

Assessing Learning after Generative AI Support: A Framework for Arabic Language Education

Dr. Mahmoud Abd ElFattah Ibrahim Essa*

Academy of Language Studies, Universiti Teknologi MARA (UiTM), Terengganu, Malaysia

*Corresponding Author

DOI: <https://doi.org/10.47772/IJRISS.2026.1026EDU0584>

Received: 10 September 2026; Accepted: 15 September 2026; Published: 24 September 2026

ABSTRACT

With support, generative artificial intelligence (GenAI) can make a language-learning task appear more successful. However, this visible improvement does not necessarily reflect what the learner can produce independently. This conceptual paper thus poses a more specific assessment question: When can the gains be considered as learning in Arabic language education in the context of GenAI-supported activity? To answer that question, the framework is based on research on the performance after support, cognitive offloading and engagement with AI, performance-learning differences, retention, transfer, and Arabic or second-language assessment. It identifies 5 categories of evidence: Supported Performance, Immediate Independent Performance, Delayed Retention, Near Transfer, and Far Transfer. These are viewed from three related perspectives: independence from AI assistance, temporal persistence, and task/context generalization. The framework in practice distinguishes between what a learner can do with GenAI, what remains after the support is withdrawn, what remains after a justified delay, and what can be used when the task changes. The five design considerations that are added to the Assessment-Validity Matrix are: AI availability, assessment timing, task similarity or distance, language output, and baseline or comparison conditions. The Arabic language target remains evident, including the lexical, grammatical, morphological, orthographic, sentence-level, and connected-production outcomes. Five propositions are presented for further empirical testing. The categories are not required to be presented in a specific order: time and task distance are independent of each other, and a study might only include the conditions necessary for the claim. The framework is not an empirically established model, but rather a conceptual specification that helps to ensure that the evidence generated by the assessment design remains close to the learning claims.

Keywords: Generative AI, post-support learning, delayed retention, learning transfer, Arabic language education

INTRODUCTION

The use of generative AI tools has been growing in importance in second-language (L2) pedagogy, where they play important roles in written composition, feedback generation, text revision, and interactive practice (Hong & Shin, 2025; Hwang et al., 2025; Lukešová & Jennings, 2026; Moeiniasl, 2026). More specifically, Hwang et al. (2025) point out that it can be used for initial brainstorming to advanced language polishing and error correction. These automatic interventions often improve the quality of the final product but also make it harder to see what the learner has really contributed.

It is very important to clarify this ambiguity because increased task performance does not necessarily mean actual learning acquisition. AI can overestimate performance results without the acquisition of independently verifiable knowledge (Bassner et al., 2026). In this regard, Jin et al. (2026) distinguish between tasks that need to be carried out immediately after the intervention and those that require retention and knowledge transfer. Performance is assessed as soon as the AI support is removed, rather than inferred from supported performance. Temporal delay testing is required for true persistence, while task variation is required for adaptability. For example, Namaziandost and Çakmak (2025) compared immediate and delayed

vocabulary acquisition, and Baalousha and Malovrh (2023) compared immediate and four-week delayed testing for Palestinian Arabic learners.

This diagnostic challenge is especially strong in the field of Arabic L2 teaching. While AI-generated drafts can be structurally sound, with correct syntax, lexical selection, morphology, and orthography, learners may not perform as well when it comes to autonomous writing. Empirical research that integrates GenAI interventions, scaffolding reduction, delayed testing, and transfer assessment in the context of Arabic, however, is limited. The proposed framework explicitly defines five categories of evidence: Supported Performance, Immediate Independent Performance, Delayed Retention, Near Transfer, and Far Transfer, to help fill this conceptual gap.

The paper addresses three research questions:

RQ1. How can Supported Performance, Immediate Independent Performance, Delayed Retention, Near Transfer, and Far Transfer be conceptually distinguished in the assessment of Arabic language learning after GenAI-supported activity?

RQ2. What conditions of AI availability, assessment timing, and task relation should be considered when designing post-support assessments to support claims about independence, temporal persistence, and generalization in Arabic language learning?

RQ3. What theoretically grounded and empirically testable propositions can be formulated concerning relationships between GenAI-supported activity, assessment conditions, and post-support independent performance, retention, and transfer in Arabic language learning?

Together, these questions shift attention from the quality of AI-assisted products to the evidential basis for claims about learning after support. Withdrawal of AI allows independence from continuing support to be examined; delayed assessment addresses temporal persistence; and changed-task assessment addresses generalization. These dimensions are not stages on a single ladder, and time and task distance need not vary together. The appropriate combination depends on the claim being made: evidence of immediate independence does not by itself establish persistence, and persistence on a familiar task does not by itself establish transfer. Claims about durable transfer therefore require evidence that addresses independence, temporal persistence, and generalization together.

From Ai-Assisted Performance to Evidence of Learning

The distinction developed in this paper concerns the evidence that different assessment conditions can provide, rather than whether GenAI support is educationally desirable. A successful product created with AI may reflect useful scaffolding, feedback, or learner–AI collaboration, but it does not by itself establish what the learner can perform independently. Supported performance and post-support performance therefore address different questions about learning. Evidence from programming education illustrates how assisted performance may diverge from independently measured learning (Bassner et al., 2026), while post-agent assessment distinguishes performance observed after support has been removed from performance achieved while an AI agent remains available (Jin et al., 2026). These distinctions provide a basis for separating four forms of post-support evidence from performance observed during AI-supported activity.

Supported Performance

Supported Performance refers to performance produced while GenAI remains available. It provides meaningful evidence about what the learner can accomplish with AI support, but it does not by itself establish that the same target capability is independently available once that support is removed.

Immediate Independent Performance

Immediate Independent Performance is assessed after GenAI support has been withdrawn. It provides evidence about whether the learner can perform the target capability without continuing AI assistance.

Because the assessment occurs immediately or near-immediately after the supported activity, however, it does not by itself establish that the capability persists over time (Jin et al., 2026).

Delayed Retention

Delayed Retention addresses temporal persistence by examining whether the target capability remains independently available after a justified interval without GenAI support, interpreted relative to an appropriate prior no-AI reference point. The target should remain comparable across assessments without reducing the delayed task to simple reproduction. Delayed measures have been used to examine vocabulary retention after AI-supported learning (Namaziandost & Çakmak, 2025) and longer-term retention in Arabic language learning without GenAI (Baalousha & Malovrh, 2023).

Near Transfer and Far Transfer

Transfer concerns whether a target capability can be applied under conditions that differ from the original supported activity. Near Transfer examines generalization to a relatively similar but modified task, whereas Far Transfer examines generalization under substantially different task or contextual conditions. These categories represent different ranges of generalization rather than an automatic hierarchy of evidential strength. Greater task distance therefore changes the scope of the transfer claim that can be examined; it does not by itself make the evidence stronger. Task distance is also distinct from elapsed time, so either form of transfer may be assessed immediately or after a delay. Modified writing and language tasks provide useful methodological precedents for examining transfer across changed conditions (Diab, 2015; Pust, 2026; Zhuang et al., 2025; Şimşek et al., 2025).

The five categories thus represent different kinds of assessment evidence rather than stages in a fixed sequence. A study need not include every category; the appropriate combination depends on whether the intended claim concerns supported performance, independence from AI support, temporal persistence, generalization, or some combination of these.

Why Post-Support Assessment is Necessary

Cognitive Offloading and Overreliance

Post-support assessment becomes especially relevant when GenAI performs or substantially contributes to operations that instruction is intended to develop in the learner, such as generating linguistic forms, selecting wording, diagnosing errors, or proposing revisions. Cognitive offloading is relevant to this problem because GenAI may perform part of the target-relevant cognitive work involved in completing a task (Jiang & Chen, 2026). Offloading is not inherently detrimental; external support may facilitate productive learning processes as well as reduce the amount of processing performed directly by the learner (Schorr et al., 2026). The interpretive issue arises when AI repeatedly performs operations that the learner is expected to develop independently. Research on AI dependence further suggests that patterns of reliance may be associated with learner outcomes, although cross-sectional evidence cannot establish a causal relationship between dependence and learning (Ren & Khan, 2026). Evidence that assisted performance can diverge from independently measured performance reinforces the need to distinguish successful task completion with AI from what remains available after support is withdrawn (Bassner et al., 2026).

Support Quality, Regulation, and Learner Engagement

The way learners engage with GenAI also varies substantially. Learners may verify suggestions, revise AI output, or adopt recommendations selectively (Kim, 2026), and they differ in how they regulate their use of AI during learning activities (Tao et al., 2026). Interactive AI-generated error-correction activities provide another example of how engagement with AI may be related to knowledge application and metacognitive processes (Yan et al., 2026). Other studies report learner perceptions of AI-supported learning experiences (Meishar-Tal & Amzalag, 2026) and associations between learner characteristics, engagement, and assisted writing performance (Khan & Habib, 2026).

These findings indicate that GenAI-supported activity is not a uniform condition. Differences in verification, revision, selective adoption, interaction, and regulation provide variables that future research can examine in relation to post-support outcomes. They do not, however, establish that guided or regulatory forms of use are universally superior to less constrained forms of GenAI use, nor do self-report measures or assisted performance alone establish what remains independently available after support is removed.

Framework Development Approach

The framework was developed through a structured conceptual and integrative synthesis. This approach was appropriate because the assessment problem spans bodies of literature that differ in their interventions, outcomes, and disciplinary assumptions. The evidence considered included empirical studies of AI-supported learning, conceptual work on generative learning and cognitive offloading, language-assessment research, and Arabic-specific studies using delayed assessment. Rather than treating these different designs as estimates of a common effect, the synthesis focused on recurring distinctions in the assessment evidence they provide. It was interpretive rather than statistical and was not intended to provide exhaustive coverage of the literature or a pooled estimate of GenAI effects.

Evidence Scope and Four Tracks

The evidence was organised around four complementary tracks: (1) GenAI-supported learning examined through post-support outcomes; (2) cognitive offloading and engagement with AI; (3) performance, retention, and transfer; and (4) Arabic and second-language assessment. The first included studies reporting independent, delayed, or transfer-related outcomes (Jin et al., 2026; Namaziandost & Çakmak, 2025), while the second drew attention to offloading and engagement during AI-supported activity (Jiang & Chen, 2026; Schorr et al., 2026).

The third track gathered evidence that supported the idea of performance that is distinct from independently measured learning (Bassner et al., 2026) and Dabbagh and Fake's (2026) conceptualization of GenAI as a mindtool that is designed to support, not supplant, generative learning. In the fourth track, Baalousha and Malovrh (2023) gave an Arabic-specific example of delayed assessment while Diab (2015) gave another example of repeated language assessment.

These four tracks were not intended to be an exhaustive classification of the literature but rather were used as organising perspectives for the development of the frameworks. Learner perceptions, self-report and tool output evidence may be informative for understanding the activity supported by AI, but evidence alone is not sufficient to support claims about what learners can do independently after AI support is removed, whether this capability is maintained over time, or whether it generalizes to changed conditions.

Evidence Integration

The four tracks of evidence were interpreted using the type of inference that the assessment design could reasonably support: performance while assistance was available, performance after the withdrawal of assistance, persistence after an interval, or application under changed task conditions. The three dimensions of the framework were guided by these differences: independence from AI support, temporal persistence and generalization.

The same distinctions were then applied to the five evidence categories and added to the Assessment-Validity Matrix. The framework connects the type of learning claim to the appropriate assessment conditions (such as support, timing, task relation, language output, and baseline or comparison conditions) without implying that the categories are sequential steps in a process.

Epistemic Boundaries

The framework ensures that the claims are contained within the scope of the evidence reported. The results of programming education can be used to support a general distinction between assisted performance and independently measured learning, but cannot be used to draw the same conclusion in Arabic language

education (Bassner et al., 2026). Similarly, delayed assessment in Arabic learning without GenAI illustrates the value of delayed assessment in Arabic learning, but does not prove the efficacy of a post-GenAI assessment framework (Baalousha & Malovrh, 2023).

The resulting framework is thus a testable conceptual structure for further research on Arabic language education. It is not offered as an integrated model that has been proven in practice, as a universal model of learning with GenAI, or as evidence of a causal effect.

The Post-Support Learning Assessment Framework

The Post-Support Learning Assessment Framework is not a psychometric scale, but rather a tool for decision making based on the evidence. It aims to ensure that the learning claim remains within the scope of the assessment design. Researchers can then only apply evidence categories that are pertinent to their research question. Claims about independently available capability, temporal persistence, or generalization require assessment conditions that address those dimensions, but a study focused on supported performance does not need to assess retention or transfer.

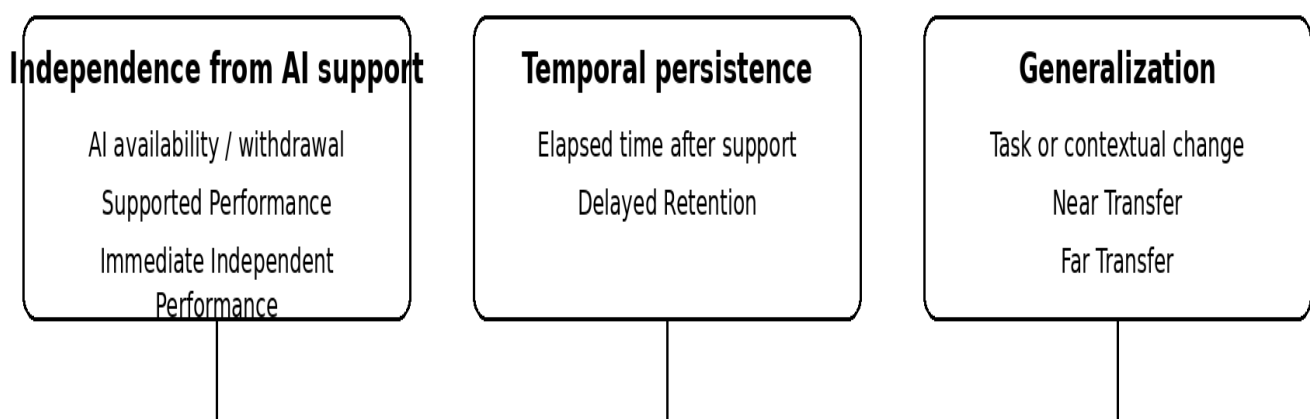
The framework is organised around three questions: Is GenAI support available during assessment? Is there any time that has passed since the supported learning activity? How different is the assessment task from the original supported task? These questions fall into three categories:

AI support independence, temporal persistence, and task/context generalization. The five evidence categories are not stages or an order of progression. There are also two types of distances: temporal distance (immediate or delayed) and task distance (similar or different) depending on the type of claim.

Figure 1. Relationship among the Three Dimensions and Five Evidence Categories

Figure 1 represents how the three dimensions relate to the five evidence categories. Withdrawal of GenAI makes independence from continuing AI assistance assessable; temporal distance provides a basis for examining persistence; and task or contextual change provides a basis for examining generalization. Near Transfer and Far Transfer represent different ranges of generalization according to task distance rather than different levels of evidential quality.

Three assessment dimensions and corresponding evidence categories



The categories are evidential conditions rather than fixed stages. Temporal distance and task distance are distinct features of assessment design.

Five Evidence Categories

Table 1 specifies the five evidence categories according to AI availability, timing, task relation, the claim supported by the assessment condition, and the primary evidential focus.

Table 1. Five Evidence Categories in the Post-Support Learning Assessment Framework

Evidence Category	AI Availability	Timing	Task Relation	What the Evidence Supports	Primary Evidential Focus
Supported Performance	Available	During supported activity	Original or directly supported task	What the learner can accomplish with GenAI support; does not by itself establish independently available capability	Assisted performance
Immediate Independent Performance	Unavailable	Immediate or near-immediate after support	Same or highly similar target	Whether the learner can perform the target capability without continuing AI assistance at that point	Independence
Delayed Retention	Unavailable	After a justified delay, with an appropriate prior reference point for interpreting persistence	Comparable target without simple reproduction	Whether independently available capability remains evident after a justified delay relative to an appropriate prior reference point	Temporal persistence
Near Transfer	Unavailable	Immediate or delayed post-support	Modified but relatively similar task or context	Whether the target capability generalizes to relatively similar changed conditions	Generalization
Far Transfer	Unavailable	Immediate or delayed post-support	Substantially different task or context while the target remains identifiable	Whether the target capability generalizes to substantially different conditions	Generalization

The categories identify different forms of assessment evidence and may be selected according to the claim being examined rather than treated as levels in a sequence.

Temporal Conditions

Timing determines whether an assessment can address temporal persistence. When persistence or retention is part of the intended claim, performance should be assessed after a justified interval with GenAI support removed. The delayed performance should be interpreted relative to an appropriate prior no-AI assessment of the same target capability. The interval should be reported and justified in relation to the construct and research context and not as a universal threshold. The design choice is supported by methodological precedents in Arabic language learning and AI-supported vocabulary learning (Baalousha & Malovrh, 2023; Namaziandost & Çakmak, 2025).

Support-Withdrawal Conditions

Availability of AI should be clearly indicated as it can be unrestricted, restricted, or withdrawn, which results

in different assessment conditions. Restricted access is not considered as withdrawal when the intended claim is about independently available capability, but may be used for the examination of fading or modified support. Researchers should therefore report which functions of the AI (if any) are available for assessment and match the claim to that condition.

Task-Distance Conditions

Task distance should not be taken for granted. Changes of topic may result in minimal difference, while changes of purpose, genre, modality, or reasoning demands may result in significant difference. Researchers should indicate what has changed and why the target capability is still identifiable and comparable between conditions. The range of generalization that can be reasonably explored is determined by the degree of task distance: Near Transfer may be supported by relatively similar changed conditions, while Far Transfer may be supported by substantially different conditions. Greater task distance alone does not necessarily provide stronger evidence; it alters the range of generalization that can be interpreted as long as the target capability is identifiable and comparable.

Assessment-Validity Matrix

The Assessment-Validity Matrix is designed to complement Table 1 by highlighting design conditions that influence the interpretation of an intended claim. The matrix does not just consider the nature of the evidence that is used as an assessment condition (as identified in Table 1), but also whether key features of the design facilitate or hinder the intended inference. It covers the availability of AI, assessment timing, similarity or distance of the tasks, language output, and baseline/comparison conditions.

The matrix is interpretive rather than additive. It assigns no numerical score, and its dimensions are not components of a hierarchy of evidential strength. Each dimension instead identifies conditions that may limit or support the particular inference a researcher intends to make.

Table 2. Assessment-Validity Matrix Dimensions and Interpretive Questions

Assessment Dimension	Condition Limiting the Claim	Condition Supporting the Intended Claim	Interpretive Question
AI Availability	GenAI remains available and may generate, correct, reformulate, or otherwise support the target performance, limiting claims about what the learner can perform independently.	GenAI is unavailable during assessment and the learner must perform the target capability without direct AI assistance, supporting a claim about independence from AI support.	To what extent does the observed performance reflect the learner's own available knowledge or skill rather than continuing AI assistance?
Assessment Timing	Performance is measured only during or immediately after AI-supported activity, limiting claims about whether the target capability persists over time.	Performance is assessed after a justified delay with AI support removed, allowing a claim about whether the target capability remains independently available after time has passed.	Does the target knowledge or skill remain independently available after a justified interval, rather than appearing only as an immediate post-support effect?
Task Similarity / Distance	Assessment closely reproduces the original supported task, prompt, or response conditions, limiting claims about generalization beyond familiar conditions.	Assessment requires application of the same identifiable and comparable target capability under meaningfully modified or substantially different conditions, supporting Near Transfer or Far Transfer claims appropriate to the degree of task	To what extent can the same target capability be applied beyond the original supported conditions, and what range of generalization does the changed task justify?

		change.	
Language Output	The assessed output differs from, obscures, or only indirectly represents the linguistic capability targeted by instruction, limiting interpretation of performance in relation to that target.	The assessment elicits observable output that remains clearly aligned with the same target linguistic capability, supporting an interpretable claim about that capability.	Does the assessment provide interpretable evidence about the specific lexical, grammatical, structural, or communicative capability that the learning claim concerns?
Baseline / Comparison Conditions	No meaningful baseline or comparison condition is available, limiting interpretation of observed change or differences in post-support performance.	An appropriate baseline and/or comparison condition provides a reference point for interpreting change or group differences alongside post-support performance.	What change has occurred relative to the relevant starting point or comparison condition, and how should that change qualify the interpretation of post-support evidence?

Table 2 therefore serves a different function from Table 1: it evaluates design features that delimit or support an intended inference rather than assigning an assessment to an evidence category.

Evidence Interpretation Rule

The central interpretation rule is that the learning claim should not extend beyond the evidence provided by the assessment design. Claims about durable transfer require evidence addressing independence from continuing AI support, temporal persistence, and generalization together. Tool capability should not be treated as learner capability: what GenAI can generate, correct, or reformulate does not by itself establish that the corresponding capability is independently available to the learner.

Application to Arabic Language Education

Arabic language education provides a useful context for applying the framework because GenAI can directly support observable linguistic capabilities such as lexical choice, grammatical and morphological control, orthographic accuracy, sentence-level production, and connected production. These capabilities can therefore be assessed under post-support conditions selected according to the specific claim being examined.

Why Arabic Requires Post-Support Measurement

In Arabic learning, GenAI may suggest lexical alternatives, correct grammatical or morphological forms, improve orthographic accuracy, reformulate sentences, or assist with connected production. When such support contributes substantially to the target performance, post-support assessment can be used to determine whether the same linguistic capability remains independently available, persists over time, or generalizes under changed conditions. The relevant assessment condition should therefore be selected in relation to the Arabic linguistic target and the intended inference rather than applied as a universal sequence.

Arabic Retention Evidence

Existing Arabic-language research provides methodological precedents for delayed assessment without directly validating the present framework. Baalousha and Malovrh (2023), for example, used immediate and four-week delayed assessment in the learning of Palestinian Arabic. Because GenAI was not involved, the study demonstrates the methodological value of examining performance beyond an immediate post-test rather than providing empirical validation of a post-GenAI framework.

Repeated and delayed language-assessment designs reported by Diab (2015) and Pust (2026) provide additional methodological precedents for examining performance across time or changed assessment

conditions. Their relevance to the present framework is methodological rather than confirmatory and does not establish the validity of the five evidence categories in Arabic GenAI-supported learning.

Illustrative Arabic Assessment Sequence

An illustrative application can be constructed around a clearly specified Arabic grammatical target. During GenAI-supported activity, learners might receive assistance in producing sentences that require the target structure. If the research question concerns independence, a subsequent task can require learners to produce new sentences using the same structure without GenAI. If temporal persistence is also relevant, the same identifiable grammatical capability can be assessed again after a justified interval without AI support.

Generalization can be examined through changed tasks while maintaining the same identifiable and comparable target capability. A Near Transfer condition might require the grammatical structure in a modified but relatively similar task, such as a different topic or sentence context. A Far Transfer condition might require the same structure under more substantially changed conditions, such as a different genre or modality. Near Transfer and Far Transfer represent different ranges of generalization rather than an evidential hierarchy. If there is a change of genre/modality, the grammatical target should be sufficiently recognizable and comparable, so that any differences in performance can be understood in terms of generalization.

Lexical choice, morphological control, orthographic accuracy, sentence level production, and connected production can all be handled in the same way. For each, the intended capability should be clearly and consistently recognisable and measurable under the assessment conditions that are relevant to the intended claim. If not, observed differences could be due to an alteration in the construct being evaluated, not a generalisation of the original capability.

This sequence is only representative and not mandatory. The framework does not specify a universal retention interval, a fixed degree of task distance, or a standard transfer task for Arabic language learning; studies should only include the assessment conditions required by the claims they are making.

THEORETICALLY GROUNDED PROPOSITIONS

The five propositions below are the translation of the framework into theoretically informed and empirically testable expectations. They are not given as facts or predictions.

P1. When a significant amount of the target-relevant cognitive load is delegated to GenAI, AI-supported performance may overstate the learner's independently available capability. Improved task performance can be accompanied by cognitive offloading and changes in regulation, but not necessarily by an increase in independently available capability (Jiang & Chen, 2026; Schorr et al., 2026; Bassner et al., 2026).

P2. Immediate Independent Performance is evidence of independence from ongoing AI support, but not temporal persistence or transfer. When the intended claim is persistence or generalization (Jin et al., 2026), delayed or changed-task conditions are necessary for independence from continued assistance to be assessed, and removing GenAI makes independence from continued assistance assessable.

P3. Delayed assessment without GenAI adds evidence about the temporal persistence of independently available capability when interpreted relative to an appropriate prior reference point; Immediate Independent Performance alone cannot establish such persistence. Temporal persistence is interpretable when performance after a justified delay is considered relative to an appropriate prior reference point, such as an immediate post-support no-AI assessment or another relevant earlier no-AI assessment of the same target capability (Namaziandost & Çakmak, 2025; Baalousha & Malovrh, 2023).

P4. Transfer evidence supports broader claims of generalization when learners can apply the same target capability under meaningfully changed task or context conditions. Near Transfer and Far Transfer represent different ranges of generalization rather than an automatic hierarchy of evidential strength.

Changed or repeated language tasks provide methodological precedents for examining performance under changed conditions (Diab, 2015; Pust, 2026; Zhuang et al., 2025; Şimşek et al., 2025).

P5. Guided or regulatory GenAI use may be associated with different post-support outcomes from unrestricted substitution-oriented use, including differences in Immediate Independent Performance, Delayed Retention, and transfer. Features such as verification, revision, selective adoption, interactive error correction, integration level, guided use, and feedback timing provide a basis for testing this expectation (Kim, 2026; Yan et al., 2026; Hong & Shin, 2025; Lukešová & Jennings, 2026; Wiboolyasarin et al., 2026). Existing evidence does not establish a universal causal advantage for guided use.

Table 3 summarizes the five propositions and their possible operationalization.

Table 3. Theoretically Grounded Propositions and Possible Operationalization

Proposition	Conceptual Expectation	Possible Operationalization	Primary Evidence
P1	AI-supported performance may provide an inflated indication of independently available capability when substantial target-relevant cognitive work is offloaded to GenAI.	Compare performance on the same identifiable target capability while GenAI support is available with performance immediately after that support is withdrawn. Where relevant, measures of offloading or engagement may be examined alongside the performance contrast.	Supported Performance and Immediate Independent Performance on a comparable target capability.
P2	Immediate Independent Performance provides evidence of independence from continuing AI support but does not by itself establish temporal persistence or transfer.	Remove GenAI before assessment and require the learner to demonstrate the same identifiable target capability immediately or near-immediately without direct AI assistance. Add delayed or changed-task assessment only when persistence or generalization is also part of the intended claim.	Immediate Independent Performance after withdrawal of GenAI support.
P3	Delayed assessment without GenAI adds evidence of temporal persistence when interpreted relative to an appropriate prior reference point; Immediate Independent Performance alone cannot establish such persistence.	Assess the same identifiable target capability after a justified delay without GenAI and interpret the delayed performance relative to an appropriate prior reference point, such as Immediate Independent Performance, a no-AI baseline assessing the same target capability, or another relevant earlier no-AI assessment.	Delayed Retention evidence interpreted relative to an appropriate prior reference point for assessing temporal persistence.
P4	Transfer evidence supports broader claims of generalization when the same target capability can be demonstrated under meaningfully changed task or context conditions. Near Transfer and Far Transfer represent different ranges of generalization rather than an automatic hierarchy of	Assess the same identifiable and comparable target capability under one or more meaningfully changed task or contextual conditions. The degree and form of task distance should be specified so that the resulting performance can be interpreted in relation to the range of generalization being examined.	Near Transfer and/or Far Transfer performance without AI support, according to the generalization claim being tested, with the target capability remaining identifiable and comparable across assessment

	evidential strength.		conditions.
P5	Guided or regulatory GenAI use may be associated with different Immediate Independent Performance, Delayed Retention, and transfer outcomes from unrestricted substitution-oriented use, with the direction and magnitude of these differences requiring empirical testing.	Compare guided or regulatory GenAI conditions with unrestricted substitution-oriented conditions while holding the target capability and relevant post-support assessment conditions as comparable as possible. Assess the post-support outcome or outcomes required by the specific claim being tested.	Immediate Independent Performance, Delayed Retention, and/or transfer outcomes across contrasted GenAI-use conditions, as relevant to the claim being tested.

Note. P1–P5 are theoretically grounded propositions for future testing, not established causal findings.

DISCUSSION

The framework emphasizes that an AI-assisted product, delayed performance, and performance under changed task conditions are evidence of non-interchangeable interpretations of learning. Post-support assessment thus does not negate, but rather extends, the research literature on AI-assisted performance, by making explicit what can and cannot be supported under various assessment circumstances.

As regards RQ1 and RQ2, the key point is that independence from AI support, temporal persistence, and generalization are conceptually different dimensions that are relevant for different claims. A learner might show evidence of Immediate Independent Performance but not show evidence of persistence to a similar task, or show evidence of persistence to a similar task but not show evidence of Immediate Independent Performance. Validity of the assessment, in turn, depends on the design being aligned with the inference being made, not on the evidential hierarchy of the various types of post-support evidence.

RQ3 builds on this logic by considering characteristics of activity supported by GenAI as variables for empirical investigation instead of causal mechanisms. The influence of cognitive offloading, regulation, learner engagement, support configuration, and prompt design on subsequent independence, temporal persistence, or generalization can be investigated. The research on selective adoption and revision, self-regulated AI use, structured or feedback-oriented support, and prompt design shows significant differences in the configuration and experience of using GenAI-supported activity (Kim, 2026; Tao et al., 2026; Guo et al., 2024; Hong & Shin, 2025; Moeiniasl, 2026; Algobaei & Alzain, 2026). This variation offers a foundation for examining the relationship between the use of GenAI and post-support outcomes, but does not imply a universal causal benefit of any specific type of GenAI use.

A broader implication is that tool capability and learner capability should remain analytically separate. Evidence concerning the accuracy, sophistication, or statistical performance of an AI system describes properties of the tool and cannot by itself establish that the corresponding linguistic capability is independently available to the learner (Almarashdi et al., 2026). In Arabic language education, maintaining this distinction allows GenAI-supported performance and post-support learner capability to be examined without presuming either beneficial or detrimental effects on learning.

IMPLICATIONS FOR RESEARCH AND ASSESSMENT DESIGN

The scope of assessment should follow the intended claim. No single study is required to include all post-support dimensions, and assessment components should be selected according to the inference being examined. Evidence addressing independence from AI support, temporal persistence, and generalization together is required only when the intended claim combines these dimensions, as in a claim about durable transfer.

Other design choices should likewise reflect the intended inference. When a claim concerns change from a starting point or differences between groups or conditions, an appropriate baseline and/or comparison condition should provide the relevant reference for interpretation. Retention intervals should be reported and justified in relation to the construct and research context rather than treated as universal thresholds; delayed Arabic-language assessment and AI-supported vocabulary research provide methodological precedents for this principle (Baalousha & Malovrh, 2023; Namaziandost & Çakmak, 2025). Claims of temporal persistence should also be interpreted relative to an appropriate prior no-AI assessment of the same target capability. Transfer designs should specify the form and degree of task distance and explain why the same target capability remains identifiable and comparable across assessment conditions.

Reporting should specify GenAI availability during assessment, assessment timing relative to the supported activity, task relation or distance, the target language output or capability being measured, and baseline or comparison conditions where applicable. Labels for outcomes should describe what was actually measured, such as vocabulary retention, grammatical accuracy, writing performance, knowledge application, or metacognition, and not just learning. Conditions of delayed or changed tasks should be added only if necessary for the intended claim.

LIMITATIONS AND FUTURE RESEARCH

There is no universal retention time, task distance, or minimum performance criterion specified in the framework. These operational decisions are dependent on the construct being evaluated and the context of the research. The framework therefore remains a conceptual structure that has not been validated empirically.

The evidence base is also limited with regard to claims for its use in Arabic language education. Evidence of the use of GenAI-supported learning and withdrawal, delayed retention, or transfer is scarce and specific to Arabic. Evidence from non-Arabic, non-language, and other disciplinary contexts can be used to inform conceptual distinctions, but not to establish equivalent effects and/or the direction and magnitude of such effects in the case of Arabic language learning.

Future research should explore the relationship between target relevant cognitive offloading and discrepancies between supported and post supported performance; between Immediate Independent Performance and Delayed Retention; and between task distance and the range of generalization that can be interpreted while the target capability is still identifiable and comparable. It should also evaluate if there are differences in post-support outcomes as a function of different GenAI use configurations, and if the framework is interpretable across linguistic targets, learner populations, and assessment contexts. The framework can be supported, refined, revised or challenged by empirical findings.

CONCLUSION

This paper differentiates between AI-assisted performance and evidence that a target capability is available to the learner without the assistance of AI. The assessment conditions that are relevant to the learning claims are different: temporal distance is relevant to temporal persistence, task distance is relevant to generalization. Evidence for claims of durable transfer must cover all three aspects: independence from AI support, temporal persistence, and generalization.

The framework does not claim that GenAI enhances or diminishes the capability to learn Arabic language skills that are available, persistent, and generalizable without the use of GenAI. It is a model which is conceptual and testable, not an empirically validated model. It is based on the idea that the learning claim must be contained within the scope of the assessment design.

Statements and Declarations

Funding: This research received no external funding.

Conflicts of Interest: The author declares no conflict of interest.

Data Availability: This conceptual framework paper generated no new empirical data. All sources used in developing the framework are cited in the manuscript.

REFERENCES

1. Algobaei, F., & Alzain, E. (2026). Prompt engineering for non-native English learners: A generative AI approach to personalised language feedback. *Social Sciences & Humanities Open*, 13, 102341. <https://doi.org/10.1016/j.ssaho.2025.102341>
2. Almarashdi, H. S., AlArabi, K., Abu Khurma, O., & Fidalgo, P. (2026). The illusion of knowledge? Investigating the statistical precision of ChatGPT and Gemini and their capacity to foster enduring students' conceptual learning. *Social Sciences & Humanities Open*, 14, 102902. <https://doi.org/10.1016/j.ssaho.2026.102902>
3. Baalousha, H., & Malovrh, P. A. (2023). Guided induction and deductive instruction: Analyzing efficiency versus depth of processing in the L2 acquisition of Palestinian Arabic. *Ampersand*, 11, 100140. <https://doi.org/10.1016/j.amper.2023.100140>
4. Bassner, P., Lenk-Ostendorf, B., Beinstingel, R., Wasner, T., & Krusche, S. (2026). Less stress, better scores, same learning: The dissociation of performance and learning in AI-supported programming education. *Computers and Education: Artificial Intelligence*, 10, 100537. <https://doi.org/10.1016/j.caeai.2025.100537>
5. Dabbagh, N., & Fake, H. (2026). Generative AI (GenAI) as a mindtool that supports generative learning (GL). *Computers and Education: Artificial Intelligence*, 11, 100626. <https://doi.org/10.1016/j.caeai.2026.100626>
6. Diab, N. M. (2015). Effectiveness of written corrective feedback: Does type of error and type of correction matter? *Assessing Writing*, 24, 16–34. <https://doi.org/10.1016/j.asw.2015.02.001>
7. Guo, K., Pan, M., Li, Y., & Lai, C. (2024). Effects of an AI-supported approach to peer feedback on university EFL students' feedback quality and writing ability. *The Internet and Higher Education*, 63, 100962. <https://doi.org/10.1016/j.iheduc.2024.100962>
8. Hong, S., & Shin, Y. K. (2025). Effects of three levels of AI integration on second language academic writing: Evaluating restricted, guided, and free use of ChatGPT. *System*, 134, 103820. <https://doi.org/10.1016/j.system.2025.103820>
9. 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>
10. Jiang, Y., & Chen, H. (2026). AI-based cognitive offloading by students in higher education in China. *Computers in Human Behavior: Artificial Humans*, 8, 100327. <https://doi.org/10.1016/j.chbah.2026.100327>
11. Jin, Y., Yang, K., Martinez-Maldonado, R., Gašević, D., & Yan, L. (2026). The agency gap in AI-supported writing: How reactive and proactive agent designs shape multimodal reasoning. *Computers & Education: Artificial Intelligence*, 11, 100655. <https://doi.org/10.1016/j.caeai.2026.100655>
12. Khan, M. S. H., & Habib, R. B. (2026). Exploring the role of learner aptitude and engagement in ChatGPT-assisted writing tasks: Evidence from Bangladeshi undergraduates. *Computers in Human Behavior Reports*, 22, 101099. <https://doi.org/10.1016/j.chbr.2026.101099>
13. Kim, E. (2026). From AI anxiety to strategic regulation: How university students transform generative AI into a strategic learning resource. *Computers and Education: Artificial Intelligence*, 10, 100622. <https://doi.org/10.1016/j.caeai.2026.100622>
14. Lukešová, A., & Jennings, P. J. (2026). AI-assisted language learning: ChatGPT's impact on mastering English tenses at the university level. *Social Sciences & Humanities Open*, 13, 102532. <https://doi.org/10.1016/j.ssaho.2026.102532>
15. Meishar-Tal, H., & Amzalag, M. (2026). Mind the gap: Perceived academic self-efficacy, creativity, and critical thinking with and without ChatGPT. *Thinking Skills and Creativity*, 62, 102246. <https://doi.org/10.1016/j.tsc.2026.102246>

16. Moeiniasl, H. (2026). Investigating the impact of ChatGPT-assisted self-assessment on college students' writing development: Insights from diverse linguistic backgrounds. *Assessing Writing*, 69, 101061. <https://doi.org/10.1016/j.asw.2026.101061>
17. Namaziandost, E., & Çakmak, F. (2025). Impact of AI-generated storytelling vs. gamified learning on vocabulary retention and engagement in CALL environments. *Computers and Education: Artificial Intelligence*, 9, 100505. <https://doi.org/10.1016/j.caeai.2025.100505>
18. Pust, D. (2026). Evolving animated grammar: Testing two types of grammar animations for developing L2 grammatical competence. *Linguistics and Education*, 94, 101547. <https://doi.org/10.1016/j.linged.2026.101547>
19. Ren, Y., & Khan, S. U. (2026). Generative AI in music learning: A mixed-method study of skill development, AI dependence, and the moderating role of AI literacy. *Acta Psychologica*, 270, 107648. <https://doi.org/10.1016/j.actpsy.2026.107648>
20. Schorr, I., Bardach, L., Bühler, B., & Kasneci, E. (2026). Learning-to-learn in the age of generative AI: A scoping review and conceptual framework. *Computers and Education: Artificial Intelligence*, 10, 100575. <https://doi.org/10.1016/j.caeai.2026.100575>
21. Şimşek, A. C., Anders, G., Göth, J., Specht, L., & Huff, M. (2025). Is ChatGPT a good study companion? The role of AI-generated summaries and reflective prompts in learning from educational videos. *Computers and Education: Artificial Intelligence*, 9, 100512. <https://doi.org/10.1016/j.caeai.2025.100512>
22. Tao, L., Song, Y., & Fu, J. (2026). Exploring students' self-regulated learning behavioural patterns and perceptions in an English speaking task within a generative AI-supported immersive VR environment. *Computers & Education*, 242, 105515. <https://doi.org/10.1016/j.compedu.2025.105515>
23. Wiboolyasar, W., Chomchuen, F., Kamonsawad, R., Boonprawes, S., Wiboolyasar, K., & Jinowat, N. (2026). Timing is everything: Investigating the impact of immediate, delayed, and combined AI-generated feedback on L2 writing proficiency. *Social Sciences & Humanities Open*, 13, 102790. <https://doi.org/10.1016/j.ssaho.2026.102790>
24. Yan, S., Yan, Y., Wang, S., Chen, S., Li, Y., & Ai, Z. (2026). Learning with AI, Not from AI: A quasi-experimental comparison of AI-assisted learning with versus without an interactive AI-generated error-correction phase on knowledge application and metacognition. *International Journal of Educational Research*, 139, 103047. <https://doi.org/10.1016/j.ijer.2026.103047>
25. Zhuang, Y., Zhao, R., Xie, Z., & Yu, P. L. H. (2025). Enhancing language learning through generative AI feedback on picture-cued writing tasks. *Computers and Education: Artificial Intelligence*, 9, 100450. <https://doi.org/10.1016/j.caeai.2025.100450>