Scientific Research Publishing

Open Journal of Social Sciences, 2026, 14(11), 430-449
https://www.scirp.org/journal/jss
ISSN Online: 2327-5960
ISSN Print: 2327-5952

Teachers' Critical Thinking in GenAI-Supported Professional Development

—A Critical Narrative Review

Yuxin Dong

Faculty of Education, Beijing Normal University, Beijing, China
Email: dongyuxin010101@163.com

How to cite this paper: Dong, Y. X. (2026). Teachers' Critical Thinking in GenAI-Supported Professional Development. *Open Journal of Social Sciences, 14,* 430-449.
https://doi.org/10.4236/jss.2026.1411024

Received: October 31, 2026
Accepted: November 20, 2026
Published: November 23, 2026

Copyright © 2026 by author(s) and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/

Open Access

Abstract

Generative artificial intelligence (GenAI) is increasingly integrated into teachers' professional development, yet claims regarding critical thinking (CT) often conflate distinct constructs and forms of evidence. This critical narrative review examines how teachers' CT is conceptualised and enacted within GenAI-supported professional development, what design conditions support or constrain it, and how existing studies establish the contribution of GenAI to reported outcomes. Synthesising current empirical research and relevant reviews, this review argues that improvements in teaching artefacts or isolated CT dimensions do not consistently indicate broader CT development. Furthermore, comparative studies differ substantially in whether they contrast GenAI with human support or evaluate an additional scaffold within an AI-supported activity. The review advocates interpreting these outcomes in close relation to their operational definitions, task demands, and comparison conditions. It concludes by proposing professional development designs that render teachers' critical reasoning—when accepting, revising, or rejecting AI contributions—more visible, alongside research that rigorously distinguishes supported task performance from sustained professional learning.

Keywords

Critical Thinking, Generative Artificial Intelligence, Teacher Professional Development, Teacher Education

1. Introduction

1.1. GenAI and Teachers' Professional Development

Generative artificial intelligence (GenAI) has been investigated as a resource for

teachers' professional development, spanning activities such as instructional design, digital storytelling, and online knowledge construction (Aydın Yıldız, 2026; Lehtinen et al., 2026; Li et al., 2024; Sun & Huang, 2025). Effective professional development requires sustained support, active collaboration, and opportunities for teachers to reflect on their practices (Borko, 2004; Darling-Hammond & McLaughlin, 2011). In GenAI-supported professional activities, teachers engage with generated explanations, suggestions, and feedback while completing instructional and learning tasks (Moorhouse et al., 2024; Zhang & Wang, 2025). Teachers' overall perceptions of AI as a tool to support their practice have also been surveyed (Chounta et al., 2022). For example, Li et al. (2024) compared ChatGPT-supported learning with support from in-service teachers in an integrated science, technology, engineering, and mathematics (iSTEM) course. Sun and Huang (2025) examined collaborative lesson planning supported by GenAI among pre-service teachers. GenAI-supported professional development therefore encompasses different activities and forms of scaffolding (Kohnen et al., 2025; Moorhouse et al., 2024), which need to be examined in different contexts. Applied to this review, these perspectives encourage attention to how teachers cooperate with GenAI in fact and what evidence demonstrates their professional learning, rather than treating mere access to a tool as a sufficient description of professional development (Chan & Tsi, 2024; Kohnke et al., 2026).

1.2. Why Teachers' Own Critical Thinking Matters in Gen-AI Support Situations

Teachers' critical thinking (CT) has been examined in teacher education long before the emergence of widely available GenAI (Huang & Sang, 2023; Pithers & Soden, 2000; Toncel Herrera & Jiménez Sierra, 2026). Yuan et al. (2022) review EFL teachers' perceptions and engagement with CT, providing a related teacher education context. More recent empirical studies illustrate why this focus requires careful interpretation in GenAI contexts. Wan and Gu (2026) found advantages for concept mapping combined with GenAI on several digital storytelling outcomes, but not a greater improvement in CT tendencies. Similarly, Li et al. (2024) reported a significant between-group difference in systematicity, while the overall post-test difference in CT was not significant. These findings raise the question of what exactly develops when teachers perform well on an AI-supported task (Kasneci et al., 2023; Kosmyna et al., 2025). They also show why the quality of a teaching product and the development of the teacher's CT should be analysed separately (Inie et al., 2026; Liao et al., 2025).

This critical narrative review examines teachers' CT within GenAI-supported initial preparation and continuing professional development. It includes studies of explicit CT outcomes and related professional practices when their data concern evaluation, justification, or revision. The purpose is to compare conceptualisations, designs, and evidence rather than estimate an overall GenAI effect. Both pre-service and in-service contexts are considered, while differences in professional role are re-

tained in the interpretation. Mixed-role samples are identified where relevant, including the participants working outside education in Zhang and Wang (2025).

The review addresses three research questions:

RQ1: How is teachers' critical thinking conceptualised and enacted in GenAI-supported professional development?

RQ2: What professional development designs and conditions support or constrain teachers' critical thinking when engaging with GenAI?

RQ3: What outcomes for teachers' critical thinking are reported in GenAI-supported professional development, and how does the evidence establish the contribution of GenAI to these outcomes?

2. Teachers' Critical Thinking in the Context of GenAI

2.1. Critical Thinking in Teachers' Professional Work

Facione's (1990) Delphi report distinguishes critical thinking (CT) skills—ranging from analysis and interpretation—from the dispositions towards using them. Halpern (1998) similarly connects skills with dispositions and metacognitive monitoring, while addressing transfer across domains. Ennis (1989) clarifies debates about the subject specificity of CT and the role of relevant subject knowledge, while identifying unresolved conceptual and empirical questions. These distinctions matter when comparing a teacher's reported willingness to consider different views (a disposition) with the reasoning visible in a particular task (a skill) (Pithers & Soden, 2000). They provide conceptual guidance for this review rather than evidence that GenAI inherently develops CT (Liao et al., 2025).

In this review, teachers' CT refers to their purposeful examination of information, reasons, and alternatives when making professional judgements. This working definition draws on Facione (1990) and is applied to instructional contexts represented in the empirical studies (Huang & Sang, 2023). Professional reasoning also requires relevant domain knowledge. Shulman's (1986) account of teacher knowledge, particularly pedagogical content knowledge, helps explain why judging a proposed activity involves more than deciding whether its language is clear. Relevant activity must also be considered in relation to subject content and how that content can be taught (Borko, 2004). Teachers' CT is enacted when they critically evaluate and adapt pedagogical strategies in their daily professional work (Liao et al., 2024).

2.2. Critical Thinking and Related Constructs in GenAI Research

The examined studies operationalise thinking in different ways, then explore broader challenges in CT education (Barta et al., 2022; Batdı et al., 2024; Felix et al., 2025). Li et al. (2024) describe their questionnaire as assessing CT skills, but the instrument is based largely on a disposition inventory and self-reported outcomes. Debates over adapting versus adopting CT education in the East Asian cultural sphere also further indicate that conceptualisations of CT are context-dependent (Wang & Wu, 2023). Wan and Gu (2026) explicitly assess CT tendencies through a six-item questionnaire. Lehtinen et al. (2026), by comparison, interpret CT

through prompting and collaborative activity patterns. These measures concern related but distinguishable aspects of thinking (Arseven & Bal, 2025; Liao et al., 2025; Raitskaya & Tikhonova, 2025).

In GenAI-supported research, Zhang and Wang (2025) analyse information sharing, analysis, negotiation, modification, application, and evaluation within course discussions. This framework makes some reasoning practices visible, but it is not an independent measure of general CT ability (Inie et al., 2026). Similarly, reflection is relevant when it contains identifiable evaluation or justification (Wei et al., 2025); a reflective report cannot be assumed to demonstrate those processes solely because it is labelled reflective (Jegstad et al., 2025). This review therefore retains the original terms used by each study and examines how the reported evidence relates to teachers' CT.

3. Review Approach

3.1. Literature Identification and Selection

This critical narrative review drew on database searches and reference tracing to identify literature on teachers' critical thinking in GenAI-supported professional learning. Scopus and Web of Science were searched on 8 September 2026 for literature published through 31 August 2026. The main Boolean searches combined terms for generative AI, teachers, and critical thinking. A supplementary search in each database added terms for reviews and other evidence syntheses. The results in both databases were filtered to English, and only English-language publications were eligible at screening. Reference tracing was used to identify foundational literature on critical thinking and teacher development. The **Appendix** presents the database-specific search strings and the limits applied to the searches.

Priority was given to empirical studies examining teachers' own thinking in professional preparation, development, or inquiry. Studies addressing only school students' CT, generic technology acceptance, or administrative efficiency were used as contextual literature rather than direct evidence of teachers' CT. Related studies were considered when their data concerned evaluation, justification, or revision, while their original constructs were retained. Both pre-service and in-service contexts were included.

3.2. Synthesis and Evidence Appraisal

For each empirical study appraised in detail, the synthesis recorded the participants and setting, the professional task, the role of GenAI, the comparison condition where applicable, the CT construct and its indicators, and the findings relevant to the review questions. Findings were organised around three questions concerning how CT was conceptualised and observed, how professional learning was designed, and what CT outcomes were reported. Within each area, findings were compared by evidence type, including self-report, observed interaction, and assessed performance. Construct clarity and data source guided the interpretation of what each finding could support. Comparison conditions and task comparabil-

ity informed attribution, while study duration and follow-up informed claims about persistence and transfer.

An episode was treated as evidence of enacted CT when the record showed that participants applied a relevant criterion, weighed alternatives on an identifiable basis, or gave a reason for accepting, revising, or rejecting a suggestion. Prompt reformulation, feedback uptake, and artefact revision were recorded as task activities when the reasoning behind them was unclear. Retrospective descriptions of such reasoning were treated as self-reported CT practices. Observed instances of reasoned evaluation supported claims about CT practice within the task. Claims about development required evidence of change on a measure aligned with the CT outcome in question.

Table 1 identifies the empirical studies appraised in detail from full texts. The studies were selected for their relevance to teachers' CT in GenAI-supported professional activities and for the complementary evidence provided by their designs and measures. They include studies of digital storytelling and peer feedback that directly assessed pre-service teachers' self-reported CT tendencies (Chen, 2026; Wan & Gu, 2026; Zhou et al., 2026), alongside studies that documented evaluative practices or related professional activities. Reviews and conceptual sources informed interpretation. Where a review discussed a primary study also examined here, both sources were treated as referring to the same empirical evidence. Positive, non-significant, and mixed findings were retained. The synthesis did not pool effect estimates.

Table 1. Selected studies of teachers' critical thinking in GenAI-supported professional activities.

Study	Comparison Condition	GenAI's Role	CT Measurement/Evidence Type	Key Findings and Methodological Boundaries
Li et al. (2024)	GenAI vs. Human Expert	AI as a partner providing feedback in iSTEM tasks	Self-report CT questionnaire (CCTDI-based); reflections	ChatGPT group showed higher systematicity; overall CT difference was not significant. Small sample limits wider claims.
Wan & Gu (2026)	GenAI + Scaffold vs. GenAI alone	AI as a content generation tool for digital storytelling	CT tendency questionnaire; story rubric	Product quality improved with scaffolding, but CT tendency growth did not exceed the GenAI-only group. (Highlights scaffold vs. tool effects).
Lehtinen et al. (2026)	No control group (Process mining)	AI as a collaborative partner in lesson planning	Video analysis, process mining (prompting & activity sequences)	The records document prompt refinement and attempts to corroborate information. Episodes support an inference of reasoned evaluation when the basis for the group's judgement is visible. Coding frequencies did not differ significantly between groups with higher- and lower-rated lesson plans.

Continued

Zhang & Wang (2025)	GenAI package vs. Conventional learning	AI integrated into a knowledge-building instructional model	Epistemic network analysis (ENA); course performance	Higher analysis coding, but lower negotiation/evaluation proportions. Confounded by the whole instructional package (tool + guidance).
Sadidi & Prestel (2026)	Pre-post test (No separate control)	AI as a generator of physics teaching content	Criterion-based reflection on AI outputs	Criterion-based critiques were observed. Perceived quality and helpfulness decreased descriptively; the reported changes were not statistically significant. CT-skill gains were not established.
Korucu-Kış & Bakla (2026)	Qualitative case study	AI as a resource in EFL practicum	Interviews; interaction records; reflective essays	Participants perceived pedagogical benefits, but recorded interactions were mostly single-turn prompts with limited follow-up or critical interrogation. Perceived usefulness did not establish demonstrated CT development.
Chen (2026)	CLEAR-guided GenAI storytelling vs. conventional GenAI storytelling	Content generation within a structured prompting framework	Six-item self-report CT tendency questionnaire. Interviews provided supplementary accounts of evaluation practices.	The CLEAR group had higher baseline-adjusted CT tendency scores (partial $\eta^2 = .086$). The comparison concerns the added CLEAR framework within GenAI-supported storytelling.
Zhou et al. (2026)	GenAI-supported vs. conventional peer feedback in an online peer assessment task	Feedback on pre-service teachers' peer comments	Self-reported CT tendencies. Feedback quality, feedback uptake, and artefact revision were assessed as separate outcomes.	The GenAI-supported condition showed more favourable outcomes for feedback quality, uptake, artefact revision, and self-reported CT tendencies. Application of self-feedback positively moderated the relationship with CT tendencies. The intervention was brief, and the analysis did not account for potential clustering.

Note. CT = critical thinking; GenAI = generative artificial intelligence; iSTEM = integrated science, technology, engineering, and mathematics. The studies are selected for relevance and complementary evidence, not ranked by citations.

4. Conceptualisations and Enactments of Teachers' Critical Thinking

4.1. Critical Thinking and Professional Judgement

The distinction between CT skills and dispositions is particularly visible in quan-

titative studies (Facione, 1990; Pithers & Soden, 2000). Li et al. (2024) used a modified questionnaire with items asking participants about their willingness to revise opinions, listen to different views, and address challenges. Given the instrument's CCTDI basis and self-report format (Li et al., 2024), this review interprets these responses as CT-related orientations rather than independent demonstrations of reasoning skill.

Wan and Gu (2026) provide a more task-specific operationalisation. Their CT tendency questionnaire asks participants about checking the accuracy of story content, assessing new information, and understanding different perspectives during digital storytelling. The task connection makes the measure relevant to the intervention. However, the responses still concern participants' accounts of their thinking. The questionnaire and digital story rubric therefore address different outcomes, even though both are collected within the same activity (Barta et al., 2022).

Professional judgement is most visible when teachers connect evaluation to instructional purposes (Shulman, 1986). Liao et al. (2024) identify expert teachers' strategies for justifying pedagogical decisions. This review uses those strategies as a conceptual reference when examining pre-service teachers' professional judgement. In Lehtinen et al. (2026), participants refined conceptual and pedagogical prompts while designing lessons on sustainable development. Their responses to GenAI were examined in relation to the lesson-planning task. Across these studies, this review distinguishes a reported orientation towards CT from the use of judgement in a particular professional activity (Huang & Sang, 2023). These distinctions follow the different forms of evidence rather than imposing one measure on all studies.

4.2. Critical Thinking in Evaluating and Revising GenAI Outputs

Lehtinen et al. (2026) examined how groups used GenAI and other online resources during a 60-minute collaborative lesson-planning task. The recorded sequences show conceptual and pedagogical requests, prompt refinement, and attempts to corroborate information. Where a sequence shows how a group checked a claim against a source or justified a choice in relation to a pedagogical goal, it provides evidence of reasoned evaluation at the group level. Cheah et al. (2026) likewise relate prompting practices to participants' task-related and professional justifications. The content of these exchanges helps distinguish episodes that reveal the basis for a decision from prompt changes whose rationale is unclear.

Related studies extend the task contexts. Drushlyak et al. (2025) asked 32 pre-service mathematics teachers to inspect ChatGPT-generated solutions, identify mistakes, and examine corrections. This task connects evaluation to disciplinary reasoning, a relationship considered here in light of debates about subject knowledge and CT (Ennis, 1989). Sadidi and Prestel (2026) engaged 39 prospective physics teachers in evaluating generated teaching content using explicit criteria. Their qualitative findings document criterion-based critiques, while changes in per-

ceived quality and helpfulness were not statistically significant. Therefore, in this review, evidence of evaluation practices has been distinguished from evidence of CT-skill development.

4.3. Relationships between Conceptualisation and Observed Practice

The relationship between a broad CT claim and its indicators needs particular attention in knowledge-building research. Zhang and Wang (2025) added information evaluation to a framework covering several discussion activities. Their examples include brief evaluative comments, and their network analysis represents co-occurrences among coded activities. The coding identifies evaluative contributions within the discussion, and the network analysis shows how these contributions co-occurred with other activities. Individual contributions were treated as evidence of reasoned evaluation only where the reported examples showed a criterion, a reasoned comparison, or a justification. This interpretation reflects the study's coding framework and analytical unit (Zhang & Wang, 2025) and the need to align CT claims with the available indicators (Liao et al., 2025).

Lehtinen et al. (2026) offer a different relationship between conceptualisation and evidence. Their recordings preserve activity sequences, but the analysis focuses on groups, while individual differences in prior knowledge and CT are acknowledged as limitations. Wan and Gu (2026) collect repeated individual questionnaire responses, but these do not directly show the reasoning behind particular design decisions. Lehtinen et al. (2026) and Wan and Gu (2026) provide complementary evidence about observed group activity and individual reported tendencies. This review treats these as different analytical perspectives rather than interchangeable demonstrations of one CT outcome.

5. Professional Development Designs and Conditions

5.1. Professional Tasks and Scaffolds for Critical Engagement

The central empirical studies embed GenAI in tasks with specific professional purposes (Kohnen et al., 2025; Sun & Huang, 2025). Chen (2026) compared CLEAR-guided GenAI storytelling with conventional GenAI storytelling, while Wan and Gu (2026) examined the addition of concept mapping to a GenAI-supported storytelling task. Li et al. (2024) required participants to learn about iSTEM teaching and prepare a teaching project. Wan and Gu (2026) used digital storytelling tasks supported by ERNIE Bot and an animation editor. Lehtinen et al. (2026) organised collaborative lesson planning, while Zhang and Wang (2025) linked course learning to knowledge-building discussions. These tasks create different occasions for evaluation and reasoning, so their outcomes need to be interpreted in relation to their specific pedagogical demands.

Organising critical engagement through task structure predates the GenAI era. Yu and Zin (2023) review different CT-oriented adaptations of problem-based learning and the outcomes reported for those adaptations. Zhou et al. (2026) com-

pared GenAI-supported and conventional peer feedback in a quasi-experimental online peer assessment task involving 199 pre-service teachers. In the GenAI-supported condition, the tool commented on the feedback participants drafted for their peers. The original technology-enhanced learning design artefacts were outside the scope of the tool's review. The mediating role of GenAI self-regulation in students' CT and problem-solving suggests that support conditions should also include a self-regulatory dimension (Zhou et al., 2024). Wan and Gu (2026) provide a direct comparison of an additional scaffold. Two intact classes, comprising 62 pre-service teachers, were assigned to concept mapping combined with GenAI or GenAI-supported storytelling without concept mapping. The integrated approach supported several dimensions of story quality, but its CT tendency growth did not exceed that of the comparison group. In this study, the added concept-mapping scaffold improved several product dimensions without producing greater growth in reported CT tendencies.

Zhang and Wang (2025) combined GenAI use with a staged instructional model and initial instructor demonstrations of prompting. Their experimental condition therefore involved organised support for learning with GenAI. In this review, this is treated as a comparison between implemented instructional models. Because tool use, prompting guidance, and staged support were introduced together, the comparison does not isolate the contribution of each component (Zhang & Wang, 2025).

5.2. Human Guidance and Teacher Knowledge in Collaboration

Li et al. (2024) make the source of support central to their design. Eleven student teachers worked with ChatGPT, while twelve worked with in-service teachers. The ChatGPT group performed better on process learning tasks and reported lower cognitive load. The final teaching proposal difference narrowly missed conventional statistical significance. In Zhou et al. (2026), the association between GenAI-supported peer feedback and self-reported CT tendencies was stronger among participants who applied more self-feedback. These results suggest different patterns across learning activities, rather than a consistent advantage for either partner across all professional outcomes (Kosmyna et al., 2025).

Collaboration also needs to be considered as something participants actively enact (Pahi et al., 2024). Lehtinen et al. (2026) formed groups and provided an open planning task, but participation in a group did not establish that all members exercised the same judgement. Their group-level analysis cannot identify how responsibility for checking content was distributed between members. Similarly, their discussion of prior knowledge as a potentially important condition was not confirmed by the reported analysis. Knowledge should therefore remain a plausible contextual explanation, rather than a measured moderator in that study.

Shulman's (1986) account of pedagogical content knowledge provides a conceptual reason to examine how teachers judge subject content and its representation for learners. Borko (2004) situates professional learning within the activities

and contexts in which teachers participate. These perspectives inform the questions used in this review; they do not establish that experience, school conditions, or prior knowledge caused the outcomes in the GenAI studies.

5.3. Constraints on Professional Learning with GenAI

The digital storytelling study shows that support for producing a stronger artefact can coexist with a less favourable pattern of CT tendency growth (Wan & Gu, 2026). The authors suggest that concept mapping may have introduced additional cognitive demands. However, their study does not directly establish this explanation. This review therefore distinguishes the observed growth pattern from the authors' proposed cognitive-load explanation, which was not directly tested in Wan and Gu's study (2026).

The knowledge-building study also contains an important tension. Zhang and Wang (2025) report a much higher proportion of information-analysis coding in the GenAI condition. However, they also show lower proportions of information negotiation and information evaluation in that condition. These descriptive findings do not support an interpretation that every form of higher-order discussion increased together, but rather point to a redistribution of observed activities, whose educational value must be judged in relation to the task and the coding framework (Inie et al., 2026). Furthermore, an EFL practicum study provides a related contextual example. Korucu-Kış and Bakla (2026) followed eleven pre-service teachers over twelve weeks using interviews, interaction records, and reflective essays. Although participants described iterative inquiry, most recorded questions were single-turn requests. Their findings illustrate why perceived usefulness and reported engagement need to be examined alongside actual interaction evidence (Cheah et al., 2026; Chiu, 2024).

6. Reported Outcomes and Evidence for GenAI's Contribution

6.1. Reported Outcomes for Teachers' Critical Thinking

Li et al. (2024) found that overall self-reported CT scores increased within the ChatGPT group. At post-test, the overall score did not differ significantly between the ChatGPT group and the group supported by in-service teachers. The systematicity subscale favoured the ChatGPT group. This finding concerns one dimension of self-reported CT, while the overall between-group comparison remains inconclusive.

Two digital storytelling studies reported different patterns for self-reported CT tendencies. In Wan and Gu (2026), adding concept mapping to GenAI-supported storytelling improved several dimensions of story quality. The time-by-group interaction for CT tendencies was significant, with scores increasing more in the GenAI-only group, although the concept-mapping group had higher descriptive means. In Chen (2026), the CLEAR-guided GenAI group scored higher on the CT tendency questionnaire than the conventional GenAI group after adjustment for

baseline scores (adjusted means = 4.40 and 4.22, respectively, $F = 8.42$, $p < .01$, partial $\eta^2 = .086$). The two studies examined different scaffolds in separate comparisons, which limits direct conclusions about their relative effectiveness.

In a quasi-experimental study of 199 pre-service teachers, Zhou et al. (2026) found that GenAI-supported peer feedback yielded more favourable self-reported CT tendencies than conventional peer feedback. The study also reported higher feedback quality, greater feedback uptake, and greater improvement in technology-enhanced learning design artefacts. The association between GenAI-supported peer feedback and CT tendencies was stronger at higher levels of self-feedback application. This moderation result identifies a condition under which the association was stronger. The study did not examine how participants in the two conditions reasoned through individual feedback decisions, so the uptake and revision outcomes offer limited evidence about the judgement behind those decisions. Zhang and Wang (2025) reported higher course performance and different patterns of knowledge-building activity under their GenAI-assisted instructional model. The GenAI condition showed a higher proportion of contributions coded as information analysis and lower proportions coded as information negotiation and evaluation. These results describe how coded activities were distributed during discussion. Their relevance to teachers' CT is further qualified by the mixed occupational backgrounds of the postgraduate participants.

Across these studies, the reported outcomes address different aspects of professional learning. Questionnaire scores indicate how participants reported their CT tendencies, artefact assessments capture the quality of their work, and discussion codes describe the activities observed during a task. Evidence of change in teachers' reasoning depends on whether the measures capture the criteria, comparisons, or justifications participants used to make decisions.

6.2. Evidence for the Contribution of GenAI

The comparison conditions determine what can be attributed to each intervention. Li et al. (2024) compared ChatGPT support with support from in-service teachers. Both groups in Wan and Gu (2026) and Chen (2026) used GenAI, so their comparisons concern the addition of concept mapping and the CLEAR framework, respectively, within GenAI-supported storytelling. Zhou et al. (2026) compared GenAI-supported with conventional peer feedback, focusing on the addition of AI-generated feedback on participants' peer comments. The brief intervention limits conclusions about persistence, and the analysis did not account for potential clustering. Each comparison therefore addresses a specific form of support within a particular professional learning task.

Zhang and Wang (2025) include a conventional learning comparison, but the experimental condition combines tool use, prompting demonstrations, staged support, and feedback arrangements (Darling-Hammond & McLaughlin, 2011). Their results are most defensibly attributed to that implemented model. Lehtinen et al. (2026) provide detailed process evidence but no experimental comparison of

AI and non-AI conditions. High- and low-performing groups were selected on the basis of lesson-plan quality, with middle-scoring groups excluded from that comparison. These boundaries follow from the study's purpose and analytic unit, rather than diminish the value of process evidence.

6.3. Evidence of Sustained and Transferable Development

The central empirical studies provide different observation windows, from a one-hour collaborative task to course-based interventions (Lehtinen et al., 2026; Li et al., 2024; Wan & Gu, 2026). A longer intervention offers more opportunity to observe participation, but duration alone does not demonstrate that thinking persists beyond the intervention (Desimone, 2009). None of these three designs, as reported, establishes delayed CT performance in a new professional task without the original support (Halpern, 1998). Zhang and Wang (2025) include an application-and-transfer stage within their instructional model. It concerns applying course knowledge in learning activities and should not be interpreted as an independent demonstration of durable CT transfer. Drushlyak et al. (2025) report gains when participants repeat the same individual work one month later, without a separate control group. This review distinguishes repeated-task improvement from evidence of CT performance in a new professional situation.

Drawing on Halpern's (1998) attention to transfer, this review proposes that future GenAI studies examine whether teachers can explain and apply their judgements in unfamiliar situations. This is a recommendation, not a demonstrated outcome of the examined evidence base. For the studies appraised in detail here, the available findings document supported performance and participation more directly than sustained, independent CT development (Drushlyak et al., 2025; Lehtinen et al., 2026; Li et al., 2024; Wan & Gu, 2026; Zhang & Wang, 2025).

7. Discussion

7.1. Teachers' Critical Thinking in Interactions with GenAI

Across the central empirical studies, GenAI occupies different positions in professional activity: a learning partner in iSTEM preparation, a production resource in digital storytelling, a conversational resource in collaborative planning, and part of a staged knowledge-building model (Lehtinen et al., 2026; Li et al., 2024; Wan & Gu, 2026; Zhang & Wang, 2025). The review interprets these differences as critical to understanding CT. Teachers are not encountering the same cognitive demand when checking a proposed explanation, negotiating a lesson plan, or choosing how to revise a digital story (Inie et al., 2026).

The AI perspective also concerns where reasoning is visible. Lehtinen et al. (2026) record prompts and activity sequences, while Wan and Gu (2026) assess the final story and reported tendencies. In the present studies, neither a finished product nor a prompt count alone is treated as sufficient evidence of the teacher's reasoning. The analysis instead considers how an AI contribution relates to the teacher's subsequent judgement, drawing on the different evidence types in

Lehtinen et al. (2026) and Wan and Gu (2026). What was checked, why it was revised, and which decision remained with the teacher are questions for analysis, rather than conclusions inferred from tool use alone.

This interpretation also qualifies claims about reduced thinking. Li et al. (2024) found lower reported cognitive load alongside better process-task performance in the ChatGPT condition. These measures do not show whether the reduction concerned unproductive effort or the removal of important reasoning work (Kosmyna et al., 2025). Conversely, Wan and Gu's (2026) product improvements do not show that all thinking was delegated to AI. The review therefore avoids treating efficiency as either proof of CT growth or proof of cognitive dependence (Raitskaya & Tikhonova, 2025).

7.2. Implications for Professional Development with GenAI

The reviewed studies support designing professional learning around tasks that require judgement, while leaving open which scaffold works best for which outcome (Borko, 2004; Lehtinen et al., 2026; Wan & Gu, 2026). Based on this synthesis, it is recommended to ask teachers to preserve their reasons for accepting, modifying, or rejecting generated material alongside the final product. The purpose of this recommendation is to make the relationship between tool support and teacher judgement more accessible to feedback and analysis. It extends the process-oriented approach illustrated by Lehtinen et al. (2026), but its effectiveness as a complete intervention remains to be tested. Meanwhile, Explicit evaluation criteria provide one possible approach. Sadidi and Prestel (2026) show the relevance of criterion-based evaluation, although their reported changes primarily concern perceptions. Disciplinary checking in Drushlyak et al. (2025) offers a related task design. These examples justify examining evaluation in PD, without establishing that one procedure will improve all dimensions of CT (Barta et al., 2022).

Human support should also be planned in relation to the activity's purpose. Li et al. (2024) do not demonstrate that ChatGPT can replace expert teachers across professional tasks. Their comparison instead provides grounds for examining where immediate AI support and human expertise contribute differently (Pahi et al., 2024). Consistent with Desimone's (2009) distinction between PD experiences and subsequent outcomes, programme evaluation should specify whether it seeks better immediate planning, changed dispositions, or sustained reasoning in practice. Teacher education must systematically integrate AI literacy and ethical frameworks to foster this professional judgement (Celik, 2023; Kohnen et al., 2025; Moorhouse et al., 2024). Teacher educators have also reported concerns about AI preparedness, ethics, and institutional policy in teacher education programmes (Massouti et al., 2026).

7.3. Research Priorities and Review Limitations

The first priority for future studies is making alignment between CT claims and

their evidence more visible. The difference between a disposition questionnaire, a coded discussion contribution, and a group activity sequence is consequential in the central empirical studies (Lehtinen et al., 2026; Li et al., 2024; Wan & Gu, 2026; Zhang & Wang, 2025). Future research could combine these sources while retaining their distinct meanings (Liao et al., 2025). A second priority is comparison designs that separate tool access, instructional scaffolds, and human guidance (Helal et al., 2026). Meanwhile, more delayed assessment across new tasks, including conditions, is also in need.

The participant base also requires more precise reporting. Pre-service samples dominate the central empirical studies, while the postgraduate sample in Zhang and Wang (2025) contains participants from both educational and non-educational employment (Huang & Sang, 2023). Pre-service science teachers' intentions to use GenAI have also been surveyed separately (Ramnarain et al., 2025), underscoring the need for more precise reporting of sample characteristics. Findings cannot simply be transferred to experienced school teachers. For future studies, reporting professional role and teaching experience would make the relevance of outcomes easier to judge (Yuan et al., 2022).

This review has a selective, interpretive scope and does not estimate the prevalence of particular findings across the entire field. Its conclusions are limited to the studies examined and to the evidence available for appraisal. Detailed methodological conclusions are drawn only where the relevant full text was examined. The selection boundary and any supplementary updating are reported in Section 3.1.

8. Conclusion

The reviewed evidence presents teachers' CT through self-reported tendencies, situated professional evaluation, and collaborative knowledge-building processes. GenAI-supported tasks create occasions for these practices, but the reported outcomes differ across designs and measures (Lehtinen et al., 2026; Li et al., 2024; Wan & Gu, 2026; Zhang & Wang, 2025). Based on these comparisons, this review interprets product quality, efficiency, and selected CT-related outcomes separately. Claims about GenAI's contribution are limited by the comparison condition and the form of evidence in each study (Lehtinen et al., 2026; Li et al., 2024; Wan & Gu, 2026; Zhang & Wang, 2025).

The principal conclusion of this review is that teachers' CT should be examined through their reasoned use of GenAI contributions and through evidence of what they subsequently learn. Among the studies appraised in detail here, supported activity and task performance are documented more directly than sustained CT transfer (Drushlyak et al., 2025; Lehtinen et al., 2026; Li et al., 2024; Wan & Gu, 2026; Zhang & Wang, 2025). Building on these findings and the importance of transfer discussed by Halpern (1998), this review recommends making teachers' judgements visible while assessing immediate AI-assisted performance separately from longer-term professional learning.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

Arseven, T., & Bal, M. (2025). Critical Literacy in Artificial Intelligence Assisted Writing Instruction: A Systematic Review. *Thinking Skills and Creativity, 57,* Article ID: 101850. https://doi.org/10.1016/j.tsc.2025.101850

Aydın Yıldız, T. (2026). Exploring the Impact of ChatGPT on Improving 21st-Century Skills for Future English Teachers during Lesson Planning. *Computers in the Schools, 43,* 302-325. https://doi.org/10.1080/07380569.2024.2429534

Barta, A., Fodor, L. A., Tamas, B., & Szamoskozi, I. (2022). The Development of Students Critical Thinking Abilities and Dispositions through the Concept Mapping Learning Method—A Meta-Analysis. *Educational Research Review, 37,* Article ID: 100481. https://doi.org/10.1016/j.edurev.2022.100481

Batdı, V., Elaldı, Ş., Özçelik, C., Semerci, N., & Özkaya, Ö. M. (2024). Evaluation of the Effectiveness of Critical Thinking Training on Critical Thinking Skills and Academic Achievement by Using Mixed-Meta Method. *Review of Education, 12,* e70001. https://doi.org/10.1002/rev3.70001

Borko, H. (2004). Professional Development and Teacher Learning: Mapping the Terrain. *Educational Researcher, 33,* 3-15. https://doi.org/10.3102/0013189x033008003

Celik, I. (2023). Towards Intelligent-TPACK: An Empirical Study on Teachers' Professional Knowledge to Ethically Integrate Artificial Intelligence (AI)-Based Tools into Education. *Computers in Human Behavior, 138,* Article ID: 107468. https://doi.org/10.1016/j.chb.2022.107468

Chan, C. K. Y., & Tsi, L. H. Y. (2024). Will Generative AI Replace Teachers in Higher Education? A Study of Teacher and Student Perceptions. *Studies in Educational Evaluation, 83,* Article ID: 101395. https://doi.org/10.1016/j.stueduc.2024.101395

Cheah, Y. H., Hsu, H., Lu, J., & Alsaka, R. (2026). Preservice Teachers' Prompting Practices and Decision-Making: Exploring Pathways from Talking to Reasoning with Generative Artificial Intelligence. *Journal of Research on Technology in Education.* https://doi.org/10.1080/15391523.2026.2704778

Chen, C. Q. (2026). Fostering AI Literacy and Critical Thinking in Pre-Service Teachers through Gai-Based Digital Storytelling within the CLEAR Framework. *Journal of Computers in Education, 13,* 653-678. https://doi.org/10.1007/s40692-025-00372-z

Chiu, T. K. F. (2024). A Classification Tool to Foster Self-Regulated Learning with Generative Artificial Intelligence by Applying Self-Determination Theory: A Case of ChatGPT. *Educational Technology Research and Development, 72,* 2401-2416. https://doi.org/10.1007/s11423-024-10366-w

Chounta, I., Bardone, E., Raudsep, A., & Pedaste, M. (2022). Exploring Teachers' Perceptions of Artificial Intelligence as a Tool to Support Their Practice in Estonian K-12 Education. *International Journal of Artificial Intelligence in Education, 32,* 725-755. https://doi.org/10.1007/s40593-021-00243-5

Darling-Hammond, L., & McLaughlin, M. W. (2011). Policies That Support Professional Development in an Era of Reform. *Phi Delta Kappan, 92,* 81-92. https://doi.org/10.1177/003172171109200622

Desimone, L. M. (2009). Improving Impact Studies of Teachers' Professional Development: Toward Better Conceptualizations and Measures. *Educational Researcher, 38,* 181-199.

https://doi.org/10.3102/0013189x08331140

Drushlyak, M., Lukashova, T., Shamonia, V., & Semenikhina, O. (2025). ChatGPT-Based Simulation Helps to Develop the Pre-Service Mathematics Teachers' Critical Thinking. *International Journal of Instruction, 18,* 153-172. https://doi.org/10.29333/iji.2025.1819a

Ennis, R. H. (1989). Critical Thinking and Subject Specificity: Clarification and Needed Research. *Educational Researcher, 18,* 4-10. https://doi.org/10.3102/0013189x018003004

Facione, P. A. (1990). *Critical Thinking: A Statement of Expert Consensus for Purposes of Educational Assessment and Instruction.* Research Findings and Recommendations. https://eric.ed.gov/?id=ED315423

Felix, S. M., Lønnum, M., Lykknes, A., & Staberg, R. L. (2025). Teachers' Understanding of and Practices in Critical Thinking in the Context of Education for Sustainable Development: A Systematic Review. *Education Sciences, 15,* Article 824. https://doi.org/10.3390/educsci15070824

Halpern, D. F. (1998). Teaching Critical Thinking for Transfer across Domains: Disposition, Skills, Structure Training, and Metacognitive Monitoring. *American Psychologist, 53,* 449-455. https://doi.org/10.1037/0003-066x.53.4.449

Helal, M. Y. I., Elgendy, I. A., Albashrawi, M. A., Dwivedi, Y. K., Al-Ahmadi, M. S., & Jeon, I. (2026). The Impact of Generative AI on Critical Thinking Skills: A Systematic Review, Conceptual Framework and Future Research Directions. *Information Discovery and Delivery, 54,* 690-704. https://doi.org/10.1108/idd-05-2025-0125

Huang, J., & Sang, G. (2023). Conceptualising Critical Thinking and Its Research in Teacher Education: A Systematic Review. *Teachers and Teaching, 29,* 638-660. https://doi.org/10.1080/13540602.2023.2212364

Inie, N., Kristensen, A. M., Misfeldt, M., & Dalsgaard, P. (2026). A Critical Review of Critical Thinking in HCI Research aboutAI. *International Journal of Human-Computer Interaction*. https://doi.org/10.1080/10447318.2026.2722522

Jegstad, K. M., Heggernes, S. L., Jøsok, E., Ryen, E., Svanes, I. K., & Tørnby, H. (2025). Approaches to Critical Thinking in Primary Education Classrooms: A Systematic Review. *Educational Research Review, 48,* Article ID: 100711. https://doi.org/10.1016/j.edurev.2025.100711

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F. et al. (2023). ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. *Learning and Individual Differences, 103,* Article ID: 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kohnen, A. M., Newell, B., Boada, D., Breil, B. T., Fabulich, L., Miller, M. et al. (2025). From University to Classroom: A Collaborative Framework for Generative AI in Middle School Education. *Middle School Journal, 56,* 11-21. https://doi.org/10.1080/00940771.2025.2557791

Kohnke, L., Zou, D., Lai, C., & Gu, M. M. (2026). Preparing Pre-Service Teachers for Responsible Generative AI Use: Curriculum Implications for Ethics, Privacy, and AI Literacy. *Computers and Education: Artificial Intelligence, 10,* Article ID: 100617. https://doi.org/10.1016/j.caeai.2026.100617

Korucu-Kış, S., & Bakla, A. (2026). Exploring Pre-Service EFL Teachers' Perceptions, Use Cases and Questioning Practices of ChatGPT during Practicum: A Critical Digital Pedagogy Perspective. *Instructional Science, 54,* Article No. 33. https://doi.org/10.1007/s11251-026-09783-6

Kosmyna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X. H., Beresnitzky, A. V., Braunstein, I., & Maes, P. (2025). Your Brain on ChatGPT: Accumulation of Cognitive

Debt When Using an AI Assistant for Essay Writing Task. *arXiv.Org.* https://www.proquest.com/publiccontent/docview/3304336749?pq-origsite=primo&sourcetype=Working%20Papers

Lehtinen, A., Channa, F., Näykki, P., Ahlström, E., & Hiljanen, M. (2026). Generative AI for Collaborative Learning: Fostering Critical Thinking in Teacher Education. *Journal of Computer Assisted Learning, 42,* e70256. https://doi.org/10.1002/jcal.70256

Li, T., Ji, Y., & Zhan, Z. (2024). Expert or Machine? Comparing the Effect of Pairing Student Teacher with In-Service Teacher and ChatGPT on Their Critical Thinking, Learning Performance, and Cognitive Load in an Integrated-Stem Course. *Asia Pacific Journal of Education, 44,* 45-60. https://doi.org/10.1080/02188791.2024.2305163

Liao, W., Liu, M., Wang, Z., & Qin, K. (2024). Chinese Expert Teachers' Critical Thinking Strategies for Professional Growth. *Professional Development in Education, 50,* 1089-1103. https://doi.org/10.1080/19415257.2022.2097290

Liao, W., Liu, W., & Yuan, R. (2025). How Is Teachers' Critical Thinking Defined, Approached, Measured, and Evaluated in Empirical Studies? A Methodological Review. *Thinking Skills and Creativity, 57,* Article ID: 101795. https://doi.org/10.1016/j.tsc.2025.101795

Massouti, A., Baroudi, S., & Shaya, N. (2026). Exploring AI Integration in United Arab Emirates' Teacher Education Programmes: The Voice of Educators Leading the Change. *Cambridge Journal of Education, 56,* 383-401. https://doi.org/10.1080/0305764x.2026.2651759

Moorhouse, B. L., Wan, Y., Wu, C., Kohnke, L., Ho, T. Y., & Kwong, T. (2024). Developing Language Teachers' Professional Generative AI Competence: An Intervention Study in an Initial Language Teacher Education Course. *System, 125,* Article ID: 103399. https://doi.org/10.1016/j.system.2024.103399

Pahi, K., Hawlader, S., Hicks, E., Zaman, A., & Phan, V. (2024). Enhancing Active Learning through Collaboration between Human Teachers and Generative AI. *Computers and Education Open, 6,* Article ID: 100183. https://doi.org/10.1016/j.caeo.2024.100183

Pithers, R. T., & Soden, R. (2000). Critical Thinking in Education: A Review. *Educational Research, 42,* 237-249. https://doi.org/10.1080/001318800440579

Raitskaya, L., & Tikhonova, E. (2025). Enhancing Critical Thinking Skills in Chatgpt-Human Interaction: A Scoping Review. *Journal of Language and Education, 11,* 5-19. https://doi.org/10.17323/jle.2025.27387

Ramnarain, U., Ogegbo, A. A., Penn, M., Ojetunde, S., & Mdlalose, N. (2025). Pre-Service Science Teachers' Intention to Use Generative Artificial Intelligence in Inquiry-Based Teaching. *Journal of Science Education and Technology, 34,* 1272-1285. https://doi.org/10.1007/s10956-024-10159-z

Sadidi, F., & Prestel, T. (2026). Impact of Criterion-Based Reflection on Prospective Physics Teachers' Perceptions of ChatGPT-Generated Content. *Discover Education, 5,* Article No. 57. https://doi.org/10.1007/s44217-025-01071-8

Shulman, L. S. (1986). Those Who Understand: Knowledge Growth in Teaching. *Educational Researcher, 15,* 4-14. https://doi.org/10.3102/0013189x015002004

Sun, Y., & Huang, Y. (2025). Exploring the Role of Generative AI in Collaborative Lesson Planning for Pre-Service Teachers. *Journal of Computer Assisted Learning, 41,* e70098. https://doi.org/10.1111/jcal.70098

Toncel Herrera, N. R., & Jiménez Sierra, Á. A. (2026). Developing Critical Thinking in English Teacher Training Programs—A Systematic Review. *Frontiers in Education, 11,* Article 1698530. https://doi.org/10.3389/feduc.2026.1698530

Wan, P., & Gu, X. (2026). Designing an Integrated Concept Mapping and Generative AI

Approach to Develop Pre-Service Teachers' Digital Storytelling Project Outcomes and Critical Thinking. *British Educational Research Journal, 52,* 1313-1338. https://doi.org/10.1002/berj.70060

Wang, Y., & Wu, Z. (2023). Adapting or Adopting? Critical Thinking Education in the East Asian Cultural Sphere: A Systematic Integrative Review. *Thinking Skills and Creativity, 49,* Article ID: 101330. https://doi.org/10.1016/j.tsc.2023.101330

Wei, X., Wang, L., Koszalka, T. A., Lee, L., & Liu, R. (2025). Enhancing Pre-Service Teachers' Reflective Thinking Skills through Generative AI-Assisted Digital Storytelling Creation: A Three-Dimensional Framework Analysis. *Computers & Education, 235,* Article ID: 105356. https://doi.org/10.1016/j.compedu.2025.105356

Yu, L., & Zin, Z. M. (2023). The Critical Thinking-Oriented Adaptations of Problem-Based Learning Models: A Systematic Review. *Frontiers in Education, 8,* Article 1139987. https://doi.org/10.3389/feduc.2023.1139987

Yuan, R., Liao, W., Wang, Z., Kong, J., & Zhang, Y. (2022). How Do English-as-a-Foreign-Language (EFL) Teachers Perceive and Engage with Critical Thinking: A Systematic Review from 2010 to 2020. *Thinking Skills and Creativity, 43,* Article ID: 101002. https://doi.org/10.1016/j.tsc.2022.101002

Zhang, H., & Wang, Q. (2025). How Could Genai Work on In-Service Teachers' Knowledge Building Process? An Empirical Study Based on Epistemic Network Analysis. *International Journal of Educational Technology in Higher Education, 22,* Article No. 47. https://doi.org/10.1186/s41239-025-00544-y

Zhou, J., Zhao, N., Gao, X., & Han, X. (2026). The Effects of Generative AI -Supported Peer Feedback on Preservice Teachers' Learning Outcomes, Feedback Uptake, and Critical Thinking. *Thinking Skills and Creativity, 62,* Article ID: 102236. https://doi.org/10.1016/j.tsc.2026.102236

Zhou, X., Teng, D., & Al-Samarraie, H. (2024). The Mediating Role of Generative AI Self-Regulation on Students' Critical Thinking and Problem-Solving. *Education Sciences, 14,* Article 1302. https://doi.org/10.3390/educsci14121302

Appendix

Database Search Strategies

Scopus and Web of Science were searched on 8 September 2026 for literature published through 31 August 2026. The main searches combined three sets of terms relating to generative artificial intelligence (GenAI), teachers, and critical thinking. The supplementary searches added a fourth set of terms to identify reviews and other evidence syntheses within the same topic. An English-language filter was applied in both databases, and only English-language publications were eligible at screening.

The search strings below were reconstructed from the search terms described in the review. The publication cut-off is reported separately because the search strings do not encode a day-level date limit.

1) Scopus main search

```
TITLE-ABS-KEY(
  ("generative AI" OR "generative artificial intelligence" OR GenAI
   OR ChatGPT OR "large language model*")
  AND
  (teacher* OR "preservice teacher*" OR "pre-service teacher*"
   OR "in-service teacher*" OR "student teacher*" OR "teacher educator*")
  AND
  ("critical thinking" OR "critical reasoning" OR "critical evaluation")
)
AND PUBYEAR < 2027
AND LIMIT-TO(LANGUAGE, "English")
```

2) Scopus supplementary search for reviews and evidence syntheses

```
TITLE-ABS-KEY(
  ("generative AI" OR "generative artificial intelligence" OR GenAI
   OR ChatGPT OR "large language model*")
  AND
  (teacher* OR "preservice teacher*" OR "pre-service teacher*"
   OR "in-service teacher*" OR "student teacher*" OR "teacher educator*")
  AND
  ("critical thinking" OR "critical reasoning" OR "critical evaluation")
  AND
  (review* OR "meta-analysis" OR "meta-analyses" OR "evidence synthesis")
)
AND PUBYEAR < 2027
AND LIMIT-TO(LANGUAGE, "English")
```

3) Web of Science main search

```
TS=(
  ("generative AI" OR "generative artificial intelligence" OR GenAI
   OR ChatGPT OR "large language model*")
  AND
```

```
   (teacher* OR "preservice teacher*" OR "pre-service teacher*"
    OR "in-service teacher*" OR "student teacher*" OR "teacher educator*")
   AND
   ("critical thinking" OR "critical reasoning" OR "critical evaluation")
)
AND LA=(English)
```

4) Web of Science supplementary search for reviews and evidence syntheses

```
TS=(
   ("generative AI" OR "generative artificial intelligence" OR GenAI
    OR ChatGPT OR "large language model*")
   AND
   (teacher* OR "preservice teacher*" OR "pre-service teacher*"
    OR "in-service teacher*" OR "student teacher*" OR "teacher educator*")
   AND
   ("critical thinking" OR "critical reasoning" OR "critical evaluation")
   AND
   (review* OR "meta-analysis" OR "meta-analyses" OR "evidence synthesis")
)
AND LA=(English)
```