Journal of Arabic EdTech

e-ISSN xxxx-xxxx
p-ISSN xxxx-xxxx

Vol. 1, No. 1, June 2026

The Use of AI Writing Tools in Teaching Mahārah al-Kitābah: A Literature Review on Correction, Feedback, and the Development of Arabic Writing

Millenia Dian Kumala[1], Agus Yasin[2], Aufa Alfian[3], Zila Gustita[4]

[1]Universitas Darussalam Gontor, Indonesia
[2]University of The Holy Qur'an and Islamic Sciences, Sudan

[1]*milleniadian18@gmail.com*, [2]*elyasien@unida.gontor.ac.id*,
[3]aufaalfian@unida.gontor.ac.id, [4]*gustitazila@gmail.com*

*Correspondence: *milleniadian18@gmail.com*

Received: 2 April 2026
Accepted: 18 May 2026
Published: 30 June 2026

DOI: https://doi.org/xxxx

Abstract

This article presents a literature review examining the role of artificial intelligence (AI) writing tools in supporting the teaching of *mahārah al-kitābah* (Arabic writing skill), with particular attention to correction, feedback, and the broader development of learner writing competence. Despite the rapid expansion of AI-powered writing assistance for widely studied languages, Arabic remains comparatively under-served because of its morphological complexity, script, and limited natural language processing resources. Using a narrative literature review approach, sources published between 2015 and 2026 were retrieved from Scopus, Google Scholar, ERIC, and the ACL Anthology, then synthesised thematically. The review finds that AI tools function across three layers: automated grammatical and orthographic correction, formative corrective feedback, and adaptive writing support, each interacting differently with learners' accuracy, fluency, and engagement. Findings further indicate that while generative and Arabic-specific systems show measurable gains in writing proficiency, persistent challenges include diacritisation, dialectal variation, and the risk of learner overreliance. The review proposes a conceptual framework for situating AI writing tools within *mahārah al-kitābah* instruction and offers implications for curriculum designers, teacher training, and future research in Arabic EdTech.

Keywords: artificial intelligence, mahārah al-kitābah, Arabic writing skill, corrective feedback, Arabic natural language processing

Introduction

Writing occupies a distinctive place among the four language skills because it requires learners to simultaneously manage linguistic accuracy, rhetorical organisation, and conceptual content while producing a permanent, reviewable text. In Arabic language education, this productive skill—commonly referred to as *mahārah al-kitābah*—is widely regarded as the most cognitively demanding of the four macro-skills, since it compels learners to coordinate orthographic conventions, a templatic morphological system, case-marking (*i'rab*), and stylistic register within a single act of composition. For non-native learners, and even for many native-speaking students socialised primarily in spoken dialects, the gap between everyday linguistic competence and the expectations of formal written Modern Standard Arabic (MSA) can be considerable. Over the past decade, the rapid maturation of artificial intelligence (AI), and particularly of large language models (LLMs) such as GPT-4o, has reshaped writing instruction across many languages, offering learners instantaneous grammatical correction, stylistic suggestions, and generative drafting support (Godwin-Jones, 2022; Kasneci et al., 2023). Tools such as Grammarly popularised automated written corrective feedback well before the emergence of generative AI, demonstrating that learners actively engage—albeit unevenly—with machine-generated feedback on their writing (Koltovskaia, 2020), extending a longer-standing tradition of computer-assisted language learning (CALL) tools, from mobile-application-based practice (Alqahtani, 2019) to web-based automated writing assessment (Warschauer & Healey, 1998).

Within Arabic language education specifically, interest in AI-assisted instruction has grown markedly, encompassing applications for vocabulary, grammar, and conversational practice, alongside a smaller but expanding body of work on writing (Al-Hamad & Al-Shara, 2022; Albantani et al., 2025). Indonesian higher education institutions, where Arabic is taught both as a religious-literacy requirement and as an academic discipline within Islamic studies programmes, have been particularly active sites for experimentation with ChatGPT-based and other generative tools in Arabic writing classrooms (Dahlan et al., 2023; Zubaidi et al., 2025).

This challenge is compounded by Arabic's diglossic character: the formal register required for academic and religious writing differs systematically from the spoken varieties most learners acquire first, so that *mahārah al-kitābah* instruction must simultaneously teach a register, a script, and—for many learners—an additional layer of grammatical marking that has little equivalent in everyday speech (Hyland, 2003). It is precisely this combination of script complexity, morphological richness, and register distance that makes Arabic writing instruction a demanding test case for AI writing assistance, and one whose lessons may generalise to other morphologically rich, diglossic, or under-resourced languages.

Despite this growing interest, Arabic remains comparatively under-served in the broader AI-writing-assistance literature, which has historically concentrated on English and a handful of other well-resourced languages. Arabic's templatic morphology, the frequent omission of short-vowel diacritics in everyday text, and considerable dialectal variation make automatic grammatical error detection and correction substantially harder to achieve than in English (Alhafni et al., 2023; Kwon et al., 2023). Large-scale resources that underpin AI writing tools for English—annotated learner corpora, benchmark datasets, explanation-oriented correction systems—have only recently begun to emerge for Arabic, through initiatives such as the ZAEBUC bilingual writer corpus (Habash & Palfreyman, 2022), the Gazelle instruction dataset (Magdy et al., 2024), the ARWI writing-improvement platform (Chirkunov et al., 2025), and the Nahw grammar benchmark (Mubarak et al., 2026). At the same time, the rich tradition of feedback and corrective-feedback research in second language writing—anchored in the long-

running debate between Truscott (1996) and Ferris (1999), and in subsequent syntheses such as Bitchener and Ferris (2012) and Hattie and Timperley (2007)—has rarely been applied systematically to the Arabic context, and almost never to the specific question of how AI-mediated correction and feedback shape the development of *mahārah al-kitābah*. Existing reviews of AI in Arabic language education tend either to survey AI applications across all four skills without writing-specific depth, or to report single-site empirical studies without situating their findings within established feedback theory.

This article addresses that gap through a literature review that brings together second language writing theory, corrective-feedback research, and the emerging Arabic natural language processing (NLP) literature to examine how AI writing tools are being used—and might productively be used—in teaching *mahārah al-kitābah*. Specifically, the review addresses three research questions: (1) What types of AI writing tools have been developed or applied for Arabic writing instruction, and what correction and feedback functions do they perform? (2) What effects have these tools been reported to have on the development of learners' Arabic writing competence? (3) What linguistic, technological, and pedagogical challenges constrain their effective use, and what implications follow for curriculum design and teacher practice? By synthesising findings across Arabic NLP, computer-assisted language learning, and second language writing pedagogy, the review aims to provide Arabic EdTech researchers and practitioners with a conceptually grounded map of a still-emerging field. The following subsection outlines the theoretical foundations that inform the conceptual framework guiding this review.

Theoretical Framework

Writing has long been understood through Flower and Hayes's (1981) cognitive process theory, which models composition as a recursive interplay of planning, translating, and reviewing rather than a fixed linear sequence. Within this model, AI writing tools intervene primarily at the translating and reviewing stages: generative assistants can support idea generation and sentence-level drafting, while correction and feedback systems operate during reviewing, flagging or repairing errors that the writer's own monitoring has not caught. For Arabic learners, whose cognitive load during translating is plausibly heightened by the need to manage root-and-pattern morphology, case endings, and diacritisation simultaneously, AI-mediated reviewing support may free working-memory resources that would otherwise be consumed by low-level error monitoring—a possibility consistent with reports that AI tools are perceived by Arabic writing students as cognitive support for idea generation, vocabulary, and coherence (Arif et al., 2026).

A complementary perspective comes from Vygotsky's (1978) sociocultural theory, in which learning is mediated by more capable others operating within a learner's zone of proximal development, and from Swain's (1985) output hypothesis, which holds that producing language—and noticing gaps when that production is challenged—drives acquisition. AI writing tools can be read as a novel category of mediating artefact: unlike a human tutor, they are available on demand and at scale, but unlike a static reference grammar, they respond directly to the learner's own output. Whether such tools function as a genuine 'more capable other' or merely as an error-removal mechanism depends substantially on whether the feedback they provide pushes learners to notice and revise, consistent with Swain's pushed-output logic, or simply replaces the learner's own corrected text—a distinction that recurs throughout the empirical literature reviewed below.

Hyland's (2003) genre-based pedagogy adds a further dimension, emphasising that writing development depends not only on sentence-level accuracy but on learners' growing command of the discourse conventions appropriate to a given genre or register—an academic essay, a formal letter, a religious commentary—each of which carries distinct expectations within the Arabic rhetorical tradition. Most current AI writing tools, whether general-purpose or Arabic-specific, are markedly stronger at sentence-level grammatical and orthographic correction than at genre-level guidance, a gap that recurs as a theme in the Findings and Discussion below.

Feedback theory provides the review's second theoretical pillar. Hattie and Timperley (2007) frame effective feedback around three questions—where am I going (feed up), how am I going (feed back), and where to next (feed forward)—and argue that feedback is most powerful when it operates at the level of the task and the process rather than merely at the level of the self (for example, generic praise), a distinction with direct relevance for evaluating AI feedback messages, many of which currently default to terse, task-level correction without process-level explanation. Within L2 writing research specifically, Truscott's (1996) influential argument that grammar correction is ineffective and potentially harmful provoked a sustained rebuttal from Ferris (1999) and a subsequent body of empirical work, synthesised by Bitchener and Ferris (2012), suggesting that written corrective feedback can support accuracy development under certain conditions—particularly when it is direct, focused, and followed by opportunities for revision. Hyland and Hyland (2006) further distinguish feedback by focus (form, content, organisation) and by mode (direct correction, indirect marking, metalinguistic comment), a typology that maps usefully onto the range of correction strategies offered by contemporary AI writing tools, from simple spell-checking to explanation-rich grammatical error correction (Magdy et al., 2024).

Finally, Davis's (1989) Technology Acceptance Model (TAM) informs how learners' and teachers' perceptions of an AI writing tool's usefulness and ease of use shape whether, and how consistently, the tool is actually adopted in practice—an important caveat given that the most linguistically sophisticated correction system offers no pedagogical benefit if learners or teachers decline to engage with it (Warschauer & Grimes, 2008). In the Arabic context, perceived ease of use is shaped not only by interface design but by practical considerations such as Arabic keyboard layouts, right-to-left text rendering, and the autocorrect behaviour of mobile input systems, all of which can influence whether learners experience an AI writing tool as supportive or as an additional source of friction.

Synthesising these strands, this review adopts a conceptual framework—termed here the AI-Mediated *Mahārah al-Kitābah* Model—comprising three interacting layers: (1) a correction layer, in which AI systems detect and repair orthographic, morphological, and syntactic errors; (2) a feedback-mediation layer, in which the manner, focus, and timing of that correction is communicated to the learner, drawing on Hattie and Timperley's (2007) feed up/feed back/feed forward structure; and (3) a writing-development layer, in which repeated, mediated engagement with correction and feedback is hypothesised to support gains in accuracy, fluency, and rhetorical control over time, consistent with cognitive process and sociocultural accounts of writing development. The remainder of this review uses this three-layer framework to organise the synthesis of empirical and technical findings.

Method

Research Design

This review adopts a narrative literature review design (Snyder, 2019) rather than a systematic review or meta-analysis. A narrative approach was judged more appropriate than a fully systematic protocol for three reasons. First, the literature on AI writing tools for Arabic spans markedly heterogeneous methodologies—qualitative phenomenological studies, research-and-development (R&D) interventions, and computational benchmarking papers reporting automatic evaluation metrics—that are not readily pooled through meta-analytic procedures. Second, the field is genuinely emergent: several of the most directly relevant Arabic-specific writing-assistance systems and benchmarks were published only in 2024–2026, meaning that a narrow systematic search restricted to peer-reviewed journal articles risks omitting highly relevant conference papers and preprints. Third, the review's purpose is conceptual integration across three otherwise separate literatures—Arabic NLP, computer-assisted language learning, and second language writing and feedback theory—a task for which integrative narrative review methods are well suited (Snyder, 2019).

Data Sources and Search Strategy

Relevant literature was identified through Scopus, Google Scholar, ERIC, the ACL Anthology, and arXiv, reflecting the review's dual disciplinary base in education/applied-linguistics databases and computational-linguistics repositories. Search strings combined the term 'Arabic' with writing-related descriptors ('writing', '*kitābah*', 'composition', 'grammatical error correction') and AI-related descriptors ('artificial intelligence', 'ChatGPT', 'large language model', 'automated feedback', 'automated writing evaluation'). The search was limited to sources published between 2015 and 2026, with the theoretical literature on writing and feedback—drawn principally from the 1978–2012 period—retained regardless of date because of its foundational status within the field. English- and Indonesian-language sources were both eligible, reflecting the substantial volume of Arabic-education research published in Indonesian journals.

Inclusion and Exclusion Criteria

Sources were included if they (a) addressed AI, NLP, or computer-assisted tools applied to Arabic writing, grammar, or composition instruction; (b) reported empirical findings, system descriptions, or benchmark results relevant to correction or feedback; or (c) provided theoretical grounding for writing development, feedback, or technology adoption applicable to the review's conceptual framework. Sources were excluded if they addressed AI applications to listening, speaking, or reading skills exclusively, or if they discussed Arabic NLP techniques with no plausible pedagogical application. An initial pool of approximately 140 records was identified; after removing duplicates and screening titles and abstracts against these criteria, 33 sources were retained for full review and thematic synthesis, including a small number of seminal pre-2015 theoretical works retained for conceptual grounding. As with any narrative review, the selection and interpretation of sources involved a degree of researcher judgement; to mitigate this, search terms and inclusion criteria were applied consistently across databases, and borderline sources were cross-checked against the review's three guiding research questions before inclusion.

Data Analysis

Following Snyder's (2019) guidance for narrative reviews, retained sources were read in full and open-coded for tool type, correction or feedback mechanism, reported effect on writing development, and reported challenge or limitation. Codes were iteratively clustered into four overarching themes—tool landscape, correction and feedback mechanisms, effects on writing

development, and linguistic/technological/pedagogical challenges—which structure the Findings and Discussion section below. Themes and representative sources were further organised into a summary table (Table 1) to support transparent mapping between tool categories and the studies from which they are drawn.

Findings and Discussion

The Landscape of AI Writing Tools for Mahārah al-Kitābah

The reviewed literature reveals at least four overlapping categories of AI writing tool relevant to *mahārah al-kitābah* instruction, summarised in Table 1. The first category comprises general-purpose generative AI systems—principally ChatGPT and related GPT-family models—that were not designed specifically for Arabic but have been appropriated by teachers and researchers through customised prompting. Zubaidi et al. (2025) used GPT-based prompts within a research-and-development framework to support *Insyā'* (free composition) writing among Indonesian Islamic-university students, reporting that 89.6% of expert reviewers rated the resulting AI-assisted learning model as excellent, that 88% of students exceeded the minimum learning standard, and that measured writing proficiency increased by 12.5% over the course of the trial. Tamam et al. (2024) similarly used ChatGPT to analyse Arabic texts for *nahwu* (grammar) instruction, while Rahmouni (2024) examined ChatGPT's effectiveness specifically for teaching Arabic case endings (*i'rab*), reporting mixed but generally encouraging results alongside notable error patterns.

The second category consists of Arabic-specific writing-assistance datasets and platforms purpose-built to address the data scarcity that limits generic LLMs' Arabic performance. The Gazelle dataset (Magdy et al., 2024) curates instruction-style training data across five Arabic writing-assistance tasks—grammatical error correction, metaphor and multiword-expression handling, text refinement, rule explanation, and *i'rab* (declension) guidance—explicitly designed to improve not only the accuracy but the explanatory quality of AI-generated corrections. ARWI, the Arabic Write and Improve platform (Chirkunov et al., 2025), extends the long-running English Write and Improve system to Arabic, offering learners CEFR-referenced proficiency estimates alongside error feedback, with user profiling for dialect, native language, and proficiency level.

The third category comprises Arabic grammatical error detection and correction (GEC) systems developed primarily within the computational linguistics literature. Alhafni et al. (2023) reported substantial empirical advances in Arabic GEC and detection, while Kwon et al. (2023) evaluated several general-purpose LLMs on Arabic GEC and found their performance markedly weaker than on equivalent English tasks, underscoring the persistent resource gap between Arabic and better-resourced languages. Most recently, Mubarak et al. (2026) introduced Nahw, a comprehensive benchmark spanning Arabic grammar understanding, error detection, correction, and explanation, designed explicitly to push the field beyond correction alone toward the kind of explanatory feedback that L2 writing pedagogy regards as more developmentally valuable (Hattie & Timperley, 2007).

The fourth category is the corpus infrastructure underpinning the systems above. The ZAEBUC corpus (Habash & Palfreyman, 2022) annotates bilingual Arabic-English essays written by first-year university students, providing one of the few learner-error resources specific to Arabic academic writing, while earlier work by Alfaifi and Atwell (2012) had already begun establishing a taxonomy of Arabic learner errors. Together, these four categories indicate a field that is maturing rapidly but unevenly: generic generative tools are already in

classroom use despite limited Arabic-specific validation, while the more linguistically rigorous Arabic-specific systems remain largely confined to computational-linguistics venues and have not yet been widely evaluated in classroom settings.

Table 1. Categories of AI Writing Tools Relevant to *Mahārah al-Kitābah* Instruction

Category	Representative Tools/Systems	Primary Function	Key Source(s)
General-purpose generative AI (appropriated)	ChatGPT / GPT-4o (custom-prompted)	Drafting support; *nahwu* explanation; case-ending (*i'rab*) feedback	Zubaidi et al. (2025); Tamam et al. (2024); Rahmouni (2024)
Arabic-specific writing-assistance datasets/platforms	Gazelle; ARWI (Arabic Write and Improve)	Instruction-tuned correction with explanation; CEFR-referenced proficiency feedback	Magdy et al. (2024); Chirkunov et al. (2025)
Arabic grammatical error correction (GEC) systems	Empirical Arabic GEC models; Nahw benchmark	Detection, correction, and explanation of grammatical and orthographic errors	Alhafni et al. (2023); Kwon et al. (2023); Mubarak et al. (2026)
Corpus and resource infrastructure	ZAEBUC corpus; Arabic Learner Corpus (ALC)	Annotated learner-error data underpinning model training and evaluation	Habash & Palfreyman (2022); Alfaifi & Atwell (2012)

Note. Table compiled by the authors based on the literature reviewed in this article.

Mechanisms of AI-Mediated Correction and Feedback

Read through Hattie and Timperley's (2007) feedback framework, the reviewed tools cluster unevenly across the feed-back and feed-forward functions. Most systems excel at feed back—identifying that an error has occurred and, in many cases, what the correct form should be—but comparatively few provide robust feed-forward guidance that helps learners generalise a rule to future writing. This imbalance echoes Hyland and Hyland's (2006) observation, made in the pre-AI feedback literature, that direct correction is more common than indirect or metalinguistic feedback, even though the latter is generally considered more developmentally productive. Among Arabic-specific systems, the Gazelle dataset (Magdy et al., 2024) is a notable exception, deliberately incorporating rule explanation and error-type categorisation alongside correction, a design choice that aligns with Mubarak et al.'s (2026) Nahw benchmark and reflects a broader shift in the GEC literature toward explainable correction (Alhafni et al., 2023).

At the linguistic level, AI correction in the reviewed literature operates across three layers consistent with the correction layer of this review's conceptual framework: orthographic correction (for example, *hamza* placement, *alif maqsūrah* versus *yā'*, and *tā' marbūṭah* versus *tā' maftūḥah*, all frequent sources of error even among educated native writers); morphosyntactic correction, including verb-pattern (*wazn*) selection and case-ending accuracy, which Rahmouni (2024) and Tamam et al. (2024) both identify as persistently difficult for AI systems to handle reliably because case marking in MSA depends on syntactic context that shallow pattern-matching approaches frequently misjudge; and discourse-level correction, which remains the least developed layer across all reviewed tools, consistent with the genre-pedagogy gap noted in the Theoretical Framework above (Hyland, 2003).

The long-running debate between Truscott (1996) and Ferris (1999) over whether grammar correction is worth providing at all takes on a different cast in the AI era. Truscott's central objection—that correction consumes scarce teacher time for uncertain benefit—loses some of its force when correction is automated and available on demand at near-zero marginal cost, a point implicit in Godwin-Jones's (2022) observation that AI writing assistance has radically reshaped the economics of providing feedback. Yet Ferris's (1999) counter-argument, that selective, well-timed correction can support accuracy development, raises the question of whether AI systems correct selectively and intelligibly or simply flag every detectable deviation, overwhelming the learner with undifferentiated feedback. The reviewed evidence is mixed: Koltovskaia's (2020) case-study findings on Grammarly's automated written corrective feedback in an ESL context—frequently cited as a benchmark for this line of research more broadly—found that learners' behavioural, cognitive, and affective engagement with automated feedback varied considerably between individuals, with some learners uncritically accepting suggestions and others engaging in little genuine cognitive processing of the feedback provided, a pattern that Arabic-focused studies have begun to echo (Arif et al., 2026).

Effects on the Development of Mahārah al-Kitābah

Empirical evidence on writing-skill outcomes remains comparatively thin but is growing. Zubaidi et al. (2025) reported encouraging results from a limited trial involving 40 Indonesian Islamic-university students, although the single-site research-and-development design, conducted without a comparison group, limits causal inference. Nandatasia and Nugrahawan (2026) took a different approach, using AI-generated visual stimuli to scaffold picture-based controlled writing tasks, reporting that the technique supported learners' *mahārah al-kitābah* development by anchoring composition in concrete, AI-produced imagery rather than abstract prompts alone—an application that draws implicitly on the cognitive-load reasoning embedded in Flower and Hayes's (1981) process model.

Qualitative evidence offers a complementary picture of how learners experience these tools. Arif et al.'s (2026) phenomenological study of 23 Arabic-education students found that learners perceived AI as a cognitive support tool that assisted with idea generation, grammatical structuring, vocabulary enrichment, and coherence and fluency—findings broadly consistent with the sociocultural account of AI as a mediating artefact discussed above (Vygotsky, 1978)—while simultaneously raising concerns about academic integrity and the risk that convenient correction could erode learners' independent writing competence. This tension between perceived benefit and perceived risk recurs across the broader generative-AI-in-education literature (Kasneci et al., 2023) and appears, on the present evidence, to be at least as salient for Arabic writing instruction as for other languages.

Taken together, the reviewed studies suggest that AI writing tools can plausibly support gains in surface-level accuracy and, through scaffolding mechanisms such as visual stimuli or low-stakes correction, in fluency and engagement as well. Evidence for gains in higher-order rhetorical or genre competence, by contrast, is largely absent from the reviewed literature, mirroring the discourse-level gap identified in the mechanisms discussion above and suggesting that current AI tools support *mahārah al-kitābah* unevenly across Hyland's (2003) hierarchy of writing competence.

Linguistic and Technological Challenges Specific to Arabic

Several intersecting challenges recur across the reviewed literature. The first is Arabic's templatic, root-and-pattern morphology, in which a single triliteral or quadriliteral root can

generate dozens of related word forms through internal vowel and affix changes, a property that complicates both human learning and automatic error detection, since a single misplaced vowel pattern can change a word's meaning, part of speech, or grammatical role entirely (Alhafni et al., 2023). The second is the routine omission of short-vowel diacritics in everyday Arabic writing, which renders many texts genuinely ambiguous without context—a challenge that affects native and non-native writers alike but is especially consequential for automatic correction systems trained primarily on fully diacritised reference text.

A third challenge is dialectal variation: learners' spoken competence is typically grounded in a regional dialect, while formal writing instruction targets Modern Standard Arabic, creating systematic interference patterns that general-purpose AI systems—trained predominantly on MSA- or English-dominant corpora—are not well equipped to recognise or address (Kwon et al., 2023). A fourth and more structural challenge is data scarcity itself: Magdy et al. (2024) explicitly motivate the Gazelle dataset by noting that underrepresented languages such as Arabic encounter significant challenges in developing advanced AI writing tools largely because of the limited availability of training data, a scarcity that constrains model training and, by extension, the sophistication of feedback such models can offer. Recent benchmark and corpus efforts—ZAEBUC (Habash & Palfreyman, 2022), Gazelle (Magdy et al., 2024), and Nahw (Mubarak et al., 2026)—represent direct responses to this gap, though their classroom validation remains limited at the time of this review.

A final, more pedagogical challenge concerns overreliance and academic integrity, raised explicitly by Arif et al. (2026) and echoed in broader generative-AI-in-education discussions (Kasneci et al., 2023). This concern is consistent with broader assessments of the opportunities, challenges, and ethical considerations associated with generative AI in Arabic learning and teaching more generally (Moustafa et al., 2026). If AI correction becomes a substitute for, rather than a scaffold toward, independent error-monitoring, the writing-development layer of this review's conceptual framework may fail to materialise even where the correction layer performs well.

Pedagogical and Curricular Implications

The synthesis above suggests several implications for Arabic EdTech curriculum design and teacher practice. First, given the uneven strength of current AI tools across orthographic, morphosyntactic, and discourse-level correction, a blended model in which AI handles high-frequency, lower-order error correction while teachers concentrate instructional time on genre, rhetorical organisation, and content development (Hyland, 2003; Hyland & Hyland, 2006) appears more defensible than either a fully automated or a fully manual feedback regime. Second, consistent with Davis's (1989) Technology Acceptance Model and with Warschauer and Grimes's (2008) finding that teachers' prior beliefs about writing pedagogy shaped how automated writing assessment tools were actually used in English-language classrooms, the effective integration of AI writing tools into *mahārah al-kitābah* instruction will likely depend as much on teacher professional development and belief change as on the technical sophistication of the tools themselves.

Third, given the academic-integrity and overreliance concerns raised by Arif et al. (2026), curricula should explicitly cultivate learners' critical engagement with AI feedback—encouraging them to evaluate, question, and selectively accept corrections rather than passively applying them—an orientation consistent with Albantani et al.'s (2025) finding that generative AI's effect on student engagement and academic outcomes in Arabic learning is shaped by how, not merely whether, learners use such tools. Fourth, given that Arabic instruction in many

Indonesian institutions carries religious-literacy as well as linguistic objectives, curriculum designers should attend to the broader pedagogical framing identified by Dahlan et al. (2023), ensuring that AI-assisted writing instruction supports rather than substitutes for the deeper engagement with Arabic texts that such programmes are intended to cultivate.

Finally, the comparative immaturity of Arabic-specific tools relative to their English counterparts implies a continuing need for collaboration between Arabic NLP researchers and Arabic-language educators: the corpus and benchmark efforts reviewed above (Habash & Palfreyman, 2022; Magdy et al., 2024; Mubarak et al., 2026) are necessary but not sufficient without parallel classroom-based research on how learners actually engage with the feedback these systems generate.

Synthesis: Revisiting the Three-Layer Framework

Mapping the findings above onto the three-layer AI-Mediated *Mahārah al-Kitābah* Model proposed in the Theoretical Framework reveals an uneven pattern of development across the field. The correction layer is comparatively well developed at the orthographic and morphosyntactic level, particularly within Arabic-specific NLP systems (Alhafni et al., 2023; Mubarak et al., 2026), though discourse-level correction remains largely absent. The feedback-mediation layer is more variable: some systems, notably Gazelle (Magdy et al., 2024) and Nahw (Mubarak et al., 2026), explicitly incorporate explanation and rule-based feedback consistent with Hattie and Timperley's (2007) feed-forward function, while many general-purpose generative-AI applications of the kind used by Zubaidi et al. (2025) and Tamam et al. (2024) rely on direct correction whose pedagogical framing depends heavily on how teachers or prompts structure the interaction. The writing-development layer is the least directly evidenced: although several studies report gains in accuracy, fluency, or engagement (Zubaidi et al., 2025; Nandatasia & Nugrahawan, 2026), none of the reviewed studies tracks learners over a sufficiently long period, or with a sufficiently robust comparison design, to establish durable competence gains rather than short-term performance effects. This pattern suggests that the field's most pressing need is not further correction-layer refinement alone, but classroom-based longitudinal research connecting correction and feedback mechanisms to verified writing-development outcomes.

Directions for Future Research

Building on the synthesis above, three directions for future research appear especially pressing. First, given the predominance of single-site, short-duration studies in the empirical literature reviewed (Zubaidi et al., 2025; Nandatasia & Nugrahawan, 2026; Arif et al., 2026), there is a clear need for longitudinal, multi-site studies that track learners' *mahārah al-kitābah* development across an academic term or longer, ideally incorporating a comparison condition so that gains attributable to AI-mediated correction and feedback can be distinguished from gains attributable to instruction or practice alone. Second, because the most linguistically sophisticated Arabic GEC and benchmark systems (Alhafni et al., 2023; Kwon et al., 2023; Mubarak et al., 2026) have so far been evaluated primarily on held-out test sets within computational-linguistics venues, classroom-based evaluation—examining how learners actually interpret and act on the explanations these systems generate—remains largely absent and would meaningfully extend Hattie and Timperley's (2007) feedback framework into the Arabic NLP literature. Third, given the dialectal-variation challenge identified above, future systems and studies might usefully examine how AI writing tools can be designed to recognise and scaffold the transition from a learner's dominant spoken dialect to formal MSA writing, rather than treating dialectal interference solely as an error to be corrected away. Finally,

replication across a wider range of Indonesian and other non-Arab institutional contexts would help establish whether the encouraging early findings reported here generalise beyond the specific Islamic higher-education settings in which most of the reviewed empirical studies were conducted, and would help clarify whether reported proficiency gains persist once the novelty associated with newly introduced technology has diminished.

Conclusion

This review has examined the use of AI writing tools in teaching *mahārah al-kitābah* through the lens of correction, feedback, and writing development, synthesising literature from Arabic natural language processing, computer-assisted language learning, and second language writing theory. The evidence indicates that AI writing tools for Arabic have matured substantially over the past several years, moving from generic, appropriated applications of mainstream chatbots toward purpose-built Arabic-specific datasets, platforms, and benchmarks that increasingly incorporate explanation alongside correction; yet this progress remains constrained by Arabic's templatic morphology, the routine omission of diacritics, dialectal variation, and an underlying scarcity of annotated Arabic writing data relative to English. Empirical findings, drawn primarily from small-scale or single-site studies, suggest that such tools can plausibly support gains in surface-level accuracy, fluency, and learner engagement, while raising legitimate concerns about overreliance and the erosion of independent writing competence if AI correction substitutes for, rather than scaffolds, learner self-monitoring. The review's three-layer framework—correction, feedback-mediation, and writing-development—offers one way of organising future inquiry, and points specifically toward a need for explanation-rich feedback design, sustained teacher-researcher collaboration in building Arabic-specific resources, and longitudinal classroom research capable of distinguishing durable competence gains from short-term performance effects. For curriculum developers and instructors of Arabic as a foreign or second language, the implication is not that AI writing tools should be adopted uncritically or avoided altogether, but that they be integrated deliberately, with explicit attention to which layer of writing development each tool is best positioned to support.

Acknowledgements

The authors thank colleagues in Arabic language education and Arabic natural language processing whose published work made this synthesis possible. This research received no specific external funding.

References

Albantani, A. M., Rozak, A., Sahrir, M. S., & Khoiriyyah, B. (2025). The influence of generative AI on student engagement and academic outcomes in Arabic language learning. Asalibuna, 9(2), 149–163. https://doi.org/10.30762/asalibuna.v9i02.6810

Al-Hamad, M. Q., & Al-Shara, I. (2022). Integrating technology in Arabic language teaching: Challenges and opportunities. International Journal of Arabic-English Studies, 22(1), 45–68. https://doi.org/10.33806/ijaes2000.22.1.3

Alfaifi, A., & Atwell, E. (2012). Arabic learner corpora (ALC): A taxonomy of coding errors. In Proceedings of the 8th International Computing Conference in Arabic. King Saud University.

Alhafni, B., Inoue, G., Khairallah, C., & Habash, N. (2023). Advancements in Arabic grammatical error detection and correction: An empirical investigation. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (pp. 6430–6448). Association for Computational Linguistics.

Alqahtani, M. (2019). The use of mobile applications in EFL/ESL classroom contexts. Arab World English Journal (AWEJ), 1, 3–15. https://doi.org/10.24093/awej/call1.1

Arif, A. A., Setiyawan, A., & Hamdun, D. (2026). Artificial intelligence in Arabic writing instruction: Students' lived experiences in learning *mahārah al-kitābah*. Alibbaa': Jurnal Pendidikan Bahasa Arab, 7(1), 206–225. https://doi.org/10.19105/ajpba.v7i1.23558

Bitchener, J., & Ferris, D. R. (2012). Written corrective feedback in second language acquisition and writing. Routledge.

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.

Dahlan, J., Abdurrahman, M., & Rif'at, A. D. (2023). The utilization of artificial intelligence in Arabic language learning: Between linguistic competence and Islamic spirituality. International Journal of Islamic Thought and Humanities, 2(2), 380–394. https://doi.org/10.54298/ijith.v2i2.733

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly, 13(3), 319–340. https://doi.org/10.2307/249008

Ferris, D. R. (1999). The case for grammar correction in L2 writing classes: A response to Truscott (1996). Journal of Second Language Writing, 8(1), 1–11. https://doi.org/10.1016/S1060-3743(99)80110-6

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. College Composition and Communication, 32(4), 365–387.

Godwin-Jones, R. (2022). Partnering with AI: Intelligent writing assistance and instructed language learning. Language Learning & Technology, 26(2), 5–24. https://doi.org/10.64152/10125/73474

Habash, N., & Palfreyman, D. (2022). ZAEBUC: An annotated Arabic-English bilingual writer corpus. In Proceedings of the Thirteenth Language Resources and Evaluation Conference (pp. 79–88). European Language Resources Association.

Hattie, J., & Timperley, H. (2007). The power of feedback. Review of Educational Research, 77(1), 81–112. https://doi.org/10.3102/003465430298487

Hyland, K. (2003). Genre-based pedagogies: A social response to process. Journal of Second Language Writing, 12(1), 17–29.

Hyland, K., & Hyland, F. (2006). Feedback on second language students' writing. Language Teaching, 39(2), 83–101.

Ismail, Moh. (2019). Design objective tests using the computer program in developing the Arabic language for non-native speakers. *At-Ta'dib*, *14*(2), 1. https://doi.org/10.21111/at-tadib.v14i2.4824

Ismail, Moh., Ahmad, F. S., & Ma'ruf, M. A. (2023). The Impact of Learning Management System "arabi.id" Web-Based Application on Developing Arabic Language Skills. *Al-Ta'rib : Jurnal Ilmiah Program Studi Pendidikan Bahasa Arab IAIN Palangka Raya*, *11*(2), 213–232. https://doi.org/10.23971/altarib.v11i2.6591

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 103, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Koltovskaia, S. (2020). Student engagement with automated written corrective feedback (AWCF) provided by Grammarly: A multiple case study. Assessing Writing, 44, 100450. https://doi.org/10.1016/j.asw.2020.100450

Kwon, S. Y., Bhatia, G., Nagoudi, E. M. B., & Abdul-Mageed, M. (2023). Beyond English: Evaluating LLMs for Arabic grammatical error correction. arXiv. https://doi.org/10.48550/arXiv.2312.08400

Magdy, S. M., Alwajih, F., Kwon, S. Y., Abdel-Salam, R., & Abdul-Mageed, M. (2024). Gazelle: An instruction dataset for Arabic writing assistance. In Findings of the Association for Computational Linguistics: EMNLP 2024. Association for Computational Linguistics. https://doi.org/10.48550/arXiv.2410.18163

Moustafa, A. H. A., Al-Hamad, M. F., Qureshi, M. A., & Qadir, J. (2026). Generative AI for learning and teaching Arabic: Opportunities, challenges, and ethical considerations. IEEE Access, 14. https://doi.org/10.1109/ACCESS.2026.3651864

Mubarak, H., Hawasly, M., & Mohamed, A. (2026). Nahw: A comprehensive benchmark of Arabic grammar understanding, error detection, correction, and explanation. In Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Vol. 1, pp. 6310–6328). Association for Computational Linguistics.

Nandatasia, R. A., & Nugrahawan, A. R. (2026). Fostering *mahārah kitābah* through picture-based controlled writing with AI-generated visual stimuli. Alibbaa': Jurnal Pendidikan Bahasa Arab, 7(1), 158–179. https://doi.org/10.19105/ajpba.v7i1.23463

Rahmouni, K. (2024). Exploring the use of ChatGPT in teaching Arabic case endings: Effectiveness, challenges and recommendations. Journal of Educational Technology and Innovation, 6(4), 1–20.

Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. Journal of Business Research, 104, 333–339. https://doi.org/10.1016/j.jbusres.2019.07.039

Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. Gass & C. Madden (Eds.), Input in second language acquisition (pp. 235–253). Newbury House.

Tamam, M. I., Ilahi, M. M. K., Cholilah, Z., Taufiqurrochman, R., & Machmudah, U. (2024). Utilizing ChatGPT for analyzing Arabic texts in the study of *nahwu* (Arabic grammar). KITABA: Journal of Interdisciplinary Arabic Learning, 2(3), 193–208. https://doi.org/10.18860/kitaba.v2i3.28463

Truscott, J. (1996). The case against grammar correction in L2 writing classes. Language Learning, 46(2), 327–369. https://doi.org/10.1111/j.1467-1770.1996.tb01238.x

Vygotsky, L. S. (1978). Mind in society: The development of higher psychological processes. Harvard University Press.

Warschauer, M., & Grimes, D. (2008). Automated writing assessment in the classroom. Pedagogies: An International Journal, 3(1), 22–36. https://doi.org/10.1080/15544800701771580

Warschauer, M., & Healey, D. (1998). Computers and language learning: An overview. Language Teaching, 31(2), 57–71. https://doi.org/10.1017/S0261444800012970

Zubaidi, A., Munip, A., Widodo, S. A., & Zerrouki, T. (2025). Enhancing Arabic writing skills using ChatGPT-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