International Journal of Arts and Humanities Studies
ISSN: 2754-4311
DOI: 10.32996/ijahs
Journal Homepage: al-kindipublisher.com/index.php/ijahs/index

Beyond Better Essays: Generative AI, Evidence Evaluation, and Critical Thinking in L2 English Writing: A Source-Audited Critical Integrative Review

Nehad Salah Mohammed Ramadan[1] and Asjad Ahmed Saeed Balla ✉[2]

[1] Department of Foreign Languages, Faculty of Education, Mansoura University, Egypt; Department of English Language and Literature, College of Languages and Humanities, Qassim University, Saudi Arabia

[2] Department of English Language, Faculty of Education, Gedarif University, Sudan; Department of English Language and Literature, College of Languages and Humanities, Qassim University, Saudi Arabia

Corresponding Author: Asjad Ahmed Saeed Balla**, E-mail:** A.Balla@qu.edu.sa

ABSTRACT

Generative artificial intelligence (GenAI) can support second-language (L2) writers in planning, drafting, revising, and receiving feedback. But the fact that it can write flowing prose does not prove it can reason independently or rely on dependable sources. This critical integrative review explores what source-audited empirical evidence supports about GenAI, L2 English writing, critical thinking, and evidence evaluation. The evidence set comprises 20 bibliographically verified publications. The researcher confirmed that six were tertiary EFL/ESL English academic-writing studies; 13 were closely related L2/EFL writing studies but lacked one strict scope detail in the accessible primary record; and one reading-based critical-thinking study was retained only as contextual evidence. She extracted only facts supported by publisher records, full texts where available, and Crossref metadata. This review is not systematic, and we make no claims about reproducible database-search logs, dual screening, or pooled effect estimation. Across the audited source record, evidence more often concerned immediate assisted revision, feedback, engagement, and process outcomes than independent transfer. Results for critical thinking were mixed and depended on the instructional design and outcome assessed. No study in the evidence set directly assessed citation accuracy, lateral reading, or a similar objective outcome of evidence evaluation. The review thus distinguishes between textual quality, argumentative reasoning, and evidential quality. It argues that GenAI-supported writing pedagogy should require source verification, claim-evidence justification, and AI-free transfer assessment, rather than using polished, AI-assisted text as a proxy for epistemic learning.

KEYWORDS

Generative AI; L2 writing; academic writing; critical thinking; evidence evaluation; source verification; AI literacy

ARTICLE INFORMATION

ACCEPTED: 10 October 2026 **PUBLISHED:** 10 November 2026 **DOI:** 10.32996/ijahs.2026.6.5.5

1. Introduction

Generative artificial intelligence (GenAI) has expanded the support available to L2 writers. This differs from earlier tools that focused on spelling, grammar, translation, or automated corrective feedback. Today's large language models can provide ideas, produce alternative formulations, rewrite passages, and respond to prompts during a longer writing process. These affordances may diminish linguistic barriers and widen access to iterative feedback, but they also raise questions of factual reliability, authorship, learner dependence, and the division of cognitive labour between writer and system (Barrot, 2023; UNESCO, 2023; Warschauer et al., 2023).

So, the key question isn't just whether GenAI is able to make an essay look better. In academic writing, writers are expected to make claims, support them with evidence, interpret sources, and make sound choices about evidence. These activities are embedded in the social and disciplinary practices that constitute academic literacy (Lea & Street, 1998). In L2 contexts, reading and writing are closely connected through summarisation, synthesis, paraphrase, and source-based composing (Grabe & Zhang, 2016; Plakans & Gebril, 2012). Textual fluency, evidential soundness, and argumentative coherence are related but distinct achievements.

The distinction matters when a system can produce plausible content quickly. Metrics like cohesion, syntactic complexity, and lexical sophistication do not determine whether an essay's claims are warranted or its sources are accurate; they can explain expert

Copyright: © 2026 the Author(s). This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) 4.0 license (https://creativecommons.org/licenses/by/4.0/). Published by Al-Kindi Centre for Research and Development, London, United Kingdom.

judgments of writing quality (McNamara et al., 2010). Argumentative writing requires reasons, counterarguments, and explicit warrants for claims rather than fluent presentation (Wingate, 2012). Furthermore, as highlighted in recent pedagogical explorations within the Saudi context, integrating digital tools and AI applications necessitates a rigorous balance between optimizing technological opportunities and mitigating risks, ensuring that foundational critical thinking remains central to language education (Elmahdi, AbdAlgane, & Balla, 2024; Hago Elmahdi, Allehyani, & AbdAlgane, 2025).

The need for verification is not abstract. Students struggle to assess online information and explain why they consider it credible (McGrew et al., 2018). Expert fact-checkers use lateral reading: they leave a site, search for independent information about it, and return with a more contextualized judgment (Wineburg & McGrew, 2019). These behaviours can be taught and assessed as source-based practices (Britt & Aglinskas, 2002). However, fluent language-model output may be ungrounded (Ji et al., 2023). In a model- and prompt-specific audit, Walters and Wilder (2023) identified fabricated citations in 55% of GPT-3.5 references and 18% of GPT-4 references, and numerous real citations contained substantive bibliographic errors. Similarly, an audit in the medical domain reported fabricated and inaccurate references under the conditions it tested (Bhattacharyya et al., 2023).

This review raises the question of what the current auditable empirical record supports regarding GenAI-assisted L2 English writing, critical thinking, and evidence evaluation. Its contribution is deliberately cautious. Instead of giving a percentage synthesis based on an uncertain scope classification in the denominator, it distinguishes (a) studies that are clearly within a strict tertiary English academic-writing scope, (b) related studies where a scope element could not be verified from accessible primary materials, and (c) contextual studies outside the writing scope. This way, the researcher avoids making statements that cannot be attributed to a source.

2. Conceptual Framework

2.1 Three qualities of writing

Three analytically separable qualities are identified by this review. Linguistic or textual quality relates to accuracy, cohesion, lexis, syntax, and organization. Argumentative quality concerns the relations among claims, reasons, counterarguments, and warrants, while evidential, or epistemic, quality concerns the trustworthiness of sources, the correctness of citations, and whether evidence actually supports the claims it is meant to support. The second and third forms cannot be inferred from the first.

Critical thinking involves argument and evidence, rather than merely fluent expression. Critical thinking involves evaluation, inference, explanation, and metacognitive control (Facione, 1990), along with coordination of knowledge claims with sources and justification (Chinn et al., 2011). To develop critical thinking, effective instruction embeds substantive work, dialogue, authentic problems, and mentoring (Abrami et al., 2015). Moreover, frameworks designed for culturally responsive language learning emphasize the direct integration of critical thinking and advanced technologies in writing classrooms to scaffold these epistemic competencies (Elmahdi, AbdAlgane, Hamid, Balla, & Ibrahim, 2025). Thus, the review considers critical-thinking outcomes direct only when a study measures them, not when learners merely report feeling more reflective.

2.2 Why verification must be visible

A typical writing rubric might reward content, organization, language, and apparent support, but fail to address whether a source is cited, a claim is accurately attributed, or the evidence fits the claim. In a GenAI setting, a writer might be presented with a seemingly plausible list of sources, a seemingly coherent rationale, or a polished paragraph, none of which has been checked. Evidence evaluation must be made observable through tasks such as checking authoritative records, corroborating claims, explaining the selection of sources, and documenting reasons for accepting or rejecting an AI suggestion.

2.3 Cognitive offloading as a mechanism hypothesis

Cognitive offloading is using external representations or actions to alter a task's information-processing demands (Risko & Gilbert, 2016). Anticipated use of external information may change what people remember leading them to retain where to find information rather than the information itself (Sparrow et al., 2011). Easy offloading may also improve immediate task performance at the expense of later memory, although salient learning goals can help reduce this trade-off (Grinschgl et al., 2021). These findings do not offer direct evidence on GenAI-supported L2 writing. They offer a mechanism hypothesis: when instructional designs reward only assisted products, writers might outsource monitoring and justification instead of practicing them.

3. Review Approach and Source Audit

3.1 Review type and evidential boundaries

The paper is a critical integrative review, not a systematic review. It does not include a prospectively registered protocol, exhaustive database retrieval, independently duplicated screening, a PRISMA flow diagram, or a meta-analysis. PRISMA reporting principles remain useful as a transparency benchmark (Page et al., 2021), and SWiM is a useful warning against unsupported vote counting when studies are heterogeneous (Campbell et al., 2020). Citing these frameworks here does not imply that this review meets all of their requirements.

The core evidence set includes 20 publications relevant to GenAI, L2/EFL writing, feedback, critical thinking, or academic writing. All bibliographic records were checked against their DOI-linked publisher records and Crossref metadata. Where the publisher provided full text, we used it to confirm participants, setting, task, design, and outcomes. If only an abstract or record was available,

extraction was limited to the evidence available. The synthesis does not go beyond the detail provided in the readily available source record.

3.2 Scope classification

The strict focal scope is tertiary EFL/ESL/L2 writers using GenAI in English academic writing. The researcher confirmed that six publications met all of these criteria from accessible primary materials. Thirteen were retained as related evidence because one scope element—usually the academic-writing genre, the tertiary status, or the language of the written product—was not explicit in the materials accessed. She reclassified one publication as contextual because it evaluated critical thinking in an English-reading course rather than a writing task. This is a record-level audit result, not an estimate of the field. Table 1 presents the eligibility criteria, and Table 2 summarizes the classification of the 20 publications.

Table 1. Eligibility Criteria for Empirical Candidate Studies

Criterion	Inclusion	Exclusion
Publication period	1 January 2020–4 September 2026	Before 1 January 2020 or after 4 September 2026
Publication type	Peer-reviewed empirical journal articles	Editorials, commentaries, purely conceptual papers for core synthesis
Population	University/tertiary EFL, ESL, or L2 English writers	L1-only writing populations
Language	Academic writing produced in English as an L2	Studies focused exclusively on writing in another target language
GenAI exposure	Direct learner use of ChatGPT/LLM/GenAI or empirical investigation of GenAI-assisted L2 writing	Conventional automated writing evaluation without generative models
Writing context	Academic, argumentative, integrated, source-based, thesis, or comparable university writing	General conversation/vocabulary tasks without writing
Outcome relevance	Writing performance/process, CT/reasoning, evidence/source evaluation, agency/critical literacy	Technology acceptance alone without a writing or cognitive outcome
Evidence	Original quantitative, qualitative, or mixed-methods data	Reviews included only as contextual evidence

Note. These criteria govern the provisional empirical core, not conceptual and review sources used to inform the broader integration. AI acting only as a rater does not by itself establish learner assistance; where learners' writing and feedback conditions were unavailable, the decision was left unclear. The supplementary workbook records eligibility decisions, accessible source versions, and outstanding questions.

Table 2: Source-audited classification of the 20 bibliographically verified publications

Classification	Publications	Use in this review
Confirmed strict-scope core (n = 6)	Mahapatra (2024); Song and Song (2023); Wang (2024); Tran (2025); Alsadoon (2026); Zhang et al. (2025).	Used for claims specifically about tertiary EFL/ESL English academic writing. Claims remain limited to the measured outcome in each study.
Related writing evidence (n = 13)	Tsai et al. (2024); Hwang et al. (2025); Cengiz et al. (2025); Zare et al. (2025); Yin et al. (2025); Zou and Huang (2024); Yan and Zhang (2024); Muñoz Muñoz et al. (2025); Asadi et al. (2025); Yan (2023); Huang and Wang (2025); Moorhouse et al. (2025); Shen and Chen (2025).	Used only for the supported construct (for example, feedback uptake, engagement, or reported perceptions). The unverified scope detail is not inferred.
Contextual evidence (n = 1)	Liang and Wu (2024).	Not counted as a writing study. Used only to illustrate that ChatGPT exposure in a tertiary EFL reading course did not automatically yield a critical-thinking test gain.

Note. "Confirmed" refers to what was established in accessible primary materials, not to overall study quality or causal certainty. Full study-level source records and access notes are provided in Supplementary File S1. Liang and Wu's contextual classification is provisional because the accessible materials do not establish whether the study assigned a qualifying writing task.

3.3 Appraisal and synthesis

Appraisal was qualitative and design-sensitive. We evaluated the correspondence among the study question, design, sampling, outcome measure, comparison, and interpretation using the Mixed Methods Appraisal Tool logic (Hong et al., 2018). The researcher paid particular attention to single-group studies, brief interventions, self-report outcomes, and product comparisons that did not use an unaided transfer measure. The studies use different tasks, feedback conditions, outcome measures, and timeframes. Thus, pooling the effect estimates would be indefensible, and no meta-analysis was attempted.

4. Findings

Table 3 summarizes the six strict-scope core studies identified in Table 2.

Table 3. Characteristics and Findings of the Six Confirmed Core Studies

Study	Context / design	Outcomes examined	Principal finding and limitation
Mahapatra (2024)	72 complete cases; tertiary ESL mixed-methods intervention.	Writing product / process, Self-directed action / integrity	Post-test and delayed writing scores favored the ChatGPT condition. Group-size reports conflict; test assistance was unclear.
Song & Song (2023)	50 tertiary EFL learners; 12-week group comparison.	Writing product / process, Essay reasoning / argument, Self-directed action / integrity	Higher post-test writing and motivation scores were reported for the ChatGPT condition; test assistance and several reporting details are unclear.
Wang (2024)	79 pre-survey and 61 valid post-survey responses; EFL writing journals.	Writing product / process, Self-directed action / integrity, Learner source-use practices	Learners reported selective use of AI and translation tools. Findings are largely self-reported and lack a comparison group.
Tran (2025)	14 tertiary EFL learners; four essays and two feedback sequences.	Writing product / process, Self-directed action / integrity	AI-, teacher- and self-initiated revisions were documented. Revision frequency is not evidence of independent proficiency gain.
Alsadoon (2026)	60 female university EFL students; observed argumentative-writing task.	Writing product / process, Self-directed action / integrity	Prompt logs and text-overlap proxies documented patterns of use. Associations and measurement inconsistencies limit interpretation.
Zhang et al. (2025)	42 tertiary EFL learners; eye tracking, logic tests and offline essay revision.	Writing product / process, Essay reasoning / argument, Self-directed action / integrity	Attention measures were associated with logic outcomes; the single-group design does not establish causality or transfer.

Note. The six studies are the strict-scope core within the 20-publication set. The 13 related studies and one contextual study remain part of the broader integrative review, as classified in Table 2. The source-supported design details and limitations are recorded in Supplementary File S1.

Table 4 maps the outcomes reported in the six core studies.

Table 4. Outcome Domains in the Six Confirmed Core Studies

Study	Writing	Essay reasoning	Action / integrity	Source use	Objective source evaluation
Mahapatra (2024)	Y	NI	Y	NI	NI
Song & Song (2023)	Y	Y	Y	NI	NI
Wang (2024)	Y	NI	Y	Y	NI
Tran (2025)	Y	NI	Y	NI	NI
Alsadoon (2026)	Y	NI	Y	U	NI
Zhang et al. (2025)	Y	Y	Y	NI	NI

Note. Y = the accessible source reports a relevant outcome or measure; NI = not identified in the sufficiently complete texts inspected; U = unclear because the available evidence is insufficient. Writing includes assessed or analyzed products or processes; reasoning covers directly assessed or analyzed essay reasoning or argument; action/integrity includes reported choices, self-initiated revision, or integrity outcomes and does not imply validated autonomy; source use covers documented sourcing or checking; objective evaluation covers assessed source credibility, citation accuracy, lateral reading, or verification. Zhang et al. included a scored sample-to-general-claim reasoning item, not a separate source-verification measure. Codes describe reporting, not effect direction or magnitude. See Supplementary File S1 for supporting evidence and access limitations.

4.1 Assisted products and revision processes

Several studies reported immediate benefits for assisted writing products or feedback-supported revision. Tsai et al. (2024) found that EFL English majors who revised essays with ChatGPT scored higher, and the abstract reported gains in vocabulary, grammar, organization, and content. In a confirmed tertiary ESL academic-writing intervention, Mahapatra (2024) reported higher post-test and delayed post-test scores for the ChatGPT-feedback condition. The article's group-size reporting is internally inconsistent. The ChatGPT-assisted group showed higher post-test writing and motivation results on IELTS Academic Writing Tasks (Song & Song, 2023). The publisher's 2026 correction fixed a contradictory sentence about the results (Song & Song, 2026).

These results do not support the stronger claim that GenAI reliably develops unaided writing ability. The studies differed in the source of feedback, task, comparison condition, and assessment timing. This distinction matters, as an edited text reflects what a learner can do with assistance rather than what the learner can do independently later. Direct evidence in the source-audited record of AI-free transfer is limited.

This process supports further product findings. Cengiz et al. (2025) found that students accepted language-use feedback more often than content or organization feedback. Hwang et al. (2025) found that EFL learners tended to use general prompts, and surface features improved more than higher-order ones. Yan and Zhang's (2024) four-case study identified considerable differences in metacognitive regulation and feedback engagement. Tran (2025) established that in an academic-writing course, AI-plus-teacher feedback generated more revisions than either source alone; however, revision frequency does not necessarily indicate overall improvement in writing proficiency. These two studies suggest that task design and feedback uptake mediate the effect rather than a uniform tool effect.

4.2 Critical thinking and argumentative reasoning

Evidence for critical thinking is mixed and should be considered in light of the measures used. Yin et al. (2025) describe a one-group undergraduate EFL intervention that improved critical thinking in writing. This finding is retained without an inferred sample size, task description, or effect magnitude, as the source audit did not have access to the full methods and results.

Zhang et al. (2025) conducted a confirmed tertiary EFL English argumentative-writing study using a GPT-4-powered logic-learning chatbot, eye tracking, logic tests, essay revision scores, and interviews. The reported analyses indicated that more frequent attention to GPT responses was associated with delayed logic-knowledge scores, while longer attention to learner prompts predicted lower post-writing scores. The authors did not observe a straightforward improvement in argumentative writing logic. The pattern better aligns with the view that critical engagement with AI outputs is an instructional process to be designed and measured, rather than a default consequence of chatbot access.

Liang and Wu (2024) reported no statistically significant pre/post change on a critical-thinking test following a ChatGPT-based program in an English Reading III course. Although the study does not provide evidence about writing, it does point to a gap between learners' positive perceptions and measurable improvement on a critical-thinking instrument.

4.3 Agency, integrity, and critical AI literacy

Qualitative and mixed-methods studies suggest that learners can find GenAI both enabling and unsettling. Wang (2024) reported on Japanese university EFL learners' changing, self-reported use of machine translation and GenAI across planning, drafting, and editing. No independently scored writing-gain outcome is provided in the paper. Moorhouse et al. (2025) found that postgraduate

L2 writers reported empowerment but limited critical awareness of the way the tools work. Graduate EFL writers used ChatGPT for language editing, knowledge inquiry, and seeking inspiration, with varying degrees of critical engagement (Shen & Chen, 2025).

A complementary behavioural perspective is provided by Alsadoon (2026). The study used screen recordings, prompt and text reuse metrics, surveys, and interviews in a timed university EFL argumentative-writing task. Its findings characterise constructive, hybrid, and substitutive use patterns and an association between increased copying and decreased integrity-belief scores. The study is not evidence that ChatGPT has improved writing. Instead, it supports examining integrity, policy awareness, and revision behavior alongside the final text.

These results align with AI-literacy frameworks that distinguish between technical operation and responsible evaluation and use (Long & Magerko, 2020; Ng et al., 2021). In writing pedagogy, agency should be understood as accountable decision-making: learners should be able to say what a system suggested, what they accepted or rejected, what they checked, and why. Furthermore, incorporating reflective practices and leveraging generative AI within culturally grounded teaching environments fosters deep instructor and learner agency, safeguarding against passive reliance on automated texts (Elmahdi, Daffaallah, Mohamed, Tayebelasmaa, & Babekir, 2026; Elmahdi, Daffaallah, Balla, Mohamed, & Tayebelasmaa, 2026).

4.4 Evidence evaluation remains unmeasured

None of the studies in the source-audited evidence set examined citation accuracy, source-credibility judgement, lateral reading, or any other objective evidence-evaluation outcome. Some studies describe learners asking artificial intelligence systems for examples or information. Some discuss integrity or accuracy. Those observations do not indicate that learners consulted sources or assessed the match between claims and evidence. This absence of a direct measure is significant, as source-based writing depends on the ability to interpret, integrate and responsibly attribute external material (Leijten et al., 2019).

The audit therefore supports a narrow conclusion: Current studies more often document assisted text, feedback uptake, engagement, perceptions, or process traces than the verification practices that establish evidential quality. This conclusion is based on the source-audited set and is not to be generalised as a prevalence estimate for all GenAI-and-writing research. Table 5 summarizes the source-supported findings and their interpretive limits.

Table 5: Source-supported findings and limits relevant to the review question

Outcome area	What the source-audited record supports	Interpretive limit
Assisted products and feedback	Immediate revised-product, test-score, and feedback-uptake benefits are reported in several studies, including Tsai et al. (2024), Mahapatra (2024), Song and Song (2023), Cengiz et al. (2025), and Tran (2025).	Conditions, comparators, and outcomes differ. Assisted performance does not establish unaided transfer.
Reasoning and argument	Yin et al. (2025) report improvement in critical thinking in writing; Zhang et al. (2025) report process–outcome associations in a GPT-4 logic-learning intervention.	The studies do not justify a general causal claim that GenAI improves critical thinking.
Agency and integrity	Wang (2024), Moorhouse et al. (2025), Shen and Chen (2025), and Alsadoon (2026) document perceived agency, awareness, and behavioural or ethical tensions.	Much of this evidence is qualitative or self-reported; it is not a proficiency outcome.
Evidence evaluation	No direct measure of citation accuracy, source credibility, or lateral reading was identified in the accessible source-audited record.	This is a record-level finding for the audited set, not a field-wide prevalence claim.

5. Discussion

5.1. Better assisted text is not sufficient evidence of better epistemic judgment

An asymmetrical synthesis is the most defensible. GenAI could help with local language revision, feedback uptake, draft efficiency, and learner engagement. Though they don't replace reasoning or evidential judgment, these outcomes are educationally valuable, especially when learners need language support quickly. Unsupported claims, fabricated citations, or misattributed evidence can still appear in a well-structured, apparently accurate paragraph. Accordingly, the quality of assisted writing should be disaggregated, rather than inferred from a single holistic score.

If the task is to produce a polished response, the tool can quickly provide the language and structure. If it involves learners questioning outputs, comparing sources, justifying revisions, and completing a delayed AI-free task, more epistemic work remains with the learner. This asymmetry also helps interpret mixed findings. The latter architecture is a more meaningful test of learning. It connects assessment with long-term capability rather than with a single assisted performance (Boud & Falchikov, 2006).

6. Conclusion

GenAI can help L2 writers generate and revise text, but a more attractive assisted essay does not show that a writer is becoming a better evaluator or justifier of knowledge. In the source-audited record, evidence of immediate textual feedback, engagement, and process outcomes was more apparent than evidence of independent reasoning or evidence evaluation. Critical-thinking results were mixed and varied by educational activity and measure. No objective evaluation of source verification was identified.

The practical implication is not to ignore GenAI. Instead, place GenAI within an architecture of accountable literacy. Writing pedagogy and research need to differentiate between textual, argumentative, and evidential quality; require learners to verify and justify; record the conditions of AI use; and assess what learners can accomplish without the system. These decisions enable technology to support language development without supplanting the epistemic work academic writing is meant to foster.

Declarations

Reporting status: This manuscript is a critical integrative review and does not claim systematic review compliance or meta-analytic estimation.

AI assistance: The submitted manuscript reports AI assistance with source checking, drafting source-linked interpretation, grammar, and editing.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Nehad Salah Mohammed Ramadan · ORCID iD: 0009-0005-1712-5349

Asjad Ahmed Saeed Balla · ORCID iD: 0000-0001-9387-8653

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

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. *Review of Educational Research*, *85*(2), 275–314. https://doi.org/10.3102/0034654314551063

[2] Alsadoon, R. (2026). Perceptions vs. practices: Academic integrity and actual ChatGPT use among EFL students. *Frontiers in Psychology*, *17*, Article 1796737. https://doi.org/10.3389/fpsyg.2026.1796737

[3] Asadi, M., Ebadi, S., & Mohammadi, L. (2025). The impact of integrating ChatGPT with teachers' feedback on EFL writing skills. *Thinking Skills and Creativity*, *56*, Article 101766. https://doi.org/10.1016/j.tsc.2025.101766

[4] Barrot, J. S. (2023). Using ChatGPT for second language writing: Pitfalls and potentials. *Assessing Writing*, *57*, Article 100745. https://doi.org/10.1016/j.asw.2023.100745

[5] Bhattacharyya, M., Miller, V. M., Bhattacharyya, D., & Miller, L. E. (2023). High rates of fabricated and inaccurate references in ChatGPT-generated medical content. *Cureus*, *15*(5), Article e39238. https://doi.org/10.7759/cureus.39238

[6] Boud, D., & Falchikov, N. (2006). Aligning assessment with long-term learning. *Assessment & Evaluation in Higher Education*, *31*(4), 399–413. https://doi.org/10.1080/02602930600679050

[7] Britt, M. A., & Aglinskas, C. (2002). Improving students' ability to identify and use source information. *Cognition and Instruction*, *20*(4), 485–522. https://doi.org/10.1207/S1532690XCI2004_2

[8] Campbell, M., McKenzie, J. E., Sowden, A., Katikireddi, S. V., Brennan, S. E., Ellis, S., Hartmann-Boyce, J., Ryan, R., Shepperd, S., Thomas, J., Welch, V., & Thomson, H. (2020). Synthesis without meta-analysis (SWiM) in systematic reviews: Reporting guideline. *BMJ*, *368*, l6890. https://doi.org/10.1136/bmj.l6890

[9] Cengiz, B. C., Bilki, Z., Ataş, A. H., & Celik, B. (2025). Exploring second language writers' engagement with ChatGPT feedback: Revision behaviors and perceptions. *System*, *134*, Article 103837. https://doi.org/10.1016/j.system.2025.103837

[10] 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

[11] Conijn, R., Martinez-Maldonado, R., Knight, S., Buckingham Shum, S., Van aes, L., & van Zaanen, M. (2022). How to provide automated feedback on the writing process? A participatory approach to design writing analytics tools. *Computer Assisted Language Learning*, *35*(8), 1838–1868. https://doi.org/10.1080/09588221.2020.1839503

[12] Elmahdi, O. E. H., AbdAlgane, M., & Balla, A. A. S. (2024). AI's Role in EFL: Optimizing Opportunities while Mitigating Risks. *International Journal of Linguistics, Literature and Translation*, *7*(10), 21–37. https://doi.org/10.32996/ijllt.2024.7.10.4

[13] Elmahdi, O. E. H., AbdAlgane, M., Hamid, F. A., Balla, A. A. S., & Ibrahim, I. Z. A. (2025). Integrating critical thinking and technology in Saudi EFL classrooms: A framework for culturally responsive language learning. *International Journal of English Language Studies*, *7*(2), 150–165. https://doi.org/10.32996/ijels.2025.7.2.4

[14] Elmahdi, O. E. H., Daffaallah, N. A. M., Mohamed, H. H. A., Tayebelasmaa, M. E. E., & Babekir, A. H. S. (2026). Pedagogy before platforms: EFL teachers' digital pedagogical practices, learner agency, and contextual constraints in Saudi higher education. *British Journal of Teacher Education and Pedagogy*, *5*(10), 34–46. https://doi.org/10.32996/bjtep.2026.5.10.3

[15] Elmahdi, P. O. E. H., Daffaallah, N. A. M., Balla, A. A. S., Mohamed, H. H. A., & Tayebelasmaa, M. E. E. (2026). Reflective Practice and Generative AI Use in Culturally Grounded Teaching among Saudi EFL Teachers. *Omdurman Islamic University Journal for Human Sciences*, *22*(2), 622–636. https://doi.org/10.52981/oiujhs.v22i2.3998

[16] Facione, P. A. (1990). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction: Research findings and recommendations [The Delphi Report]*. American Philosophical Association. https://eric.ed.gov/?id=ED315423

[17] Grabe, W., & Zhang, C. (2016). Reading–writing relationships in first and second language academic literacy development. *Language Teaching*, *49*(3), 339–355. https://doi.org/10.1017/S0261444816000082

[18] Grinschgl, S., Papenmeier, F., & Meyerhoff, H. S. (2021). Consequences of cognitive offloading: Boosting performance but diminishing memory. *Quarterly Journal of Experimental Psychology*, *74*(9), 1477–1496. https://doi.org/10.1177/17470218211008060

[19] Hago Elmahdi, O. E., Allehyani, F. S., & AbdAlgane, M. (2025). Exploring the Ethical Implications and Cultural Considerations in Using AI Technologies for EFL Education in the Saudi Context. *Nsukka Journal of the Humanities*, *33*(2).

[20] Hong, Q. N., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M.-P., Griffiths, F., Nicolau, B., O'Cathain, A., Rousseau, M.-C., Vedel, I., & Pluye, P. (2018). The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. *Education for Information*, *34*(4), 285–291. https://doi.org/10.3233/EFI-180221

[21] Huang, Y., & Wang, D. (2025). Can ChatGPT serve as a writing collaborator? Insights from Chinese EFL learners. *System*, *133*, Article 103775. https://doi.org/10.1016/j.system.2025.103775

[22] Hwang, M., Jeens, R., & Lee, H.-K. (2025). Exploring learner prompting behavior and its effect on ChatGPT-assisted English writing revision. *The Asia-Pacific Education Researcher*, *34*(3), 1157–1167. https://doi.org/10.1007/s40299-024-00930-6

[23] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, *55*(12), 1–38. https://doi.org/10.1145/3571730

[24] Lea, M. R., & Street, B. V. (1998). Student writing in higher education: An academic literacies approach. *Studies in Higher Education*, *23*(2), 157–172. https://doi.org/10.1080/03075079812331380364

[25] Leijten, M., Van Waes, L., Schrijver, I., Bernolet, S., & Vangehuchten, L. (2019). Mapping master's students' use of external sources in source-based writing in L1 and L2. *Studies in Second Language Acquisition*, *41*(3), 555–582. https://doi.org/10.1017/S0272263119000251

[26] Liang, W., & Wu, Y. (2024). Exploring the use of ChatGPT to foster EFL learners' critical thinking skills from a post-humanist perspective. *Thinking Skills and Creativity*, *54*, Article 101645. https://doi.org/10.1016/j.tsc.2024.101645

[27] 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

[28] Mahapatra, S. (2024). Impact of ChatGPT on ESL students' academic writing skills: A mixed methods intervention study. *Smart Learning Environments*, *11*, Article 9. https://doi.org/10.1186/s40561-024-00295-9

[29] McGrew, S., Breakstone, J., Ortega, T., Smith, M., & Wineburg, S. (2018). Can students evaluate online sources? Learning from assessments of civic online reasoning. *Theory & Research in Social Education*, *46*(2), 165–193. https://doi.org/10.1080/00933104.2017.1416320

[30] McNamara, D. S., Crossley, S. A., & McCarthy, M. P. (2010). Linguistic features of writing quality. *Written Communication*, *27*(1), 57–86. https://doi.org/10.1177/0741088309351547

[31] Molenaar, I. (2022). Towards hybrid human–AI learning technologies. *European Journal of Education*, *57*(4), 632–645. https://doi.org/10.1111/ejed.12527

[32] Moorhouse, B. L., Wan, Y., Wu, C., Wu, M., & Ho, T. Y. (2025). Generative AI tools and empowerment in L2 academic writing. *System*, *133*, Article 103779. https://doi.org/10.1016/j.system.2025.103779

[33] Muñoz Muñoz, B. C., Nassaji, H., & Bello Carrillo, F. I. (2025). ChatGPT-generated versus human direct corrective feedback on L2 writing. *System*, *134*, Article 103805. https://doi.org/10.1016/j.system.2025.103805

[34] 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

[35] Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*, n71. https://doi.org/10.1136/bmj.n71

[36] Perkins, M., Furze, L., Roe, J., & MacVaugh, J. (2024). The Artificial Intelligence Assessment Scale (AIAS): A framework for ethical integration of generative AI in educational assessment. *Journal of University Teaching and Learning Practice*, *21*(6). https://doi.org/10.53761/q3azde36

[37] Plakans, L., & Gebril, A. (2012). A close investigation into source use in integrated second language writing tasks. *Assessing Writing*, *17*(1), 18–34. https://doi.org/10.1016/j.asw.2011.09.002

[38] Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences*, *20*(9), 676–688. https://doi.org/10.1016/j.tics.2016.07.002

[39] Shen, Y., & Chen, L. (2025). *"Critical chatting" or "casual cheating": How graduate EFL students utilize ChatGPT for academic writing. Computer Assisted Language Learning*. Advance online publication. https://doi.org/10.1080/09588221.2025.2479141

[40] Song, C., & Song, Y. (2023). Enhancing academic writing skills and motivation: Assessing the efficacy of ChatGPT in AI-assisted language learning for EFL students. *Frontiers in Psychology*, *14*, Article 1260843. https://doi.org/10.3389/fpsyg.2023.1260843

[41] Song, C., & Song, Y. (2026). Corrigendum: Enhancing academic writing skills and motivation: Assessing the efficacy of ChatGPT in AI-assisted language learning for EFL students. *Frontiers in Psychology*, *17*, Article 1917101. https://doi.org/10.3389/fpsyg.2026.1917101

[42] Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. *Science*, *333*(6043), 776–778. https://doi.org/10.1126/science.1207745

[43] Swiecki, Z., Khosravi, H., Chen, G., Martinez-Maldonado, R., Lodge, J. M., Milligan, S., Selwyn, N., & Gašević, D. (2022). Assessment in the age of artificial intelligence. *Computers and Education: Artificial Intelligence*, *3*, Article 100075. https://doi.org/10.1016/j.caeai.2022.100075

[44] Tran, T. T. T. (2025). Enhancing EFL writing revision practices: The impact of AI- and teacher-generated feedback and their sequences. *Education Sciences*, *15*(2), Article 232. https://doi.org/10.3390/educsci15020232

[45] Tsai, C.-Y., Lin, Y.-T., & Brown, I. K. (2024). Impacts of ChatGPT-assisted writing for EFL English majors: Feasibility and challenges. *Education and Information Technologies*, *29*(17), 22427–22445. https://doi.org/10.1007/s10639-024-12722-y

[46] UNESCO. (2023). *Guidance for generative AI in education and research*. https://unesdoc.unesco.org/ark:/48223/pf0000386693

[47] 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

[48] Wang, Y. (2024). Cognitive and sociocultural dynamics of self-regulated use of machine translation and generative AI tools in academic EFL writing. *System*, *126*, Article 103505. https://doi.org/10.1016/j.system.2024.103505

[49] Warschauer, M., Tseng, W., Yim, S., Webster, T., Jacob, S., Du, Q., & Tate, T. (2023). The affordances and contradictions of AI-generated text for writers of English as a second or foreign language. *Journal of Second Language Writing*, *62*, Article 101071. https://doi.org/10.1016/j.jslw.2023.101071

[50] Wineburg, S., & McGrew, S. (2019). Lateral reading and the nature of expertise: Reading less and learning more when evaluating digital information. *Teachers College Record*, *121*(11), 1–40. https://doi.org/10.1177/016146811912101102

[51] Wingate, U. (2012). "Argument!" Helping students understand what essay writing is about. *Journal of English for Academic Purposes*, *11*(2), 145–154. https://doi.org/10.1016/j.jeap.2011.11.001

[52] Yan, D. (2023). Impact of ChatGPT on learners in a L2 writing practicum: An exploratory investigation. *Education and Information Technologies*, *28*(11), 13943–13967. https://doi.org/10.1007/s10639-023-11742-4

[53] Yan, D., & Zhang, S. (2024). L2 writer engagement with automated written corrective feedback provided by ChatGPT: A mixed-method multiple case study. *Humanities and Social Sciences Communications*, *11*, Article 1086. https://doi.org/10.1057/s41599-024-03543-y

[54] Yin, X., Dou, K., & Han, J. (2025). An AI-assisted critical thinking intervention to enhance undergraduate EFL learners' writing proficiency. *Studies in Educational Evaluation*, *86*, Article 101480. https://doi.org/10.1016/j.stueduc.2025.101480

[55] Zare, J., Ranjbaran Madiseh, F., & Derakhshan, A. (2025). *Generative AI and English essay writing: Exploring the role of ChatGPT in enhancing learners' task engagement. Applied Linguistics*. Advance online publication. https://doi.org/10.1093/applin/amaf045

[56] Zhang, R., Zou, D., Xie, H., & Wang, F. L. (2025). Engagements with GPT responses and learner prompts in ChatGPT-based learning of English argumentative writing logic and their impacts. *Computers and Education: Artificial Intelligence*, *9*, Article 100489. https://doi.org/10.1016/j.caeai.2025.100489

[57] Zou, M., & Huang, L. (2024). The impact of ChatGPT on L2 writing and expected responses: Voice from doctoral students. *Education and Information Technologies*, *29*(11), 13201–13219. https://doi.org/10.1007/s10639-023-12397-x