International Journal of Linguistics, Literature and Translation
ISSN: 2617-0299 (Online); ISSN: 2708-0099 (Print)
DOI: 10.32996/ijllt
Journal Homepage: www.al-kindipublisher.com/index.php/ijllt

IJLLT
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Managing AI reliability in L2 writing: Designing a systematic framework

Saad Aljebreen
Department of English Language and Literature, College of Languages and Humanities, Qassim University, Saudi Arabia
Corresponding Author: Saad Aljebreen, **E-mail**: smtierie@qu.edu.sa

| ABSTRACT

Despite the widespread institutionalized application of AI tools as learning aids, the onus to manage and check the reliability of AI output, especially in writing, lies with the learners. This study conceptualizes reliability management in AI-assisted L2 writing as epistemic trust calibration. Based on this conceptualization, L2 writers need to continually evaluate the reliability of AI output for a given purpose, helping them decide whether to accept, revise, verify or reject such output. Thus, this paper proposes an epistemic trust calibration framework which integrates five theoretical constructs through a trust checking loop linking AI suggestions to writer judgment and outcomes. The framework also provides three operational tools: a risk taxonomy that identifies lower and high-risk zones, a five-step verification routine for high-risk AI output, and a reliability management rubric that operationalizes calibration as a developmental competence. This framework can be utilized to show how AI can boost learning when writers calibrate trust and verify high risk AI output. The study concludes with pertinent recommendations for future research directions.

| KEYWORDS

AI-assisted L2 writing; reliability management; generative AI; Critical AI literacy

| ARTICLE INFORMATION

ACCEPTED: 30 April 2026 **PUBLISHED:** 14 May 2026 **DOI:** 10.32996/ijllt.2026.5.1.10

Introduction

Generative artificial intelligence (AI) tools such as ChatGPT are being increasingly integrated into the education process at higher education institutions given their potential to improve second language (L2) writing by assisting learners to brainstorm ideas, write drafts, make text revisions, and get immediate feedback. However, improving text quality with AI assistance does not necessarily lead to permanent writing development. The challenge is not whether learners rely on using AI output during the writing process but rather how they manage the reliability of AI output during the writing process such as when making grammatical choices, taking rhetorical decisions, or supporting claims with evidence.

Research shows that learners rely heavily on AI during the early stages of the writing process, especially for brainstorming, and later they use it for refinement (Hwang et al., 2025). Moreover, there is significant variation in the way L2 writers engage with generative AI feedback and how they handle it in revision (Cengiz et al., 2025; Yan & Zhang, 2024). At least one study has concluded that writer engagement and trust reflect whether feedback enhances learning more than surface editing of text (Ranalli, 2021). Whatever the purpose and outcome of the integration of AI into the education process, it has opened the doors for deeper engagement with feedback allowing for further scrutiny of evidence and writing tools that go beyond error correction.

Feedback in the age of AI is different from the teacher's feedback in the sense that L2 learners view the teacher as the expert and take their feedback for granted. With the advent of AI and its use in L2 learning and teaching the concept of feedback needs redefinition. The feedback that L2 learners receive from AI may not be taken for granted but rather should be minutely scrutinized. As per studies and even the disclaimer of some of these tools, AI sometimes hallucinates and makes up content which may not necessarily be true, valid, or accurate (Chelli et al., 2024; Watson, 2024).

At the same time, many studies shed light on the affordances of AI tools, their impact on performance outcomes, or users' attitudes towards them (Aljabr & Al-Ahdal, 2024; Alsaweed & Aljebreen, 2024; Cengiz et al., 2025; Hwang et al., 2025; Moorhouse et al., 2025; Shi et al., 2025; Yan & Zhang, 2024; Yeung, 2025; Zhou & Wang, 2026). These findings highlight the need to investigate whether AI tools can lead to deep learning or only superficial skill enhancement, which is a perceptible gap in available literature which this study aims to fill. It proposes an epistemic trust calibration framework which integrates five theoretical constructs through a trust checking loop linking AI suggestions to writer judgment and outcomes along with three operational tools to identify low and high-risk zones, steps for validating high-risk AI output, and a reliability management rubric that operationalizes calibration as a developmental competence.

Theoretical foundations:

Epistemic vigilance

Epistemic vigilance theory suggests that there is always the risk of misinformation during the communication process and to rule this out humans have evolved cognitive mechanisms (Sperber et al., 2010). In the case of the communication of L2 learners with AI tools, the other 'interlocutor' is an algorithm which might be substantially fluent in language but not necessarily reliable. This places on the L2 user the responsibility to meticulously verify AI output.

Trust and appropriate reliance

Research indicates that overtrust on AI output may lead to decreased monitoring which in turn may cause error propagation, on the other hand undertrust limits user gains from valid AI assistance (Lee & See, 2004). The framework proposed in this paper carries out this logic to L2 writing decisions by treating trust as a dynamic, contingent, and task-specific process. This process is enacted by L2 writers' choices to accept, revise, verify, or reject AI output during AI-assisted writing tasks. These steps are followed by a step involving documenting and reflecting which also tracks calibration over time.

Critical AI literacy and agency in L2 writing

Recent research on L2 writing suggests that critical AI literacy evolves around awareness, positionality, strategic interaction, and evaluation of AI affordances (Wang & Wang, 2025). The framework proposed in this paper i.e., the reliability-management framework builds on this conceptualization by specifying the micro-decision mechanism that L2 writers can use to turn critical practices into text and learning outcomes, building on the foundational work by Ng et al. (2021). In an effort to define AI literacy, they conceptualize the concept of AI literacy as having four aspects that need to be considered: "*know and understand, use and apply, evaluate and create, and ethical issues*." However, before users learn to apply the framework proposed here, it is imperative for L2 learners to develop digital literacy to attain adequate awareness of the best practices of using AI tools for writing, especially concerning the maintenance of their voice and agency. This resonates with Darvin (2025) which argues that L2 learners need critical digital literacies which will allow them to see how AI tools favor certain ways of thinking and writing. This could lead L2 learners to take for granted AI's output and lose their own voice and stance.

Literature review

AI has great potential to enhance all language skills, especially writing. L2 learners can use AI at all stages of the writing process. Research shows that writers use AI more frequently in the early stages when they compose, ideate, and plan their writing and later, when they refine their writing (Hwang et al., 2025). Regarding the framework suggested in this paper, the early use of AI in the writing process can affect the selection and refinement of the topic, the arguments that are embedded in the text and the plausibility of the claims used in the piece of writing. Thus, the framework can give writers insight into calibration needed in each stage of the writing process. For example, at the ideation stage vigilance is needed against meaning drift or inappropriate stance of the AI model where the framework can be utilized by users to stay on track.

Research also shows that writers engage with ChatGPT feedback in various ways in terms of revision behaviors and perceptions (Cengiz et al., 2025). This suggests that the impact of AI feedback depends on how L2 learners interpret and integrate it in the process and product (Cengiz et al., 2025; Yan & Zhang, 2024). This perspective aligns with Automated Writing Evaluation (AWE) research which shows that engagement is mediated by perceived accuracy and trust, and whether writers would treat feedback as a learning resource or as a shortcut to producing error-free text (Ranalli, 2021). The framework suggested in this study

reinterprets the variation in engaging with AI as differences are natural in users' calibration competence. L2 writers who justify and verify their decisions would likely turn feedback into learning compared to those who accept AI output without much verification.

It is worth noting that studies that compared AI generated feedback with human feedback showed that AI generated direct corrective feedback might be effective and sometimes better than human direct corrective feedback in some dimensions (Muñoz et al., 2025). However, even if AI feedback enhances performance, reliability does not appear to be uniform across feedback types. This makes the framework suggested in this paper of substantial value since it specifies reliability boundaries and provides a roadmap to gauge the credibility of AI output and ways of dealing with that output. The framework suggests that writers should not deal with AI feedback as a single stable entity but as a dynamic construct that needs to be constantly analyzed and verified.

From a different perspective, comparative research has shown that Large Language Models when compared to classic AWE systems, are able to provide discourse level rewrites of the text and can also generate model texts; this can have an impact on rhetorical organization, voice, and stance (Lu & Zeng, 2025). Experimental research also showed that LLMs like ChatGPT can lead to improvement in writing performance as well as could affect motivational constructs such as the ideal L2 writing self (Shi et al., 2025). Such findings support the claim that the framework presented in this paper puts forward, that is, managing reliability of AI output can also enhance agency preservation and independence for L2 writers.

One of the challenges of using AI for writing is that it sometimes generates imaginary or inaccurate references (termed 'hallucination') (Chelli et al., 2024). This can be a dilemma especially in academic writing where writers cite references to support or build arguments. Some research has shown that academic work submitted by students sometimes includes AI generated sources (Watson, 2024). This can have serious repercussions in academic writing as it can lead to proliferation of fake sources and their subsequent circulation in academics. These findings highlight the need for a verification-first approach like the one this study suggests. The research cited above could also lay the foundation for the risk taxonomy that this paper proposes which treats citations, claims, and paraphrase integrity as high-risk zones.

Buçinca et al. (2021) investigated why users frequently over-rely on AI decision support tools in a controlled experiment and concluded that cognitive forcing interventions significantly reduced overreliance on AI explanations. By cognitive forcing interventions they mean designs that force users to engage analytically with AI output. They indicated that simply adding explanations to AI output does not reduce overreliance. Using dual-process theory, they argue that users do not normally engage analytically with AI recommendations. They believe that users develop general heuristics about when to follow AI suggestions.

The above overview of studies shows that the crucial aspect is not whether AI can help improve or retard L2 writers' skills. The key point is that the outcomes of using AI to help L2 learners depend on their judgement and evaluation of AI output. It boils down to how they judge the reliability of AI output and verify it which points to the need for a theoretical framework that can potentially assist L2 learners in engaging with AI output, evaluating evidence, and ensuring the veracity of the produced content.

Designing a framework for managing reliability in AI-assisted L2 writing

Key constructs of the framework

The suggested framework is built on five theoretical constructs that L2 learners need to develop calibration competence: (a) task-specific epistemic trust, (b) calibration accuracy i.e. appropriate reliance, (c) verification repertoire, (d) decision transparency, and (e) agency preservation. These constructs are operationalized through four action choices that L2 writers enact when they deal with AI output (accept, verify, revise, reject), followed by a documentation and reflection step to track calibration over time. The framework also provides three operational tools: a risk taxonomy that identifies low and high-risk zones, a five-step verification procedure for high-risk AI output, and a reliability management rubric. The framework brings together in its design epistemic vigilance (Sperber et al., 2010), appropriate reliance in automation (Lee & See, 2004), and critical AI literacy in L2 writing (Wang & Wang, 2025).

By task-specific epistemic trust we imply the L2 learner's willingness to view AI output as dependable and therefore aid their writing exercise. Calibration accuracy means the alignment between learners' trust and the actual reliability of the output. Verification repertoire includes the different epistemic moves that L2 writers can use such as, contrasting outputs, consulting external resources, and triangulating claims. Decision transparency refers to the ability to explain, justify, and document AI-related decisions. Finally, agency preservation refers to L2 learners' ability to maintain authority and control over claims, meaning, and evidence rather than depending solely on AI assistance.

The writing loop

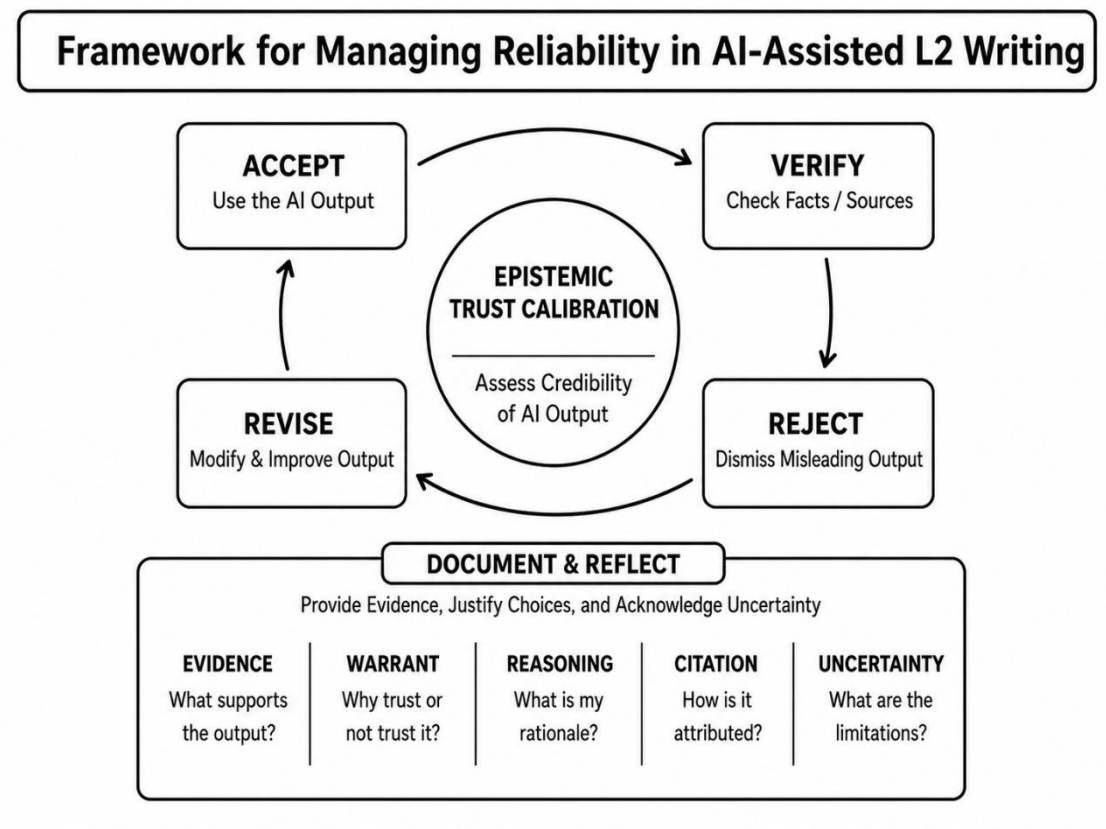

Figure 1. Framework for Managing Reliability in AI-Assisted L2 Writing

Figure 1 shows the writing loop suggested by the framework. It starts with the learners' writing goal and leads to outcomes: Writing goal → AI output → trust judgment → action choice (accept / revise / reject / verify) → document and reflect → outcomes (text quality and learning). The suggested framework considers L2 learners' choice of action as the starting point where calibration becomes observable and teachable. For instance, two learners might have the same AI output, but one might verify a claim and adapt their language to maintain their stance, while the other might unquestioningly accept the AI feedback. This process of orchestrating the way learners treat and react to AI feedback and output can be taught explicitly to them so that they become aware of the different choices at their disposal to make their writing better. The framework can be regarded as a cognitive forcing mechanism (Buçinca et al., 2021) that forces learners to engage analytically with AI-output rather than accepting it due to its fluency. The *document and reflect* step of the framework attempts to compel L2 writers to articulate evidence, warrant reasoning, and cross-check citations before trusting and incorporating AI output into their writing.

Risk taxonomy

The framework suggests that calibration should be risk-sensitive since AI reliability is not stable across different writing functions. Thus, there are low-risk zones which include spelling errors, grammatical errors, and phrasing alternatives. On the other hand, there are high-risk zones which include factual claims, citations, references, paraphrasing integrity, stance and argument logic. In these latter areas, fluent output can sometimes hide epistemic weaknesses (Chelli et al., 2024; Watson, 2024).

AI output brings with it some concerns and risks for L2 learners such as the extent to which they trust that output. AI can sometimes be wrong and make mistakes; it is known to hallucinate at times. The framework suggests that the risk taxonomy turns these concerns into actionable classroom routines. For instance, in the case of a high-risk zone task, verification and decision documentation should take place and in the case of a low-risk zone the emphasis should be on meaning preservation and appropriateness checks. Since academic writing can sometimes be an uneasy task for L2 learners, clarifying and layering risks for them can help reduce overreliance on AI and support their agency (Moorhouse et al., 2025).

Extending current models

The suggested framework builds on and extends two established lines of research efforts. The framework builds on Lee and See's (2004) theory of trust in automation. It borrows from their model the core assumption that trust needs to be calibrated and that overtrust and undertrust are regarded as failure modes. The key difference is that Lee and See's framework was developed for man-machine interaction and was not built for textual, rhetorical, or developmental aspects of writing. The framework proposed in this paper also builds on Wang and Wang's APSE model. It shares a similar commitment to critical AI literacy in L2 writing that includes considering awareness, strategy, positionality, and evaluation. However, the APSE model focuses on describing competencies rather than specifying procedures. It also does not explain how learners can differentiate between AI outputs of varying risks, nor accounts for the development of learners' calibration competence over time. These factors are accounted for in the proposed framework.

Following from the above discussion, the proposed framework contributes three operational elements that these two models do not account for. First, it introduces a risk taxonomy that distinguishes low risk zones from high risk zones. Second, it introduces a specific system for verification, which is a five step verification procedure for high risk AI output: (1) Validate Existence, (2) Exact Match, (3) Relevance, (4) Integrity, and (5) Footprint (see Figure 2 for details). Third, the framework provides a reliability management rubric that operationalizes calibration as a developmental concept across various levels. This contribution extends the above two models and turns the concepts of trust calibration and critical AI literacy into a framework for L2 writing that can be taught and assessed in L2 classrooms.

The framework suggested in this paper complements other models such as critical AI literacy frameworks like the APSE model suggested by Wang & Wang (2025). Moreover, the suggested framework pinpoints and specifies the micro-decision mechanism that connects literacy to outcomes. Thus, it can be argued that some components of the APSE model – 'critical strategies' and 'critical evaluation' can help expand the L2 writers' verification repertoire and improve calibration accuracy (Wang & Wang, 2025).

Implications

It is speculated that calibration accuracy of AI output might be a good predictor for learning outcomes compared to the frequency of using AI for writing. L2 learners use AI tools in various ways during the writing process. So, the frequency of using AI to prepare for writing or to polish the product does not show the instances when better learning could happen compared to mere copying of the AI output to finish the writing task. In L2 writing the ideation and polishing stages have different epistemic risks depending on the learner's trust judgment and action choices which will give different outcomes (Hwang et al., 2025; Ranalli, 2021). Thus, in AI-assisted writing, educators should investigate the distribution of learners' actions (accept/ revise /reject /verify) and the documentation they engage in during these actions, and not rely solely on the frequency of AI use while writing.

It is also worth noting that overtrust of AI output for L2 writing can badly affect the writing outcome in high-risk zones (claims/citations/stance). In such cases, students cannot solely depend on the fluency of the AI system since AI sometimes produces hallucinated or inaccurate citations and references which may get masked in the tool's fluent language, misleading the L2 learner, even leading them to believe that since the output is fluent it can be true and end up using these fabricated resources. When there is no access to disciplinary plausibility cues in L2 writing this problem would potentially happen (Chelli et al., 2024; Watson, 2024; Moorhouse et al., 2025).

However, it should also be noted that undertrust can affect writers' learning since it can limit the uptake of helpful feedback and revision from AI tools. It seems that L2 learners' perceived trust can influence the extent to which they engage effectively with AI feedback. If learners take a distrust stance toward all the AI output that they engage with, it will prevent them from using the scaffolding for self-editing (Ranalli, 2021). For L2 writing a good practice would be to frame reliability management as a calibrated procedure and use guided activities to compare AI output with external resources which will also enhance selective reliance.

Previous research has shown that engagement with ChatGPT feedback reflected various revision behaviors in L2 writers (Cengiz et al., 2025). The model suggested by this paper predicts that learners who have a wider repertoire of verification behaviors will be less likely to use incorrect or inappropriate ChatGPT suggestions and will integrate feedback strategically (Cengiz et al., 2025; Yan & Zhang, 2024; Wang & Wang, 2025). The best practice would be to code verification moves (dictionary checks, contrast prompting, source triangulation) and check association with revision quality and delayed transfer.

Studies investigating L2 writers' critical AI literacy in AI-assisted writing have shown the importance of strategic interaction and evaluation when using AI models for writing. For example, Wang & Wang (2025) indicated that learners should seek alternatives and rationales that allow them to compare output and detect inconsistencies and keep better alignment between reliance and

reliability. Thus, the implication here is that learners should go beyond a single prompt when using AI models and opt to use more structured prompting mechanisms that should include alternatives, rationales and limits. It would be advisable that learners reflect on why choose certain output suggestions.

It is also worth noting that when learners write they should think about how AI can affect their arguments. Research shows that AI can influence content and argument trajectories. If L2 learners use and reuse shallow ideas from AI output without evaluation, it could lead them to end up with a weak product in need of polishing instead of focusing on enhancing the arguments the content attempts to put forward (Hwang et al., 2025). For this reason, the proposed framework has steps for verification and checkpoints to evaluate counterarguments, trace missing evidence, and justify whether AI output aligns with their arguments in the writing task at hand.

From another perspective, research on AI literacy pedagogy has shown the importance of reflective practice (Hakim, 2025). When L2 writers are required to justify their decisions, whether they accept or reject AI output when writing, they keep their ownership of meaning and develop metacognitive control which would hopefully help them progress beyond AI-supported contexts. The implication of this for L2 educators and learners is that they can include decision notes along with drafts so that there is some kind of transparency which can be rewarded as part of the classroom practice. They will expose to teachers the decision processes that learners go through to select the appropriate AI output relevant to their writing task since this process is usually hidden and teachers do not see how learners reached their decisions.

One of the features of AI models is that they provide discourse-level rewrites and can also provide model texts that reshape learners' rhetorical organization and voice which can impact the quality of their writing. This in turn will broaden the reliability judgements they have to make (Lu & Zeng, 2025; Shi et al., 2025; Ranalli, 2021). This entails that learners need to understand the difference between low-risk editing support and high-risk rhetorical and evidential decisions that could affect their stance and voice. In other words, learners need to be trained in how to preserve their voice and stance when writing with the assistance of AI tools.

From an empowerment perspective, using AI tools for writing assistance can be enriching for learners. However, research has shown that this empowerment might lack critical awareness, which can lead to unreflective adoption of AI output (Moorhouse et al., 2025). This shows the importance of encouraging educators to make reliability management an explicit learning objective and educating their learners about the importance of verification of AI output and not blindly trusting it. This also shows that false empowerment can make learners trust whatever output is generated by AI models given the fluency of most models. Empirical evidence shows that it is of utmost importance that learners locate and validate sources before citing them (Chelli et al., 2024; Watson, 2024). To add data to these claims, Walters & Wilder (2023) found that GPT-4 fabricates citations at a rate of 18%. The conclusion we draw here is that the process of verifying sources should be explicit and learners should be guided on how to search databases and verify the sources.

From another perspective, learners can be taught certain practices that support verification of AI output. The framework suggested by this paper shows that focusing on identifying risks and using the reliability management rubric might increase verification behaviors by learners and allow them to preserve their agency over time (Hakim, 2025; Han, 2024; Wang & Wang, 2025). This is in line with what Darvin (2025) argues for in terms of developing a critical perspective so that L2 learners become aware of how AI output shapes the way they think and write. Educators can, therefore, evaluate and gauge the use of the reliability management rubric before and after training learners on how to use it to develop their verification behaviors to reflect the change in learners' practices and better outcomes of AI-assisted writing.

Pedagogical considerations

The framework that this paper proposes reconsiders AI-assisted writing pedagogy as verification-aware authorship. L2 learners can use AI tools for writing support but they should also retain accountability for meaning and evidence. This potentially could lead to a shift in practice so that there will be explicit instruction of reliability management of AI output. That is, learners need to know when to accept output, and when to verify it first. This is supported by research on AI literacy which encourages the use of practical repertoires and reflective norms (Hakim, 2025).

In practice, learners can apply the proposed framework through three pedagogical moves. The first move would be to check whether the task is low-risk or high-risk before they seek AI assistance. If the task at hand is high-risk then they need to go for verification and documentation of the AI output. The second move would be to ask AI for alternatives and rationales. They should compare outputs and check for limitations. The third move is that learners locate and validate sources before accepting citations claimed by the AI tool. This is especially needed in evidence-linked tasks.

System for high-risk zones

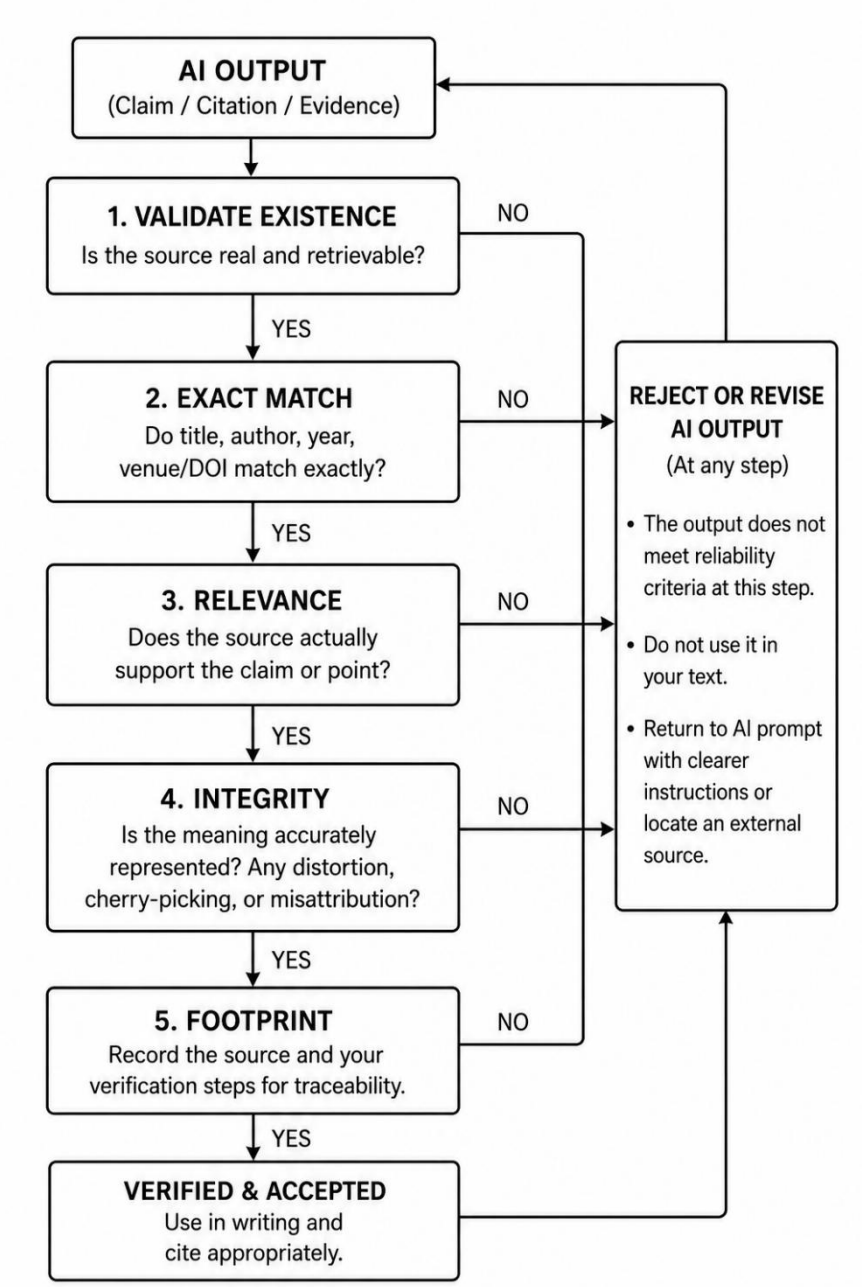

Figure 2. System for high-risk zones

This system is a manifestation of the 'corroborate' element in the AI literacy framework suggested by Warschauer et al. (2023). It asserts that L2 writers need to corroborate the accuracy of AI output during the writing and revision processes. Empirical research has shown that source evaluation must be conducted outside the source itself rather than check from within (Wineburg & McGrew, 2019). They have shown that expert fact checkers reach accurate levels of credibility judgements by reading laterally while non-experts read vertically and are misled by surface features. The first three steps of the above routine adopt this principle and apply it to AI output. Instead of only judging the AI generated citation by looking at its claim, L2 learners need to check the existence of the sources outside the AI tool that generated the citation. Additionally, McGrew et al. (2018) have shown that students at all levels perform poorly at evaluating online sources when using only their own judgement or surface features, highlighting the need to explicitly teach them the verification procedure that can help make better judgement. Moreover, Brodsky et al. (2021) have shown that verification competence can be developed via explicit instruction using an online module on lateral reading which improved students' use of lateral reading strategies from 12.2% to 52.5%. The reliability management

rubric proposed in this paper attempts to implement the same concept to AI-assisted writing by viewing calibration accuracy as a developmental competence.

Assessment rubric for reliability management

The five constructs of the framework proposed in this study are operationalized as a developmental rubric (Table 1) designed for use as a research instrument and as a pedagogical assessment tool. It aims to make calibration competence visible and traceable during writing sessions. In AI-assisted writing, learners' writing behaviors are not similar but actually vary across individuals and tasks. This entails that assessment should consider calibration competence of the learners rather than how they use AI tools for writing. Mere use of AI tools does not clearly uncover the details of how learners deal with AI output in their writing. The reliability management rubric shown in Table 1 attempts to operationalize the proposed framework across five dimensions: Task-specific epistemic trust, calibration accuracy, verification repertoire, decision transparency, agency preservation. Assessment of these dimensions is supported by evidence from research that shows that engagement and trust affect learning from automated feedback (Ranalli, 202)

The four rubric levels track the calibration accuracy dimension which is the degree of alignment between L2 writer's trust judgement and the actual reliability of AI output across risk zones and task types. This dimension is built on Lee and See's (2004) overview of trust in automation. They showed that the central failure modes are overtrust and undertrust in relation to the system's actual reliability. The rubric below expands this reasoning to L2 writing contexts by specifying how calibration might look like when the system is 'generative AI' and the judgement is whether to accept, revise, verify, or reject the AI output.

The four levels of the rubric reflect a scale to measure calibration accuracy. Each level represents the relationship between the learner's judgement and the AI output. At Level 1 – Uncalibrated L2 writers show light engagement or systematic alignment between trust and reliability. At Level 2 – Developing, L2 writers start to differentiate trust depending on the output type. They can correctly distrust output with errors but still can easily overtrust fluent output. At Level 3 – Calibrated, L2 writers seek evidence and identify risks and scrutinize AI output by carefully checking citations, claims, and stance agreement. At Level 4 – Strategic, writers plan and anticipate things. They design their prompts carefully before they generate AI output. They continually monitor their trust judgments throughout the writing process.

The five dimensions of the rubric: *task-specific epistemic trust, calibration accuracy, verification repertoire, decision transparency, and agency preservation* reflect the overall calibration accuracy. Each dimension shows an aspect of the learner's behavior. First, *task-specific epistemic trust* shows how trust judgments vary across risk zones and tasks. Second, *calibration accuracy* shows how these judgements reflect the actual reliability of AI output. Third, *verification repertoire* shows the checks that the L2 writers perform to test their judgements. Fourth, *decision transparency*, shows how L2 writers articulate their reasoning when they make their choices (accept, revise, reject). Fifth, *agency preservation* reflects whose reasoning and voice influence the final text that learners end up with. Since the rubric has varying levels, a writer might be calibrated on a certain dimension to the exclusion of the others. This will visibly show a useful representation of their calibration accuracy.

Level	Task-specific epistemic trust	Calibration accuracy	Verification repertoire	Decision transparency	Agency preservation
1 — Uncalibrated	Treats all AI output the same. Trust is fixed (uniformly high or uniformly low) and does not shift between low-risk tasks (e.g., spelling) and high-risk tasks (e.g., citations, claims).	Trust judgments rarely match actual reliability. The writer accepts inaccurate output and rejects accurate output at similar rates; errors are not concentrated in any predictable zone.	No verification, or verification limited to re-reading for fluency. AI suggestions are accepted or rejected without external checks.	Accept/revise/reject decisions are unstated, intuitive, or justified only by surface features ("it sounds right").	Final text reproduces AI wording; the writer's stance, structure, and lexical choices are not visibly distinguishable from the AI's.
2 — Developing	Trust varies between low-risk and high-risk tasks, but the boundary is misplaced — the writer often treats content-level output (claims, evidence) as low-risk because it sounds confident.	Trust matches reliability for surface-level output (grammar, spelling) but not for content-level output. Errors cluster in claims, citations, and stance, where the writer continues to overtrust fluent text.	Verification occurs but is irregular and limited to one strategy (e.g., a single search). Checks are performed when the writer happens to notice something odd, not systematically by risk zone.	Some accept/revise/reject decisions are justified, but rationales are partial, generated only when prompted, and often constructed after the fact rather than during decision-making.	Some sections are reworded into the writer's voice; others remain in AI phrasing. The writer's stance is present but unevenly maintained across the text.
3 — Calibrated	Trust varies systematically by risk zone. The writer applies stricter scrutiny to citations, evidence, and stance than to spelling and phrasing, and can articulate why.	Trust matches reliability across most output types, including content-level output. Remaining errors are isolated rather than systematic; the writer correctly identifies the high-risk zones in their own work.	Verification is routine for high-risk output and uses at least two strategies (e.g., database search + cross-prompt comparison). Citations are checked for existence and accuracy as a default practice.	All consequential accept/revise/reject decisions are justified during the writing process, with task-relevant reasoning that names the evidence considered.	The writer's voice and argument structure dominate the final text. AI output is adapted, reordered, or paraphrased to fit the writer's purpose; no extended passages remain in unmodified AI phrasing.
4 — Strategic	Trust requirements are anticipated before generating output. The writer designs prompts, selects tools, and plans verification steps based on the risk zone of the task at hand.	Trust closely matches reliability across diverse and unfamiliar tasks. The writer monitors their own trust judgments and revises them when discrepancies appear, including catching their own miscalibrations mid-task.	Verification is systematic and triangulated: existence, accuracy, relevance, and integrity of meaning are checked across multiple sources, with explicit protocols for citation handling.	Decisions are documented in real time with named evidence, articulated reasoning, and stated uncertainty. The writer can reconstruct why each AI contribution was accepted, revised, or rejected.	AI is used as a subordinate tool. The writer originates the argument, selects what to retain, and treats AI suggestions as candidate material rather than as text. Authorial accountability is explicit and consistent.

Table 1. Reliability-Management Rubric for AI-assisted L2 writing

Scores from the rubric could be analyzed along with learners' process data such as chat logs and drafts to gain better understanding of calibration and check if improvements in calibration could predict transfer of learning to writing tasks which are not supported by AI (Hwang et al., 2025; Ranalli, 2021).

Conclusion

This paper proposes a framework for reliability management in AI-assisted L2 writing. The framework can potentially lead to better learning when L2 learners calibrate their trust of AI output and spot high-risk AI outputs. This in turn can provide L2 learners with a principled way on how to treat AI output and make them think carefully rather than use the output blindly. It is hoped that the proposed framework will provide L2 learners with a disciplined approach that carefully allows them to finetune and verify AI output during their writing tasks. The paper attempts to operationalize the verification of AI output in L2 writing as a trackable and teachable system by providing a concrete sequence of observable behaviors as shown in the verification system.

Recommendations

Based on the framework proposed in this study, it is imperative for institutions and educators to sensitize the learners to the potential of AI tools to mislead them in their learning journey. Educators themselves need in-service training and support in robust use of AI in learning while keeping its risks at bay. AI use as an educational aid needs to be institutionalized since this will ensure checks and measures are in place to ensure healthy and optimum use of the same.

Limitations and direction for future research

This paper proposes a reliability management framework purely on theoretical basis. This shows that the next step would be to investigate the proposed framework empirically. A validation study could examine whether the rubric reliably shows differences among L2 writers at different calibration levels.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.
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]. Aljabr, F. S., & Al-Ahdal, A. A. M. H. (2024). Ethical and pedagogical implications of AI in language education: An empirical study at Ha'il University. *Acta Psychologica*, *251*,1-8
[2]. 10.1016/j.actpsy.2024.104605

[3]. Alsaweed, W., & Aljebreen, S. (2024). Investigating the accuracy of ChatGPT as a writing error correction tool. *International Journal of Computer-Assisted Language Learning and Teaching, 14*(1), 1–18. https://doi.org/10.4018/IJCALLT.364847
[4]. Brodsky, J. E., Brooks, P. J., Scimeca, D., Galati, P., Todorova, R., & Caulfield, M. (2021). Associations between online instruction in lateral reading strategies and fact-checking COVID-19 news among college students. *AERA Open, 7*. https://doi.org/10.1177/23328584211038937
[5]. Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction, 5*(CSCW1), Article 188. https://doi.org/10.1145/3449287
[6]. Cengiz, B. C., Bilki, Z., Ataş, A. H., & Çelik, B. (2025). Exploring second language writers' engagement with ChatGPT feedback: Revision behaviors and perceptions. *System, 134*, 103837. https://doi.org/10.1016/j.system.2025.103837
[7]. Chelli, M., Descamps, J., Lavoué, V., Trojani, C., Azar, M., Deckert, M., Raynier, J.-L., Clowez, G., Boileau, P., & Ruetsch-Chelli, C. (2024). Hallucination rates and reference accuracy of ChatGPT and Bard for systematic reviews: Comparative analysis. *Journal of Medical Internet Research, 26*, e53164. https://doi.org/10.2196/53164
[8]. Darvin, R. (2025). The need for critical digital literacies in generative AI-mediated L2 writing. *Journal of Second Language Writing, 67*, 101186. https://doi.org/10.1016/j.jslw.2025.101186
[9]. Hakim, A. (2025). Implementing AI literacy teaching in university-level L2 writing instruction: Exploring one pedagogical approach. *TESOL Journal, 16*(3), e70050. https://doi.org/10.1002/tesj.70050
[10]. Han, Z. (2024). ChatGPT in and for second language acquisition: A call for systematic research. *Studies in Second Language Acquisition, 46*(2), 301–306. https://doi.org/10.1017/S0272263124000111
[11]. Hwang, H., Chang, X., & Sun, J. (2025). Generative AI is useful for second language writing, but when, why, and for how long do learners use it? *Journal of Second Language Writing, 69*, 101230. https://doi.org/10.1016/j.jslw.2025.101230
[12]. Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392
[13]. Lu, D. J., & Zeng, Y. (2025). Exploring the use of ChatGPT-generated model texts as a feedback instrument: EFL students' text quality and perceptions. *Innovation in Language Learning and Teaching*. https://doi.org/10.1080/17501229.2025.2525341
[14]. 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
[15]. Moorhouse, B. L., Wan, Y., Wu, C., Wu, M., & Ho, T. Y. (2025). Generative AI tools and empowerment in L2 academic writing. *System, 133*, 103779. https://doi.org/10.1016/j.system.2025.103779
[16]. Muñoz, B. C., Nassaji, H., & Bello Carrillo, F. I. (2025). ChatGPT-generated versus human direct corrective feedback on L2 writing. *System, 134*, 103805. https://doi.org/10.1016/j.system.2025.103805
[17]. 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*, 100041. https://doi.org/10.1016/j.caeai.2021.100041
[18]. Ranalli, J. (2021). L2 student engagement with automated feedback on writing: Potential for learning and issues of trust. *Journal of Second Language Writing, 52*, 100816. https://doi.org/10.1016/j.jslw.2021.100816
[19]. Shi, H., Chai, C.-S., Zhou, S., & Aubrey, S. (2025). Comparing the effects of ChatGPT and automated writing evaluation on students' writing and ideal L2 writing self. *Computer Assisted Language Learning*. https://doi.org/10.1080/09588221.2025.2454541
[20]. Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind & Language, 25*(4), 359–393. https://doi.org/10.1111/j.1468-0017.2010.01394.x
[21]. Walters, W. H., & Wilder, E. I. (2023). Fabrication and errors in the bibliographic citations generated by ChatGPT. *Scientific Reports, 13*, 14045. https://doi.org/10.1038/s41598-023-41032-5

[22]. Wang, C., & Wang, Z. (2025). Investigating L2 writers' critical AI literacy in AI-assisted writing: An APSE model. *Journal of Second Language Writing, 67*, 101187. https://doi.org/10.1016/j.jslw.2025.101187

[23]. 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*, 101071. https://doi.org/10.1016/j.jslw.2023.101071

[24]. Watson, A. P. (2024). Hallucinated citation analysis: Delving into student-submitted AI-generated sources at the University of Mississippi. *The Serials Librarian, 85*(5–6), 172–180. https://doi.org/10.1080/0361526X.2024.2433640

[25]. 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

[26]. 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*, 1086. https://doi.org/10.1057/s41599-024-03543-y

[27]. Yeung, S. (2025). University students' engagement with generative AI-supported automated writing evaluation feedback. *Journal of Second Language Writing*. Advance online publication.

[28]. Zhou, X., & Wang, Y. (2026). University students' writing feedback literacy in the AI era: The interplay of generative AI acceptance, writing anxiety and writing self-efficacy in Chinese EMI educational contexts. *European Journal of Education, 61*(1), e70478. https://doi.org/10.1111/ejed.70478