Hybrid AI-Augmented Assessment in University English L2 Classrooms: A Tripartite Framework for Teacher Judgement, AI Evidence, and Learner Agency

Saif Al Baimani

Preparatory Studies Centre (PSC), University of Technology and Applied Sciences

DOI: https://doi.org/10.47772/IJRISS.2026.1026EDU0507

Received: 02 August 2026; Accepted: 07 August 2026; Published: 20 August 2026

ABSTRACT

Generative artificial intelligence (GAI) has entered English L2 classrooms faster than assessment frameworks have adapted, leaving teachers to make consequential evaluative decisions without theoretical guidance. Existing accounts document what automated evaluation does well; few specify how its use can remain compatible with the validity-argument tradition on which defensible assessment rests. This conceptual paper addresses that gap. Synthesising recent empirical evidence on AI-mediated writing assessment, the paper proposes the Hybrid AI-Augmented Assessment (HAIAA) model, which distributes evaluative work across three interdependent components. Automated Diagnostic Evaluation confines AI to rapid formative diagnosis of surface features; Human Interpretive Assessment reserves judgements of argument, rhetoric, and cultural appropriateness for instructors; Learner Reflective Verification makes learners' documented engagement with both feedback streams an assessable warrant of authorship. Operating as a recursive cycle rather than a sequence, the components protect the inference from score to construct while restoring the learner to the centre of the feedback process. Implications are drawn for classroom practitioners and curriculum designers, and the conditions under which the model would require empirical validation are specified.

Keywords: generative artificial intelligence; L2 writing assessment; automated writing evaluation; higher-order thinking skills; construct validity

INTRODUCTION

The arrival of large language models (LLMs) as everyday teaching tools has unsettled assumptions that shaped English L2 assessment for decades (Hyland, 2026; Moorhouse & Wong, 2025). A shift of this size usually takes years to work through the field. This one took months. Generative artificial intelligence (GAI) moved from research laboratories into ordinary classrooms, and from a niche curiosity into an institutional concern, within a single academic cycle (Crompton & Burke, 2024). A change at that pace would normally deserve a response that is both quick and careful.

L2 assessment has always been contested ground. Its theoretical foundations were built slowly across the second half of the twentieth century and consolidated in the validity-argument tradition associated with Bachman and Palmer (2010), with Kane's (2013) account of validation as the construction and appraisal of an interpretation-and-use argument, and with the extension of that tradition to technology-mediated assessment by Chapelle and her colleagues (Chapelle et al., 2015; Chapelle & Voss, 2016). The tradition rests on a simple premise: a test should measure what it claims to measure, and the conclusions drawn from a performance should be warranted by the construct behind it. GAI strains the premise exactly where it was already weakest. When a student's essay is partly their own thinking and partly machine-generated, the question of what the assessment is capturing stops being academic and becomes the central problem.

Responses to this development have so far divided into two literatures that rarely meet. One catalogues affordances, documenting gains in feedback speed, scoring consistency, and instructional reach; the other treats GAI principally as an integrity threat to be detected and policed. Neither posture serves assessment well

on its own. Warschauer et al. (2023) capture the inadequacy of both in their analysis of the contradictions GAI poses for L2 writers, among them the imitation contradiction, whereby the fluency that helps a learner communicate also obscures what the learner can do unaided, and the rich-get-richer contradiction, whereby new tools reward those already best resourced to exploit them. An assessment framework must hold both faces of the technology in view at once. That is the ground this paper attempts to occupy.

The paper does not attempt to settle the authorship dilemma, which would require the kind of longitudinal evidence the field does not yet possess. Its aims are more modest. I draw on recent work on AI and L2 assessment to map what the evidence supports on both sides of the ledger, and I use that map to propose the Hybrid AI-Augmented Assessment (HAIAA) model, a framework grounded in validity theory but built for practice. The literature reviewed here is selective rather than systematic. I have prioritised empirical studies of automated writing evaluation (AWE) and LLM-based feedback published in applied-linguistics and educational-technology venues, weighted towards work that bears directly on validity and fairness and on the learner's part in the feedback cycle, and I have preferred meta-analyses and controlled designs where they exist.

It should be noted here that HAIAA itself is a scaffold, not a prescription. It offers a way of structuring the decisions that teachers and curriculum designers now face, so that GAI is used where it demonstrably helps and kept away from the judgements it cannot yet make. This paper proceeds in three parts: a review of what AI-mediated assessment can and cannot do (Section 2), the model itself together with its theoretical grounding and validity argument (Section 3), and finally recommendations for practice and curriculum design (Section 4).

LITERATURE REVIEW

The Shifting Ground of L2 Assessment

The assessment of L2 writing has always depended, in the end, on human judgement of a performance against a construct (Weigle, 2002), and technological innovation has repeatedly reorganised how that judgement is exercised. AWE tools such as Grammarly, e-rater, and Pigai, built on natural language processing, formed the first wave, and their use in L2 writing instruction is well documented (Dikli & Bleyle, 2014; Ding & Zou, 2024; Ranalli, 2021). Over two decades these systems developed from flagging surface errors towards identifying features of structure and style (Ranalli et al., 2017), and a substantial body of research has examined their validity, their impact on writing development, and the factors shaping how learners and instructors utilise them in different contexts for varied purposes (Shi & Aryadoust, 2024). The effects are not trivial. Zhai and Ma's (2023) meta-analysis of twenty-six primary studies found a large positive effect of AWE feedback on writing quality ($g = 0.861$), with greater benefits for EFL and ESL learners than for native-speaker writers. Consistent with this, Wei et al. (2023), in a randomised controlled trial with 190 Chinese EFL learners, found that AWE-informed instruction produced measurable gains across IELTS writing criteria relative to a control condition.

LLMs have widened the integration of technology into writing feedback. Hyland (2026) describes the arrival of GAI as a paradigm shift in how L2 writing is taught and assessed, and reviews of AI in academic writing now identify contributions running from idea generation and structuring through literature synthesis to editing and ethical compliance (Khalifa & Albadawy, 2024). Yet the pattern within these affordances is telling, where AI feedback is strongest precisely where human judgement adds the least, at the level of surface form. Godwin-Jones (2022) anticipated the consequence before the current generation of tools matured. As machine writing assistance grows more capable, he argued, the pressing question is no longer efficiency but the relationship between learners and their tools. If the machine catches the errors, supplies the vocabulary, and reshapes the syntax, whose ability is the test measuring? The question cuts straight to construct validity, which Bachman and Palmer (2010) treated as the cornerstone of any assessment: scores must license warranted conclusions about underlying language ability. That licence is harder to defend now that even trained applied linguists struggle to distinguish AI-generated from human-written academic prose (Casal & Kessler, 2023).

The release of ChatGPT in November 2022 was a change in kind, not merely in degree. Earlier AWE systems worked largely by matching patterns against corpora and flagging errors against rules. LLMs generate text that is contextually coherent, rhetorically structured, and at times close to what a skilled human writer would

produce. Kohnke et al. (2023) observe that this shift from tool to collaborator reshapes the pedagogical relationship itself: the student no longer simply receives feedback from a system but negotiates meaning with an agent able to produce, critique, and rewrite entire passages. Huang et al. (2023), synthesising several hundred studies of AI in language education, reached a compatible conclusion, namely that AI has moved beyond the administration of assessment and into the very learning processes assessment is meant to capture.

What AI Does Well in L2 Assessment

A practical treatment of AI-mediated assessment should start from what the evidence shows, and the literature supports a clear if still limited account. The first and best-documented strength is immediacy. Feedback theory has long held that comments are most effective when they address the gap between current and desired performance while the task is still live in the writer's mind (Hattie&Timperley, 2007; Sadler, 1989). Traditional written feedback, whether from a teacher or a peer, arrives days or weeks after submission, whereas AWE and LLM-based tools respond in seconds. In a mixed-methods study of EFL learners working with AWE-supported instruction, Sari (2024) found that immediacy was the affordance learners mentioned most often, and that they linked it to higher self-efficacy, lower writing anxiety, and more frequent revision. Escalante et al. (2023) reported a similar picture, with learners rating AI feedback as comparable to instructor feedback in clarity and usefulness and, for some error types, preferable. Speed matters most in the high-frequency interactions where teacher time runs out fastest.

The second strength is consistency. Even expert human raters vary: fatigue, cultural assumptions, individual habits of interpretation, and the order in which essays arrive all skew judgements (Lu et al., 2025; Pack et al., 2024; Wang et al., 2025). Pack et al. (2024) compared LLM-generated scores against human raters across a corpus of EFL learner essays and found the LLMs more internally consistent than pairs of human raters on several dimensions. Work on automated essay scoring points the same way while marking its limits. Mizumoto and Eguchi (2023), who used a GPT model to score all 12,100 essays in the TOEFL11 corpus, found a level of accuracy and reliability sufficient to support, though not to replace, human evaluation, and showed that agreement improved when model scores were combined with measured linguistic features. Follow-up work suggests the same holds for accuracy judgements in L2 writing research (Mizumoto et al., 2024). The third strength is scalability. Where the teacher-to-student ratio makes individual feedback unworkable, AI offers a plausible remedy. Su et al. (2023) showed as much in an argumentative-writing study in which ChatGPT generated iterative feedback across drafts: learners who revised with AI support produced essays rated significantly higher on cohesion and argument development than those of a control group. Jeon et al. (2023), reviewing speech-recognition chatbots for language learning, likewise identified individualised response at scale as a defining feature of these environments.

There are also signs that newer generative tools can reach further up the cognitive ladder than older systems could. Song and Song (2023) found that ChatGPT-assisted instruction improved organisational coherence and argumentative depth as well as surface accuracy, areas previously thought out of reach for automation. Yan (2025), synthesising evidence across several productive-skill domains, concluded that generative AI can scaffold higher-order writing processes when prompted well, though the effect is heavily prompt-dependent and does not transfer reliably across task types. Cautious optimism is the right reading. Scaffolding higher-order work under favourable conditions is not the same as evaluating it dependably, and the latter remains contingent on human intervention.

Limitations and Concerns in AI-Mediated L2 Assessment

In several respects, AI's weaknesses are the mirror image of its strengths. It is consistent, scalable, and fast when evaluating surface features. It is considerably less reliable when feedback requires contextual interpretation, evaluative judgement, attention to higher-order meaning, or sensitivity to the sociocultural conditions in which language is used (Jiang et al., 2020; Lee, 2023; Ranalli, 2021). The weak evaluation of higher-order thinking skills is the most serious of these limits. Bloom's revised taxonomy (Anderson & Krathwohl, 2001) places analysis, evaluation, and creation at the heart of academically meaningful language use, yet these are precisely the areas where automated evaluation performs worst. Chapelle et al. (2015) argued that validity arguments for automated assessment are hardest to sustain when the target construct involves

pragmatic competence, rhetorical appropriateness, or culturally inflected argument. Later work on LLMs has not overturned that judgement. GPT-4 and its successors produce more coherent text than their predecessors, but their capacity to judge the originality of an argument, the ethics of a rhetorical choice, or the intercultural fit of a stance remains narrowly constrained (Kuteeva & Andersson, 2024; Ou et al., 2024; Saricaoglu & Bilki, 2025).

Voss et al. (2023), debating assistive technologies in language assessment, raise a problem that goes beyond construct coverage: whatdoes it mean to assess language ability in an environment where AI augmentation is pervasive and largely invisible? Permitting test-takers to use AI risks an assessment that no longer measures the construct it claims to; prohibiting it drifts the assessment further from the conditions under which people now actually use language. There is no clean way out. The authors describe this as a gap between theory and practice that current frameworks were not built to bridge.

Bias and equity add an ethical layer, and here the evidence has hardened. Liang et al. (2023) evaluated seven widely used GPT detectors on essays known to be human-written and found that more than half of the TOEFL essays by non-native writers were misclassified as AI-generated, while accuracy on essays by native-speaker school students remained near perfect; the mechanism, they showed, is that detectors reward the linguistic variability that proficient but cautious L2 writers tend to lack. Jiang et al. (2024), studying detection across roughly 8,000 essays in a large-scale writing assessment, confirmed the same systematic bias against non-native writers. The consequences are direct in high-stakes settings, where detection is increasingly deployed as an integrity measure, and they compound the rich-get-richer contradiction that Warschauer et al. (2023) identify: the learners most likely to be falsely accused are those with the fewest resources to contest the accusation. Darvin (2025) approaches the same terrain from a critical sociolinguistic angle, arguing that debate around AI-mediated assessment conceals ideological choices that deserve to be named: which varieties of English, which rhetorical conventions, and which ways of knowing are built into LLM training data, and what follows when such systems judge learners whose repertoires diverge from that normative centre. Zhang and Hyland (2018, 2022) had flagged a related pattern earlier, observing that automated systems tend to reward conventional academic prose while penalising the rhetorical departures that often mark expert writing in non-Anglo contexts. These are not side issues; they go to the fairness of any AI-mediated assessment and, hence, to its validity.

Finally, there is learner dependency, and its two faces are worth distinguishing. Ranalli (2021) found that sustained use of automated feedback can produce what he called trust issues: a combination of over-reliance on AI for revision decisions with lingering doubt about whether the advice is pedagogically sound. Koltovskaia's (2020) multiple-case study makes the point concrete. Of her two focal students, the more advanced writer rejected half of Grammarly's feedback as inaccurate when independent coding had judged all of it correct, while the less proficient writer accepted every suggestion, including the roughly one quarter that was wrong, because she lacked the linguistic knowledge to evaluate them. Under-reliance and over-reliance are thus twin failures of the same underlying capacity: evaluative judgement. Shen et al. (2023), in a process-product analysis of EFL learners working with AWE feedback, found in the same vein that computer-generated feedback increased surface-level revision but did not reliably push learners to rethink argument or content. The value of AI feedback therefore depends heavily on learner metacognition, and this is something an assessment framework must address directly rather than assume.

A final qualification concerns where this evidence comes from. The strongest quantitative findings reviewed here derive from comparatively well-resourced university settings, principally Chinese and other East Asian EFL programmes (Shen et al., 2023; Su et al., 2023; Wei et al., 2023), and from large standardised corpora such as TOEFL11 (Mizumoto & Eguchi, 2023); Zhai and Ma's (2023) meta-analytic pool is weighted towards the same regions. Results produced under those conditions presuppose stable connectivity, institutional tool licences, and instructors with time to mediate automated feedback. None of this can be assumed in low-resource settings, where free tool tiers, intermittent access, and very large classes alter both what AI feedback costs and what it displaces (Miao & Holmes, 2023). The rich-get-richer contradiction that Warschauer et al. (2023) identify operates between institutions as well as within classrooms. Claims about AI-mediated

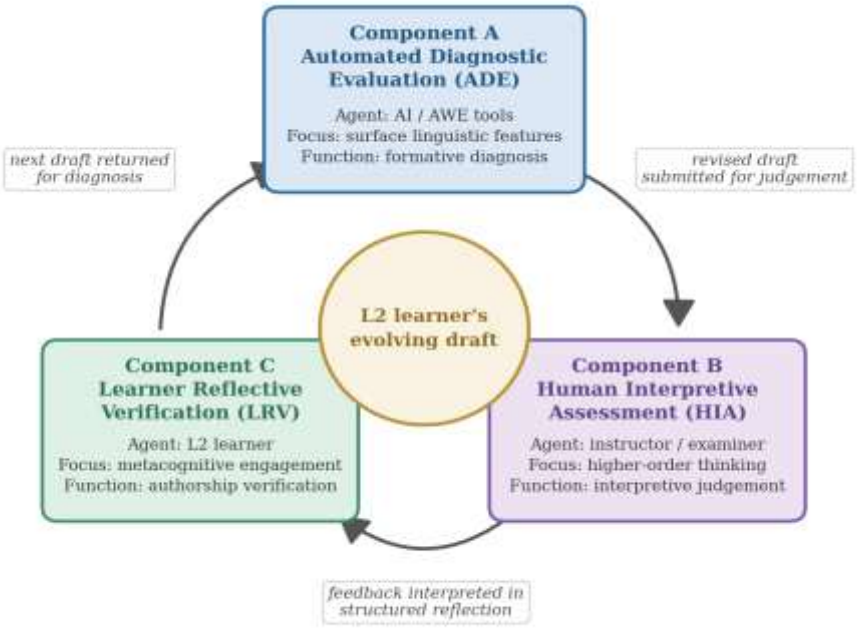
assessment therefore travel less readily than the technology itself, and the framework proposed in Section 3 is offered with that constraint in view.

Towards a Principled Framework: The Hybrid AI-Augmented Assessment (HAIAA) Model

The question facing L2 assessment is not whether to admit AI but how to do so in ways that are construct-valid, fair, and pedagogically productive. The HAIAA model organises integration around three distinct but interdependent components, each matched to the strengths and limits of the agent responsible for it: Automated Diagnostic Evaluation (ADE), Human Interpretive Assessment (HIA), and Learner Reflective Verification (LRV). The model's theoretical debts are twofold. From formative assessment theory it takes the principle that feedback exists to close the gap between current and reference performance (Black & Wiliam, 1998; Sadler, 1989); from the validity-argument tradition it takes the requirement that every link in the chain from performance to interpretation be defensible (Kane, 2013). Figure 1 represents the model as the cycle it is; Table 1 specifies the components; and Section 3.4 sets out the validity argument the arrangement supports.

HAIAA is not the first attempt to give GAI a regulated place in assessment, and its contribution is best stated against the nearest alternatives. The AI Assessment Scale of Perkins et al. (2024) grades the permitted extent of AI use across the levels of a task, while the two-lane arrangement developed at the University of Sydney separates secured assessment of learning from open assessment for and as learning (Bridgeman & Liu, 2024). Both are frameworks for governing access, in that they specify when, and how much, AI a learner may use; both also share the conviction that assessment validity, rather than cheating detection, is the proper organising concern (Dawson et al., 2024). Chapelle et al. (2015), conversely, mount a validity argument for a single automated system operating in a fixed diagnostic role. HAIAA takes a different object. It regulates the division of evaluative labour within one task cycle, so that machine, teacher, and learner each warrant a distinct inference, and it makes the learner's documented engagement assessable evidence in its own right. The novelty, hence, lies in the tripartite allocation rather than in the hybridity itself: so far as I am aware, no existing framework assigns the authorship warrant to a structured learner role positioned alongside machine diagnosis and human judgement. In the terms proposed by Corbin et al. (2025), HAIAA is a structural rather than a discursive response, because the warrant it offers does not depend on learners obeying instructions about tool use.

Figure 1 The HAIAA Cycle



Automated Diagnostic Evaluation

ADE assigns to AI what Feng et al. (2025) call the lower-tier features of L2 writing: grammatical accuracy, lexical density and sophistication, syntactic complexity, and the use of cohesive devices. These are the features AI handles most reliably, and also the ones whose manual marking consumes the most teacher time. ADE functions as a triage step. It surfaces the surface-level profile of a text, returns immediate diagnostic feedback, and frees instructors for the interpretive and relational work only humans do better. In Sadler's (1989) terms, ADE supplies the learner with rapid, repeatable information about the gap between the current draft and the reference standard at the linguistic level, which is exactly the register of feedback that formative assessment research has shown to raise achievement when it is timely and specific (Black & Wiliam, 1998; Hattie&Timperley, 2007). Within HAIAA, ADE feedback is provisional and iterative, never summative; its job is to support revision, not to certify achievement. This restriction is deliberate. Confining automated output to formative use keeps the weakest link in the inferential chain, the machine's judgement, away from the decisions that carry consequences for learners.

Human Interpretive Assessment

HIA places the instructor in charge of the higher-order dimensions: argument quality, rhetorical purpose, critical reasoning, cultural and contextual fit, and the integration of evidence. Bachman and Palmer (2010) insisted that the chain of inference from performance to construct is always a human act, and Kane (2013) makes the requirement precise: an interpretation-and-use argument must state the inferences on which a score interpretation depends and back each with evidence, and the high-inference links, those running from observed performance to claims about reasoning, rhetoric, and communicative competence, are the ones automated systems cannot yet warrant. HIA keeps those links under professional control. Drawing on the ADE profile, instructors read learners' texts as whole communicative acts, concentrating on what Bloom's revised taxonomy (Anderson & Krathwohl, 2001) treats as the most demanding work: analysing rhetorical structure, weighing how well an argument holds, and judging how a learner reworks source material in their own voice. The empirical signs here are encouraging. Moorhouse and Kohnke (2024) found that teachers given AI pre-processing of student texts felt more confident in their summative judgements and spent more of their feedback time on content and argument, which is exactly the shift HIA is designed to institutionalise. HIA also carries the model's equity function: because a human reader makes every consequential judgement, the biases documented in automated scoring and detection (Jiang et al., 2024; Liang et al., 2023) are subject to professional scrutiny rather than passed through unexamined.

Learner Reflective Verification

LRV takes on what Darvin (2025) and Kohnke et al. (2024) identify as the least developed element of AI-integrated assessment: the learner's own role in interpreting and acting on feedback. The component rests on the premise that assessment teaches as much as it measures. It asks learners to reflect, in a structured log, on the feedback they receive from machine and human sources, to record the revisions they made and why, and to weigh AI suggestions against their own sense of what they were trying to say. Two research traditions supply the design rationale. Literature on feedback literacy holds that benefiting from feedback is itself a learned capacity, comprising the ability to appreciate feedback, to make sound judgements about one's own work, and to act on what is received (Carless & Boud, 2018). Work on self-regulated learning shows that the setting of goals, monitoring of progress, and adjustment of strategy distinguish learners who improve from those who merely comply (Zimmerman, 2002). LRV gives both capacities a place in the assessment itself. Pack and Maloney (2024), writing on the ethics and pedagogy of AI in TESOL, argue that learner agency in the feedback cycle is a necessary condition if AI-mediated assessment is to support language development rather than allow learners to bypass it. Koltovskaia's (2020) cases show what its absence looks like in practice. LRV renders the argument operational: the structured reflective protocol makes a learner's engagement with feedback an assessable part of the performance itself. The authorship warrant this yields must nonetheless be claimed with care. A reflective log is itself a text, and a learner prepared to outsource an essay could outsource the reflection with it, so LRV evidence should be read as corroborative rather than probative. Its force depends on triangulation with records and interactions that are harder to counterfeit. Version histories and process logs

of the kind writing researchers use to reconstruct composing behaviour (Leijten & Van Waes, 2013) anchor the log in observable revision activity, and a brief oral defence of the log, conducted within the HIA conference, allows dialogic follow-up in which the instructor probes whether the learner can account for the decisions the log reports. Interactive oral formats of this kind have a documented record of deterring outsourced work in higher education (Sotiriadou et al., 2020). Within HAIAA, then, LRV supplies the documentary trace and HIA the dialogic check; neither alone certifies authorship. A further caution is in order. Assessed reflection introduces its own validity question, since the quality of a reflective log is partly a function of proficiency and of a learner's willingness to perform reflection for a grade. HAIAA therefore treats LRV evidence as an index of engagement and authorship, not as a further measure of writing ability, and any implementation should keep the two scores apart.

The HAIAA Cycle and Its Validity Argument

The three components work in a loop, not a line, as Figure 1 shows. A learner receives ADE feedback on a draft, revises, submits for HIA evaluation, writes an LRV reflection explaining how they read both streams of feedback, and returns to ADE for a further diagnostic pass. The loop embodies a consistent finding of the feedback literature: feedback promotes learning when learners engage with it reflectively and repeatedly across drafts, not when it arrives once at the end (Hattie & Timperley, 2007; Zhang & Hyland, 2018, 2022). In a typical writing course, the cycle maps onto familiar structures. A multi-draft assignment might weight the HIA judgement most heavily in the final grade, attach a smaller weight to the LRV log, and attach none to ADE output, which serves the process rather than the product. Table 2 illustrates one such scheme for a multi-draft argumentative essay, together with sample prompts for the reflective log; the weightings are indicative rather than prescribed, and their calibration across proficiency levels is one of the empirical questions noted in Section 5.

Framed as an interpretation-and-use argument (Kane, 2013; Chapelle et al., 2015), the division of labour is itself the warrant. The inference from score to construct is threatened when automated judgement extends beyond the features it measures well. HAIAA blocks that extension structurally by restricting ADE to low-inference surface description and reserving all high-inference interpretation for HIA. The inference from performance to authorship, which GAI has made newly fragile, is supported by LRV, whose documented revision rationale, corroborated where the stakes warrant it by the oral defence described in Section 3.3, provides evidence that the learner, not the tool, made the consequential decisions. The inference from score to fair treatment, threatened by the detection and scoring biases reviewed in Section 2.3, is protected by routing every consequential judgement through a human reader answerable for it. HAIAA is thus not a scoring scheme. It is an assessment ecology intended to create the conditions under which feedback serves learning while the resulting judgements remain defensible.

Table 1 Specifications of the Three HAIAA Components

Component A: Automated Diagnostic Evaluation (ADE)	**Component B: Human Interpretive Assessment (HIA)**	**Component C: Learner Reflective Verification (LRV)**
Primary agent: AI / AWE tools **Construct focus:** surface linguistic features **Key indicators:** grammatical accuracy; lexical density and range; syntactic complexity; cohesive-device use **Function:** formative / diagnostic **Role in cycle:** triage and revision	**Primary agent:** instructor / examiner **Construct focus:** higher-order thinking skills **Key indicators:** argumentative coherence; critical reasoning depth; rhetorical appropriateness; evidence integration **Function:** interpretive / summative	**Primary agent:** L2 learner **Construct focus:** metacognitive engagement **Mechanism:** structured reflection log **Output:** annotated revision rationale **Function:** agency and authorship

scaffold	**Role in cycle:** construct-critical judgement	verification **Protocol:** review, evaluate, justify, revise, re-engage

Table 2 An Illustrative HAIAA Assessment Scheme for a Multi-Draft Argumentative Essay

Element	**Weighting**	**Assessed evidence**
ADE diagnostic reports	0% (formative only)	Automated reports attached to each draft; they inform revision but are not graded.
HIA judgement of the final essay	70%	Argument quality, critical reasoning, rhetorical appropriateness, and evidence integration, judged against the course rubric.
LRV reflective log	20%	Revisions recorded with reasons; AI suggestions accepted, adapted, or rejected with justification. The log is assessed for quality of engagement, not writing ability.
Oral defence of the log (within the HIA conference)	10%	A short dialogic follow-up in which the learner accounts for two or three logged decisions selected by the instructor.
Sample reflection prompts: Which piece of AI feedback did you reject, and on what grounds? Which revision made in response to your instructor's comments changed your argument most, and how? Where did an AI suggestion conflict with what you were trying to say, and how did you resolve it? What would you now do differently at the drafting stage?		

Note. Percentages refer to the final grade for a multi-draft argumentative essay in a semester-length EFL writing course; ADE remains formative throughout, for the reasons given in Section 3.1.

RECOMMENDATIONS

Classroom Practitioners

Putting HAIAA into practice asks teachers to reconceive their role, not as a source of authority competing with AI but as the interpretive agent whose expertise matters most precisely where AI is weakest. The recommendations that follow derive from the evidence reviewed in Section 2 and are scaled to the everyday constraints of EFL teaching. First, teachers ought to use AI feedback tools for diagnostic and formative purposes rather than summative ones. Both Zhai and Ma (2023) and Wei et al. (2023) found that AWE delivers its most consistent gains when woven into iterative drafting, not bolted on at the end. In practice, this might mean asking learners to produce a first draft with AI diagnostic support, document the feedback received and the revisions made, and only then submit for human assessment. Such sequencing captures the speed and reach of AI while protecting the construct-critical role of human judgement.

Second, instructors should treat learner AI literacy, and the feedback literacy beneath it, as teaching goals in their own right. Darvin (2025) argues persuasively that uncritical AI use entrenches dependence rather than building agency, and Barrot (2023) documents how easily the pitfalls of ChatGPT-assisted writing outrun its potentials when learners lack evaluative criteria. A learner who accepts AI revisions without grasping why ends up with a polished product but little of the underlying competence the assessment is meant to measure; Koltovskaia's (2020) over-reliant case is the cautionary example. The remedy is instructional, not merely regulatory. Carless and Boud (2018) show that the capacity to judge one's own work and act on feedback can be taught, and Teng (2024) recommends setting aside explicit instructional time for learners to interrogate, challenge, and selectively reject AI suggestions; both recommendations map directly onto the LRV component.

Third, teachers should use AI feedback as a prompt for professional reflection rather than a substitute for judgement. Moorhouse and Kohnke (2024) found that teachers who treated AI diagnostics as a starting point for their own thinking, not a verdict to relay, developed more construct-sensitive feedback over time.

Curriculum Designers

The curricular implications of AI-mediated assessment may be the furthest-reaching and the least worked out in the current literature. If assessment drives curriculum, then changing how we assess changes what curricula contain and how they are sequenced. Three moves follow from the framework. Curriculum designers should first build AI literacy explicitly into EFL programmes as a graduate competency. Warschauer et al. (2023) offer a ready curricular anchor in their five-part framework, in which learners are taught to understand how generative tools work, access them equitably, prompt them effectively, corroborate their output against other sources, and incorporate what survives scrutiny into their own writing; a curriculum aligned with HAIAA would add the metacognitive strategies needed to reconcile automated and human feedback. Designers should, secondly, give greater assessable weight to higher-order cognitive processes, redesigning tasks so that argument, critical synthesis, and reflective positioning carry the marks. Thirdly, they should consider embedding multimodal and process-portfolio assessment as a structural counterweight to the text-product orientation that GAI mimics most easily. Process portfolios are more defensible on construct grounds, and they build the self-regulatory capacities on which sustained L2 development depends (Shen et al., 2023; Zimmerman, 2002).

CONCLUSION AND LIMITATIONS

How should English L2 assessment respond to GAI? No single answer will hold because the technology is neither simple nor stable. The evidence does, however, support a small set of principled distinctions that should survive the tools' evolution. AI belongs where its strengths are demonstrated: immediate, consistent, scalable evaluation of surface features. Human judgement remains irreplaceable where the constructs matter most: argument, critical reasoning, rhetorical awareness, and culturally situated communicative competence. Learner agency, finally, is a precondition for the legitimacy of any assessment system, AI-mediated or otherwise.

The HAIAA model offered here is a thinking tool, not a finished architecture, and its limits should be stated plainly. As a conceptual proposal it awaits empirical validation, and the research agenda is specifiable. The workload implications of HIA for teachers with large classes need cost-and-benefit study; the reliability with which LRV logs can be judged, and the washback of assessing reflection need classroom investigation. Further, the weighting of components in final grades needs comparative trial across proficiency levels and task types. The model is also schematic, and implementations will need local adaptation of the kind Dimova et al. (2022) describe for locally developed tests, since the balance among the three components must answer to institutional constraints and learner populations that vary widely. Transfer to low-resource settings, for the reasons set out in Section 2.3, should be demonstrated rather than assumed (Miao & Holmes, 2023). What the framework provides is a pair of commitments. First, AI and human evaluative capacities should complement rather than compete with each other. Second, a learner's reflective engagement with evaluation is itself an educative act deserving a place in the system. My own experience as an EFL instructor suggests that teachers do not want AI removed from assessment. Rather, they need a defensible account of where it belongs, and HAIAA is offered as a step towards that account.

REFERENCES

1. Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives. Longman.
2. Bachman, L. F., & Palmer, A. S. (2010). Language assessment in practice: Developing language assessments and justifying their use in the real world. Oxford University Press.
3. Barrot, J. S. (2023). Using ChatGPT for second language writing: Pitfalls and potentials. Assessing Writing, 57, 100745. https://doi.org/10.1016/j.asw.2023.100745

4. Black, P., & Wiliam, D. (1998). Assessment and classroom learning. Assessment in Education: Principles, Policy&Practice, 5(1), 7–74. https://doi.org/10.1080/0969595980050102
5. Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. Assessment&Evaluation in Higher Education, 43(8), 1315–1325. https://doi.org/10.1080/02602938.2018.1463354
6. Casal, J. E., & Kessler, M. (2023). Can linguists distinguish between ChatGPT/AI and human writing? A study of research ethics and academic publishing. Research Methods in Applied Linguistics, 2(3), 100068. https://doi.org/10.1016/j.rmal.2023.100068
7. Chapelle, C. A., Cotos, E., & Lee, J. (2015). Validity arguments for diagnostic assessment using automated writing evaluation. Language Testing, 32(3), 385–405. https://doi.org/10.1177/0265532214565386
8. Chapelle, C. A., & Voss, E. (2016). 20 years of technology and language assessment in Language Learning & Technology. Language Learning&Technology, 20(2), 116–128. https://hdl.handle.net/10125/44464
9. Corbin, T., Dawson, P., & Liu, D. (2025). Talk is cheap: Why structural assessment changes are needed for a time of GenAI. Assessment&Evaluation in Higher Education, 50(7), 1087–1097. https://doi.org/10.1080/02602938.2025.2503964
10. Crompton, H., & Burke, D. (2024). The educational affordances and challenges of ChatGPT: State of the field. TechTrends, 68, 380–392. https://doi.org/10.1007/s11528-024-00939-0
11. 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
12. Dawson, P., Bearman, M., Dollinger, M., & Boud, D. (2024). Validity matters more than cheating. Assessment&Evaluation in Higher Education, 49(7), 1005–1016. https://doi.org/10.1080/02602938.2024.2386662
13. Dikli, S., & Bleyle, S. (2014). Automated essay scoring feedback for second language writers: How does it compare to instructor feedback? Assessing Writing, 22, 1–17. https://doi.org/10.1016/j.asw.2014.03.006
14. Dimova, S., Yan, X., & Ginther, A. (2022). Local tests, local contexts. Language Testing, 39(3), 341–354. https://doi.org/10.1177/02655322221075014
15. Ding, L., & Zou, D. (2024). Automated writing evaluation systems: A systematic review of Grammarly, Pigai, and Criterion with a perspective on future directions in the age of generative artificial intelligence. Education and Information Technologies, 29(11), 14151–14203. https://doi.org/10.1007/s10639-023-12402-3
16. Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. International Journal of Educational Technology in Higher Education, 20(1), 57. https://doi.org/10.1186/s41239-023-00425-2
17. Feng, H., Li, K., & Zhang, L. J. (2025). What does AI bring to second language writing? A systematic review (2014–2024). Language Learning&Technology, 29(1), 1–27. https://hdl.handle.net/10125/73629
18. Godwin-Jones, R. (2022). Partnering with AI: Intelligent writing assistance and instructed language learning. Language Learning&Technology, 26(2), 5–24. https://hdl.handle.net/10125/73474
19. Hattie, J., & Timperley, H. (2007). The power of feedback. Review of Educational Research, 77(1), 81–112. https://doi.org/10.3102/003465430298487
20. Huang, X., Zou, D., Cheng, G., Chen, X., & Xie, H. (2023). Trends, research issues and applications of artificial intelligence in language education. Educational Technology&Society, 26(1), 112–131. https://doi.org/10.30191/ETS.202301_26(1).0009
21. Hyland, K. (2026). Tomorrowland? Technical innovation and feedback on writing. RELC Journal, 57(1), 233–241. https://doi.org/10.1177/00336882251357675
22. Jeon, J., Lee, S., & Choe, H. (2023). Beyond ChatGPT: A conceptual framework and systematic review of speech-recognition chatbots for language learning. Computers&Education, 206, 104898. https://doi.org/10.1016/j.compedu.2023.104898
23. Jiang, L., Yu, S., & Wang, C. (2020). Second language writing instructors' feedback practice in response to automated writing evaluation: A sociocultural perspective. System, 93, 102302. https://doi.org/10.1016/j.system.2020.102302

24. Jiang, Y., Hao, J., Fauss, M., & Li, C. (2024). Detecting ChatGPT-generated essays in a large-scale writing assessment: Is there a bias against non-native English speakers? Computers&Education, 217, 105070. https://doi.org/10.1016/j.compedu.2024.105070
25. Kane, M. T. (2013). Validating the interpretations and uses of test scores. Journal of Educational Measurement, 50(1), 1–73. https://doi.org/10.1111/jedm.12000
26. Khalifa, M., & Albadawy, M. (2024). Using artificial intelligence in academic writing and research: An essential productivity tool. Computer Methods and Programs in Biomedicine Update, 5, 100145. https://doi.org/10.1016/j.cmpbup.2024.100145
27. Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for language teaching and learning. RELC Journal, 54(2), 537–550. https://doi.org/10.1177/00336882231162868
28. Kohnke, L., Zou, D., & Moorhouse, B. L. (2024). Technostress and English language teaching in the age of generative AI. Educational Technology&Society, 27(2), 306–320. https://doi.org/10.30191/ETS.202404_27(2).TP02
29. Koltovskaia, S. (2020). Student engagement with automated written corrective feedback (AWCF) provided by Grammarly: A multiple case study. Assessing Writing, 44, 100450. https://doi.org/10.1016/j.asw.2020.100450
30. Kuteeva, M., & Andersson, M. (2024). Diversity and standards in writing for publication in the age of AI—Between a rock and a hard place. Applied Linguistics, 45(3), 561–567. https://doi.org/10.1093/applin/amae025
31. Lee, I. (2023). Problematising written corrective feedback: A Global Englishes perspective. Applied Linguistics, 44(4), 791–796. https://doi.org/10.1093/applin/amad038
32. Leijten, M., & Van Waes, L. (2013). Keystroke logging in writing research: Using Inputlog to analyze and visualize writing processes. Written Communication, 30(3), 358–392. https://doi.org/10.1177/0741088313491692
33. Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. Patterns, 4(7), 100779. https://doi.org/10.1016/j.patter.2023.100779
34. Lu, Y., Liles, X., & Ma, X. (2025). GenAI and human assessments of L2 Chinese writing: Interrater reliability and rater bias. Assessing Writing, 66, 100989. https://doi.org/10.1016/j.asw.2025.100989
35. Miao, F., & Holmes, W. (2023). Guidance for generative AI in education and research. UNESCO.
36. Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. Research Methods in Applied Linguistics, 2(2), 100050. https://doi.org/10.1016/j.rmal.2023.100050
37. Mizumoto, A., Shintani, N., Sasaki, M., & Teng, M. F. (2024). Testing the viability of ChatGPT as a companion in L2 writing accuracy assessment. Research Methods in Applied Linguistics, 3(2), 100116. https://doi.org/10.1016/j.rmal.2024.100116
38. Moorhouse, B. L., & Kohnke, L. (2024). The effects of generative AI on initial language teacher education: The perceptions of teacher educators. System, 122, 103290. https://doi.org/10.1016/j.system.2024.103290
39. Moorhouse, B. L., & Wong, K. M. (2025). Generative artificial intelligence and language teaching. Cambridge University Press. https://doi.org/10.1017/9781009618823
40. Ou, A. W., Khuder, B., Franzetti, S., & Negretti, R. (2024). Conceptualising and cultivating critical GAI literacy in doctoral academic writing. Journal of Second Language Writing, 66, 101156. https://doi.org/10.1016/j.jslw.2024.101156
41. Pack, A., & Maloney, J. (2024). Using artificial intelligence in TESOL: Some ethical and pedagogical considerations. TESOL Quarterly, 58(2), 1007–1018. https://doi.org/10.1002/tesq.3255
42. Pack, A., Maloney, J., Barrett, A., & Escalante, J. (2024). Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability. Computers and Education: Artificial Intelligence, 6, 100234. https://doi.org/10.1016/j.caeai.2024.100234
43. 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
44. 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

45. Ranalli, J., Link, S., & Chukharev-Hudilainen, E. (2017). Automated writing evaluation for formative assessment of second language writing: Investigating the accuracy and usefulness of feedback as part of argument-based validation. Educational Psychology, 37(1), 8–25. https://doi.org/10.1080/01443410.2015.1136407
46. Sadler, D. R. (1989). Formative assessment and the design of instructional systems. Instructional Science, 18(2), 119–144. https://doi.org/10.1007/BF00117714
47. Sari, E. (2024). The impact of automated writing evaluation on English as a foreign language learners' writing self-efficacy, self-regulation, anxiety, and performance. Journal of Computer Assisted Learning, 40(3), 1012–1025. https://doi.org/10.1111/jcal.13004
48. Saricaoglu, A., & Bilki, Z. (2025). The capacity of ChatGPT-4 for L2 writing assessment: A closer look at accuracy, specificity, and relevance. Annual Review of Applied Linguistics, 45, 253–273. https://doi.org/10.1017/S0267190525100160
49. Shen, C., Shi, P., Guo, J., Xu, S., & Tian, J. (2023). From process to product: Writing engagement and performance of EFL learners under computer-generated feedback instruction. Frontiers in Psychology, 14, 1258286. https://doi.org/10.3389/fpsyg.2023.1258286
50. Shi, H., & Aryadoust, V. (2024). A systematic review of AI-based automated written feedback research. ReCALL, 36(2), 187–209. https://doi.org/10.1017/S0958344023000265
51. 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, 1260843. https://doi.org/10.3389/fpsyg.2023.1260843
52. Sotiriadou, P., Logan, D., Daly, A., & Guest, R. (2020). The role of authentic assessment to preserve academic integrity and promote skill development and employability. Studies in Higher Education, 45(11), 2132–2148. https://doi.org/10.1080/03075079.2019.1582015
53. Su, Y., Lin, Y., & Lai, C. (2023). Collaborating with ChatGPT in argumentative writing classrooms. Assessing Writing, 57, 100752. https://doi.org/10.1016/j.asw.2023.100752
54. Teng, M. F. (2024). A systematic review of ChatGPT for English as a foreign language writing: Opportunities, challenges, and recommendations. International Journal of TESOL Studies, 6(3), 36–57. https://doi.org/10.58304/ijts.20240304
55. Voss, E., Cushing, S. T., Ockey, G. J., & Yan, X. (2023). The use of assistive technologies including generative AI by test takers in language assessment: A debate of theory and practice. Language Assessment Quarterly, 20(4–5), 520–532. https://doi.org/10.1080/15434303.2023.2258384
56. Wang, Y., Huang, J., Du, L., Guo, Y., Liu, Y., & Wang, R. (2025). Evaluating large language models as raters in large-scale writing assessments: A psychometric framework for reliability and validity. Computers and Education: Artificial Intelligence, 9, 100481. https://doi.org/10.1016/j.caeai.2025.100481
57. 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
58. Wei, P., Wang, X., & Dong, H. (2023). The impact of automated writing evaluation on second language writing skills of Chinese EFL learners: A randomized controlled trial. Frontiers in Psychology, 14, 1249991. https://doi.org/10.3389/fpsyg.2023.1249991
59. Weigle, S. C. (2002). Assessing writing. Cambridge University Press.
60. Yan, X. (2025). Generative AI for the teaching, learning, and assessment of productive skills: An evidence-based approach to understanding its real impact. TESOL Quarterly. https://doi.org/10.1002/tesq.70043
61. Zhai, N., & Ma, X. (2023). The effectiveness of automated writing evaluation on writing quality: A meta-analysis. Journal of Educational Computing Research, 61(4), 875–900. https://doi.org/10.1177/07356331221127300
62. Zhang, Z. V., & Hyland, K. (2018). Student engagement with teacher and automated feedback on L2 writing. Assessing Writing, 36, 90–102. https://doi.org/10.1016/j.asw.2018.02.004
63. Zhang, Z. V., & Hyland, K. (2022). Fostering student engagement with feedback: An integrated approach. Assessing Writing, 51, 100586. https://doi.org/10.1016/j.asw.2021.100586
64. Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. Theory into Practice, 41(2), 64–70. https://doi.org/10.1207/s15430421tip4102_2