.586	26.070	< .001	0.221
Residual	3915.685	92	42.562	—	—	—

Note. Partially summed squares are used to compare the full model to a model which has that term omitted; they need not add up to the full corrected sum. Intercept omitted. The two covariating variables occur at the same time.

The adjusted GenAI-minus-control contrast was 6.82 points, 95% CI [4.17, 9.47], F(1, 92) = 26.070, p < .001, partial eta squared = 0.221. The full model R^2 was 0.582. The null was rejected in the study sample, the comparison is still significant with Holm adjustment in the two outcome-specific tests (adjusted p < .001). Partial eta squared defines the group term as compared to that group term and the residual variation, not the fraction of the overall learning attributed to the intervention.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 09:2026
DOI: 10.5281/zenodo.22771708

4.7 Epistemic Autonomy: Research Question 2

Table 8: Full Adjusted Epistemic-Autonomy Model

Source	Partial SS	df	MS	F	p	Partial η²
CT pretest	0.047	1	0.047	0.124	0.725	0.001
EA pretest	32.366	1	32.366	84.765	< .001	0.480
Group	14.726	1	14.726	38.567	< .001	0.295
Residual	35.129	92	0.382	—	—	—

Note. Partial sums of squares compare the complete model to one which ignores that term; they do not have to add up to the corrected one. Intercept omitted. The two covariates at the baseline are present together. The adjusted GenAI-minus-control contrast was 0.79 points, 95% CI [0.53, 1.04], F(1, 92) = 38.567, p < .001, partial eta squared = 0.295. The full model R^2 was 0.627. Corresponding null was rejected in the study sample, comparison is significant with Holm adjustment in the two outcome-specific tests (adjusted p < .001). Partial eta squared is used to explain the group term, as compared to the group term in addition to variation, rather than the amount of total learning that has been caused by the intervention.

4.8 Adjusted Means, Precision, and Robustness

Table 9: Estimated Marginal Means At Pooled Baseline Means

Outcome	Condition	Adjusted M	SE	95% CI
CT	Control	21.24	0.943	[19.36, 23.11]
CT	GenAI	28.05	0.943	[26.18, 29.93]
EA	Control	2.88	0.089	[2.71, 3.06]
EA	GenAI	3.67	0.089	[3.49, 3.85]

Note. Covariates are held at the overall sample means. These are fitted group means, not predictions for a new student. Their individual confidence intervals are not the confidence interval for their difference.

Table 10: Sensitivity of Adjusted Group Contrasts to HC3 Standard Errors

Outcome	SE estimator	Difference	SE	95% CI	p
CT	Conventional	6.82	1.335	[4.17, 9.47]	< .001
CT	HC3 robust	6.82	1.386	[4.07, 9.57]	< .001
EA	Conventional	0.79	0.126	[0.53, 1.04]	< .001
EA	HC3 robust	0.79	0.133	[0.52, 1.05]	< .001

Note. HC3 is applied to change the uncertainty estimates but not the fitted coefficient. The t (92) is used to calibrate intervals/test in both versions; the inference in HC3 is approximate. Sensitivity p values are not supplementary confirmatory tests, but rather descriptive and unadjusted.

The direction of both contrasts was not any different and the HC3 intervals did not include zero. This helps to stabilize the numbers to the preferred variance estimator. It fails to

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 09:2026
DOI: 10.5281/zenodo.22771708

mention classroom dependence, omitted variables or measurement error. No post hoc power calculation was done and a priori sample-size justification is not stated.

5. DISCUSSION

The discussion has been structured according to the three research questions. The response analyses of 96 students revealed that the adjusted posttest scores in GenAI condition were higher in critical thinking and epistemic autonomy, and that there was a significant multivariate group difference. Results are understood within the context of the applied verification-based instructional approach and study design and measurement evidence limitations**.**

5.1 Discussion of the First Research Question: Critical Thinking

The first research question was answered by exploring how the GenAI-based instructional strategy influences student's critical thinking upon controlling the baseline of critical thinking and epistemic autonomy. The analysis revealed a statistically significant adjusted group difference, $F(1, 92) = 26.07$, $p < .001$ with a partial eta squared of.221. The experimental and the control group had adjusted posttest means of 28.05 and 21.24 respectively. The difference of 6.82, 95% CI [4.17, 9.47] between the adjusted means of the experimental group and the control group shows that the experimental group had higher scores in critical-thinking following the two-baseline measures that were held constant.

This finding can be explained within the framework of the organization of the applied instruction strategy. The experimental group students had to consider AI-generated answers as tentative statements that had to be confirmed. They contrasted information with external sources, discovered assumptions that lacked support, explored the connection between evidence and conclusions, created counterarguments, and articulated how AI recommendations were accepted or rejected. The activities are directly related to the essential critical-thinking processes of analysis, evaluation, inference, and explanation. This finding aligns with the hypothetical educational usefulness of the reasoning activities constructions around GenAI and not the generation of AI-generated content (Facione, 1990**).**

The result is in agreement with the meta-analysis of Abrami et al. (2015) who concluded that the development of critical thinking is more robust in case of instruction that involves dialogue, realistic problems, and direct practice of reasoning. These aspects were implemented in the strategy by peer critique, assessment of realistic AI-generated assertions, and justification of changes repeatedly. The outcome can also be aligned with the ICAP model since the students had to go beyond the process of receiving or modifying information and be productive through producing explanations and interactive through justifying their decisions upon peer discussion (Chi and Wylie, 2014).

The difference could also be explained using argument mapping. It may be helpful to structure an AI response into claims, evidence, warrants, assumptions, and counterarguments to highlight weaknesses in reasoning. This procedure can assist

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 09:2026
DOI: 10.5281/zenodo.22771708

students to differentiate a persuasive writing style and a logically backing conclusion. The past studies revealed that argument mapping could be helpful to achieve better performance in critical thinking, in case learners are provided with clear instructions and given a number of chances to implement the approach (Dwyer et al., 2012; Dwyer et al., 2014).

The outcome must also be taken into consideration with the results provided by Essien et al. (2024). Their analysis of postgraduate business learners has shown that the most obvious enhancement related to AI text generators was at lower levels of the Bloom taxonomy. The current teaching plan was aimed at getting beyond information memorization and understanding and demanded the evaluation of the source, counterargumentation, analysis of uncertainty and rational revision. The adjusted difference that was observed is in line with the possible value of such a structured approach. The contribution of GenAI is compared to the impact of the critical-thinking pedagogy implemented in the intervention.

In the study population, the adjusted difference was large, and statistical significance alone does not allow establishing the educational significance. Independent instrument-validation evidence, blinded scoring, interrater agreement, and a predetermined criterion of meaningful improvement would add to interpretation. Other tasks in terms of transfers that are performed without a close assistance of GenAI would be useful to identify whether students are capable of applying the reasoning processes on their own.

5.2 Discussion of the Second Research Question: Epistemic Autonomy

The second research question was on the impact of the strategy on epistemic autonomy when both baseline outcomes are controlled. The adjusted group difference was found to be statistically significant, $F(1, 92) = 38.57$, $p < .001$ partially eta squared=.295. The mean of the experimental and control groups in the adjusted posttest was 3.67 and 2.88 respectively. The adjusted difference of 0.79 point, 95% CI [0.53, 1.04] indicates that the experimental group scores higher in epistemic-autonomy in the study sample.

This outcome is directly related to the main objective of the strategy to keep the learner accountable to knowing during the usage of GenAI. Students had to specify what they had to know, set standards of a satisfactory response, identify the sources that were worthwhile, keep track of what they did not know, and explain the ultimate decision. GenAI provided potential answers and alternative views, yet, it was up to the learner to determine whether such contributions were made to proper epistemic standards. Such a trend explains why autonomy is seen as competent and justified dependence instead of the denial of any external aid (Barzilai and Chinn, 2018).

This finding can be directly explained by the AIR framework. Epistemic performance is a way of interaction between goals, ideals and trusted processes. In the adopted plan, students have identified an epistemic purpose, e.g., to ascertain the credibility of a claim; used an ideal, e.g., evidence adequacy or credibility of source; and a process, e.g., to verify the claim by means of independent sources. Continuous interaction with these

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 09:2026
DOI: 10.5281/zenodo.22771708

elements can enhance students to better control their ability to form and defend knowledge (Chinn et al., 2011).

The outcome can also be explained by self-regulated learning. The process of intervention handling involved planning and then consulting GenAI, checking the quality of generated response, regulating the process of verification and revision, and self-reflection about the causes of the final decision. Such moves can inhibit the instant adoption of a plausible answer and leave the decision making to the control of the learner. Reflection logs and AI-use records can additionally facilitate this process because they can make students explicit in their judgments and revisions (Muis, 2007; Zimmerman, 2002).

The outcome introduces a new dimension to the works that have focused on the attitudes of students towards GenAI. Students can find AI helpful, convenient, and supportive yet become dependent on its suggestions. Positive perceptions cannot thus necessarily reflect epistemic autonomy. The epistemic-autonomy outcome is concerned with whether the learners judge assertions, gauge trust, and maintain responsibility to conclusions, which is conceptually different than satisfaction or readiness to use the technology (Chan and Hu, 2023; Long and Magerko, 2020). The greater partial eta squared on epistemic autonomy is not to be construed as an indication that the strategy influences autonomy more powerfully in comparison with critical thinking. The two results have varied scales, item structures and error variances. Moreover, the measure of autonomy is a self-report composite and it does not directly show autonomous behavior. The future research ought to involve self-report scores with evidence of behavior such as identifying fabricated citations, use of independent sources, proper rejection of unsupported claims, and behavior in a new task without the aid of AI.

5.3 Discussion of the Third Research Question: Combined Development of Both Outcomes

The third research question was to see whether the strategy influenced posttest critical thinking and epistemic autonomy jointly post adjustment of both pretests. The multivariate analysis produced Wilks' lambda = .674, $F(2, 91) = 21.99$, $p < .001$, with Pillai's trace = .326. This finding suggests that there is a statistically significant adjusted group difference in the overall outcome vector. The independent outcome tests determined that the two adjusted contrasts were statistically significant.

The overall outcome is theoretically significant since critical thinking and epistemic autonomy are similar yet different issues of responsible learning in GenAI. Critical thinking is about quality of analysis, evaluation of evidence, inference and explanation. Epistemic autonomy deals with the ability of the learner to control the goals, criteria, procedures and accountability of achieving the knowledge claims. In order to engage in effective GenAI-supported learning, students should be able to analyze an argument, and they also need to take responsibility in determining whether to believe or to reformulate the argument (Facione, 1990; Barzilai and Chinn, 2018).

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 09:2026
DOI: 10.5281/zenodo.22771708

The learning sequence was intended to facilitate the relating of the two outcomes in the same learning activities. By confirming an AI-generated statement, students train on evaluating evidence, and they also determine standards that should be applied to accept the statement. They examine the robustness of reasoning when they build up a counterargument and are independent of the framing used in the model. When they record a revision, they justify the quality of the new conclusion, and own the process in which that conclusion was arrived at. The multivariate finding is thus harmonized with the theoretical synthesis of critical-thinking training, epistemic learning and self-regulated learning.

AI-literacy frameworks that incorporate the aspects of evaluation, ethical judgment, and human agency also support this interpretation. AI literacy is not just being familiar with how to use a tool or create effective prompts. It involves knowledge of system constraints, analysis of output, awareness of the need to verify and responsible decisions regarding use. The co-construction of critical thinking and epistemic autonomy thus is a valuable learning outcome of responsible GenAI teaching (Long and Magerko, 2020; Ng et al., 2021).

The multivariate outcome does not lead to a conclusion that the two constructs are the same and that enhancement in one led to enhancement in the other. A closer look at the connections between the changes in the two outcomes and the connection of them to the process measures would help understand how they were developed. Late posttests would assist in finding out if the gains are maintained with the removal of instructional supports.

When combined, the results demonstrate positive adjusted posttest differences that are related to the adopted GenAI-based instructional strategy. The critical-thinking outcome is related to reasoning performance, the epistemic-autonomy outcome is related to reported responsibility in knowledge judgments, and the multivariate outcome is an analysis of the two outcomes together. The results are in favor of the future research on a variety of course sections, active comparison condition, more substantial measurement evidence, fidelity of implementation documented, and independent transfer testing.

6. CONCLUSION

The research involved an analysis of an eight-week GenAI-based teaching plan based on pretest and posttest feedback of 96 undergraduate students. The analyses of the three research questions revealed that there were positive differences of critical thinking and epistemic autonomy (adjusted by baseline) and statistically significant multivariate group difference of the two outcomes.

The experimental condition students had better adjusted posttest scores compared to the control condition students. These results characterize the observed group differences and they should be understood within the confines of the study design and measurement evidence. The article presents a realized teaching cycle and a study of learning outcomes to explore the possibility of integrating GenAI into learning without impairing the role of

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 09:2026
DOI: 10.5281/zenodo.22771708

students in judgment. The plan was to do source verification, argument analysis, counterargument, and reflective revision. Critical thinking is about the quality of the reasoning of the students, and epistemic autonomy is about the control of the students over the standards, the evidence and the decisions in view of which the conclusions are drawn. These complementary results were linked with the help of the strategy that demanded learners to assess AI-generation responses and explain their acceptance, rejection, or reworking. The results highlight the importance of assessing GenAI by the reasoning and responsibility to judgment of students, as well as the quality of products that students produce.

The basis of generalization would be enhanced by replication in settings, greater validation evidence, and documented implementation. There are still delayed evaluations and activities done with the help of AI, which are needed to see whether the observed changes are sustainable, transferable development.

Future Research Recommendations

- Pre-register the primary outcomes, exclusion criteria and analysis plan in replication studies and record the relevant ethics approvals and consent process.
- Prolong the test of the Critical Thinking Performance Test and Epistemic Autonomy Scale to independent samples and trained raters.
- Use various course sections and where possible random assignment on a student or classroom level.
- Add an active comparison group which would be treated to the same pedagogy of critical-thinking without GenAI.
- Include delayed follow-up and no-AI transfer tasks to assess retention and independent performance.
- Collect process information including prompt records, source-verification records, revision histories, and blinded ratings of final arguments.
- Test the presence of moderating effects of AI literacy, academic discipline, language proficiency, or baseline epistemic beliefs in the intervention effect.

References

1) Abrami, P. C., Bernard, R. M., Borokhovski, E., Waddington, D. I., Wade, C. A., & Persson, T. (2015). Strategies for teaching students to think critically: A meta-analysis. Review of Educational Research, 85(2), 275–314. https://doi.org/10.3102/0034654314551063

2) Abrami, P. C., Bernard, R. M., Borokhovski, E., Wade, A., Surkes, M. A., Tamim, R., & Zhang, D. (2008). Instructional interventions affecting critical thinking skills and dispositions: A stage 1 meta-analysis. Review of Educational Research, 78(4), 1102–1134. https://doi.org/10.3102/0034654308326084

3) Andreucci-Annunziata, P., Riedemann, A., Cortés, S., Mellado, A., del Río, M. T., & Vega-Muñoz, A. (2023). Conceptualizations and instructional strategies on critical thinking in higher education: A systematic review of systematic reviews. Frontiers in Education, 8, Article 1141686. https://doi.org/10.3389/feduc.2023.1141686

4) Aslett, K., Sanderson, Z., Godel, W., Persily, N., Nagler, J., & Tucker, J. A. (2024). Online searches to evaluate misinformation can increase its perceived veracity. Nature, 625(7995), 548–556. https://doi.org/10.1038/s41586-023-06883-y

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 09:2026
DOI: 10.5281/zenodo.22771708

5) Barzilai, S., & Chinn, C. A. (2018). On the goals of epistemic education: Promoting apt epistemic performance. Journal of the Learning Sciences, 27(3), 353–389. https://doi.org/10.1080/10508406.2017.1392968

6) Barzilai, S., & Chinn, C. A. (2024). The AIR and Apt-AIR frameworks of epistemic performance and growth: Reflections on educational theory development. Educational Psychology Review, 36, Article 91. https://doi.org/10.1007/s10648-024-09927-5

7) Butler, H. A. (2012). Halpern Critical Thinking Assessment predicts real-world outcomes of critical thinking. Applied Cognitive Psychology, 26(5), 721–729. https://doi.org/10.1002/acp.2851

8) Butler, H. A., Dwyer, C. P., Hogan, M. J., Franco, A., Rivas, S. F., Saiz, C., & Almeida, L. S. (2012). The Halpern Critical Thinking Assessment and real-world outcomes: Cross-national applications. Thinking Skills and Creativity, 7(2), 112–121. https://doi.org/10.1016/j.tsc.2012.04.001

9) Chan, C. K. Y. (2023). A comprehensive AI policy education framework for university teaching and learning. International Journal of Educational Technology in Higher Education, 20, Article 38. https://doi.org/10.1186/s41239-023-00408-3

10) Chan, C. K. Y., & Hu, W. (2023). Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. International Journal of Educational Technology in Higher Education, 20, Article 43. https://doi.org/10.1186/s41239-023-00411-8

11) Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. Educational Psychologist, 49(4), 219–243. https://doi.org/10.1080/00461520.2014.965823

12) Chinn, C. A., Buckland, L. A., & Samarapungavan, A. (2011). Expanding the dimensions of epistemic cognition: Arguments from philosophy and psychology. Educational Psychologist, 46(3), 141–167. https://doi.org/10.1080/00461520.2011.587722

13) Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2024). Chatting and cheating? Ensuring academic integrity in the era of ChatGPT. Innovations in Education and Teaching International, 61(2), 228–239. https://doi.org/10.1080/14703297.2023.2190148

14) Crompton, H., & Burke, D. (2023). Artificial intelligence in higher education: The state of the field. International Journal of Educational Technology in Higher Education, 20, Article 22. https://doi.org/10.1186/s41239-023-00392-8

15) DeVellis, R. F., & Thorpe, C. T. (2021). Scale development: Theory and applications (5th ed.). SAGE Publications. https://collegepublishing.sagepub.com/products/scale-development-5-269114

16) Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., . . . Wright, R. (2023). Opinion paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. International Journal of Information Management, 71, Article 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642

17) Dwyer, C. P., Hogan, M. J., & Stewart, I. (2012). An evaluation of argument mapping as a method of enhancing critical thinking performance in e-learning environments. Metacognition and Learning, 7(3), 219–244. https://doi.org/10.1007/s11409-012-9092-1

18) Dwyer, C. P., Hogan, M. J., & Stewart, I. (2014). An integrated critical thinking framework for the 21st century. Thinking Skills and Creativity, 12, 43–52. https://doi.org/10.1016/j.tsc.2013.12.004

19) Ennis, R. H. (1985). A logical basis for measuring critical thinking skills. Educational Leadership, 43(2), 44–48.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 09:2026
DOI: 10.5281/zenodo.22771708

20) Essien, A., Bukoye, O. T., O'Dea, X., & Kremantzis, M. (2024). The influence of AI text generators on critical thinking skills in UK business schools. Studies in Higher Education, 49(5), 865–882. https://doi.org/10.1080/03075079.2024.2316881

21) Facione, P. A. (1990). Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction. Research findings and recommendations. The California Academic Press. https://eric.ed.gov/?id=ED315423

22) Farrokhnia, M., Banihashem, S. K., Noroozi, O., & Wals, A. (2024). A SWOT analysis of ChatGPT: Implications for educational practice and research. Innovations in Education and Teaching International, 61(3), 460–474. https://doi.org/10.1080/14703297.2023.2195846

23) Grassini, S. (2023). Shaping the future of education: Exploring the potential and consequences of AI and ChatGPT in educational settings. Education Sciences, 13(7), Article 692. https://doi.org/10.3390/educsci13070692

24) Halpern, D. F. (1998). Teaching critical thinking for transfer across domains: Disposition, skills, structure training, and metacognitive monitoring. American Psychologist, 53(4), 449–455. https://doi.org/10.1037/0003-066X.53.4.449

25) Józsa, K., Oo, T. Z., Borbélyová, D., & Zentai, G. (2023). Exploring the accuracy and consistency of a school readiness assessment tool for preschoolers: Reliability, validity and measurement invariance analysis. Journal of Intelligence, 11(10), Article 189. https://doi.org/10.3390/jintelligence11100189

26) 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, Article 102274. https://doi.org/10.1016/j.lindif.2023.102274

27) Lim, W. M., Gunasekara, A., Pallant, J. L., Pallant, J. I., & Pechenkina, E. (2023). Generative AI and the future of education: Ragnarök or reformation? A paradoxical perspective from management educators. International Journal of Management Education, 21(2), Article 100790. https://doi.org/10.1016/j.ijme.2023.100790

28) Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. Education Sciences, 13(4), Article 410. https://doi.org/10.3390/educsci13040410

29) Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (pp. 1–16). Association for Computing Machinery. https://doi.org/10.1145/3313831.3376727

30) Miao, F., & Holmes, W. (2023). Guidance for generative AI in education and research. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000386693

31) Muis, K. R. (2007). The role of epistemic beliefs in self-regulated learning. Educational Psychologist, 42(3), 173–190. https://doi.org/10.1080/00461520701416306

32) Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. Computers and Education: Artificial Intelligence, 2, Article 100041. https://doi.org/10.1016/j.caeai.2021.100041

33) Steyvers, M., Tejeda, H., Kumar, A., Belem, C., Karny, S., Hu, X., Mayer, L. W., & Smyth, P. (2025). What large language models know and what people think they know. Nature Machine Intelligence, 7(2), 221–231. https://doi.org/10.1038/s42256-024-00976-7

34) Tlili, A., Shehata, B., Adarkwah, M. A., Bozkurt, A., Hickey, D. T., Huang, R., & Agyemang, B. (2023). What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education? Smart Learning Environments, 10, Article 15. https://doi.org/10.1186/s40561-023-00237-x

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 09:2026
DOI: 10.5281/zenodo.22771708

35) UNESCO. (2021). Recommendation on the ethics of artificial intelligence. https://unesdoc.unesco.org/ark:/48223/pf0000381137

36) Walters, W. H., & Wilder, E. I. (2023). Fabrication and errors in the bibliographic citations generated by ChatGPT. Scientific Reports, 13, Article 14045. https://doi.org/10.1038/s41598-023-41032-5

37) Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. Theory into Practice, 41(2), 64–70. https://doi.org/10.1207/s15430421tip4102_2

Appendix A. Instrument Blueprints and Item Content

A1. Critical Thinking Performance Test Blueprint

Table A1: Critical-Thinking Test Blueprint

Domain	Items	Required response	Scoring focus
Analysis	1-7	Identify claims, reasons, assumptions, and relationships	Accuracy and completeness of structural analysis
Evidence evaluation	8-14	Judge credibility, relevance, sufficiency, and corroboration	Use of explicit evidential criteria
Inference and explanation	15-20	Draw a warranted conclusion and explain uncertainty	Logical warrant, alternatives, and calibration

Note. *The critical-thinking instrument comprised 20 items across three domains.*

Items presented short scenarios and required written or multi-response justification. Responses were scored using the 0–2 rubric described in Section 3.6.1. Parallel-form development, rater training, blinding, and interrater reliability are not documented in this report.

A2. Epistemic Autonomy Scale: Item Content

I decide what evidence would be sufficient before accepting an answer.

I compare important claims with sources that are independent of the AI system.

I can explain why I trust one source more than another.

I revise my position when stronger evidence contradicts it.

I distinguish my own judgment from wording suggested by an AI system.

I can state what remains uncertain after completing a research task.

I reject an AI-generated answer when its reasoning does not meet my criteria.

I seek expert guidance when a question exceeds my competence.

I keep responsibility for the final conclusion even when AI assistance is used.

The nine statements above describe item content related to evidence evaluation and responsibility for judgment. They do not constitute the complete 18-item instrument. The full wording, scoring key, and validation evidence are not reproduced in this appendix.

Appendix B. Statistical Model Record

The analysis included 96 student records. Each primary outcome model contained an intercept, both pretest scores, and instructional condition. Residual df = 96 − 4 = 92; the joint slope-test residual df = 90; the single-predictor multivariate F uses df (2, 91). The reported analyses included conventional and HC3 standard errors, with Holm adjustment across the two primary outcome contrasts.