



International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: <https://www.svedbergopen.com/>



Research Paper

Open Access

Beyond Proficiency-Matched Support: A Multidimensional Framework for Calibrating Generative AI Scaffolding in Arabic L2 Writing

Mahmoud Abd ElFattah Ibrahim Essa*

Academy of Language Studies, Universiti Teknologi MARA (UiTM), Terengganu, Malaysia. E-mail: mahmouud@uitm.edu.my

ABSTRACT

This article proposes a multi-dimensional instructional-design framework for scaffolding Arabic L2 writing using a structured integrative conceptual synthesis of scaffolding research, GenAI/L2 writing studies, evidence on Arabic L2 writing challenges, and Arabic writing-technology precedents. The framework suggests three diagnostic inputs: Learner Proficiency (P), Arabic-Specific Linguistic Demand (L), and Task/Stage Demand (T). Their joint consideration informs five distinct support decisions: explicitness, scope, initiative, verification, and fading. The framework separates language proficiency from GenAI literacy, linguistic demand from the overall Arabic difficulty, task stage from task complexity, and considers engagement, self-regulation, prior knowledge, and GenAI literacy as boundary conditions instead of further core axes. The suggested relationships are conceptual design propositions rather than empirically proven interactions or dosage rules. Further studies are needed to validate P-L-T relationships and support configurations across different Arabic learner groups, linguistic targets, and writing tasks.

Keywords: Generative artificial intelligence; Arabic L2 writing; adaptive scaffolding; instructional design; writing pedagogy

Introduction

Generative artificial intelligence (GenAI) has expanded the range of support available during second-language writing. Learners can now request ideas, explanations, reformulations, feedback, or more direct textual assistance at different points in the writing process. Research has consequently moved beyond asking whether GenAI can assist writing toward examining how the design of that assistance shapes learner activity. Studies of personalised chatbots, agency-supporting writing interfaces, and reactive versus proactive AI agents show that the educational consequences of GenAI depend not only on the quality of generated text, but also on how support is initiated, structured, evaluated, and incorporated into the learner's own work (Guo & Li, 2024; Jin et al., 2026; Kim et al., 2026). From an instructional-design perspective, the central problem is therefore one of calibration: determining what kind of assistance is appropriate, how explicit it should be, and when it should increase, decrease, or be withdrawn.

Recent research has begun to address this problem directly. Qi et al. (2026), for example, designed a scaffolding model for EFL writing that relied on AI to provide support in an implicit to more explicit manner, allowing students to choose the level of support based on their perceived needs, and gradually reducing the level of support as their ability increased. This work is an important empirical example of graduated and proficiency-sensitive AI support. It also shows that proficiency matching alone is no longer sufficient as a conceptual contribution. However, the conditions that could impact the usefulness of a specific type of AI support are not limited to learner proficiency. However, Jin et al. (2026) demonstrated that the requirements imposed on learners by reactive and proactive AI designs are intertwined with learner capabilities, including GenAI literacy, and warned against making blanket judgments of which interaction design is better. Their results also differentiate between the level of support and where it is placed, suggesting that intensity of support and the structure of interaction are two different design choices.

This is especially significant in the case of L2 writing in Arabic, where the difficulty to be supported is not even. Arabic learners can face difficulties at various levels of their language learning and may require different types of support across these levels. In a study of university learners of Arabic, Brosh (2015) reported a variety of spelling problems associated with phonological, morphological and orthographic awareness. Similar problems involving definiteness, word choice, conjunctions and prepositions, agreement, case marking, and word order have also been reported in learner-corpus research on non-native Arabic writing (Zaghouni et al., 2015).

These findings make it difficult to treat Arabic writing difficulty as a homogeneous condition. A learner who can plan the content of a text but has difficulties with morphological agreement is a different learner than a learner who is unable to formulate sentence structure and/or who needs assistance in organising ideas in a longer text.

Existing Arabic writing technologies already reflect part of this differentiation. ARWI, for example, combines Arabic grammatical error detection and correction with a prompt database designed for different proficiency levels and automated scoring aligned with proficiency standards (Chirkunov et al., 2025). Research on ChatGPT-supported Arabic writing has also shown the practical importance of human checking of linguistic and contextual output, particularly where grammatical or lexical accuracy cannot be assumed (Zubaidi et al., 2025). These developments establish that both proficiency sensitivity and Arabic-specific linguistic support are feasible design concerns. They do not, however, provide an integrated instructional logic for deciding how learner proficiency should be considered together with the specific linguistic demand of a writing problem and the demands of the writing task or stage.

Across the evidence reviewed for the present article, these dimensions are more often treated separately than as interacting inputs to a single calibration decision. Proficiency-sensitive scaffolding focuses primarily on how much or how explicit support should be for learners with different levels of capability. Human-AI interaction research focuses on who initiates and regulates the exchange. Arabic writing research identifies language-specific error patterns and verification needs. Writing-process research, meanwhile, distinguishes support for planning, formulation, feedback, and revision. Each strand is useful, but none by itself specifies how an instructor or educational AI design should respond when these conditions occur together. The design problem is therefore not simply whether a learner is "beginner" or "advanced," nor whether the AI should be more or less proactive. It is how support should be adjusted when learner proficiency, the linguistic locus of difficulty, and the immediate writing demand jointly shape what the learner can productively do.

This article addresses that problem by proposing a multidimensional framework for calibrating GenAI scaffolding in Arabic L2 writing. The framework is organised around three jointly considered inputs: **Learner Proficiency**, **Arabic-Specific Linguistic Demand**, and **Task/Stage Demand**. These inputs are used to inform five support decisions: the **explicitness**, **scope**, **initiative**, **verification**, and **fading** of GenAI assistance. The purpose is not to prescribe a fixed amount of AI support for a given proficiency level. Rather, the framework provides an instructional-design logic for asking what support is needed for a particular learner, for a particular linguistic problem, at a particular point in the writing process.

Accordingly, the article addresses the following question: **How should GenAI scaffolding in Arabic L2 writing be calibrated jointly to learner proficiency, Arabic-specific linguistic demand, and the demands of the writing task/stage?** The contribution is conceptual. The proposed relationships are derived from the synthesis of evidence on adaptive scaffolding, GenAI-supported L2 writing, and Arabic L2 writing difficulties; they are not presented as empirically validated effects. The framework is intended to make those relationships explicit and testable, thereby providing a basis for subsequent empirical work comparing fixed and calibrated forms of GenAI support in Arabic language education.

Literature Review

GenAI scaffolding in L2 writing is not a single form of assistance. Current designs range from learner-configured chatbots and direct human-AI collaboration to AI-supported peer review, teacher-mediated feedback, and deliberately structured interfaces. These arrangements differ in what the AI is allowed to do, when it intervenes, who initiates the interaction, and how much evaluative responsibility remains with the learner. Treating them as a unitary category of "AI support" therefore obscures instructional differences that are central to calibration (Alyasin & Shah, 2026; Godwin-Jones, 2024).

Learner-directed and personalised uses illustrate the breadth of possible support, but also its limits. Guo and Li (2024) showed that learners could configure chatbots for functions such as idea generation, outlining, language checking, customisation, and translation, while Huang and Wang (2025) found that upper-intermediate learners used ChatGPT across idea development, evidence, language, grammar, and genre-related guidance. In both cases, broad functionality did not by itself resolve how much support was pedagogically appropriate or how well that support matched what learners could productively use. Huang and Wang also reported difficulty processing large volumes of AI output and occasions when the chatbot failed to understand learner needs. Personalisation is therefore evidence of adaptable functionality, not yet a complete model of instructional calibration.

More bounded designs make this distinction clearer. Guo et al. (2024) embedded AI within peer review so that learners first evaluated a peer's text and then used AI to refine their comments, whereas Han and Li (2024) placed ChatGPT behind teacher mediation, with instructors adapting AI-generated feedback before students revised. These configurations allocate responsibility differently, yet both show that GenAI can scaffold an

intermediate process rather than replace the learner's or teacher's evaluative role. Their instructional significance lies less in the presence of AI than in how the intervention redistributes evaluation, feedback, and revision across participants.

A related design problem concerns learner regulation. Kim et al. (2026) tested interface features intended to support agency during GenAI-assisted argumentative writing. Interaction patterns associated with the argument-outline feature suggested more selective use of AI-generated material, but the authors also cautioned that writing proficiency and other learner characteristics had not been controlled. Jin et al. (2026) make the distinction between support amount and conversational initiative more explicit. Reactive systems can be used to involve learners in more of the interaction while proactive systems can sequence attention, elicit reflection, and provide more directive guidance. When the initiative of a system is not aligned with the capability of the learner, their agency-gap account indicates that either under-support or over-direction is possible.

This adaptive logic has been expanded recently to graduated writing support. This is important because the system can be flexible in its function without being pedagogically adaptive; the latter needs to be supported to adapt to the learner's needs, not just provide alternatives. The model designed by Qi et al. (2026) was a student-driven model where students can progress from implicit to more explicit assistance as they encounter more difficulties, and the assistance is meant to fade as their skills increase. Their results offer a first-hand empirical example of the scaffolding of EFL writing through AI that is graduated based on student proficiency. This is significant conceptually in that it is no longer enough to just match support to learner need or learner proficiency.

In all of these studies, the question is not if GenAI can help with writing, but how. Support can vary in explicitness, initiative, timing, human mediation, and the amount of cognitive work retained by the learner. The evidence reviewed here does not establish that the same configurations will produce equivalent outcomes in Arabic L2 writing, and the studies differ substantially in population, task, and design. Nor do they identify one configuration that should be treated as a universal default across learner profiles or writing purposes. They nevertheless provide the broader instructional-design basis for moving from generic AI availability toward calibration of support.

Adaptivity and Expertise

The case for adaptive scaffolding is not new and is well established before the advent of GenAI. Van de Pol et al. (2010) list three characteristics of scaffolding: contingency, fading, and transfer of responsibility, and say that scaffolding is contingent upon the learner's needs, must fade when it is no longer needed, and must transfer increasing control back to the learner. The related warning comes from expertise-reversal research. Advice for the less experienced learners can become repetitive or restrictive for the more experienced learners (Kalyuga, 2007; Kalyuga & Renkl, 2010). These ideas are not intended to suggest that the more competent students require less support, but rather that the effectiveness of the support varies according to how well the task fits the instructional form and the knowledge of students.

Supportive evidence of this conditionality is provided by L2 evidence. In a quasi-experimental grammar-animation study, Pust (2026) found the relative effectiveness of two teaching techniques as a function of learners' background knowledge, but the differences between the two were not consistently large. The benefit of the result is that it brings the prior-knowledge question into language learning, but does not imply a universal optimal support format for L2 proficiency, either. The concepts of domain knowledge, task-specific knowledge, and general proficiency have not been fully unified.

GenAI-supported writing introduces an additional layer of complexity to the question of adaptation (Ou et al., 2024). Jin et al. (2026) differentiate between GenAI literacy and learner agency as well as the initiative embedded in the interface itself. Their quantitative analyses failed to demonstrate that proactive scaffolding mitigated lower GenAI literacy or that the slope of the GenAI literacy was different between the designs. Qualitative evidence also indicated that guidance could be helpful when learners needed support or prompting, but less useful when tasks were perceived as easy or when learners were already capable of directing the interaction independently. The correct inference then is a calibration hypothesis only, not a confirmed causal mechanism. It also provides an understanding of the reasons for not equating initiative with intensity of support.

Other learner variables also make it difficult to employ language competence as a full measure of readiness. Khan and Habib (2026) explored the explained variance of aptitude and engagement in writing using ChatGPT and reported a modest variance, suggesting the possible influence of other factors, including first-language literacy, self-regulation, prompting ability, digital literacy, and prior writing skills. While their study is exploratory and does not provide any prescriptive rules, it shows that engagement, aptitude, and proficiency are different features of the learner. Even when the same technological scaffold is provided to all learners, Kim et al. (2026) warn that proficiency and other learner factors can impact how agency-supporting interface features are utilized.

This distinction is especially relevant for GenAI literacy, as the ability to utilize the tool is different from judging language usage. A learner could be fluent in the L2, but lack the skills to provide effective prompts, to assess the output from the AI, or to identify when the AI is providing inappropriate support. Another learner might be a very good user of GenAI, but not have the linguistic tools to evaluate a morphosyntactic or lexical suggestion. Thus, the preparation to use AI for writing is multidimensional and not a single component that can be summarized by a language name. The goal of using a proficiency descriptor may not be met by relying exclusively on the descriptor, since it may be useful for instructional planning, but it may not be enough to represent what a learner can initiate, evaluate or verify independently when using AI for writing. Moreover, proficiency is an important input at the learner level in the current framework as it is pedagogically interpretable and directly applicable to L2 teaching. However, it is not viewed as a comprehensive indicator of levels of support need. Prior knowledge, GenAI literacy, engagement, self-regulation, and task-specific capability are considered as modifiers or boundary conditions instead of new axes in the model. They do not turn adaptive scaffolding into a determinate rule or translate findings about expertise reversal or individual differences into rules.

Task/Stage Dependence

The appropriateness of GenAI support also depends on what the learner is doing at a particular point in the writing process. Planning, drafting, evaluation, and revision involve different kinds of cognitive, linguistic, and metacognitive work. A support function that is useful during idea generation may be unnecessary or intrusive during revision, while feedback that requires an existing draft cannot serve the same function during initial planning. The relevant design question is therefore not only how much support is provided, but which part of the writing process is supported and which work remains cognitively active for the learner.

Evidence across L2 writing shows that GenAI already occupies different positions in the writing cycle. Huang and Wang (2025) describe student-directed use across idea generation, argument development, evidence, language support, grammar, and genre guidance. Guo et al. (2024) situate AI in the context of peer evaluation, where learners first generate feedback and then revise it using AI, and Han and Li (2024) locate ChatGPT in the context of teacher-mediated feedback with subsequent learner revision. The purpose and design of these studies vary, but taken together, they demonstrate that the educational role of GenAI is different depending on where it fits into the process and who has the evaluative responsibility. They do not suggest that any one stage of writing is better suited for AI intervention.

The most comparable evidence is from Lai et al. (2026) who compared three task offloading configurations in a draft-feedback-revision cycle. One condition was delegating GenAI to draft, and learners to evaluate and revise; another condition was keeping the learner to draft, but delegating evaluation to GenAI; a third condition was keeping the learner to draft, evaluate, and revise. The results yielded varied patterns of writing performance, cognitive involvement, metacognitive awareness, and self-efficacy, offering a direct insight into where AI can make a difference in the process. The study should not, however, be turned into a universal prescription for drafting with AI first. It was conducted in one university setting, with mainly intermediate-high level learners, with argumentative writing and with a limited number of offloading configurations. The defensible inference is more limited: Distributing various stages of the writing process between AI and learner may change the learning context in which writing takes place.

Planning-oriented support provides a complementary example. Guo and Li (2024) show that learners can use GenAI for idea generation and outlining, while Kim et al. (2026) show that an argument-outline feature can structure activity without necessarily generating the final prose. Structured process-oriented designs similarly distribute GenAI use across distinct writing activities rather than treating writing as a single undifferentiated task (Kong et al., 2024). These findings suggest that support can externalise some organisational work while preserving substantial learner decision-making. They do not establish planning as the optimal site for intervention, but they make clear that the pedagogical effect of a scaffold depends partly on what cognitive work it leaves with the learner.

Task demand further complicates any stage-based prescription. Jin et al. (2026) argue that task complexity and disciplinary expectations shape how much conversational initiative a learner can productively manage; proactive guidance may be useful when the task requires connections or structured prompting yet redundant in easier situations. This means that stage and task complexity should not be collapsed into a single variable. Two learners may both be revising, for example, while facing very different demands depending on genre, evidence requirements, linguistic difficulty, or disciplinary conventions.

The literature consequently supports treating **Task/Stage Demand (T)** as an independent calibration input. T identifies both the learner's current location in the writing process and the cognitive demands of the work required at that point. It does not determine the appropriate support by itself, because the same stage can involve different learner capabilities and different linguistic problems. Its function is instead to constrain what kind of assistance is likely to preserve productive engagement and appropriate distribution of cognitive work.

This conclusion remains a conceptual synthesis across the reviewed L2/EFL evidence rather than a claim that any fixed stage sequence or task-offloading arrangement has been validated for Arabic L2 writing.

Arabic-Specific Writing Demands

Arabic L2 writing does not present a single, uniform linguistic difficulty. Direct evidence from learner writing points to problems across orthographic, morphological, lexical, syntactic, and discourse-related levels. Brosh (2015) documents recurrent spelling difficulties linked to phonological, morphological, and orthographic awareness, while Zaghouni et al. (2015) identify problems involving definiteness, lexical choice, conjunctions and prepositions, agreement, case marking, and word order. These studies do not constitute a complete taxonomy of Arabic L2 writing difficulty, and their populations should not be treated as representative of all learners. They nevertheless establish that “difficulty in Arabic writing” cannot be treated as a single undifferentiated condition.

Morphological and inflectional evidence makes this heterogeneity more concrete. Khamaiseh et al. (2024) report difficulties with Arabic verb-pattern morphology, while Kadai et al. (2025) identify problems in present-tense formation, including morphological marking and written verb forms. Such evidence does not justify a fixed AI response whenever a morphological error occurs. It does show, however, that a learner may require support for a specific formal operation even when other aspects of the writing task are manageable. From an instructional-design perspective, a morphologically localised problem may therefore call for a different scope or degree of explicitness than assistance that generates or reformulates an entire sentence.

Lexical-semantic difficulty raises a different instructional problem because formal correction alone may not resolve it. Daud et al. (2022) identify lexical choice in context, use of verbs and particles, and word- and sentence-level construction as recurring difficulties among Malay learners, while Zaghouni et al. (2015) similarly document lexical-choice and function-word problems. In such cases, supplying a replacement form may not explain why one option fits the intended context better than another. The possible need for contrast, contextual explanation, or alternative formulations is a conceptual design inference from this evidence, not an empirically validated prescription for GenAI feedback.

Difficulty can also extend beyond local form. Ismail and Che Mat (2026) report weaknesses involving syntax, idea cohesion, and paragraph development among university learners of Arabic in Malaysia. Their findings are context-specific, but they reinforce the distinction between local linguistic problems and higher-level textual demands. Two learners with broadly similar proficiency may therefore face different immediate writing problems: one may struggle with agreement or verb morphology while another can construct accurate sentences but needs support in connecting ideas across a paragraph. Proficiency remains relevant, but it does not by itself locate the linguistic work that is currently difficult or indicate whether the problem is local, sentence-level, or discourse-level.

This distinction becomes especially important when GenAI is used as a scaffold. Zubaidi et al. (2025) show both useful Arabic writing support and continuing need for human checking of linguistic and contextual output, while Chirkunov et al. (2025) demonstrate that proficiency-sensitive prompting and grammatical error detection can be combined within Arabic writing technology. Neither line of work validates a calibration rule or establishes an optimal response for a given error type. They do, however, support treating **Arabic-Specific Linguistic Demand (L)** as the locus and nature of the linguistic work the learner cannot yet perform independently at a given moment. L is therefore not a general measure of how “difficult Arabic” is, but a diagnostic input into support design.

The Unresolved Calibration Problem

Research on GenAI-supported writing now provides several partial logics for calibration. Support can be personalised, graduated, proactive or reactive, teacher-mediated, and positioned at different points in the writing process (Jin et al., 2026; Kim et al., 2026; Qi et al., 2026). These developments show that assistance can be adjusted in amount, initiative, timing, and form rather than held constant. What they do not yet provide is a single instructional logic for deciding which adjustments are appropriate when learner, linguistic, and task conditions occur together.

A single calibration dimension leaves important parts of that decision unresolved. Proficiency-sensitive scaffolding helps identify differences in learner readiness but does not specify the linguistic locus of a current problem or the cognitive demands of the writing stage. Task- and stage-sensitive approaches identify where AI enters the writing process and what work remains with the learner, but stage alone does not determine whether the immediate difficulty is morphological, lexical-semantic, syntactic, or discourse-level (Lai et al., 2026). Conversely, Arabic learner-writing research can identify where linguistic difficulty occurs, but diagnosis alone does not determine how much initiative, explanation, or verification a particular learner requires at a particular point in the writing process (Brosh, 2015; Zaghouni et al., 2015).

The unresolved problem is therefore not the absence of adaptivity, language sensitivity, or process-sensitive support as separate ideas. It is the lack of an integrated calibration logic that brings them into the same design

decision. On the basis of the evidence reviewed here, the present article consequently proposes the joint consideration of **Learner Proficiency (P)**, **Arabic-Specific Linguistic Demand (L)**, and **Task/Stage Demand (T)**. This formulation is a conceptual synthesis, not an empirically established interaction model. The current evidence does not determine whether the three inputs carry equal weight, how they interact statistically, or whether particular combinations reliably produce superior learning outcomes. Their role at this stage is diagnostic: together they capture different aspects of the instructional situation that a one-dimensional calibration rule would leave unaddressed.

The next design question is therefore what these inputs should influence. If P captures learner-level resources, L identifies the linguistic work that remains difficult, and T locates that difficulty within the demands of the writing activity, calibration must concern more than the overall quantity of AI help. It must specify how assistance is configured. The proposed framework therefore uses P, L, and T to inform five support decisions: **explicitness**, **scope**, **initiative**, **verification**, and **fading**. These relationships are presented as design propositions for empirical examination and refinement rather than as validated prescriptions. Before specifying those relationships, the next section explains the structured integrative synthesis through which the framework was developed.

Methods

Framework Development Approach

This article adopts a structured integrative conceptual synthesis to develop an instructional-design framework for calibrating GenAI scaffolding in Arabic L2 writing. The purpose of the method is not to estimate effects, test hypotheses, or provide an exhaustive review of the literature, but to make the conceptual derivation of the framework transparent and traceable. The approach therefore combines theoretically relevant and empirically informative sources according to the design functions they can legitimately support. It is best understood as a structured integrative evidence approach for conceptual framework development rather than as a systematic review, scoping review, meta-analysis, or empirical study. Its methodological value lies in showing how distinct evidence strands were compared, bounded, and combined to address the calibration problem identified in the literature review.

The evidence corpus was organised by analytical role rather than by publication type alone. Foundational scaffolding and adaptive-instruction literature provided the conceptual basis for contingency, fading, transfer of responsibility, and sensitivity to learner expertise. Contemporary GenAI and L2-writing studies were used to examine variation in support functions, learner regulation, conversational initiative, task allocation, and graduated assistance. Direct Arabic L2 studies were treated differently: their role was to establish the heterogeneity and locus of linguistic difficulty in Arabic writing, including orthographic, morphological, lexical, syntactic, and discourse-level demands. The targeted, proficiency-sensitive and verification-relevant forms of support were found to be technically and pedagogically feasible in Arabic contexts, as shown by Arabic writing-technology and AI precedents. For transferable design reasoning, therefore, wider EFL/L2 evidence was used but not as evidence that the same effects would occur in Arabic L2 learners.

The synthesis was carried out by mapping the source to the claim. The analysis identified what evidence from the studies supported, what was context-specific or untested, and what instructional implication could be drawn without going beyond the evidence for each evidence strand. The learner-capability strand reinforced the importance of proficiency and highlighted that proficiency is not synonymous with readiness for AI-supported writing. The linguistic problem itself was found to be variable in locus and type in Arabic-specific studies, and task- and process-oriented studies revealed that support needs are also influenced by the learner's position in the writing process and the cognitive demands of the task. These partial calibration logics were then checked for overlaps and incompleteness. Their combination led to the three diagnostic inputs proposed in this article: **Learner Proficiency (P)**, **Arabic-Specific Linguistic Demand (L)**, and **Task/Stage Demand (T)**. The joint $P \times L \times T$ formulation is therefore a conceptual synthesis introduced by the present framework, not an empirical interaction discovered in the literature.

Claim strength was controlled by distinguishing among three levels of inference. First, direct evidence-supported claims were limited to findings that could be attributed to the relevant empirical or theoretical literature. Second, conceptual inferences were used where multiple evidence strands jointly supported a bounded design conclusion, such as the insufficiency of one-dimensional calibration. Third, framework propositions were reserved for relationships introduced by the present article, including the joint use of P, L, and T to inform support decisions. The evidence base does not establish the relative weighting of these inputs, their statistical interaction, or the effectiveness of the proposed framework. Nor does evidence from wider EFL/L2 contexts constitute direct empirical validation for Arabic L2 writing. These limits are retained deliberately so that the framework functions as a testable instructional-design proposition rather than as a validated prescription.

Results and Discussion

From Diagnostic Inputs to Calibration Logic

The framework starts with a diagnostic, not dosage-based perspective on GenAI scaffolding. The literature reviewed above suggests that the usefulness of support is limited by the learner's capability, the linguistic difficulty of the task, and the conditions of the task, but none of these factors alone is enough. The proposed calibration logic thus regards Learner Proficiency (P), Arabic-Specific Linguistic Demand (L), and Task/Stage Demand (T) as collectively taken as an input to a design decision. The term “jointly considered” does not mean there is a statistical interaction, equal weighting or a fixed causal sequence. It implies that the meaning of the instructional value of one input is understood in the context of the other two inputs prior to choosing the type of assistance.

This moves away from the one-size-fits-all “more” or “less” AI support continuum. The linguistic locus of difficulty may differ between two learners at a similar proficiency level, and between a learner at the same level when shifting from planning to revision, or from a familiar to a more challenging task. The framework thus poses three questions before making any support decision: What linguistic resources can the learner currently mobilise? What Arabic specific linguistic work is not supported? What does the learner want to do now in the writing process? This logic is shown in Figure 1 as three diagnostic inputs that feed into a calibration process which leads to five support decisions, instead of as three separate checklists.

The Three Calibration Inputs

Learner Proficiency (P) refers to the linguistic resources the learner can mobilise for the writing work at hand. It includes relevant lexical, grammatical, syntactic, and discourse-level capability, but it is not a complete model of learner readiness. GenAI literacy, engagement, self-regulation, aptitude, and prompting skill may influence how effectively a learner uses assistance, yet they are conceptually distinct from language proficiency. P therefore functions as a principal learner-level input: it helps estimate what linguistic work the learner may be able to perform independently, but it cannot by itself determine what support is needed.

Arabic-Specific Linguistic Demand (L) refers to the locus and nature of the Arabic linguistic work that the learner cannot yet perform independently at a given moment. The relevant demand may be orthographic, morphological, lexical-semantic, syntactic, or discourse-level. The purpose of L is not to assign Arabic an overall difficulty score or to create a fixed error taxonomy. Its value is diagnostic. A learner who struggles with verb morphology presents a different design problem from one who can form accurate sentences but has difficulty maintaining lexical appropriateness or paragraph cohesion. The Arabic evidence base supports this heterogeneity, but it does not establish a validated mapping from each linguistic problem to a particular AI response.

Task/Stage Demand (T) captures both the learner's current location in the writing process and the cognitive work required at that point. Planning, drafting, evaluation, and revision are distinct instructional sites, but stage alone is insufficient because task complexity, genre, evidence requirements, and disciplinary expectations can alter what the learner must do within the same stage. T therefore concerns not only where support enters the writing process but also what cognitive and metacognitive work should remain with the learner. It is distinct from L: two learners may be revising at the same stage while facing different linguistic demands, and the same linguistic problem may require different support depending on whether it occurs during formulation, feedback, or final revision.

Joint Interpretation of P, L, and T

The three inputs are intended to constrain one another diagnostically rather than to operate as independent predictors. P without L can indicate general linguistic capability but cannot identify the current locus of difficulty. L without P can identify the linguistic problem but cannot indicate how much of that problem the learner can manage independently. T without P and L can locate the activity in the writing process but cannot determine what form of linguistic assistance is appropriate. The framework therefore treats calibration as a relational judgement across the three inputs.

This relational view avoids a fixed hierarchy. Proficiency is not always the dominant consideration, linguistic demand is not always the most important, and task/stage does not automatically override learner capability. Their relative importance may vary across instructional situations, but the present evidence does not justify assigning numerical weights or an ordering rule. Table 1 makes this derivation explicit by distinguishing what each evidence strand establishes from what remains unvalidated.

Five Support Decisions

There is no single support score for the three inputs. They provide five independent design choices that describe the configuration of assistance.

Explicitness is related to the directness of the linguistic or writing information that is needed in the system. Support can be a cue, question or contrast, or a direct explanation, model or correction. The primary

justification for the different levels of explicitness is the scaffolding and graduated-support literature, and the heterogeneity of Arabic language suggests that the different problem loci may need different types of explanation. A learner who has a restricted control of a certain morphological operation might need more direct instruction than a learner who can self-correct after a prompt. The framework does not, however, mandate that lower proficiency always need explicit support, nor does it specify that advanced learners always need implicit cues. Explicitness is a design implication to be tested in the context of P, L and T.

The scope of an intervention is the extent of text or language addressed. Support may be given to a single form, word, phrase, sentence, paragraph or larger organisational problem. This is particularly the case of L: a narrow intervention may be required for a local morphological problem, while a wider one may be needed for cohesion or paragraph development. Planning support can be at an organisational level and revision support can be more local, so T is also relevant. Scope should not be conflated with strength. Not all interventions are necessarily better, and there is no evidence that there is a fixed relationship between error category and intervention size.

Initiative refers to who initiates and orchestrates the interaction. Reactive support requires the learner to recognise uncertainty and ask for help, while pro-active support can provide prompts, focus or ask questions. As demonstrated by the work of Jin et al. (2026) and other agency-oriented studies, it is important to differentiate between initiative and level of assistance. A proactive system can provide short prompts, whereas a reactive system can provide a lot of information after a learner request. P and T can affect the amount of self-initiation that is manageable and L can affect the learner's ability to recognise that a problem exists. The framework does not presume that proactive designs are better for lower-proficiency learners or that reactive designs are less autonomy-supportive.

Verification is the evaluative effort needed prior to accepting or using AI output. It can include learner checking, contrastive comparison, prompted justification, teacher mediation, or reference to an external source. Verification is especially important in Arabic-focused work, as the linguistic and contextual appropriateness of fluent output cannot be taken for granted. P is what the learner can verify, L is what linguistic knowledge needs to be verified, and T is the stakes and time for verification. The verification is not the same as automatic rejection of AI output nor the automatic correction by the teacher in all cases. It serves to maintain evaluative control of uptake.

Fading refers to the progressive decrease or rearrangement of support when the learner is more able to complete the required task without support. Fading does not necessarily mean that access to the tool is removed in a GenAI environment. It can include shifting from direct explanation to prompting, from proactive to more learner-driven interaction, from paragraph-level support to local support, and from externally prompted checking to learner-driven checking. Fading should therefore be sensitive to performance in a specific linguistic and task setting rather than to an overall proficiency label. A learner could need less support for morphology and yet need significant support for discourse organisation. There is no recommended schedule for fixed fading.

Non-equivalences and Boundary Conditions

The framework has a number of differences that help prevent it from being too general a learner model. But proficiency is not GenAI literacy: linguistic competence and the capacity to prompt, assess, and manage AI interaction are not synonymous. L is not the overall difficulty of Arabic, but rather the linguistic demand that is being made in the immediate context of unsupported linguistic work. Task stage is not equivalent to task complexity: revising a simple sentence is not the same as revising an evidence-heavy argumentative paragraph. Who starts or continues support is not the amount of support: the amount of information provided at the end is not dependent upon who initiates or continues the support. Fading is not "withdrawal" but "redistribution" of responsibility. Verification is not a rejection of the output of AI – it is an evaluative control over whether and how suggestions are adopted.

Prior knowledge, engagement, and self-regulation are therefore not considered as new axes of the framework, but rather as boundary conditions or modifiers of GenAI literacy. They can change the way P, L and T are operationalized in practice but adding each as a separate dimension would make the model too wide-ranging to account for learner readiness, and would be beyond the scope of the evidence for the present model.

Illustrative Calibration and Empirical Propositions

Its diagnostic value becomes clearer when superficially similar learners or tasks call for different support configurations. Table 2 provides three deliberately non-prescriptive scenarios.

These examples illustrate differentiation rather than prescribe fixed responses. Their purpose is to make the framework empirically testable: future studies can manipulate or observe combinations of P, L, and T and examine whether different configurations of explicitness, scope, initiative, verification, and fading produce different patterns of learner engagement, independent performance, or writing quality. The present article does not predict which configuration will be optimal, nor does it assume linear relations among the inputs and

support decisions. Its contribution is to specify a set of instructional-design propositions that can be examined, challenged, and refined through empirical work.

Pedagogical Implications

The framework is not only a design architecture, but also has a direct pedagogical implication: GenAI support should be considered a programmable instructional tool, not a fixed support level tied to a proficiency level. Three questions need to be asked in tandem before choosing a form of support: what linguistic resources can the learner currently mobilise (P), what Arabic-specific linguistic work is still difficult (L), and what the learner is expected to do at the current writing stage or task (T)? This moves away from the general statement “lower-proficiency learners require more AI support” and towards a more precise diagnostic judgement of the type of support that is needed for the current writing challenge.

P, L, and T may be viewed as a set of diagnostic considerations prior to decisions regarding explicitness, scope, initiative, verification, and fading in practice. While proficiency can help estimate what language information a learner is able to process independently, it should not replace diagnosis. Once linguistic demand has been identified, the support target becomes more specific, such as a local morphological or lexical-semantic difficulty, which can be the basis for a more focused response than a problem related to paragraph cohesion or organisation. Task/stage demand is another constraint that asks the question: What is the learner planning, drafting, evaluating, or revising, and what cognitive work should be left to the learner? The pedagogical significance of the framework is that it is not a fixed set of answers, but a set of questions that can be asked together.

The five support decisions can then be adjusted conditionally. When a learner is unable to use indirect cues, there may be more explicitness but it should be related to the linguistic and task context, not to proficiency. The scope can be limited if the problem is limited in scope, and it can be wider if the writing task is wider. The balance between learner-led and more proactive support can be adjusted according to the learner’s capacity to initiate interaction productively. Verification can be supported by having limited resources available for assessing appropriateness of Arabic language or context, and fading can be initiated when the learner starts to do the same relevant work independently. These decisions are not to be seen as a fixed learner profile but might vary with linguistic goals, tasks and stages.

The framework does not prescribe a fixed mapping between CEFR and AI, or a numerical “dosage” of support, but allows for flexible support planning in classroom and materials design. The framework can be used by teachers to create prompts, feedback routines, AI-supported writing activities, or interface settings that will maintain learner responsibility while focusing on the current area of difficulty. It is a provisional and reflective use: it is a set of dimensions that can help when making decisions about instruction, but it does not yet tell us which of its configurations will lead to the best learning outcomes. Pedagogical implementation should thus be supported by the observation of the learner’s response and be accompanied by the willingness to adjust support based on the learners’ performance.

Limitations and Future Studies

This article develops an evidence-informed conceptual framework rather than an empirically validated model. Its three inputs—Learner Proficiency (P), Arabic-Specific Linguistic Demand (L), and Task/Stage Demand (T)—were derived through integrative synthesis of theoretical, L2/GenAI, and Arabic-specific evidence, but their relationships have not been tested as a formal model. The current literature does not establish the relative weighting of P, L, and T, whether one input should take priority under particular conditions, or whether specific combinations of the three reliably require particular levels of explicitness, scope, initiative, verification, or fading. Future experimental and quasi-experimental studies should therefore test selected framework relationships rather than the entire model at once, for example by comparing support configurations while holding linguistic demand or task stage constant across learners with different proficiency levels.

A second limitation concerns the evidence base. Arabic-specific studies establish important variation in orthographic, morphological, lexical, syntactic, and discourse-level difficulty, but this literature remains less extensive than the broader EFL/L2 and educational-technology evidence used to inform scaffolding, initiative, and task-allocation principles. Transfer from those wider contexts is therefore theoretical rather than direct evidence of effectiveness in Arabic L2 writing. Validation should consequently include Arabic learners across different proficiency ranges, linguistic difficulty loci, genres, and writing stages. Studies could compare whether the same support decision operates differently when, for example, a local morphological problem is contrasted with a discourse-level problem, or when the same linguistic difficulty occurs during drafting rather than revision.

A third limitation is that the framework specifies a decision architecture without yet validating the learner outcomes associated with its use. Future work should compare alternative configurations of explicitness, scope, initiative, verification, and fading and examine not only supported task performance but, where methodologically appropriate, whether learners can subsequently perform the targeted linguistic or writing

work with less assistance. Longitudinal designs would be especially useful for examining whether recalibration and fading correspond to growing independent control over particular writing processes. Such studies should not assume that one configuration will generalise across all learners or tasks; rather, they should test which relationships hold, where they fail, and whether additional boundary conditions such as GenAI literacy or self-regulation need to be incorporated into later versions of the framework.

Conclusion

The central issue of this article is not whether GenAI can assist L2 Arabic writing but how it can be set up to assist L2 Arabic writing when the capability of the learners, the difficulty of the language, and the demands of the writing do not correspond in the same way in each situation. Previous studies have identified a number of partial calibration logics such as proficiency-sensitive support, adaptive scaffolding, differences in AI initiation, stage-specific task allocation, and Arabic-specific writing challenges. The present article combines these strands into a multidimensional instructional-design framework that involves three inputs: Learner Proficiency (P), Arabic-Specific Linguistic Demand (L), and Task/Stage Demand (T).

The framework suggests that these inputs should be taken into account together when setting up five distinct dimensions of GenAI support: explicitness, scope, initiative, verification, and fading. It is conceptual and not empirical. It does not specify any fixed support levels, numerical dosage, or a universal priority of P, L, or T. It provides a framework to help determine the type of support problem being encountered before deciding on the nature of support. This moves the focus from “matching AI support to proficiency” to a more language-sensitive and task-sensitive view of scaffolding in L2 writing in Arabic.

The next step is to examine the framework empirically. Specific relationships between P, L, T and the five support decisions can be tested in future studies with various populations of Arabic learners, difficulty loci of language and writing tasks. This validation can help to establish some of the relationships, provide a counterpoint to others, and reveal other boundary conditions. The framework is thus conceived as an experimental design architecture that can guide future decisions and research questions in writing instruction, rather than as a complete prescription for GenAI-supported writing instruction.

Declarations

Conflict of Interest

The author declares no conflict of interest.

Funding

This research received no external funding.

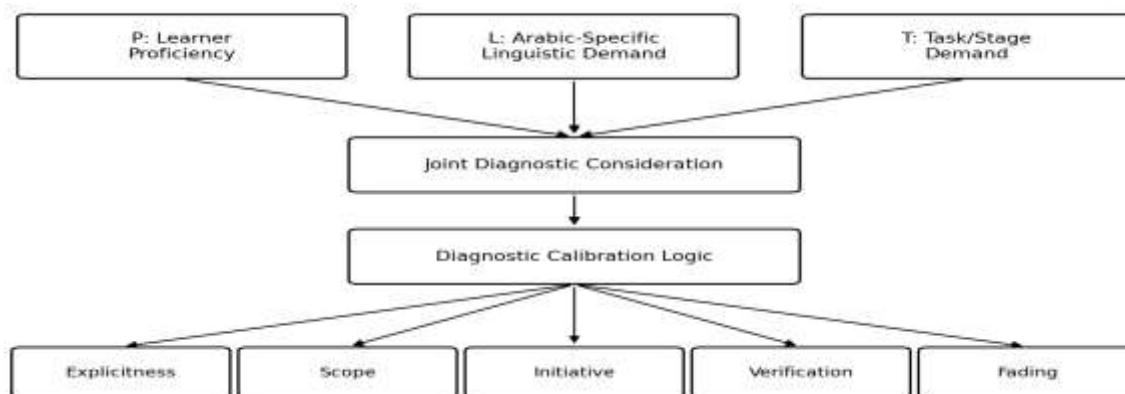
Data Availability Statement

No new data were created or analyzed in this study.

AI Statement

Generative AI was used only for language editing. The author takes full responsibility for the final content.

Figure 1. Multidimensional Calibration Framework for GenAI Scaffolding in Arabic L2 Writing



Note. The relationships shown are conceptual design propositions. The arrows represent diagnostic consideration, not validated causal or statistical pathways.

Table 1. Evidence-to-Calibration Synthesis

Evidence strand	What it establishes	What it does not establish	Calibration implication	Principal evidence base
Contingent/adaptive scaffolding	Support should respond to learner need and may fade as responsibility shifts	A fixed dosage or universal sequence	Support should be adjustable rather than constant	van de Pol et al. (2010); Kalyuga (2007); Kalyuga & Renkl (2010)
Learner capability and AI interaction	Proficiency is relevant but does not capture all readiness; initiative can be misaligned with capability	That one agent design is universally superior	P should inform support while remaining distinct from GenAI literacy and agency	Jin et al. (2026); Kim et al. (2026); Khan & Habib (2026)
Task/stage allocation	AI placement changes the cognitive work retained by the learner	That one writing stage is the optimal intervention point	T should constrain when and where support is introduced	Lai et al. (2026); Guo et al. (2024); Han & Li (2024)
Arabic linguistic heterogeneity	Arabic L2 difficulties occur at different linguistic levels	A complete taxonomy or a validated support rule for each difficulty	L should identify the linguistic locus of unsupported work	Brosh (2015); Zaghouni et al. (2015); Khamaiseh et al. (2024); Kadai et al. (2025)
Arabic AI/writing-support precedents	Targeted and proficiency-sensitive support is feasible in Arabic writing contexts	Validation of the present framework	Arabic support design should remain language-sensitive and verification-aware	Chirkunov et al. (2025); Zubaidi et al. (2025)

Table 2. Illustrative Calibration Scenarios

Scenario	Diagnostic contrast	Plausible calibration implication
Same P, different L	Two learners at similar proficiency; one has a local verb-morphology problem, the other a paragraph-cohesion problem	The first may call for narrower, more form-focused support; the second may require broader organisational guidance and different verification
Same P + L, different T	Similar proficiency and lexical-semantic difficulty, but one learner is drafting and the other is revising	Drafting may justify support for option generation and contextual contrasts; revision may place greater emphasis on evaluation and selective uptake
Different P, same L + T	Two learners revising the same morphosyntactic problem at different proficiency levels	Explicitness, initiative, and verification demands may differ even when the linguistic locus and stage are held constant

References

1. Alyasin, A., & Shah, M. A. (2026). Human-AI collaboration in academic writing: Exploring university students' agency through a sociocultural lens. *Ampersand*, 16, 100256. <https://doi.org/10.1016/j.amper.2026.100256>
2. Brosh, H. (2015). Arabic spelling: Errors, perceptions, and strategies. *Foreign Language Annals*, 48(4), 584–603. <https://doi.org/10.1111/flan.12158>
3. Chirkunov, K., Alhafni, B., Qwaider, C., Habash, N., & Briscoe, T. (2025). ARWI: Arabic Write and Improve. In *Proceedings of the Fourth Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2025)* (pp. 11–18). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2025.in2writing-1.2>
4. Daud, A. F. C., Mustapha, N. F., Toklubok, P. H. @ P., Jabar, M. A. A., & Mohamad, A. H. (2022). Problems in Arabic writing skills among Malay students in Malaysia: A review. *International Journal of Academic Research in Business and Social Sciences*, 12(1), 2028–2039. <https://doi.org/10.6007/IJARBS/v12-i1/12038>
5. Godwin-Jones, R. (2024). Distributed agency in language learning and teaching through generative AI. *Language Learning & Technology*, 28(2), 5–30. <https://doi.org/10.64152/10125/73570>
6. Guo, K., & Li, D. (2024). Understanding EFL students' use of self-made AI chatbots as personalized writing assistance tools: A mixed methods study. *System*, 124, 103362. <https://doi.org/10.1016/j.system.2024.103362>
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. Han, J., & Li, M. (2024). Exploring ChatGPT-supported teacher feedback in the EFL context. *System*, 126, 103502. <https://doi.org/10.1016/j.system.2024.103502>
9. Huang, Y., & Wang, D. (2025). Can ChatGPT serve as a writing collaborator? Insights from Chinese EFL learners. *System*, 133, 103775. <https://doi.org/10.1016/j.system.2025.103775>
10. Ismail, M. R., & Che Mat, A. (2026). Bridging the gap between perceptions and reality in Arabic writing proficiency among Malaysian higher education students: A triangulation approach. *Ijaz Arabi Journal of Arabic Learning*, 9(1), 482–489. <https://doi.org/10.18860/ijazarabi.v9i1.31513>
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. Kadai, K., Baharudin, H., Wicaksono, T., & Fitriani, N. (2025). Analysis of errors in the formation of fi'il mudari' by the University of Malaysia Sarawak students. *Ijaz Arabi Journal of Arabic Learning*, 8(2), 613–629. <https://doi.org/10.18860/ijazarabi.v8i2.30329>
13. Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. *Educational Psychology Review*, 19(4), 509–539. <https://doi.org/10.1007/s10648-007-9054-3>
14. Kalyuga, S., & Renkl, A. (2010). Expertise reversal effect and its instructional implications: Introduction to the special issue. *Instructional Science*, 38, 209–215. <https://doi.org/10.1007/s11251-009-9102-0>
15. 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>
16. Khamaiseh, M. A., Osman, N., & Elawadi, E. M. S. (2024). Morphological errors in the ten wazan for the writing skill of non-native Arabic language learners. *Ijaz Arabi Journal of Arabic Learning*, 7(1), 1–14. <https://doi.org/10.18860/ijazarabi.v7i1.26073>
17. Kim, S., So, H.-J., & Park, K. (2026). Supporting learner agency in collaborative writing with generative AI. *British Journal of Educational Technology*, 57(4), 984–1008. <https://doi.org/10.1111/bjet.70015>
18. Kong, S.-C., Lee, J. C.-K., & Tsang, O. (2024). A pedagogical design for self-regulated learning in academic writing using text-based generative artificial intelligence tools: 6-P pedagogy of plan, prompt, preview, produce, peer-review, portfolio-tracking. *Research and Practice in Technology Enhanced Learning*, 19, 030. <https://doi.org/10.58459/rptel.2024.19030>
19. Lai, C., Pan, M., Guo, K., & Cui, Y. (2026). What task to offload to GenAI in the writing feedback process? Effects of task-offloading approaches on EFL learners' writing skill development. *Computers & Education*, 252, 105675. <https://doi.org/10.1016/j.compedu.2026.105675>
20. 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>

21. 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>
22. Qi, Z., Liu, Y., Liu, W., & Xu, W. (2026). How does AI match students' proficiency levels to provide support? Student-driven AI-mediated scaffolding in EFL writing instruction. *System*. Advance online publication. <https://doi.org/10.1016/j.system.2026.104119>
23. van de Pol, J., Volman, M., & Beishuizen, J. (2010). Scaffolding in teacher–student interaction: A decade of research. *Educational Psychology Review*, 22(3), 271–296. <https://doi.org/10.1007/s10648-010-9127-6>
24. Zaghouni, W., Habash, N., Bouamor, H., Rozovskaya, A., Mohit, B., Heider, A., & Oflazer, K. (2015). Correction annotation for non-native Arabic texts: Guidelines and corpus. In *Proceedings of the 9th Linguistic Annotation Workshop* (pp. 129–139). Association for Computational Linguistics. <https://doi.org/10.3115/v1/W15-1614>
25. Zubaidi, A., Munip, A., Widodo, S. A., & Zerrouki, T. (2025). Enhancing Arabic writing skills using Chat GPT-based AI learning models: A tridimensional human-AI collaboration framework. *Indonesian Journal of Applied Linguistics*, 15(1), 87–101. <https://doi.org/10.17509/ijal.v15i1.75378>