e-ISSN: 2986-1667
EDUCATUM
Scientific Journal of Education
https://ojs.jurnalbk.com/index.php/educatum

WHY AI STRUGGLES TO UNDERSTAND ARABIC GRAMMAR: A TECHNO-PEDAGOGICAL REVIEW OF COMPUTATIONAL AND LINGUISTIC CHALLENGES

Muhammad Fariq Heemal Attruk[1*], Nik Hanan Mustapha[2]

[1,2.]Department of Arabic Language and Literature, Abdulhamid Abusulayman Kulliyyah of Islamic Revealed Knowledge and Human Sciences, International Islamic University Malaysia, Kuala Lumpur, Malaysia
Corresponding author e-mail: fariqhemal@gmail.com

Abstract. *Generative artificial intelligence is increasingly used in Arabic language education, yet its performance in understanding and processing Arabic grammar remains markedly inconsistent. This narrative review examines why AI systems struggle with Arabic grammar and explores the pedagogical implications of these computational limitations. Drawing on 26 primary studies published between 2020 and 2026, the review identifies five interconnected linguistic challenges: the non-linear root-and-pattern morphology of Arabic, the routine omission of diacritics that obscures grammatical case, extensive dialectal diversity, persistent data scarcity, and the syntactic complexity of the i'rab case system. Empirical evidence shows that even advanced models perform substantially worse on morphological and syntactic tasks than on surface-level tasks, with GPT-4o achieving only 67 percent accuracy on Arabic grammar benchmarks and Arabic-specific models scoring considerably lower. The review demonstrates that the structural features of Arabic that make natural language processing difficult are precisely the features that pose risks for learners who depend on AI without critical oversight. These risks include the formation of misconceptions, overreliance on AI-generated outputs, and the erosion of critical thinking and teacher expertise. The findings suggest that effective AI integration in Arabic grammar instruction requires a Human-in-the-Loop approach, targeted teacher training, and the development of critical AI literacy among learners.*

Keywords: *generative artificial intelligence, large language models, Arabic grammar, Nahwu, Shorof.*

Article info:

Submitted: 13, april, 2026 *Accepted: 10, june, 2026*

How to cite this article:
M. F. H. Attruk, N. H. Mustapha "Why AI Struggles to Understand Arabic Grammar: A Techno-Pedagogical Review of Computational and Linguistic Challenges", *EDUCATUM: Scientific Journal of Education*. Vol. 4, No. 2, pp. 65-76, June, 2026.

1. INTRODUCTION

The rapid advancement of generative artificial intelligence has significantly influenced language education, offering new possibilities for personalized learning, real-time feedback, and interactive practice. Tools such as ChatGPT, Gemini, and other large language models are now widely used by both teachers and learners across various linguistic contexts [1], [2]. In Arabic language education, these technologies hold particular promise. Arabic is spoken by over 400 million people across 22 countries and holds profound religious, cultural, and geopolitical significance [3], [4]. Yet despite its global importance, the integration of AI into Arabic language learning has been comparatively slow and fraught with challenges not encountered to the same degree in languages like English.

The difficulties stem primarily from the unique structural properties of Arabic. The language is characterized by a non-linear, templatic morphological system built upon consonantal roots and vocalic patterns, a system that forms the foundation of *Shorof* (Arabic morphology). A single triliteral root can generate dozens of derived forms with distinct but related meanings, creating combinatorial complexity that challenges probabilistic AI models [3], [4]. This morphological richness is further compounded by orthographic ambiguity resulting from the routine omission of diacritical marks in standard writing, which obscures grammatical case and function [3], [5]. Additionally, Arabic's diglossic nature, the coexistence of Modern Standard Arabic with numerous regional dialects, adds further complexity, as most AI systems are trained predominantly on formal Arabic and perform poorly on dialectal input [6], [7]. The scarcity of high-quality, annotated Arabic datasets further restricts model training and performance, with even the most advanced models exhibiting significant gaps in grammatical comprehension [8], [9].

Recent empirical evaluations confirm that current AI systems fall short of human-level competence in Arabic grammar. Studies have shown that GPT-4o achieves only 67 percent average accuracy on Arabic grammar tasks, while Arabic-specific models score considerably lower [8]. In tasks requiring morphological conjugation (*taṣrīf*) and syntactic analysis (*i'rab*), models frequently produce errors that would be obvious to any proficient Arabic speaker [1], [10], [11]. While many models perform well on surface-level tasks such as spelling and reading comprehension, they struggle with deeper grammatical and syntactic reasoning, often succeeding through memorization or pattern recognition rather than genuine understanding [12]. These limitations have direct implications for learners who increasingly rely on AI tools for language practice and instruction.

Despite the growing body of research on AI in Arabic language education, existing studies tend to focus either on the technical challenges of Arabic natural language processing or on the pedagogical applications of AI in classroom settings. Few have systematically examined how the structural features of Arabic that make NLP difficult are precisely the features that create risks for learners who depend on AI without critical oversight. This connection between computational limitations and pedagogical risks remains underexplored, leaving educators, learners, and technology developers without a comprehensive framework.

This review addresses that gap by asking: Why does AI struggle to understand Arabic grammar, and what are the pedagogical implications of these struggles? Drawing on a narrative literature review of recent empirical studies and computational analyses, this article examines the key linguistic challenges that hinder AI performance in Arabic grammar processing, including morphological complexity, diacritic omission, dialectal diversity, data scarcity, and syntactic intricacy. It critically evaluates how these technical limitations translate into risks for Arabic language learners and educators. By synthesizing findings from both computational linguistics and language pedagogy, this review offers a techno-pedagogical perspective that connects the structural features of Arabic to the practical realities of AI-assisted learning. It argues that effective integration of AI into Arabic grammar instruction requires not only technical improvements but also pedagogical frameworks that maintain human judgment at the center of the learning process.

2. RESEARCH METHODS

This study employed a Narrative Literature Review (NLR) approach to critically examine why generative AI struggles to understand Arabic grammar and to explore the pedagogical implications of these computational limitations. A narrative review was considered appropriate because the topic bridges computational linguistics and language pedagogy, two fields often treated separately but fundamentally interconnected in discussions of AI-assisted Arabic learning. Unlike systematic reviews that seek to answer narrowly defined questions through exhaustive procedures, narrative reviews allow researchers to synthesize diverse forms of evidence, critically interpret findings, identify emerging themes, and generate new

conceptual perspectives [13], [14]. This approach is particularly suitable for exploring broader theoretical and pedagogical implications across multiple disciplines [15], [16].

The literature search was conducted using two major academic sources: Scopus and Google Scholar. Scopus was selected for its access to peer-reviewed international research with established quality standards, while Google Scholar was used to broaden coverage and capture relevant studies that may not yet be indexed in Scopus, particularly in emerging areas such as generative AI applications in Arabic language education. Search terms included combinations of keywords such as "generative AI," "artificial intelligence," "large language models," "ChatGPT," "Arabic language learning," "Arabic grammar," "*nahwu*," "*shorof*," "Arabic NLP," "Arabic morphology," "Arabic syntax," "morphosyntactic tagging," "Arabic diacritics," and "Arabic dialect." The literature search was limited to publications from 2020 to 2026 to ensure that the review reflects recent developments in generative AI and remains relevant to current educational practices.

Following the recommendations of narrative review methodology [14], [17], a purposive and iterative selection process was employed. Studies were included if they: (1) focused on the use of artificial intelligence, large language models, or generative AI in Arabic language processing or Arabic language education; (2) discussed computational, linguistic, pedagogical, or ethical challenges related to AI and Arabic grammar (*Nahwu* and *Shorof*); and (3) were published in peer-reviewed journals, conference proceedings, or other reputable academic sources including preprints from established repositories. This process resulted in a final corpus of 24 primary studies that were considered directly relevant to the objectives of the review and formed the basis of the thematic analysis. Additional methodological and conceptual references were consulted to support the theoretical framing and interpretation of findings. Consistent with the nature of narrative reviews, the emphasis was placed on selecting studies that offered conceptual richness and relevance rather than achieving exhaustive coverage [18].

The selected studies were analyzed through thematic synthesis. Repeated reading and comparison of the literature enabled the identification of recurring concepts, areas of convergence, and critical tensions across studies. Following the recommendations of Torraco [15], the analysis moved beyond descriptive summarization by identifying gaps, tensions, and unresolved questions within the literature. The analytical framework was structured around two guiding questions: first, what computational and linguistic challenges prevent AI from accurately processing Arabic grammar; and second, how do these technical limitations translate into pedagogical risks for Arabic language learners and educators. This process generated six major themes: morphological complexity, diacritic omission, dialectal variation, data scarcity, syntactic intricacy, and pedagogical synthesis, which form the structure of the results and discussion section. Rather than presenting a definitive model, this review offers a techno-pedagogical framework that connects computational challenges to pedagogical risks, serving as a conceptual foundation for future research and practice in AI-assisted Arabic grammar instruction.

3. RESULTS AND DISCUSSION

Table 1. Arabic grammar processing challenges and their pedagogical implications

Source of Challenge	Arabic Linguistic Feature	AI Manifestation	Pedagogical Risk
Morphological complexity	Root-and-pattern system; derivational morphology; inflection for gender, number, case	LLMs struggle with gender/number tagging; errors in *taṣrīf*; models underperform in morphology (80% accuracy)	Learners internalize incorrect morphological patterns; teachers cannot assess genuine learning
Orthographic ambiguity	Omission of diacritics (*fathah, kasrah, dhammah*); homographs; multiple hamza forms	AI mispronounces words; fails to distinguish *ya'lamu/yu'allimu*; hallucinates tokens	Learners receive misleading case explanations; confuse accusative/genitive usage
Dialectal diversity	65+ dialects; diglossia (MSA vs colloquial); Arabizi	Systems trained on MSA fail on dialects; Egyptian Arabic dominates outputs	Learners exposed to culturally inappropriate translations; hybrid varieties confuse
Data scarcity	Arabic only 1% of internet; datasets lack dialects; noisy/translated data	GPT-4o 67% accuracy on Arabic grammar; ALLaM-7B 42%; synthetic data 15.8% improvement vs natural 44.8%	AI lacks reliability for pedagogical applications; instructors distrust AI outputs
Syntactic complexity	*I'rab* case system; free word order (SVO, VSO,	ChatGPT 99% accuracy but subtle particle errors; syntax	Learners receive inaccurate case

	VOS, OVS); agreement rules	most challenging (53.3% accuracy)	information; cannot identify subtle grammatical errors

The Morphological Bottleneck: Why AI Struggles with Arabic Root-and-Pattern Morphology

Arabic morphology, the study of word structure and formation known in the Arabic tradition as *Shorof*, presents one of the most significant computational challenges for artificial intelligence systems. Unlike English, which relies primarily on concatenative affixation, Arabic employs a non-linear, templatic morphological system built upon consonantal roots and vocalic patterns [3]. This system allows a single triliteral root such as "*k-t-b*" to generate dozens of derived forms with distinct but related meanings, such as *kataba* (he wrote), *kutub* (books), *maktab* (office), and *maktūb* (written) [3], [4].

This richness increases not only the number of possible word forms but also the complexity of identifying the correct grammatical representation from context. Arabic words are highly inflected for number, gender, voice, and case, with various clitics attached for conjunction, definiteness, prepositions, and pronominal possession [3], [5]. A single word can carry multiple layers of grammatical meaning through prefixes, suffixes, and infixes embedded within the root itself [4]. The productivity of derivational morphology further exacerbates the challenge, allowing the creation of active and passive participles, abstract nouns, nouns of place and time, occupational nouns, and numerous other forms through systematic pattern alternations [3]. Morphological analysis is therefore substantially more demanding for AI systems than in morphologically simpler languages.

Despite the development of sophisticated Arabic-specific language models, empirical evaluations reveal persistent limitations in morphological processing. Adel et al. [19] found that even the best-performing LLMs struggled with specific morphological features, particularly gender and number, with strict morphosyntactic bundle prediction remaining challenging. Kwon et al. [20], [21] similarly reported that Arabic's rich morphology complicated grammatical error correction, with instruction-fine-tuned models outperformed by fully fine-tuned ones.

Perhaps the most direct evidence comes from studies of *taṣrīf* (morphological conjugation) and *wazan* (morphological pattern) analysis. Karima et al. [10] compared ChatGPT and Deepseek in producing morphological conjugations for 30 triliteral augmented verbs across all major categories. ChatGPT scored 54 out of 120 points, with errors concentrated in *maṣdar mīmī*, *ism makān*, *ism zamān*, and *i'lāl* rules. Deepseek scored 61 out of 120, showing closer alignment with classical rules yet still producing no fully correct responses. Both models, the authors concluded, require human supervision to maintain adherence to classical standards [10]. The AraLingBench benchmark further confirmed these difficulties, with models consistently underperforming in morphology and high-performing models reaching only 80 percent accuracy, significantly lower than their performance on surface-level tasks [12].

These recurring failures cannot be attributed solely to insufficient training data or imperfect optimization. They also reflect deeper architectural limitations in how current LLMs represent Arabic morphology. Abdelrehim et al. [22] explain that Arabic's morphological richness creates a vast combinatorial space that probabilistic models struggle to navigate without extensive training data. Othman and Asbulah [9] emphasize that the Arabic morphological system is so complex that many rules cannot easily align with simple computational models. Saoudi and Gammoudi [7] note that morphological analysis is one of the most demanding tasks in Arabic NLP due to the language's rich morphological structure encompassing numerous allomorphs and exceptions. These findings point to a challenge rooted primarily in representing the linguistic structure of Arabic rather than in simply increasing computational scale.

These morphological limitations carry serious pedagogical risks. Abidin and Sain [23] found that generative AI fails to maintain the integrity of *Shorof*, while Karima et al. [10] demonstrated that even the best models require human supervision to produce accurate *taṣrīf*. Learners who depend on AI without verification risk internalizing incorrect morphological patterns, a risk compounded by the fact that many models succeed through memorization rather than authentic comprehension [12]. Frequent reliance on AI-generated outputs may lead learners to reproduce correct forms procedurally while still lacking the morphological reasoning needed to generate or evaluate new forms independently. Since Arabic is commonly written without diacritical marks, AI systems must also address the ambiguity created by missing short vowels.

The Diacritic Void: How Missing Short Vowels Undermine AI's Syntactic Analysis

Arabic orthography presents a fundamental challenge for AI systems through the absence of diacritical marks in standard writing. Diacritics, including short vowel marks such as *fathah*, *kasrah*, and *dhammah*, are essential for indicating pronunciation, grammatical function, and meaning, yet they are routinely omitted from most modern Arabic texts except religious and educational materials [3], [4], [5].

This omission creates pervasive ambiguity because one sequence of consonants can represent multiple distinct words with different grammatical functions. Alayba [3] illustrates this with the example of "*ktb*" without diacritics, which can be read as *kataba* (he wrote), *kutub* (books), or *kutiba* (it was written), depending on the context. Boulesnam and Boucetti [4] similarly demonstrate that the non-voweled word "*br*" supports multiple alternatives including *birr* (obedience), *barr* (land), and *burr* (wheat). This ambiguity forces both human readers and AI systems to infer meaning from context, which current AI systems still perform inconsistently.

The problem extends beyond simple vowel marking. Arabic orthography presents additional complexities including multiple forms of letters such as *Hamza* (ء, أ, إ, آ, ؤ, ئ), the phonetic similarity between *taa marbuta* (ة) and *ha* (ه) at word endings, and the absence of capitalization as a marker of proper nouns [3], [4], [5]. As Boulesnam and Boucetti [4] note, restoring or generating diacritics is a crucial additional task for Arabic NLP, yet one that current systems perform imperfectly.

Empirical studies confirm these technical limitations. Rizki et al. [24] evaluated Qalam AI, an automatic harakat detection system, and found that it systematically failed to distinguish between words identical in consonants but differing only in diacritics. Their analysis revealed that Qalam AI could not differentiate *ya'lamu* (he knows) from *yu'allimu* (he teaches) or *isytara* (he bought) from *isytari* (buy!), representing failures in *ṣarf*-based ambiguity and contextual interpretation that reflect broader challenges in Arabic NLP [24].

Al-Jarf [25] documented similar patterns in AI-narrated Arabic YouTube videos, finding that AI narrators made pronunciation errors primarily in diacritics and homographs. These errors accounted for 32 percent of all mispronunciations and led to ambiguity, altered meaning, or nonsense words. The study further found that AI failed to match pronunciation with context in 29 percent of cases and made errors in jussive verb vocalization (17%), confusion between verb and noun forms (11%), and distinctions between active and passive participles (7%). Al-Jarf [25] concluded that AI struggles because it cannot infer correct vocalization from context, unlike human readers who rely on linguistic knowledge and contextual cues. These findings show that missing diacritics affect not only pronunciation but also the identification of grammatical roles, thus reducing the accuracy of AI-based syntactic analysis.

Adel et al. [19] examined Arabic morphosyntactic tagging and dependency parsing and found that tokenization errors substantially affected performance in raw-text settings. Their analysis revealed that while specialized parsers like CamelParser made errors dominated by surface-level issues such as punctuation normalization, LLMs like Gemini3 produced errors primarily in the form of hallucinations, generating tokens that are not present in the input. These findings indicate that tokenization errors become substantially more difficult to resolve when Arabic text lacks diacritics, forcing AI systems to infer word boundaries, morphological identity, and grammatical function simultaneously.

The absence of diacritics carries significant pedagogical risks. Rahmouni [1] found that ChatGPT's explanations of Arabic case endings frequently contained misleading information because the omission of diacritics obscures the grammatical markers essential for understanding *i'rab*. Students who rely on such explanations without verification risk internalizing incorrect rules. Abidin and Sain [23] similarly reported that generative AI produces syntactic inconsistencies that could undermine curriculum quality. Moustafa et al. [6] and Samiya [26] further note that AI systems trained on datasets with limited diacritic representation fail to generate or restore diacritics accurately, creating output that is pedagogically confusing for learners.

Dialectal and Orthographic Noise: The Challenge of Arabic Variation

Arabic is not a single uniform language but encompasses a wide spectrum of regional dialects that differ significantly from one another and from Modern Standard Arabic (MSA). Khalatia and Al-Romany [5] note that Arabic has more than sixty-five different dialects due to the diversity of countries and regions, each with its own characteristics, sounds, and expressive structures. This diversity complicates machine translation because users encounter differences in morphology, vocabulary, grammatical cases, and conjugation across dialects [5].

The diglossic nature of Arabic further compounds these difficulties. Boulesnam and Boucetti [4] explain that MSA serves as the high variety used in formal contexts such as literature, education, and media, while numerous spoken colloquial dialects function as low varieties in everyday communication. Each dialect

possesses its own vocabulary and pronunciation patterns. The situation is complicated further by '*Arabizi*,' a non-standard romanization in which Arabic text is written using a mixture of Latin characters, numbers, and punctuation marks, commonly used by Arab internet users on social networks and discussion forums [4], [7]. All these varieties constitute challenges for NLP applications that must process both MSA and its different dialects [4], [7].

The consequences of dialectal diversity for AI performance are well documented. Saoudi and Gammoudi [7] observe that most existing NLP and speech systems are trained predominantly on MSA, resulting in limited generalizability and reduced performance when applied to colloquial speech. The absence of large annotated corpora representing diverse dialects further impedes the development of robust, dialect-sensitive models [7]. Moustafa et al. [6] add that Egyptian Arabic, which dominates online media, frequently shapes LLM output, leading to responses that disproportionately reflect Egyptian usage. This influence can carry over into MSA, subtly altering syntactic structures and lexical choices, thereby compromising the neutrality expected of the formal register and risking learner confusion.

Orthographic variation adds another layer of difficulty. Arabic script lacks capitalization, meaning there are no uppercase or lowercase letters to distinguish proper names from common nouns [4], [5]. The shape of Arabic letters changes depending on their position within a word, a feature that NLP systems must accommodate [4]. Alayba [3] further notes that *hamza* has multiple forms (ء, أ, إ, آ, ؤ, ئ), and *taa marbuta* (ة) has similar phonetics to *ha* (ه) at word endings, creating orthographic ambiguity that complicates computational processing. These orthographic features, combined with the absence of diacritics and the lack of standard orthographic rules for dialects, make Arabic text processing particularly challenging for AI systems [3], [4], [9].

The pedagogical implications of these dialectal and orthographic challenges are serious. Abidin and Sain [23] argue that AI tools struggle with processing diverse dialects and ensuring cultural authenticity. They note that even neural machine translation, despite progress in handling idiomatic expressions, still requires human intervention for cultural nuances. AlSbou et al. [27] report that ChatGPT struggles with dialectal variations and cultural nuances, especially in complex text types, and that translations often fail to capture the cultural depth of Arabic expressions. Alkaabi and Almaamari [2] found that instructors struggled with AI accuracy in dialect processing and cultural authenticity, and they observed that AI gave different answers in different Arabic dialects, making it difficult to determine whether students were genuinely learning Arabic.

Al-Jarf [28] provides concrete examples of these failures. In translating the English expressions "I tip my hat to you" and "Hats off to you," DeepSeek produced literal translations that were linguistically correct but culturally inappropriate for Gulf Arabic contexts, where the culturally appropriate equivalent involves raising the *'iqāl* (headband) rather than a hat. When corrected, the AI suggested an alternative that was also culturally inaccurate. This pattern, according to Al-Jarf [28], demonstrates that AI translations are often linguistically correct but culturally wrong, representing a pragmatic mismatch that has direct pedagogical consequences for learners who may adopt inappropriate expressions.

Hayad et al. [29] synthesize these findings, noting that the linguistic complexity of Arabic, including its morphology, syntax, and semantics, along with dialectal diversity and the difficulty AI faces in handling orthography and diacritics accurately, constitutes a primary challenge for AI integration into Arabic language education. They conclude that these challenges are not merely technical but also pedagogical, as they limit the reliability of AI tools for authentic language learning. Othman and Asbulah [9] emphasize that addressing these issues requires sustained research into dialect-aware architectures, the creation of expansive and inclusive datasets, and the development of hybrid computational-linguistic frameworks capable of modeling the full breadth of Arabic linguistic variation.

The Resource Scarcity Trap: Why Limited Data Undermines AI Performance

A fundamental challenge for AI systems processing Arabic lies in the scarcity of high-quality linguistic resources. Unlike English, which benefits from vast amounts of annotated text, Arabic suffers from severe data limitations that restrict model training and performance [3], [7], [9]. Othman and Asbulah [9] highlight that Arabic content on the internet constitutes only one percent of total online content, despite Arabic speakers representing five percent of the global population. This imbalance reflects limited Arabic resources compared to languages like English and a shortage of experts and scholars in Arabic language computing [5], [9].

The consequences of this resource scarcity are evident in the availability and quality of Arabic NLP datasets. Alayba [3] notes that most available Arabic NLP datasets focus on MSA or a few specific dialects, while datasets for other dialects remain scarce. Additionally, there is a shortage of well-annotated and labeled Arabic NLP datasets, as well as accurate annotation tools and guidelines. Most existing datasets have limited

quality due to data noise, particularly those retrieved from the web. Datasets also concentrate on formal domains such as politics, commerce, and news, making them unsuitable for informal or specialized tasks. Inconsistencies in labelling and annotation across datasets further complicate effective resource integration [3].

Saoudi and Gammoudi [7] observe that most Arabic question-answering datasets are derived from translations of other languages, often relying on resources like English-SQuAD, TREC, or CLEF. The authors note that translated datasets, such as Arabic-SQuAD and ARCD, contain text elements in languages other than Arabic, including unknown sub-words and characters. This issue arises from inadequate training samples translated from the English SQUAD dataset and introduces language-related complexities that limit model performance [7]. Boulesnam and Boucetti [4] similarly note that the scarcity of Arabic resources further complicates NLP efforts.

The impact of data scarcity on model performance is substantial. Mubarak et al. [8] introduced the *Nahw* benchmark to evaluate Arabic grammar understanding across multiple LLMs. They found that GPT-4o achieved only 67 percent average accuracy across all tasks, while the best-performing Arabic model, ALLaM-7B, scored just 42 percent. These figures indicate that even the most advanced models exhibit significant gaps in Arabic grammar comprehension, partly because available training data is insufficient to cover the full range of Arabic grammatical phenomena. When testing on tasks such as grammatical error detection, the best-performing medium-sized model achieved only 35.74 percent F1 score, and even GPT-4o reached only 64 percent, far below human expert performance [8]. The study also found that scores for theoretical questions were consistently higher than for practical ones, reflecting the scarcity of error-free Arabic content for training practical grammatical usage [8].

Researchers have turned to synthetic data generation as a strategy to address data scarcity. Kwon et al. [20] found that models fine-tuned on synthetic data generated by ChatGPT performed comparably to models trained on the original gold dataset, confirming the utility of ChatGPT for generating synthetic data. They noted that further fine-tuning a model on 10,000 out-of-domain examples resulted in a significant performance drop [20]. In a related study, Kwon et al. [21] showed that synthetic data generated by ChatGPT outperformed equivalent-sized gold datasets when models were further fine-tuned, but performance declined sharply when models were fine-tuned on reverse-noised datasets [21].

Abdelrehim et al. [22] developed a hybrid approach combining rule-based methods and LLMs to generate synthetic data for Arabic grammatical error correction. They found that the distribution of error classes in training data is heavily skewed, with approximately 71 percent of words containing no errors and the most common errors being unnecessary punctuation, *Hamza* errors, and *Taa Marbuta/Haa* confusion errors. Their synthetic data generation recipe, which controlled the distribution of target error types, improved performance on challenging non-orthographic and non-punctuation errors. Models trained with synthetic data performed nearly as well as state-of-the-art systems and even better when punctuation marks were excluded from evaluation [22].

Mubarak et al. [8] further demonstrated that synthetic data improves performance but does not match natural data: 10K synthetic MCQ yielded 15.8 percent improvement, while 5K natural examples yielded 44.8 percent. Natural, expert-written questions contain richer linguistic signals. Nonetheless, synthetic data can bridge knowledge gaps in low-resource scenarios. The authors also noted that a language expert rated the synthetic data 8.1 out of 10 on average, confirming its pedagogical soundness.

The limited availability of high-quality Arabic datasets also carries significant pedagogical implications. Samiya [26] emphasizes that resource scarcity restricts the training and benchmarking of AI models across various tasks, including automatic speech recognition, syntactic parsing, and semantic interpretation. This limitation is compounded by Arabic's rich morphology and orthographic ambiguity, which require highly granular annotations that are resource-intensive to produce. Hayad et al. [29] identify the lack of rich Arabic corpora as a key obstacle, noting that without adequate data, AI models cannot achieve the accuracy needed for reliable pedagogical applications. Moustafa et al. [6] explain that data scarcity poses a significant barrier for GenAI in Arabic education, particularly for models that must handle diverse dialects and complex grammatical structures. Alkaabi and Almaamari [2] observe that Arabic instructors found it challenging to evaluate AI-generated content across different dialects, partly because limited training data restricts AI's ability to produce consistently accurate outputs. While synthetic data offers a partial solution, natural, high-quality data remains essential for achieving peak accuracy.

Syntactic Complexity and the Challenge of *I'rab* Analysis

Arabic syntax presents another significant hurdle for AI systems. The language is characterized by a complex system of inflectional endings that mark grammatical case, known as *i'rab*, along with flexible word

order, intricate agreement patterns, and anaphoric relationships that require sophisticated contextual understanding [4], [9]. These features, which form the foundation of *Nahwu*, create substantial difficulties for computational processing and grammatical analysis.

One of the most challenging aspects is the system of case endings that mark the syntactic function of words. Arabic nouns and adjectives receive specific endings, such as nominative (*marfūʿ*), accusative (*manṣūb*), and genitive (*majrūr*), that indicate their grammatical role in the sentence. Mubarak et al. [8] explain that Arabic grammar establishes rules for case endings that mark syntactic roles, requires verbs to agree with subjects in person, number, and gender, and includes a wide range of particles that alter the case endings of the words they govern. These case markers are essential for understanding sentence structure and meaning, yet they are often omitted in modern written Arabic and must be inferred from context, a task that current AI systems perform imperfectly [1].

Empirical studies demonstrate that AI systems frequently struggle with accurate syntactic analysis. Tamam et al. [11] tested ChatGPT's ability to analyze Arabic grammatical structures from a classical text and found that while the AI achieved 99 percent accuracy in detecting *i'rab*, it still made mistakes in identifying certain particle categories, particularly *huruf jazm nafyi* and *huruf nashab* when an implied particle was present. These subtle errors are problematic because they appear in otherwise correct explanations, making them difficult for learners to detect [11].

Rahmouni [1] conducted a more detailed evaluation of ChatGPT's effectiveness in teaching Arabic case endings and found that despite providing detailed and personalized explanations, the AI frequently gave misleading descriptions of case functions, incorrect examples, and inaccurate information about pronouns and prepositions. ChatGPT claimed, for instance, that prepositions must agree with the nouns they govern in case, which is grammatically incorrect since prepositions in Arabic have fixed forms. It also incorrectly stated that pronouns change their endings to match specific cases, when in fact Arabic pronouns have fixed endings regardless of case. These inaccuracies directly impacted learners' quiz performance, with students frequently confusing accusative and genitive case usage after receiving misleading explanations. Notably, 77.8 percent of students expressed preference for learning from a human tutor rather than ChatGPT, recognizing the tool's limitations in providing reliable grammatical instruction [1].

The challenge extends beyond case endings to the broader domain of syntactic parsing and dependency analysis. Adel et al. [19] evaluated LLMs on dependency parsing and found that tokenization errors substantially affect performance in raw-text settings, with strict morphosyntactic bundle prediction remaining challenging. They also found significant differences in root prediction accuracy, with CamelParser achieving 66.1 percent compared to 85.5 percent for Gemini3. When root identification was excluded, CamelParser's LAS increased from 87.5 to 89.3 percent, surpassing Gemini3, whose LAS only rose marginally from 88.2 to 88.5 percent. Advanced LLMs may excel at some syntactic tasks, but their performance is uneven and depends on specific linguistic phenomena [19].

The free word order of Arabic adds another dimension of difficulty. Boulesnam and Boucetti [4] explain that Arabic is generally considered a free-word-order language, with different word order patterns such as SVO, VSO, VOS, and OVS, all considered correct and conveying the same meaning. This syntactic flexibility complicates NLP applications that must parse and generate sentences with multiple valid structural possibilities. Anaphora resolution also poses a challenge; pronominal anaphora has no meaning independently of its antecedent, while zero anaphora is easily resolved by humans but challenging for automated systems. Agreement rules further complicate processing, as adjectives follow nouns in number, gender, and case, but agreement depends on word order and can be total or partial, creating numerous exceptions [4]. Rizki et al. [24] found that Qalam AI, an automatic harakat detection system, could not distinguish between verb forms with identical consonants but different diacritics, reflecting broader challenges in Arabic NLP related to *ṣarf*-based ambiguity and contextual interpretation.

The AraLingBench benchmark provides comprehensive evidence of syntactic difficulties across 35 Arabic and bilingual LLMs. Zbib et al. [12] found that syntax was the most challenging category for all models, with even the best-performing models achieving only 53.3 percent accuracy on syntactic tasks. Grammar and morphology showed high correlation, indicating they develop as a tightly coupled subsystem, but syntax exhibited the weakest correlations with other categories. This suggests that syntactic reasoning relies on distinct mechanisms and requires compositional and hierarchical understanding that does not emerge directly from surface-level distributional patterns [12].

These syntactic limitations carry serious pedagogical consequences. Hayad et al. [29] report that AI's effectiveness is inconsistent in complex tasks such as grammatical analysis (*i'rab*), where its accuracy still varies. Abidin and Sain [23] emphasize that generative AI faces serious obstacles because of low linguistic

accuracy in maintaining the integrity of Arabic grammar (*nahwu* and *shorof*). Moustafa et al. [6] caution that AI tools trained primarily on MSA often perform poorly with dialectal input, and the dominance of Egyptian Arabic in training data can subtly alter syntactic structures. Al-Jarf [28] and Borham et al. [30] further note that AI systems often fail to capture pragmatic nuance and syntactic ambiguity, highlighting the need for human expertise in language teaching. Consistent with these findings, learners who depend on AI for grammatical instruction may receive inaccurate information about sentence structure and case assignment, potentially undermining their development of syntactic competence.

From Computational Limitations to Pedagogical Risks: A Techno-Pedagogical Synthesis

The preceding sections have documented how the structural complexity of Arabic, including its root-and-pattern morphology, diacritic ambiguity, dialectal diversity, data scarcity, and syntactic intricacy, creates substantial computational challenges for AI systems. This final section synthesizes these findings by examining how these computational limitations translate into pedagogical risks for Arabic language learners and educators, and by identifying strategies for responsible AI integration.

The first and most immediate pedagogical risk is the formation of misconceptions. When learners rely on AI tools that produce inaccurate grammatical analyses, they risk internalizing incorrect rules through repeated use. Alkaabi and Almaamari [2] found that students who used AI for basic Arabic grammar exercises often accepted AI outputs as correct, making it difficult for teachers to assess genuine learning. Rahmouni [1] documented this phenomenon directly: students who received misleading explanations of case endings from ChatGPT subsequently made errors on quizzes, confusing accusative and genitive case usage. As Al-Jarf [28] demonstrates, these inaccuracies extend beyond grammar to include culturally inappropriate translations that are linguistically correct but pragmatically wrong, errors that learners may not recognize without expert guidance.

The risk of misconceptions is particularly acute in morphology and syntax, where the consequences of error are most significant for foundational language competence. Karima et al. [10] found that ChatGPT scored only 54 out of 120 points in producing morphological conjugations (*taṣrīf*), with errors concentrated in complex forms such as *masdar mimi*, *ism makan*, and *ism zaman*, concepts essential for advanced Arabic proficiency. Learners who depend on such inaccurate outputs without verification may develop incomplete or incorrect understandings that become difficult to correct later.

The second major risk is overreliance and the erosion of critical thinking. Hayad et al. [29] identify over-dependence as a primary pedagogical challenge, noting that students increasingly use AI to complete assignments without genuine engagement. Abidin and Sain [23] warn that unsupervised reliance on LLMs for grammar content could pose serious risks to curriculum quality, as students may lose opportunities to develop their own analytical and evaluative skills. Zbib et al. [12] provide empirical evidence for this concern, showing that while models perform well on surface-level tasks, they struggle with deeper structural understanding. Learners who depend on AI may develop superficial competence without genuine linguistic mastery.

A third risk concerns the changing role of teachers and the potential erosion of human expertise. Alkaabi and Almaamari [2] found that Arabic instructors are not confident in using AI tools and need significant professional development to implement them effectively. Instructors struggled to evaluate AI-generated Arabic content across different dialects and to maintain academic integrity in student assignments. Yet as Abidin and Sain [23] argue, the need for teacher expertise is greater than ever; the TPACK framework was identified as an essential model for training teachers to combine advanced technology with strong content and pedagogical knowledge, and to critically validate AI-generated content.

The solution to these risks lies in adopting a Human-in-the-Loop (HITL) approach that centers human judgment in AI-assisted learning [29]. Alkaabi and Almaamari [2] recommend developing instructor training specifically on using GenAI tools for Arabic dialect variations. Al-Jarf [28] emphasizes training students in phonological verification, morphological analysis, lexical validation, pragmatic awareness, and bibliographic literacy, skills that enable critical evaluation of AI outputs. Moustafa et al. [6] argue for integrating AI with traditional teaching through blended learning approaches and developing ethical AI policy guidelines for language education. Samiya [26] advocates for interdisciplinary collaboration among AI, linguistics, education, and cognitive science to address the unique complexities of Arabic language learning.

The findings also reveal a persistent gap between performance on general knowledge benchmarks and true linguistic competence. Zbib et al. [12] found that high performance on ArabicMMLU does not reliably translate to strong results on linguistic benchmarks. Models heavily tuned on synthetic instruction data achieve high scores on knowledge benchmarks but underperform on linguistic ones, suggesting overreliance

on memorized or templated patterns. This gap highlights the importance of diagnostic frameworks that specifically evaluate fundamental linguistic skills.

Several limitations should be acknowledged. First, the literature search was limited to English-language publications and focused primarily on MSA, which may have excluded relevant Arabic-language studies and overlooked dialectal and Classical Arabic challenges. Second, many reviewed studies rely on self-reported data or small sample sizes, and the review does not conduct meta-analysis, limiting generalizability. Third, the rapid evolution of AI models means findings from even two-year-old studies may not fully reflect current capabilities.

Despite these limitations, this review contributes by explicitly connecting computational challenges to pedagogical risks in Arabic grammar learning. While previous studies have examined either the technical limitations of AI in Arabic NLP or the pedagogical applications in language education, this review demonstrates that the structural features of Arabic that make NLP difficult are precisely the features that pose risks for learners who depend on AI without critical oversight. Effective AI integration thus requires not only technical improvements but also pedagogical frameworks that maintain human judgment at the center of learning.

In conclusion, the integration of generative AI into Arabic grammar learning presents a complex landscape of opportunities and risks. While AI tools offer new possibilities for language practice and explanation, their computational limitations, rooted in the structural complexity of Arabic, create serious pedagogical risks including misconceptions, overreliance, and the erosion of critical thinking and teacher expertise. Addressing these risks requires a balanced approach that combines technological innovation with sound pedagogy, maintains the vital role of human supervision, and develops critical digital literacy among both teachers and learners.

4. CONCLUSIONS

This review examined why generative AI struggles to understand Arabic grammar and how these computational limitations translate into pedagogical risks. The findings reveal that AI's difficulties are rooted in five interconnected linguistic challenges: the non-linear root-and-pattern morphology of Arabic (*Shorof*), the omission of diacritics that obscures grammatical case, extensive dialectal diversity, severe data scarcity, and the syntactic complexity of the *i'rab* case system. Even the most advanced models perform substantially worse on morphological and syntactic tasks than on surface-level tasks, and learners who depend on AI without critical oversight risk internalizing incorrect patterns, receiving misleading explanations, and developing superficial competence without genuine understanding.

The central contribution of this review is its demonstration that the structural features of Arabic that make NLP difficult are precisely the features that pose risks for learners who depend on AI without critical oversight. Three major risks emerge from this review: misconceptions, overreliance, and the erosion of critical thinking and teacher expertise. The review proposes strategies for responsible AI integration, including a Human-in-the-Loop approach, targeted teacher training, and the development of critical AI literacy among learners.

Future research should pursue both quantitative and qualitative investigations. Experimental studies are needed to measure the long-term impact of AI-assisted instruction on grammatical competence and to validate instruments for measuring critical AI literacy in Arabic learning contexts. Classroom-based ethnographic studies should explore how students and teachers experience AI-assisted learning and identify effective pedagogical interventions. Ultimately, while generative AI offers unprecedented opportunities, its integration must be guided by sound pedagogy, human supervision, and critical digital literacy among both teachers and learners.

REFERENCES

[1] K. Rahmouni, "Exploring the Use of ChatGPT in Teaching Arabic Case Endings: Effectiveness, Challenges and Recommendations," *Journal of Educational Technology and Innovation*, vol. 6, no. 4, pp. 1–20, 2024, doi: 10.61414/jeti.v6i4.198.

[2] M. H. Alkaabi and A. S. Almaamari, "Generative AI Implementation and Assessment in Arabic Language Teaching," *International Journal of Online Pedagogy and Course Design (IJOPCD)*, vol. 15, no. 1, pp. 1–14, 2025, doi: 10.4018/IJOPCD.368037.

[3] A. M. Alayba, "Arabic Natural Language Processing (NLP): A Comprehensive Review of Challenges, Techniques, and Emerging Trends," *Computers (MDPI)*, vol. 14, no. 497, pp. 1–32, 2025, doi: 10.3390/computers14110497.
[4] I. Boulesnam and R. Boucetti, "Arabic Language Characteristics that Make its Automatic Processing Challenging," *The International Arab Journal of Information Technolog*, vol. 22, no. 4, pp. 814–831, 2025, doi: 10.34028/iajit/22/4/14.
[5] M. M. Khalatia and T. A. H. Al-Romany, "Artificial Intelligence Development and Challenges (Arabic Language as a Model)," *International Journal of Innovation, Creativity and Change*, vol. 13, no. 5, pp. 916–926, 2020.
[6] A. H. A. Moustafa, M. F. Al-Hamad, M. A. Qureshi, and J. Qadir, "Generative AI for Learning and Teaching Arabic: Opportunities, Challenges, and Ethical Considerations," *IEEE Access*, vol. 14, pp. 7379–7395, 2026, doi: 10.1109/ACCESS.2026.3651864.
[7] Y. Saoudi and M. M. Gammoudi, "Trends and Challenges of Arabic Chatbots: Literature Review," *Jordanian Journal of Computers and Information Technology (JJCIT)*, vol. 09, no. 03, pp. 261–288, 2023.
[8] H. Mubarak, M. Hawasly, and A. Mohamed, "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 (EACL)*, 2026, pp. 6310–6328.
[9] A. M. W. Othman and L. H. B. Asbulah, "Problems of Artificial Intelligence Applications in Digitizing the Arabic Language in the Areas of Grammar and Morphology and Methods to Resolve It," *IJAZ ARABI: Journal of Arabic Learning*, vol. 8, no. 2, pp. 721–740, 2025, doi: 10.18860/ijazarabi.v8i2.31596.
[10] S. R. Karima, T. Edidarmo, and Raswan, "A Comparison of the Accuracy of Generative Artificial Intelligence Models in Taṣrīf and the Explanation of Wazan Meanings: A Study on Their Application in Arabic Morphology (Ṣarf)," *JALSAT Journal of Arabic Language Studies and Teaching*, vol. 5, no. 2, pp. 234–250, 2025, doi: 10.15642/jalsat.2025.5.2.234-250.
[11] M. I. Tamam, M. M. K. Ilahi, Z. Cholilah, R. Taufiqurrochman, and U. Machmudah, "Utilizing Chatgpt for Analyzing Arabic Texts in the Study of Nahwu (Arabic Grammar)," *KITABA: Journal of Interdisciplinary Arabic Learning*, vol. 2, no. 3, pp. 193–208, 2024.
[12] M. Zbib *et al.*, "AraLingBench: A Human-Annotated Benchmark for Evaluating Arabic Linguistic Capabilities of Large Language Models," *arXiv*. [Online]. Available: https://arxiv.org/abs/2511.14295
[13] R. Baumeister and M. Leary, "Writing Narrative Literature Reviews," *Review of General Psychology*, vol. 1, no. 3, pp. 311–320, 1997, doi: 10.1037/1089-2680.1.3.311.
[14] R. Ferrari, "Writing Narrative Style Literature Reviews," *Medical Writing*, vol. 24, no. 4, pp. 230–235, 2015, doi: 10.1111/j.1471-1842.2009.00848.x.
[15] R. J. Torraco, "Writing Integrative Literature Reviews: Guidelines and Examples," *Human Resource Development Review*, vol. 4, no. 3, pp. 356–367, 2005, doi: 10.1177/1534484305278283.
[16] H. Snyder, "Literature Review as a Research Methodology: An Overview and Guidelines," *Journal of Business Research*, vol. 104, pp. 333–339, 2019, doi: 10.1016/j.jbusres.2019.07.039.
[17] M. N. Ahmad, "Narrative Literature Reviews in Scientific Research: Pros and Cons," *Jordan Journal of Agricultural Sciences*, vol. 21, no. 1, pp. 1–4, 2025, doi: 10.35516/jjas.v21i1.4143.
[18] M. J. Grant and A. Booth, "A Typology of Reviews: An Analysis of 14 Review Types and Associated Methodologies," *Health Information and Libraries Journal*, vol. 26, no. 2, pp. 91–108, 2009, doi: 10.1111/j.1471-1842.2009.00848.x.
[19] M. Adel, B. Alhafni, and N. Habash, "Arabic Morphosyntactic Tagging and Dependency Parsing with Large Language Models," 2026, *arXiv:* 2603.16718. [Online]. Available: https://arxiv.org/abs/2603.16718
[20] S. Y. Kwon, G. Bhatia, B. Nagoudi, and M. Abdul-Mageed, "ChatGPT for Arabic Grammatical Error Correction," 2023, *arXiv*. [Online]. Available: https://arxiv.org/abs/2308.04492
[21] S. Y. Kwon, G. Bhatia, B. Nagoudi, and M. Abdul-Mageed, "Beyond English: Evaluating LLMs for Arabic Grammatical Error Correction," in *Proceedings of the First Arabic Natural Language Processing Conference (ArabicNLP 2023)*, 2023, pp. 101–119.
[22] M. Abdelrehim, M. Torki, and N. El-Makky, "Hybrid LLM and Rule-Based Synthetic Data Generation for Arabic Grammatical Error Correction," in *2025 International Conference on Machine Intelligence and Smart Innovation (ICMISI)*, 2025, pp. 280–285. doi: 10.1109/ICMISI65108.2025.11115884.

[23] H. Abidin and Z. H. Sain, "The Transformation of Arabic Language Learning in the Digital Era: A Critical Review of Technological Innovation, Linguistic Complexity, and the Pedagogical Imperatives of TPACK (2020-2025)," *JETech: Journal of Education and Technology*, vol. 1, no. 3, pp. 97–106, 2025, doi: 10.65678/jetech.v1i3.231.
[24] R. B. Rizki, M. F. Rizal, C. Rahmawati, and M. Ali, "Qalam AI: a Study on the Potential of Automatic Harakat Detection for Arabic Sentence Learning," *Journal of Arabic Studies*, vol. 7, no. 2, pp. 285–316, 2025, doi: 10.21580/alsina.7.2.27500.
[25] R. Al-Jarf, "Pronunciation Errors in Arabic YouTube Videos Narrated by AI," *Frontiers in Computer Science and Artificial Intelligence*, vol. 2, no. 1, pp. 1–12, 2025, doi: 10.32996/jcsts.2025.2.2.1.
[26] R. Samiya, "Artificial Intelligence in Arabic Language Education: Current Applications, Challenges, and Future Perspectives," *Studies in Education Sciences*, vol. 6, no. 4, pp. 1–21, 2025, doi: 10.54019/sesv6n4-011.
[27] A. M. AlSbou, F. S. Abdullah, and A. M. Deris, "Systematic Review: The Application of ChatGPT on Arabic Language Text Processing," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 15, no. 5, pp. 4837–4847, 2025, doi: 10.11591/ijece.v15i5.pp4837-4847.
[28] R. Al-Jarf, "Specific Linguistic Questions that Artificial Intelligence (AI) Cannot Answer Accurately: Implications for Digital Didactics," *Frontiers in Computer Science and Artificial Intelligence*, vol. 4, no. 4, pp. 43–61, 2025, doi: 10.32996/fcsai.2025.4.4.4.
[29] Z. Hayad, R. W. L. Pertiwi, T. Ayumagita, and A. Safirah, "The Role, Effectiveness, and Challenges of Ai in Arabic Language Learning: A Systematic Literature Review," in *proceeding PINBA XV 2025 - IMLA Indonesia*, 2025, pp. 280–295.
[30] S. R. Borham, S. Ramli, and M. T. A. Ghani, "AI Concepts Integration in Developing E-Muhadathat Kits For Non-Arabic Speakers," *IJAZ ARABI: Journal of Arabic Learning*, vol. 7, no. 3, pp. 1215–1225, 2024, doi: 10.18860/ijazarabi.V7i3.26568.