RECONCEPTUALISING EFL ASSESSMENT IN THE AGE OF GENERATIVE AI: DEVELOPING A HUMAN-CENTRED AND AI-RESILIENT ASSESSMENT FRAMEWORK FOR EFL LEARNERS

Yusupov Bekzod
Andijan District No. 53 General Secondary School, Uzbekistan
E-mail: begginy19@gmail.com
https://doi.org/10.5281/zenodo.23246114

Abstract

The rapid development of generative artificial intelligence (GenAI) has challenged the assumptions underlying conventional English as a Foreign Language (EFL) assessment. Assessment practices that rely predominantly on final written products increasingly face difficulties in establishing whether submitted work represents learners' own linguistic knowledge, reasoning, and decision-making. This conceptual study examines how EFL assessment can be reconceptualised in response to the growing presence of GenAI and proposes a Human-Centred AI-Resilient EFL Assessment Framework. The study employs a structured conceptual synthesis of recent scholarship published primarily between 2024 and 2026 on generative AI, language assessment, EFL writing, assessment validity, teacher agency, and AI-resilient assessment. The analysis indicates that detection-oriented approaches alone provide an insufficient response to the assessment validity problem. Instead, assessment should generate multiple forms of evidence of learning through authentic task design, process documentation, learner reflection, oral verification, transparent AI-use disclosure, and professional teacher judgement. Based on this synthesis, the proposed framework integrates six dimensions: authenticity, process evidence, human agency, oral verification, responsible AI use, and multidimensional judgement. The framework shifts the central assessment question from whether AI has been used to whether the learner can demonstrate meaningful linguistic competence, reasoning, and ownership of learning. The study argues that AI-resilient assessment should not seek to eliminate technology from educational environments but should redesign assessment so that evidence of learning remains attributable to the learner. The framework provides a conceptual basis for future empirical validation in secondary-school EFL contexts.

Keywords: generative artificial intelligence, EFL assessment, AI-resilient assessment, human-centred assessment, language assessment, learner agency, authentic assessment, assessment validity

1. INTRODUCTION

Generative artificial intelligence (GenAI) has rapidly transformed the technological environment in which language teaching and assessment take place. Large language models can generate coherent texts, provide grammatical corrections, translate passages, suggest vocabulary, restructure arguments, and produce responses tailored to particular prompts. Consequently, the relationship between a learner and the final linguistic product can no longer be assumed to be straightforward. A polished essay submitted by an EFL learner may represent the learner's independent linguistic competence, extensive use of AI assistance, or a combination of human and machine contributions. This development creates a fundamental challenge for assessment systems whose validity depends on interpreting the submitted product as evidence of individual learning.

The implications extend beyond academic integrity. Language assessment has traditionally been concerned with the validity, reliability, fairness, and practicality of inferences made from learners' performances. A systematic review of 77 studies conducted by Chuang and Yan demonstrated that GenAI is reshaping these core dimensions of language assessment and that the consequences of AI

adoption vary according to assessment context, purpose, and implementation [1]. The authors specifically emphasise the need for critical rather than unselective adoption of GenAI in assessment and identify continuing challenges concerning validity, reliability, fairness, and practicality.

The problem is particularly significant in EFL education because language competence cannot be reduced to grammatical accuracy or the surface quality of a written product. Effective language use involves the ability to formulate ideas, select appropriate linguistic resources, construct meaning, respond to context, justify decisions, and communicate with an audience. When a generative system can perform some of these operations on behalf of a learner, the evidential relationship between performance and competence becomes more complex. Thus, the central assessment problem is not simply whether learners use AI, but whether assessment tasks continue to provide trustworthy evidence of the intended construct.

Recent research provides evidence that this problem is already visible in EFL contexts. Chen and Pi found that EFL teachers experience considerable difficulty identifying GenAI-generated or GenAI-assisted writing and that teachers' familiarity with GenAI and pedagogical beliefs influence their assessment responses [2]. Their findings also identify teacher agency as a critical factor in assessing GenAI-assisted writing. This finding challenges the assumption that teachers can simply distinguish human-generated from AI-generated texts through professional intuition or visual inspection.

At the same time, the growing capabilities of large language models should not lead to the conclusion that AI-assisted assessment is inherently unreliable. Yavuz, Çelik, and Yavaş Çelik examined the use of ChatGPT and Bard in rubric-based EFL essay grading and reported substantial levels of agreement between some AI-generated scores and human ratings [3]. Their results demonstrate that large language models may have considerable potential for supporting particular forms of assessment. Nevertheless, the authors also indicate the need for further adjustment when assessing more nuanced and non-objective dimensions of writing. Therefore, the question is not whether AI should replace human assessment, but how human judgement and AI capabilities can be positioned within a valid assessment architecture.

A related issue concerns the limitations of detection-oriented responses. Attempts to determine whether a text was produced by AI place the assessment system in a technological race against rapidly changing generative models. More importantly, detection does not itself demonstrate what a learner knows or can do. A detection result may indicate a possible source of text, but it does not provide direct evidence of linguistic reasoning, communicative competence, or ownership of learning. Recent scholarship on AI-resilient assessment consequently proposes a shift from detecting AI use toward redesigning assessment so that students must demonstrate reasoning, decision-making, and ownership of their learning [4]. Nikolić and Basta Nikolić conceptualise AI-resilient assessment around four major pillars: process-based documentation, oral defence, authentic task design, and transparent AI-use policies aligned with intended learning outcomes [4].

This emerging perspective is particularly relevant to EFL assessment. A learner may use an AI system to obtain suggestions, but the assessment can still require the learner to explain why particular linguistic choices were made, revise a draft, respond to unexpected questions, defend an argument orally, or reflect on how feedback changed the final product. In such a model, AI use does not automatically invalidate the assessment. Instead, the assessment is designed to produce evidence that remains attributable to the learner.

The need for such a shift is also supported by recent research on formative assessment in EFL environments. A 2026 study of GenAI integration into online formative assessment reported technical, operational, ethical, and equity-related barriers and emphasised the importance of human-centred approaches to AI-supported assessment [5]. These findings suggest that technology alone cannot

resolve assessment challenges. Teacher AI literacy, pedagogical judgement, institutional infrastructure, and transparent assessment policies must operate together.

Recent research also indicates that GenAI can be pedagogically productive when appropriately integrated into language learning. For example, research published in 2026 on GenAI-supported EFL argumentative writing investigated its relationship with writing performance, motivation, and enjoyment [6]. Similarly, research on secondary-school EFL writing has compared AI-generated and teacher feedback, demonstrating that the role of GenAI in language assessment can be examined as part of human–machine collaboration rather than only as a threat to assessment integrity [7]. These findings reinforce the argument that banning AI is not necessarily equivalent to creating valid assessment.

The emerging literature therefore reveals an important research gap. Existing scholarship has examined GenAI-assisted writing, automated scoring, teacher perceptions, AI-supported feedback, assessment validity, and AI-resilient assessment separately. However, comparatively less attention has been devoted to integrating these strands into a coherent **human-centred AI-resilient framework specifically designed for EFL learners in school contexts**. The need is not simply for another method of detecting AI-generated text but for a redesigned assessment architecture capable of generating multiple and complementary forms of evidence of learning.

Accordingly, this study addresses the following research questions:

RQ1: How does generative AI challenge the validity, authenticity, reliability, and fairness of conventional EFL assessment?

RQ2: Which assessment practices can preserve credible evidence of EFL learners' linguistic competence, reasoning, and ownership of learning in AI-mediated environments?

RQ3: How can these practices be integrated into a human-centred and AI-resilient assessment framework for EFL learners?

The purpose of this study is therefore to reconceptualise EFL assessment in the age of GenAI and to develop a theoretically grounded framework that preserves the evidential value of assessment while recognising the realities of AI-mediated learning.

The central proposition of the study is that AI-resilient assessment should not attempt to make assessment independent of AI; rather, it should make evidence of learning independent of the mere quality of an AI-produced product.

2. METHODS

2.1 Research design

The study employed a **conceptual research design based on structured literature synthesis**. This approach was selected because the purpose was not to measure the effectiveness of a previously validated intervention but to integrate recent theoretical and empirical findings into a coherent assessment framework.

The conceptual analysis was informed primarily by peer-reviewed research published between 2024 and 2026. Particular attention was given to studies addressing five interconnected areas: (1) GenAI and language assessment; (2) EFL writing and assessment; (3) validity and reliability of AI-supported assessment; (4) teacher agency and human judgement; and (5) AI-resilient and human-centred assessment design.

The methodological logic followed three stages:

Stage 1 — Problem identification. Recent research was examined to identify how GenAI disrupts the assumptions underlying product-oriented EFL assessment.

Stage 2 — Principle extraction. Recurring assessment principles were extracted from the literature, including authenticity, process evidence, transparency, human agency, oral verification, and teacher judgement.

Stage 3 — Framework synthesis. The extracted principles were integrated into a proposed Human-Centred AI-Resilient EFL Assessment Framework.

The purpose of the synthesis was therefore not to claim empirical validation of the proposed framework. Rather, it was to establish a theoretically and research-informed model suitable for subsequent empirical testing.

2.2 Source selection

The literature base prioritised recent peer-reviewed research published from 2024 onward, with particular emphasis on 2025 and 2026 publications because of the rapid evolution of GenAI technologies.

The selected literature included:

- systematic reviews of GenAI and language assessment;
- empirical studies of GenAI-assisted EFL writing;
- studies of AI-supported essay grading;
- research on secondary EFL assessment;
- research concerning teacher judgement and agency;
- conceptual studies of AI-resilient assessment;
- research on human-centred AI integration.

The 2025 systematic review by Chuang and Yan was particularly important because it synthesised 77 studies and explicitly examined reliability, validity, fairness, and practicality across assessment contexts [1]. The 2026 AI-resilient framework by Nikolić and Basta Nikolić was used as a conceptual reference point for process documentation, oral defence, authentic task design, and transparent AI-use policies [4].

2.3 Analytical procedure

The literature was analysed according to six questions:

1. What evidence of learning is produced by conventional EFL assessment?
2. How can GenAI weaken the relationship between the final product and learner competence?
3. What kinds of evidence remain difficult for GenAI to substitute for the learner?
4. What role should teachers retain in assessment?
5. How can AI use be made transparent rather than hidden?
6. How can assessment tasks generate multiple forms of evidence of learning?

The answers were synthesised into six dimensions of the proposed framework:

1. **Authenticity**
2. **Process Evidence**
3. **Human Agency**
4. **Oral Verification**
5. **Responsible AI Use**
6. **Multidimensional Judgement**

3. RESULTS

3.1 The assessment validity problem

The literature synthesis indicates that the most fundamental consequence of GenAI is not simply increased cheating risk but a weakening of the traditional relationship between the **assessment product** and the **learner construct**.

In conventional writing assessment, the final essay is treated as observable evidence from which a teacher infers the learner's linguistic competence. This inference becomes less secure when the final product can be substantially generated, reorganised, corrected, or enhanced by an external AI system.

The problem can therefore be represented as:

Traditional assessment:

Learner competence → Learner performance → Final product → Teacher judgement

Under widespread GenAI availability, the relationship may become:

Learner competence + AI capability → AI-assisted performance → Final product → Teacher judgement

The final product may consequently reflect both human and machine capabilities. The central validity problem is therefore to determine which aspects of the final performance can reasonably be attributed to the learner.

Chuang and Yan's systematic review confirms that this problem extends across reliability, validity, fairness, and practicality rather than being limited to academic integrity [1].

3.2 From product-centred to evidence-centred assessment

The synthesis suggests that EFL assessment should move from an exclusive focus on **product quality** toward a broader concept of **evidence of learning**.

The proposed transformation is:

Product-centred assessment

"How good is the final text?"

to:

Evidence-centred assessment

"What evidence demonstrates that the learner understands, can use, explain, revise, and defend the language represented in the final product?"

This distinction is crucial because GenAI can produce a high-quality product without necessarily transferring the underlying competence to the learner.

The proposed framework therefore treats the final product as **one component of evidence rather than the complete evidence of learning**.

3.3 The Human-Centred AI-Resilient EFL Assessment Framework

Based on the synthesis, this study proposes a six-dimensional framework.

Table 1. Human-Centred AI-Resilient EFL Assessment Framework

Dimension	Central question	Assessment evidence
1. Authenticity	Is the task meaningful and contextually relevant?	Real-world problem, audience, communicative purpose
2. Process Evidence	Can the learner demonstrate how the product developed?	Drafts, revisions, notes, planning, feedback responses
3. Human Agency	Can the learner make and justify independent decisions?	Individual choices, reasoning, personal interpretation
4. Oral Verification	Can the learner explain and defend the submitted work?	Speaking interview, follow-up questions, mini-defence
5. Responsible AI Use	Is AI use transparent and pedagogically appropriate?	AI-use statement, prompts, verification, reflection
6. Multidimensional Judgement	Does assessment capture more than final textual quality?	Writing, speaking, reasoning, reflection, interaction

The six dimensions are complementary rather than independent. Together, they create an assessment environment in which the learner is required to demonstrate competence through multiple forms of evidence.

3.4 Dimension 1: Authenticity

Authenticity is the first protective layer of the proposed framework.

Traditional language tasks often ask learners to produce standard essays or answer predetermined questions. Such tasks may be particularly vulnerable to GenAI because a learner can submit a polished response without demonstrating meaningful engagement with the underlying communicative problem.

Authentic assessment instead places language within a meaningful context. For example, rather than asking learners to "write an essay about environmental pollution," an authentic task might require them to prepare a short proposal for reducing plastic waste in their school, address a specific audience, compare possible solutions, and justify one recommendation.

Such a task remains assessable even when AI is permitted because the learner must demonstrate contextual understanding and decision-making.

3.5 Dimension 2: Process Evidence

Process documentation provides evidence that cannot be fully captured by the final product.

For writing tasks, teachers may collect:

- brainstorming notes;
- initial drafts;
- revisions;
- teacher or peer feedback;
- vocabulary decisions;
- structural changes;
- reflection notes.

The purpose is not to require excessive documentation for every assignment. Rather, selected high-value assessments should include sufficient process evidence to establish how the learner arrived at the final product.

This principle corresponds closely with the process-based documentation proposed in contemporary AI-resilient assessment research [4].

3.6 Dimension 3: Human Agency

Human agency refers to the learner's capacity to make meaningful decisions and take responsibility for the resulting performance.

An EFL learner may use AI to generate several possible introductions, for example, but the assessment should examine whether the learner can:

- choose an appropriate version;
- explain why it was selected;
- adapt it to the communicative context;
- identify weaknesses;
- revise the language;
- defend the final decision.

Thus, AI assistance does not automatically eliminate evidence of learning. The crucial issue is whether the learner remains an active decision-maker.

This dimension is especially important because research on EFL teachers has shown that GenAI disrupts established assessment routines and increases uncertainty concerning authorship and originality [2].

3.7 Dimension 4: Oral Verification

Oral verification is one of the strongest components of the proposed framework.

After completing a written task, the learner may be asked several short questions:

- Why did you choose this argument?
- What does this expression mean?
- Why did you change this sentence?
- What alternative vocabulary could you use?
- What was the most difficult part of the task?
- How would you explain your conclusion without referring to the written text?

The purpose is not to turn every writing assessment into a formal examination. Rather, a short oral component can provide additional evidence concerning comprehension, reasoning, vocabulary, and ownership.

The approach is consistent with the emerging AI-resilient assessment literature, which identifies oral defence as a means of shifting assessment toward demonstrable understanding and real-time reasoning [4].

3.8 Dimension 5: Responsible AI Use

A prohibition-based model assumes that AI should not be used. A responsible-use model instead distinguishes between appropriate and inappropriate uses.

For example:

AI use	Possible assessment status
Brainstorming ideas	Permitted with disclosure
Vocabulary suggestions	Permitted with verification
Grammar feedback	Permitted with learner revision
Generating the complete answer	Restricted or prohibited for independent-performance tasks
Translating the entire assignment	Restricted depending on learning objective
Submitting AI output as personal work	Not permitted
Reflecting on AI feedback	Encouraged where pedagogically appropriate

The central principle is **transparency**.

Learners should be taught that responsible AI use involves identifying when AI was used, understanding what it contributed, checking its accuracy, and retaining responsibility for the final work.

This approach is compatible with recent AI-resilient assessment scholarship, which emphasises transparent AI-use policies aligned with intended learning outcomes [4].

3.9 Dimension 6: Multidimensional Judgement

The final dimension is multidimensional assessment.

A single numerical score based on the final written product cannot adequately capture the complexity of AI-mediated EFL learning. The proposed model therefore recommends combining several evidence sources.

Table 2. Illustrative assessment distribution

Assessment component	Suggested weighting
Final language product	25%
Process evidence	20%
Oral verification	20%
Reasoning and decision-making	15%
Reflection	10%
Responsible AI-use practice	10%
Total	**100%**

These percentages are **illustrative rather than empirically validated**. They are intended to demonstrate how an assessment designer could reduce excessive dependence on the final product.

The precise weighting should depend on the intended learning outcomes, learner age, proficiency level, task type, and institutional requirements.

3.10 Assessment resilience as a design property

The framework suggests that assessment resilience should be treated as a **design property** rather than a detection technology.

An assessment is relatively resilient when:

1. the learner must engage with an authentic problem;
2. the learner's process is observable;
3. the learner makes meaningful decisions;
4. the learner can explain the submitted work;
5. AI use is transparent;
6. several forms of evidence contribute to the final judgement.

This can be represented as:

Assessment Resilience = Authenticity + Process Evidence + Human Agency + Oral Verification + AI Transparency + Multidimensional Judgement

This expression is conceptual rather than mathematical. It is intended to represent the framework's structural logic rather than to function as a psychometric equation.

4. DISCUSSION

The findings of the conceptual synthesis indicate that GenAI requires a fundamental reconsideration of what counts as evidence in EFL assessment. The central problem is not that AI makes assessment impossible. Rather, GenAI exposes a weakness that was already present in product-dominated assessment: the assumption that a high-quality product necessarily provides sufficient evidence of the learner's underlying competence.

The 2025 systematic review by Chuang and Yan provides strong support for this interpretation by demonstrating that GenAI affects multiple core assessment qualities, including validity, reliability, fairness, and practicality [1]. The present framework extends this discussion by translating those concerns into a specific EFL assessment architecture.

4.1 Why detection alone is insufficient

The literature suggests that detecting AI-generated text cannot constitute the central strategy for future EFL assessment.

First, detection focuses on the source of the product rather than the competence of the learner.

Second, learners may use AI in different ways, ranging from minor grammatical assistance to extensive content generation. A binary distinction between "AI-generated" and "human-generated" therefore oversimplifies the actual learning environment.

Third, teacher confidence in identifying AI-assisted writing is limited. Chen and Pi's 2026 study demonstrates that EFL teachers struggle to identify GenAI-generated content consistently and that their responses vary according to their experience and pedagogical beliefs [2].

The proposed framework consequently replaces the question:

"Can we detect AI?"

with:

"Can the learner demonstrate the intended learning outcome through multiple forms of evidence?"

This is a more educationally meaningful question.

4.2 Human judgement remains central

The framework does not propose the removal of AI from assessment. Nor does it propose replacing teachers with automated scoring systems.

Research on LLM-based essay grading demonstrates that AI systems can produce consistent rubric-based judgements in some dimensions [3]. This suggests that AI may have a useful supporting role in assessment.

However, consistency is not equivalent to complete validity. Language assessment involves contextual interpretation, communicative appropriateness, learner development, and pedagogical judgement. These dimensions require teachers to interpret evidence rather than simply calculate scores.

Therefore, the proposed model adopts a **human-in-the-loop** principle:

AI may support assessment; the teacher remains responsible for interpreting evidence and making consequential pedagogical judgements.

This position is also consistent with the finding that teacher agency is critical in EFL GenAI-assisted writing assessment [2].

4.3 Implications for secondary-school EFL education

The school context requires particular attention because younger EFL learners are simultaneously developing language proficiency, learning strategies, critical thinking, and digital literacy.

An assessment system that simply prohibits AI may prevent learners from developing responsible AI literacy. Conversely, an assessment system that permits unrestricted AI-generated answers may weaken opportunities for independent language development.

The proposed solution is graduated and transparent AI use.

For example, a teacher may permit AI brainstorming during one task while requiring independent writing during another. In a third task, AI may be explicitly integrated into the activity, with learners evaluated on their ability to critique and improve AI-generated language.

This creates three distinct assessment purposes:

Independent-performance assessment
→ measures what the learner can do without substantive AI assistance.

AI-supported learning assessment
→ measures how effectively the learner can use AI while maintaining human judgement.

AI-critical assessment
→ measures the learner's ability to identify, evaluate, correct, and improve AI-generated language.

Such differentiation is more pedagogically informative than treating AI use as universally acceptable or universally prohibited.

4.4 Relationship between AI resilience and authenticity

AI-resilient assessment should not be confused with "AI-proof" assessment.

An AI-proof task attempts to prevent AI from being used.

An AI-resilient task remains educationally meaningful even when AI exists.

This distinction is important because technological restrictions are temporary, whereas assessment principles can remain relevant despite technological change.

The 2026 AI-resilient assessment framework similarly conceptualises resilience around process documentation, oral defence, authentic tasks, and transparent AI policies rather than dependence on detection [4].

The present study extends this logic into EFL by adding a stronger emphasis on language-specific evidence: vocabulary choice, communicative appropriateness, linguistic reasoning, speaking performance, revision, and reflection.

4.5 Fairness and equity

Human-centred assessment must also consider unequal access to AI technologies.

Learners may differ in:

- access to advanced AI tools;
- digital literacy;
- familiarity with prompting;
- home technological resources;
- teacher support.

Consequently, allowing unrestricted AI use can unintentionally create new inequalities.

The 2026 study of GenAI in EFL formative assessment identified technical, operational, ethical, and equity barriers to implementation [5].

Therefore, a fair AI policy should not simply state "AI is allowed." It should specify:

- which tools are permitted;
- for what purposes;
- under what conditions;
- how AI use should be disclosed;
- what alternative support is available;
- which learning outcomes must be demonstrated independently.

Fairness requires **equivalent opportunities to demonstrate the intended construct**, not necessarily identical technological conditions.

4.6 The teacher's role in the new assessment ecology

The changing assessment environment requires an expanded conception of teacher expertise.

The teacher is no longer only a scorer of final products. The teacher becomes:

- assessment designer;
- evidence interpreter;
- AI-literacy facilitator;
- evaluator of learner reasoning;
- designer of authentic communicative tasks;
- moderator of AI-use policies.

This transformation does not reduce the importance of teachers. Instead, it increases the importance of professional judgement.

The framework therefore places **teacher judgement** at the end of the assessment chain, after multiple evidence sources have been collected.

5. CONCEPTUAL CONTRIBUTION OF THE STUDY

The principal contribution of this study is the development of a **Human-Centred AI-Resilient EFL Assessment Framework** that integrates six dimensions into a single assessment architecture.

Previous research has identified important individual components of the problem. Systematic language-assessment research has highlighted validity, reliability, fairness, and practicality [1]. EFL research has documented teachers' difficulties with GenAI-assisted writing [2]. Research on AI-supported grading has demonstrated potential for LLM-based scoring [3]. AI-resilient assessment scholarship has proposed process evidence, oral defence, authentic tasks, and transparent AI policies [4].

The present framework brings these strands together specifically for EFL assessment and adds a stronger emphasis on linguistic evidence and learner agency.

Its conceptual sequence can be summarised as follows:

Authentic task
↓
Learner planning and engagement
↓
Process evidence
↓
AI-supported or independent production, according to task purpose
↓
Revision and reflection
↓
Final language product
↓
Oral verification
↓
Teacher interpretation of multiple evidence sources
↓
Assessment judgement

This sequence represents a fundamental change from:

Prompt → AI-generated answer → Final score

to:

Learning process → Demonstrated competence → Evidence-informed judgement

6. CONCLUSION

Generative AI has not made EFL assessment impossible; it has made the limitations of product-centred assessment more visible. When an AI system can generate grammatically accurate and structurally sophisticated language, the final written product alone can no longer be assumed to provide sufficient evidence of individual learning.

The appropriate response is therefore not to construct an endless technological competition between AI generators and AI detectors. A more sustainable response is to redesign assessment around evidence that demonstrates what learners understand, how they reason, what decisions they make, and whether they can take ownership of their linguistic performance.

This study proposed a **Human-Centred AI-Resilient EFL Assessment Framework** comprising six interconnected dimensions: authenticity, process evidence, human agency, oral verification, responsible AI use, and multidimensional judgement. The framework reconceptualises assessment from a product-centred activity into an evidence-centred process.

The model also recognises that GenAI can have legitimate educational value. AI may support brainstorming, feedback, vocabulary development, revision, and language exploration when its use is transparent and aligned with learning objectives. The purpose of AI-resilient assessment is therefore not to exclude technology but to ensure that technology does not obscure the evidence of learning.

For secondary-school EFL education, this means that assessment designers should increasingly distinguish between independent-performance tasks, AI-supported learning tasks, and AI-critical tasks. Such differentiation can simultaneously protect assessment validity, promote responsible AI literacy, and prepare learners for authentic communication in technology-rich environments.

The proposed framework is conceptual and therefore requires empirical validation. Future research should test its effectiveness across different EFL proficiency levels, learner ages, school contexts, and assessment purposes. Quantitative studies could examine reliability and validity, while qualitative studies could investigate learner perceptions, teacher workload, agency, and fairness. Experimental research could also compare traditional product-based assessment with the proposed process-and-evidence-based model.

Ultimately, the central principle is straightforward: the goal of contemporary EFL assessment should not be to prove that a learner did not use AI; it should be to produce convincing evidence of what the learner can understand, create, explain, evaluate, and communicate**.** In this sense, AI resilience is not primarily a technological property. It is a property of thoughtful assessment design.

REFERENCES

1. Chuang, P.-L., & Yan, X. (2025). Language assessment in the era of generative artificial intelligence: Opportunities, challenges, and future directions. System, 134, 103846. https://doi.org/10.1016/j.system.2025.103846
2. Chen, X., & Pi, X. (2026). Assessing EFL students' GenAI-assisted writing: Teachers' pains, perceptions and practices. Assessing Writing, 67, 100996. https://doi.org/10.1016/j.asw.2025.100996
3. Yavuz, F., Çelik, Ö., & Yavaş Çelik, G. (2025). Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments. British Journal of Educational Technology, 56, 150–166. https://doi.org/10.1111/bjet.13494
4. Nikolić, D., & Basta Nikolić, M. (2026). Designing AI-resilient assessment in higher education: A four-pillar conceptual framework. Frontiers in Artificial Intelligence, 9, 1841682. https://doi.org/10.3389/frai.2026.1841682
5. "We cannot incorporate them because …": Challenges of integrating generative AI into online formative assessments in EFL contexts. (2026). Computers in Human Behavior Reports, 21, 100934. https://doi.org/10.1016/j.chbr.2026.100934
6. Chen, H., Rasool, U., Wang, R., & Dong, J. (2026). Integrating generative AI in EFL classrooms: A mixed-methods study on argumentative writing performance, writing enjoyment and motivation. Acta Psychologica, 268, 107216. https://doi.org/10.1016/j.actpsy.2026.107216
7. Acan, Y. G., & Liman Kaban, A. (2026). Human vs. machine feedback: Evaluating ChatGPT-4 in the assessment of secondary EFL learners' writing. Asia Pacific Journal of Educators and Education, 41(1). https://doi.org/10.21315/apjee2026.41.1.8
8. Zhang, T., Erlam, R., & Botelho de Magalhães, M. (2025). Exploring the dual impact of AI in post-entry language assessment: Potentials and pitfalls. Annual Review of Applied Linguistics. https://doi.org/10.1017/S0267190525000030
9. UNESCO. (2024). AI competency framework for teachers. UNESCO. UNESCO — AI Competency Framework for Teachers