Challenges and Opportunities in AI-Based Arabic Language Tutoring for Adult Learners

«التحديات والفرص في تعليم اللغة العربية للمتعلمين البالغين باستخدام أنظمة التدريس القائمة على الذكاء الاصطناعي»

10.35781/1637-000-203-007

Dr. Ayman Aljuraisehy*

*Associate Professor -Institute of Teaching Arabic for Non-native Speakers Islamic University, Madinah, Saudi Arabia.

Abstract:

Artificial intelligence–based tutoring systems have demonstrated significant potential to enhance personalized learning through adaptive dialogue, automated feedback, and learner-centered instructional pathways. In Arabic-medium education, however, the design and deployment of effective AI tutors remain underexplored; particularly for adult non-native learners whose linguistic and pedagogical needs differ substantially from those of native-speaking students. This paper investigates the viability of agentic AI tutors for Arabic language instruction by examining the intersection of Arabic-specific linguistic challenges, adult learning principles, and contemporary AI system design.

From a natural language processing perspective, Arabic presents persistent obstacles, including rich non-concatenative morphology, orthographic ambiguity, syntactic variability, and widespread diglossia between Modern Standard Arabic and regional dialects. These characteristics complicate language understanding, error diagnosis, and feedback generation, especially in interactive tutoring contexts involving non-native adult learners. At the system level, the adoption of agentic AI introduces additional challenges related to autonomous instructional planning, learner modeling, and decision-making under uncertainty, where pedagogical misalignment can negatively affect learning outcomes and learner trust.

The paper further emphasizes the role of continuous evaluation and human-in-the-loop feedback in constraining and refining agentic behavior, enabling iterative improvement while preserving instructional reliability and cultural sensitivity. Finally, it explores how national AI infrastructure democratization, illustrated through Saudi Arabia's Arabic-first AI initiatives can reduce development costs and accelerate the deployment of scalable, culturally grounded AI tutors. By grounding the analysis in both linguistic and pedagogical dimensions, this study offers a realistic assessment of

ISSN : 2410-1818

current limitations and outlines design considerations for deploying effective AI tutoring systems in Arabic-medium adult education.

The analysis also considers a practical question that is sometimes overlooked in discussions of educational AI i.e. what makes an AI tutor educationally useful rather than merely conversationally impressive? For adult learners, the answer depends not only on whether a system can produce fluent Arabic, but on whether it can identify why a learner is struggling, choose an appropriate explanation, adjust the level of support, and preserve a clear relationship between each interaction and the learner's longer-term goals. This distinction is particularly important in Arabic, where a grammatically acceptable response may still be inappropriate for a learner's level, register, communicative purpose, or cultural setting. The paper therefore treats tutoring quality as a combined linguistic, pedagogical, and system-design problem rather than as a language-generation problem alone.

Keywords:

Agentic AI tutoring systems, Arabic natural language processing, adult non-native language learning, human-in-the-loop evaluation, Arabic diglossia and dialectal variation.

ISSN : 2410-1818

«التحديات والفرص في تعليم اللغة العربية للمتعلمين البالغين باستخدام أنظمة التدريس القائمة على الذكاء الاصطناعي»

د. أيمن بن سويلم الجريشي

الملخص

أظهرت أنظمة التدريس القائمة على الذكاء الاصطناعي إمكانات واعدة في تعزيز التعلم الشخصي، من خلال الحوار التكيفي، والتغذية الراجعة الآلية، والمسارات التعليمية المتمحورة حول المتعلم. ومع ذلك، لا يزال تصميم أنظمة تدريس ذكية فعّالة وتوظيفها في التعليم باللغة العربية مجالًا لم يحظَ بالقدر الكافي من البحث، ولا سيما فيما يتعلق بالمتعلمين البالغين الناطقين بغير العربية، الذين تختلف احتياجاتهم اللغوية والتربوية اختلافًا جوهريًا عن احتياجات الطلاب الناطقين بالعربية لغةً أمًّا. وتبحث هذه الدراسة في جدوى توظيف أنظمة التدريس الذكية القائمة على الذكاء الاصطناعي الوكيلي (Agentic AI) في تعليم اللغة العربية، من خلال تحليل التقاطع بين التحديات اللغوية الخاصة بالعربية، ومبادئ تعلم الكبار، والتصميم المعاصر لأنظمة الذكاء الاصطناعي.

ومن منظور معالجة اللغات الطبيعية، تواجه اللغة العربية تحديات مستمرة، من أبرزها غنى نظامها الصرفي القائم على الاشتقاق غير التسلسلي، والغموض الناشئ عن نظام الكتابة، والتنوع في التراكيب النحوية، والازدواجية اللغوية الواسعة بين العربية الفصحى المعاصرة واللهجات الإقليمية. وتزيد هذه الخصائص من تعقيد عمليات الفهم الآلي للغة، وتشخيص الأخطاء، وتوليد التغذية الراجعة، ولا سيما في سياقات التدريس التفاعلي الموجهة إلى المتعلمين البالغين الناطقين بغير العربية. وعلى مستوى تصميم الأنظمة، يطرح تبنّي الذكاء الاصطناعي الوكيلي تحديات إضافية تتصل بالتخطيط التعليمي المستقل، ونمذجة المتعلم، واتخاذ القرارات في ظل عدم اليقين؛ إذ قد يؤثر عدم التوافق التربوي سلبًا في نواتج التعلم وثقة المتعلمين بالنظام.

وتؤكد الدراسة كذلك أهمية التقويم المستمر والتغذية الراجعة القائمة على إشراف الإنسان وتدخله (Human-in-the-Loop) في ضبط سلوك الأنظمة الوكيلية وتحسينه، بما يتيح تطويراً تكرارياً متواصلًا، مع المحافظة على موثوقية العملية التعليمية ومراعاة الحساسية الثقافية. وتستكشف الدراسة أيضًا كيف يمكن لإتاحة البنية التحتية الوطنية للذكاء الاصطناعي على نطاق أوسع، كما يتجلى في المبادرات السعودية الرامية إلى تطوير تقنيات ذكاء اصطناعي تعطي الأولوية للغة العربية، أن تسهم في خفض تكاليف التطوير وتسريع نشر أنظمة تدريس ذكية قابلة للتوسع ومرتكزة على السياق الثقافي المحلي. ومن خلال تأسيس التحليل على البعدين اللغوي والتربوي معًا، تقدم الدراسة تقييمًا واقعيًا للقيود والتحديات الراهنة، وتحدد مجموعة من الاعتبارات التصميمية اللازمة لتطوير أنظمة تدريس ذكية فعّالة وتوظيفها في تعليم اللغة العربية للمتعلمين البالغين.

ويتناول التحليل أيضًا سؤالًا عمليًا قد لا يحظى بالاهتمام الكافي في النقاشات المتعلقة بالذكاء الاصطناعي التعليمي، وهو: ما الذي يجعل نظام التدريس الذكي مفيداً من الناحية التعليمية، لا مجرد نظام مبهر في قدراته الحوارية؟ وبالنسبة

ISSN : 2410-1818

إلى المتعلمين البالغين، لا تتوقف الإجابة عند قدرة النظام على إنتاج لغة عربية سليمة وطليقة، بل تتجاوز ذلك إلى قدرته على تشخيص أسباب تعثر المتعلم، واختيار التفسير التعليمي الملائم، وتكييف مستوى الدعم المقدم له، والمحافظة على ارتباط واضح بين كل تفاعل تعليمي وأهداف المتعلم بعيدة المدى.

وتكتسب هذه التفرقة أهمية خاصة في تعليم اللغة العربية؛ إذ قد تكون الاستجابة صحيحة من الناحية النحوية، لكنها غير ملائمة لمستوى المتعلم، أو للسجل اللغوي المناسب، أو للغرض التواصلي، أو للسياق الثقافي الذي تجري فيه عملية التعلم. وبناءً على ذلك، تتعامل الدراسة مع جودة التدريس الذكي بوصفها قضية متكاملة تتداخل فيها الأبعاد اللغوية والتربوية والتقنية المرتبطة بتصميم الأنظمة، وليس بوصفها مجرد مسألة تتعلق بقدرة النظام على توليد اللغة.

الكلمات المفتاحية: أنظمة التدريس الذكية القائمة على الذكاء الاصطناعي الوكيلي، المعالجة الآلية للغة العربية، تعليم العربية للكبار الناطقين بغيرها، التقويم القائم على إشراف الإنسان وتدخله، الازدواجية اللغوية والتنوع اللهجي في العربية.

ISSN : 2410-1818

Introduction

Artificial intelligence–based tutoring systems have emerged as a promising application of advances in natural language processing (NLP), machine learning, and educational data analytics. These systems aim to deliver personalized instruction through interactive dialogue, automated feedback, and adaptive learning pathways. In adult education contexts in particular, AI tutors offer the potential to support self-directed learning, flexible pacing, and just-in-time instructional assistance features that align closely with established principles of adult learning theory.

The educational value of these systems becomes clearer when the realities of adult learning are considered. Adult learners often study around work, family responsibilities, travel, or other commitments, which means that access to a teacher may be irregular even when motivation is high. A conversational tutor that is available at the point of need can fill some of this gap by allowing a learner to ask a short grammar question, rehearse a workplace dialogue, review unfamiliar vocabulary, or practice a difficult sentence without waiting for the next scheduled lesson. In this sense, the strongest case for AI tutoring is not necessarily the replacement of formal instruction. It is the extension of meaningful practice beyond the classroom and the creation of more opportunities for guided interaction between lessons.

This opportunity is especially relevant to Arabic because progress often depends on repeated exposure to forms that behave differently from those familiar to speakers of many other languages. A learner may understand a rule when it is explained in class but still fail to recognize the same pattern when it appears in connected speech, in a text without short vowels, or in a sentence containing an unfamiliar root. A well-designed tutor can revisit such material in different forms and at different levels of difficulty. It can also connect an error in the present conversation to a pattern observed earlier, which gives the learner a sense that feedback is cumulative rather than isolated. That continuity is one of the main advantages promised by learner modeling, but it is also one of the areas in which design quality matters most.

Despite this progress, the development and deployment of AI tutors for Arabic-medium education remain underexplored, especially for adult learners who are non-native speakers of Arabic. From an NLP perspective, Arabic poses a set of well-documented challenges that complicate the design of robust conversational and instructional systems. These challenges include rich and non-concatenative morphology, the absence of short vowels in most written texts, high lexical ambiguity, and syntactic variability. For non-native learners, these linguistic properties significantly increase cognitive load, making accurate comprehension, pronunciation guidance, and grammatical feedback critical components of any effective tutoring system.

Arabic morphology illustrates why fluent text generation alone is not enough. Words may encode information about person, number, gender, tense, aspect,

ISSN : 2410-1818

definiteness, case, and possession through combinations of roots, patterns, prefixes, and suffixes. For an experienced speaker, many of these signals are processed automatically. For a non-native adult learner, however, a small morphological change may alter the grammatical role or intended meaning of an entire expression. An AI tutor must therefore be able to separate an error that reflects a misunderstanding of a morphological rule from one that is simply a spelling mistake, a typing shortcut, or a temporary performance error. If every deviation is treated in the same way, feedback becomes noisy and the learner may be corrected too often on minor details while more important gaps remain unaddressed.

The limited use of short-vowel marks in ordinary Arabic writing creates another instructional problem. A learner may encounter the same sequence of consonants in more than one possible reading, and the correct interpretation may depend on syntax, semantics, or wider context. Research on Arabic diacritics and L2 reading has shown their importance for reading accuracy and comprehension among learners, yet adult learners also need to become comfortable with the largely undiacritized texts they will meet outside teaching materials. The tutor must therefore know when diacritics are helpful as scaffolding and when they should be reduced. An early-stage learner may benefit from full or selective vowel marking, while a more advanced learner may need the same sentence presented without vowels and then be asked to infer the correct reading from context. This is a pedagogical choice, not simply a text-formatting choice.

The same principle applies to grammar explanation. Adult learners frequently ask explicit questions such as why a word changed ending, why a particular preposition was used, or why a verb appears in a certain pattern. A generic chatbot can provide an explanation, but an instructional system should ideally tailor that explanation to the learner's level and prior knowledge. A beginner may need a concrete example and a simple rule; an advanced learner may benefit from terminology, exceptions, and comparison with related constructions. The system should also avoid introducing several new rules at once when only one is necessary to resolve the learner's confusion. In practice, good tutoring often means deciding what not to explain yet.

The issue is further compounded by Arabic diglossia, wherein Modern Standard Arabic (MSA) is used in formal education and written communication, while regional dialects dominate spoken interaction. AI tutors intended for adult learners such as migrants, professionals, or religious and academic students must therefore negotiate a complex instructional space teaching standardized Arabic while remaining sensitive to learners' exposure to colloquial forms. Current NLP models often struggle with this duality, as training data is unevenly distributed across registers and dialects, leading to reduced accuracy in language understanding and generation tasks.

Diglossia further complicates the idea of a single correct answer. A learner who has studied MSA may hear a Saudi colleague, Egyptian television presenter, or

ISSN : 2410-1818

Levantine shopkeeper use a form that differs from the textbook without being wrong in its own communicative setting. If the tutor simply labels every non-MSA form as an error, it may leave the learner with an unrealistic picture of Arabic use. If it responds entirely in dialect, however, it may weaken the consistency of formal instruction. A more useful approach is to make register visible. The tutor can identify a form as MSA, dialectal, informal, formal, or context-dependent and then explain which alternative is most suitable for the learner's immediate goal.

This is particularly important for adult learners whose purposes for learning Arabic differ widely. A learner preparing for university study may need strong command of MSA reading and formal writing. A professional working in Riyadh may need enough spoken Saudi Arabic to understand colleagues and participate naturally in routine interaction while still using formal Arabic in written communication. A learner studying Islamic texts may have a different relationship to vocabulary, grammar, and pronunciation altogether. These learners cannot be served effectively by the same sequence of examples, correction priorities, or conversational scenarios. Personalization therefore begins with a clear representation of purpose, not only a numerical proficiency level.

Adult learners of Arabic also present pedagogical requirements distinct from those of native-speaking school-age students. They are typically goal-oriented, context-driven, and motivated by immediate applicability, such as workplace communication, academic study, or religious literacy. Consequently, an AI tutor for this population must go beyond generic language instruction and support functions such as explicit grammatical explanation, contrastive error correction, vocabulary contextualization, and adaptive scaffolding based on proficiency level and first-language interference patterns. Achieving these capabilities places substantial demands on NLP pipelines, particularly in learner modeling, error diagnosis, and feedback generation.

Goal orientation also changes the way mistakes should be handled. In a conventional exercise, the objective may be to produce a fully correct sentence. In real communication, an adult learner may instead need to complete a task, ask for clarification in a meeting, understand an instruction, introduce a topic politely, or read a passage accurately. The tutor should distinguish between errors that block the task and errors that can be addressed later. Correcting every article, agreement feature, or pronunciation detail during a role-play can interrupt fluency and increase frustration. On the other hand, ignoring a recurring error for too long can allow it to become stable. Effective tutoring requires a deliberate balance between communicative flow and corrective intervention.

First-language influence adds another layer to this balance. Learners often transfer familiar sound patterns, word order, or grammatical categories from their first language into Arabic. The resulting error may look identical on the surface across two learners while having different underlying causes. An AI tutor that retains a structured learner profile can use repeated evidence to infer which explanations are

ISSN : 2410-1818

most likely to help. It may, for example, notice that a learner repeatedly confuses gender agreement, omits the definite article in specific contexts, or struggles with sounds that do not occur in the learner's first language. The purpose of such modeling should not be to label the learner permanently, but to make instruction more responsive to patterns that would otherwise be easy to miss in short interactions.

At the same time, personalization needs limits. Learner models are probabilistic representations and can be wrong. A system may interpret hesitation as lack of knowledge when the learner is simply tired, or may attribute an error to first-language interference when the actual problem is vocabulary. For this reason, adaptive systems should leave room for the learner to correct the tutor's assumptions. Simple prompts such as asking whether an explanation was clear, whether the exercise felt too easy, or whether the learner wants a formal or conversational example can provide valuable control. The learner should experience personalization as support rather than as an invisible mechanism making unexplained decisions.

This paper investigates the viability of AI tutors for Arabic-medium education with a specific focus on adult non-native learners. It evaluates current NLP capabilities and limitations in handling Arabic language instruction, examines how these constraints interact with adult learning requirements, and identifies design considerations necessary for effective deployment. By grounding the analysis in both linguistic and pedagogical dimensions, the paper aims to provide a realistic assessment of the potential and current limitations of AI tutoring systems in this specialized educational context.

Taken together, these issues suggest that the central design question is not whether current AI can converse in Arabic, because modern systems can already do so with increasing fluency. The harder question is whether an AI tutor can sustain academically coherent behavior across time and then diagnose an error accurately, connect it to a learner model, choose a suitable intervention, evaluate whether that intervention worked, and decide what should happen next. That sequence turns a conversational model into an instructional system. It also explains why Arabic tutoring should be evaluated at the level of learning behavior and outcomes, not only through general measures of language-model fluency.

2. The Challenges of Agentic AI Composition Specific to Arabic Language Complications

One of the primary linguistic challenges in developing agentic AI tutors for Arabic lies in the language's extensive regional and sociolinguistic variation. Arabic exists as a continuum of dialects that differ significantly in phonology, vocabulary, syntax, and idiomatic usage across regions. While Modern Standard Arabic (MSA) serves as the formal medium of instruction, adult learners particularly non-native speakers are frequently exposed to dialectal forms through media, workplace interaction, and daily communication. This divergence creates ambiguity for AI

ISSN : 2410-1818

tutors, which must consistently generate pedagogically appropriate MSA while also recognizing and interpreting dialectal inputs that may not conform to standardized grammatical or lexical norms. Limited high-quality, annotated datasets covering dialectal Arabic further exacerbate these issues, reducing the reliability of language understanding and response generation in interactive tutoring contexts.

A useful way to understand this challenge is to separate recognition from instruction. The tutor first has to understand what the learner has said or written, including forms that may be incomplete, misspelled, dialectal, or influenced by another language. It then has to decide what the learner intended. Only after those steps can it determine whether correction is required and how that correction should be delivered. Errors in the first two stages can easily produce inappropriate teaching at the third. A dialect expression misread as a grammatical mistake, for example, may trigger a long correction that confuses a learner who actually used the expression correctly in an informal context.

Orthographic variation can create similar problems in written interaction. Arabic users in informal digital settings may omit hamza distinctions, vary the spelling of dialectal words, mix Arabic and Latin characters, or write according to pronunciation rather than formal orthographic conventions. These practices are not all equally relevant to a language-learning objective. A tutor must decide whether it is operating as a writing instructor, a conversational partner, or a comprehension assistant in a particular moment. If the goal is formal writing, orthographic precision may be central. If the goal is understanding a WhatsApp-style message from a colleague, rigid correction may be less useful than explaining the intended meaning and then showing the formal equivalent.

Beyond language-specific constraints, the use of agentic AI in tutoring introduces broader system-level challenges. Agentic AI tutors are expected to autonomously plan instructional actions, adapt learning trajectories, monitor learner progress, and intervene appropriately without constant human oversight. Achieving this level of autonomy requires robust learner modeling, long-horizon reasoning, and reliable decision-making under uncertainty. In educational settings, incorrect instructional choices such as inappropriate difficulty progression, misleading explanations, or misinterpretation of learner intent can negatively affect learning outcomes and learner trust.

Long-horizon instructional planning is particularly demanding because learning is not a sequence of independent questions. A tutor may have to decide whether a learner should review a weak concept before moving forward, whether a new activity should recycle previous vocabulary, or whether a persistent error deserves explicit remedial work. These decisions require memory across sessions and a representation of mastery that can change with new evidence. A single correct answer from the learner should not be treated as proof of mastery, just as a single mistake should not automatically trigger a full lesson. The system needs to accumulate evidence over time and use that evidence conservatively.

ISSN : 2410-1818

This introduces an important difference between educational agents and general task agents. In many software tasks, success can be defined by whether an action was completed, a reservation was made, a document was found, or a calculation was returned. In tutoring, the immediate action may be completed even when learning has not occurred. The tutor can explain a rule perfectly and still fail if the learner cannot apply it later. Educational agents therefore need delayed indicators of success, such as transfer to a new example, reduced recurrence of the same error, improved reading accuracy, or successful use of a form in spontaneous speech. Designing around these slower signals makes the system more complex, but it also makes the concept of adaptation educationally meaningful.

Another difficulty concerns the confidence of the model itself. Language models can produce explanations that sound clear even when an analysis is uncertain. In Arabic instruction, this is risky because learners may not yet have enough knowledge to detect an incorrect grammatical explanation, an inappropriate dialect label, or a misleading example. A tutoring architecture should therefore distinguish between high-confidence instructional routines and cases that require caution. Where uncertainty is significant, the system can provide a limited explanation, present alternatives, consult an approved knowledge source, or escalate the case for human review rather than presenting a speculative answer as settled fact.

Additionally, agentic AI tutors must balance initiative with control. Excessive autonomy may lead to over-guidance, reduced learner agency, or pedagogically unsound interventions, while insufficient autonomy limits the system's effectiveness as a personalized tutor. Ensuring alignment between agentic behavior, pedagogical objectives, and ethical constraints remains a non-trivial challenge, particularly in language education where nuanced feedback, cultural sensitivity, and learner affect play a critical role. These challenges underscore the need for carefully constrained agentic architectures when deploying AI tutors in Arabic language learning environments.

The balance between initiative and learner agency can be made concrete through interaction design. The tutor might recommend a short review after observing repeated errors, but the learner should be able to postpone it. It might suggest switching from translation exercises to dialogue practice, but it should explain why that change is being proposed. It might detect that an exercise is too difficult and reduce complexity, yet still allow the learner to attempt the original task. These small design choices preserve a sense of partnership. They also reflect an important principle in adult education and that is that the learner is not simply the object of adaptation but an active participant who brings goals, preferences, experience, and judgment to the learning process.

Cultural sensitivity is equally difficult to reduce to a fixed rule set. Arabic is used across diverse societies and institutional settings, and appropriate examples can vary according to region, age, gender, profession, and degree of formality. A tutor designed for use in Saudi Arabia may reasonably prioritize contexts and expressions

ISSN : 2410-1818

that learners are likely to encounter locally, but it should avoid implying that one local usage represents the whole Arabic-speaking world. Contextual labels and transparent explanations can help. The system can say that an expression is common in Saudi speech, provide the MSA equivalent, and note that another region may use a different form. This approach supports authenticity without turning regional variation into confusion.

Pronunciation presents a further challenge because useful feedback requires more than converting speech to text. A speech recognizer may correctly infer the intended word even when the learner's pronunciation is weak, which means recognition accuracy alone does not measure teaching quality. A pronunciation tutor needs to identify which sound or prosodic feature caused difficulty, decide whether the deviation affects intelligibility, and offer a correction the learner can act on. For Arabic, this may involve consonants unfamiliar to many learners, distinctions in vowel length, stress, emphatic sounds, or connected-speech effects. The feedback must also remain proportionate; a learner who is communicating successfully should not be overwhelmed by microscopic correction on every utterance.

For these reasons, the most credible architecture for an Arabic agentic tutor is likely to be modular rather than unrestricted. A general language model can manage dialogue and explanation, while specialized components or constrained tools support functions such as speech analysis, morphological parsing, curriculum retrieval, learner-profile updates, and assessment. Policies can define which actions the agent is allowed to take automatically and which require confirmation or review. Such an arrangement does not eliminate error, but it reduces the likelihood that every instructional decision depends on a single unconstrained generation step.

3. Continuous Evaluation and Self-Improvement through Human-in-the-Loop Feedback

The effectiveness of agentic AI tutors particularly in complex language-learning contexts such as Arabic for non-native adult learners depends critically on continuous evaluation and iterative system refinement. Static model deployment is insufficient in educational settings where learner needs, linguistic inputs, and pedagogical contexts evolve over time. Continuous evaluation mechanisms are therefore essential to ensure instructional accuracy, pedagogical alignment, and sustained learning effectiveness.

Continuous evaluation should begin before a system is released to learners. A useful development process can combine linguistic testing, pedagogical review, and scenario-based trials with representative learner profiles. Arabic specialists can examine whether generated explanations are grammatically sound and whether examples are natural. Teachers can evaluate whether the timing and amount of feedback are appropriate. Adult learners can reveal usability problems that experts may overlook, such as explanations that are technically correct but too dense, exercises that feel repetitive, or dialect labels that are not clear enough. These

ISSN : 2410-1818

perspectives answer different questions, and none can be replaced completely by automatic benchmarking.

Once deployed, evaluation should operate at several levels. At the interaction level, the system can track whether the learner understands a correction, asks for repeated explanation, abandons an exercise, or accepts a suggested review. At the learning level, it can observe whether a previously corrected pattern reappears and whether performance transfers to new contexts. At the program level, educators can review broader trends across groups of learners to identify curriculum areas where the tutor is consistently successful or consistently weak. This layered view prevents the system from equating short, smooth conversations with genuine learning.

Human-in-the-loop feedback (HLF) plays a central role in enabling adaptive self-improvement for AI tutors. By incorporating feedback from educators, language experts, and learners themselves, AI systems can identify systematic errors in language interpretation, explanation quality, and instructional decision-making. In the context of Arabic language instruction, HLF is particularly valuable for correcting misinterpretations arising from dialectal variation, ambiguous learner inputs, and culturally specific expressions that remain challenging for automated NLP systems.

Human feedback is most useful when it is structured rather than collected as an undifferentiated stream of approval and disapproval. An Arabic instructor reviewing a tutor response may need to indicate whether the problem concerns grammar, vocabulary, register, dialect identification, cultural appropriateness, explanation level, or pedagogical timing. These categories make later analysis more actionable. If many reviewers identify the same type of error, developers can determine whether the source lies in prompts, retrieval materials, model behavior, speech recognition, or the curriculum logic surrounding the model.

Learner feedback should also be treated as more than a satisfaction score. A learner may prefer an answer because it is short, even if it does not resolve the underlying misconception. Another learner may dislike a demanding exercise that is nevertheless educationally valuable. The system therefore needs to interpret preferences alongside performance evidence. Feedback such as ‘too difficult,’ ‘too much explanation,’ or ‘I wanted Saudi Arabic, not formal Arabic’ can guide personalization, while learning outcomes help determine whether those preferences are supporting progress. In this way, human-in-the-loop design is not simply about making the system agreeable; it is about making adaptation accountable.

From a system design perspective, HLF enables supervised calibration of agentic behaviors, including lesson sequencing, feedback granularity, and intervention timing. Rather than allowing unrestricted autonomy, human feedback can be used to constrain and guide agentic decision-making, ensuring that instructional strategies remain pedagogically sound and aligned with adult learning principles. This approach mitigates risks associated with erroneous guidance while preserving the benefits of adaptive, personalized instruction.

ISSN : 2410-1818

The role of the educator is especially important when the tutor begins to make decisions across sessions. Teachers or academic supervisors should be able to see why the system has changed a learner's pathway, which evidence triggered a remedial activity, and what content the tutor has prioritized or delayed. This does not require exposing every internal model calculation. It does require a practical audit trail that translates system behavior into educational terms. For example, a supervisor should be able to see that a learner was assigned additional practice because gender-agreement errors persisted across several contexts, rather than because of an unexplained model score.

Such visibility also supports governance. Institutions adopting AI tutors need clear responsibility for curriculum boundaries, approved sources, data handling, escalation pathways, and review of high-impact errors. In an Arabic learning context, governance may include rules about when dialectal material can be introduced, how religious or culturally sensitive examples are handled, and how the system responds to questions that move beyond language learning into domains requiring specialist authority. These boundaries help the tutor remain useful without allowing conversational flexibility to blur the limits of its instructional role.

Moreover, continuous evaluation supported by HLF facilitates long-term model robustness and trustworthiness. Metrics such as learner progress, error recurrence rates, and feedback satisfaction can be monitored to assess system performance over time. Insights derived from these evaluations inform targeted retraining, prompt refinement, and policy adjustment, creating a feedback-driven improvement cycle. In Arabic AI tutoring systems, such cycles are essential for progressively enhancing linguistic coverage, instructional quality, and contextual sensitivity.

Evaluation metrics should therefore combine accuracy with educational usefulness. Linguistic measures may include correct identification of grammatical errors, quality of diacritization support, dialect recognition, and consistency of register. Pedagogical measures may include retention, reduction in repeated errors, successful completion of communicative tasks, and the learner's ability to apply a concept in an unfamiliar example. System measures may include latency, stability, cost per interaction, and the rate at which cases are escalated for human review. Looking at these dimensions together can reveal trade-offs that a single metric would hide.

A tutor should also be tested for uneven performance across learner groups. Differences in first language, accent, proficiency, age, and educational background may affect both speech recognition and the kinds of errors learners produce. A system that works well for highly educated English-speaking learners of MSA may not perform equally well for a beginner whose first language uses a different script or for a professional whose priority is spoken interaction. Continuous evaluation makes these gaps visible and provides evidence for targeted improvement rather than assuming that one overall accuracy figure represents every learner.

ISSN : 2410-1818

In addition to improving model performance, human-in-the-loop feedback frameworks create a bidirectional learning process in which human supervisors and educators also benefit from interaction with the AI tutor. As agentic systems propose alternative instructional paths, error explanations, or adaptive scaffolding strategies, they can surface suggestive teaching angles that may not be immediately apparent to human instructors. Exposure to these data-driven pedagogical variations enables educators to reflect on and refine their own instructional approaches, particularly in the design of individualized learning experiences for adult learners with diverse linguistic backgrounds and proficiency profiles.

This does not mean that data automatically reveals the correct pedagogical response. Quantitative patterns need interpretation by teachers who understand the learning context. A high error rate may indicate that an item is difficult, but it may also indicate that the prompt is ambiguous, that the speech recognizer is failing on a particular accent, or that learners have not been given enough preparation. Human review remains essential precisely because educational data is meaningful only when connected to the circumstances in which it was produced.

This reciprocal learning dynamic broadens the range of personalized interventions available in Arabic language education. By analyzing AI-generated recommendations alongside learner outcomes, human supervisors can identify effective micro-strategies for addressing specific learner challenges, such as persistent grammatical errors or vocabulary gaps influenced by first-language interference. Over time, this collaborative feedback loop supports the co-evolution of both human-led and AI-mediated instruction, enhancing the overall capacity to deliver nuanced, learner-centered education at scale.

Over time, a mature human-in-the-loop process can create an institutional knowledge base around Arabic learning difficulties. Common error patterns, successful explanations, preferred examples, and effective remediation strategies can be documented and reused rather than rediscovered in isolated teacher-learner interactions. The AI tutor then becomes not only a delivery channel but also a mechanism for organizing teaching knowledge. This is particularly valuable for institutions serving large and diverse adult populations, where consistency can be difficult to maintain across instructors, courses, and schedules.

The strongest model is therefore one of supervised co-development. Educators define the learning objectives and interpret evidence; language specialists protect linguistic quality; technical teams improve the system; and learners provide the real-world interactions through which weaknesses become visible. The AI contributes scale, memory, rapid adaptation, and the ability to generate many forms of practice. Each participant has a different role, and the reliability of the tutor depends on those roles remaining connected rather than assuming that model improvement alone will solve every educational problem.

ISSN : 2410-1818

4. Optimizing the Cost of AI Tutor Ownership through National AI Infrastructure Democratization

The long-term viability of AI tutors in Arabic-medium education is closely tied to the cost of development, deployment, and maintenance of AI infrastructure. High computational requirements, dependency on large-scale language models, and the need for continuous evaluation and retraining can make AI tutoring systems prohibitively expensive for individual educational institutions. This cost barrier is particularly relevant for advanced agentic systems that rely on real-time inference, adaptive planning, and multilingual natural language processing.

The cost question extends beyond the price of model inference. Institutions also have to consider data preparation, integration with learning platforms, speech services, security controls, monitoring, curriculum development, teacher review, and ongoing technical support. A pilot can appear inexpensive when it relies on a general-purpose interface, yet production deployment may require considerably more work to ensure that learner data is protected, instructional content is governed, and performance is monitored. Total cost of ownership should therefore be assessed across the full lifecycle of the tutor rather than through the price of a single model or API call.

At the same time, AI tutoring can create forms of value that are difficult to achieve through conventional scaling alone. A human instructor has limited time for individualized repetition, especially outside scheduled teaching hours. A digital tutor can provide additional practice at marginal cost, preserve a learner's history across sessions, and support large numbers of low-stakes questions that learners may hesitate to ask in class. The relevant economic comparison is therefore not simply 'AI versus teacher.' It is whether a combined model can use teacher time more effectively while extending access to guided practice.

Within the Kingdom of Saudi Arabia, this challenge is being strategically addressed through national-level initiatives aimed at democratizing access to AI infrastructure. Under the leadership of HRH The Crown Prince Mohammed bin Salman, the establishment of Humain represents a coordinated effort to provide scalable, sovereign AI capabilities across public and private sectors. By centralizing high-performance compute, foundational models, and AI development platforms, Humain significantly reduces the marginal cost of building and operating AI-driven applications, including educational AI tutors.

National infrastructure can influence this equation by turning capabilities that would otherwise be institution-specific investments into shared services. Compute capacity, Arabic foundation models, secure hosting, and common development tooling can lower duplication across universities and training organizations. An institution can then concentrate resources on the parts that differentiate educational quality and curriculum alignment, learner experience, evaluation design, specialized Arabic content, and teacher oversight. This division of effort is particularly important

ISSN : 2410-1818

for smaller institutions that may have strong academic expertise but limited ability to build and operate large-scale AI infrastructure independently.

Sovereign infrastructure also creates a strategic option for institutions that need tighter control over data location, access policies, and model deployment. Educational records can contain detailed information about learner performance, language weaknesses, interaction histories, and sometimes voice data. These datasets have value for personalization but also create governance responsibilities. A national ecosystem that supports secure deployment can make it easier to design AI tutoring within local regulatory and institutional requirements rather than forcing each organization to solve the same infrastructure and compliance questions on its own.

For educational institutions within Saudi Arabia, this shared infrastructure model enables access to advanced AI capabilities without the need for large upfront capital investment. Foundational AI app-building components—such as pretrained Arabic language models, speech recognition systems, orchestration frameworks, and secure deployment environments—allow universities, training centers, and language institutes to focus on pedagogical innovation rather than infrastructure engineering. This shift lowers barriers to experimentation and accelerates the adoption of AI-based learning experiences.

Shared foundational components can also shorten the distance between an academic idea and a working pilot. A language institute should not need to train a foundation model before testing whether adaptive Arabic dialogue improves learning. If reliable speech recognition, Arabic language models, retrieval systems, and agent orchestration are available as building blocks, researchers can focus experiments on questions such as how often the tutor should correct a learner, how dialect should be introduced, or what type of feedback improves retention. This changes the economics of innovation by shifting effort from rebuilding basic technology toward testing educational hypotheses.

However, access to shared infrastructure does not remove the need for local capability. Institutions still need teams that understand both pedagogy and technology well enough to configure, evaluate, and govern the tutor. A technically powerful platform can produce weak educational results if the curriculum is poorly represented or if evaluation is treated as an afterthought. Capacity building should therefore include teachers who are comfortable working with AI systems, technical staff who understand educational objectives, and governance processes that bring these groups together during design rather than only after deployment.

Importantly, Humain's emphasis on Arabic-first AI development opens new possibilities for advancing Saudi Arabian Arabic utterances within educational technology. By supporting localized speech and language models that capture regional pronunciation, prosody, and usage patterns, the platform enables the creation of AI tutors that are both linguistically authentic and pedagogically aligned with learners' real-world exposure. This positions Saudi institutions to lead in the development of Arabic AI tutors that balance Modern Standard Arabic instruction

ISSN : 2410-1818

with accurate recognition and contextualization of Saudi Arabic expressions, thereby enhancing learner engagement and instructional relevance.

Localized language support is especially valuable in speech-based tutoring. A learner living in Saudi Arabia may repeatedly hear forms, pronunciations, and conversational routines that are not represented in formal textbooks. An AI tutor can use these encounters as teaching moments. It can explain the local expression, show the MSA equivalent, and indicate when each form is appropriate. This helps the learner connect classroom knowledge to the surrounding language environment rather than treating the two as separate systems. For adult learners, that connection can be highly motivating because progress becomes visible in daily life.

Collectively, the combination of infrastructure democratization, foundational AI tooling, and localized language support creates a cost-efficient and innovation-friendly ecosystem. This ecosystem not only improves the economic feasibility of AI tutors but also strengthens the Kingdom's capacity to serve as a regional leader in Arabic AI–driven education.

A national ecosystem also creates opportunities for reusable Arabic educational assets. Curated examples, error taxonomies, pronunciation datasets, dialect labels, assessment items, and teacher-reviewed explanations can potentially be developed once and adapted across programs, subject to appropriate ownership and governance. These assets may become as important as the underlying model because they carry the pedagogical and linguistic knowledge that makes a tutor trustworthy in a specific learning environment.

Humain's early AI offerings already show tangible relevance for educational contexts, extending beyond infrastructure to practical tools that can be leveraged in teaching and learning. The flagship application HUMAIN Chat released last year and powered by the ALLAM 34B Arabic large language model provides a culturally grounded conversational AI interface that understands Modern Standard Arabic as well as a wide range of regional dialects and supports seamless switching between Arabic and English. This capability can be directly utilized in language learning environments to support real-time dialogue practice, pronunciation feedback, vocabulary expansion, and explanatory dialogue tailored to adult non-native learners' needs, helping bridge gaps between formal instruction and authentic language use.

For educational adoption, however, a conversational application should be treated as a starting point rather than a complete tutor. Open-ended conversation can increase speaking confidence and expose learners to vocabulary in context, but structured learning requires additional layers, proficiency goals, curriculum sequencing, memory of prior performance, controlled feedback, and assessment. The distinction matters because users may perceive a fluent conversational system as knowledgeable enough to teach autonomously. Institutions should resist that assumption and evaluate each instructional capability separately before integrating it into formal learning pathways.

ISSN : 2410-1818

In parallel, the ongoing rollout of Humain One deployed initially across government contexts—suggests a platform capable of hosting domain-specific agents and applications that could be adapted for educational purposes, such as intelligent tutoring modules or curriculum-aligned learning companions that leverage the same underlying AI models. These applications, built on shared sovereign infrastructure, allow educational institutions to exploit mature AI capabilities without the high costs and technical burden associated with developing their own foundation models. Consequently, educators can integrate sophisticated language understanding, adaptive feedback, and culturally tuned generative AI into curricula, enhancing personalized learning experiences while minimizing ownership and operational costs.

A practical deployment model could begin with narrowly defined use cases where benefits are high and risks are manageable. Examples include guided conversation practice, vocabulary review, reading support with optional diacritics, grammar question answering from an approved curriculum, and post-lesson exercises that give teachers visibility into recurring errors. More autonomous functions, such as changing a learner's curriculum pathway or issuing high-stakes proficiency judgments, can be introduced only after sufficient evidence of reliability. This staged approach allows institutions to learn from real use without making early pilots responsible for every aspect of teaching.

A further advantage of this staged model is that it produces evidence that can guide later investment. Rather than committing immediately to a fully autonomous tutor, institutions can compare learning gains, teacher workload, learner engagement, and operating cost across clearly defined use cases. Findings from a reading-support pilot may justify deeper work on diacritization and comprehension, while results from speaking practice may reveal a greater need for accent-sensitive speech analysis or dialect handling. This evidence-based progression reduces the risk of adopting impressive technology without a clear educational return. It also gives academic leadership a stronger basis for deciding which capabilities should remain experimental, which are ready for wider deployment, and where human supervision must remain intensive. In this way, implementation becomes a learning process for the institution itself, allowing technical ambition to grow alongside pedagogical confidence, governance maturity, and demonstrated value for learners.

Integration with existing teaching practice is equally important. The tutor should not become a parallel curriculum that learners use independently of their course. Teachers need to know what the system is reinforcing, learners need consistent terminology across classroom and digital environments, and assessments should reflect the same learning objectives. When these elements are aligned, AI can extend the teacher's instructional reach. When they are not, the learner may receive contradictory explanations or spend time practicing content that does not support course progression.

ISSN : 2410-1818

From an operational perspective, the platform should support clear roles for content owners, academic reviewers, technical administrators, and data-governance teams. Changes to prompts, knowledge sources, or learner-model policies can affect instructional behavior as much as changes to the underlying model. Versioning and review processes therefore matter. An institution should be able to determine which configuration produced a particular tutor response and to roll back or correct problematic behavior when necessary. This is part of treating the tutor as an educational system rather than as a consumer chatbot.

5. Conclusion and Call to Action

The convergence of advanced AI tutoring capabilities, Arabic-first language technologies, and nationally democratized AI infrastructure presents a rare and timely opportunity for educational systems in Saudi Arabia. Through initiatives such as Humain, supported under the leadership of Mohammed bin Salman, the Kingdom has laid the technological and institutional foundations necessary to deploy scalable, cost-efficient, and culturally grounded AI-based learning solutions.

The opportunity is significant precisely because the technology and the educational need are developing at the same time. Adult learners increasingly expect digital support that is immediate, personalized, and available across devices, while Arabic language institutions possess deep expertise in grammar, curriculum design, pronunciation, and learner development. Bringing these strengths together creates the possibility of systems that are technologically advanced without becoming pedagogically shallow. The contribution of educational institutions is not simply to provide content to an AI platform; it is to define what good Arabic teaching looks like when instruction is partly mediated by an intelligent agent.

Educational institutions across Saudi Arabia are uniquely positioned to capitalize on this ecosystem to lead the global advancement of Arabic language adoption. By integrating AI tutors that combine Modern Standard Arabic instruction with accurate modeling of Saudi Arabic utterances, institutions can deliver personalized, engaging, and context-aware learning experiences for both native and non-native learners. Such systems not only address existing challenges in access, consistency, and instructional quality, but also redefine how Arabic can be taught and learned in the digital age.

To realize this potential, personalization should remain connected to clearly stated learning outcomes. The tutor should know whether it is helping a learner read accurately, participate in workplace conversation, improve formal writing, understand lectures, or access religious and academic texts. These goals can then guide the choice of register, vocabulary, correction intensity, and assessment. A system that tries to personalize without a clear instructional destination may become responsive in conversation but directionless over time. Adult learners are likely to value flexibility most when it helps them reach a goal they recognize as relevant to their lives.

ISSN : 2410-1818

Saudi institutions are also in a strong position to contribute knowledge that global language models may not capture reliably on their own. Expertise in teaching Arabic to non-native speakers, familiarity with Saudi linguistic contexts, and access to diverse learner populations can inform datasets, evaluation protocols, and tutoring strategies that are specific to Arabic rather than borrowed directly from English-language educational technology. This local academic contribution is essential if Arabic AI tutoring is to mature as a field rather than remain an adaptation of general conversational AI.

Embracing AI-powered education at this stage enables Saudi institutions to move beyond passive adoption toward active leadership shaping standards, pedagogical models, and technological benchmarks for Arabic AI education worldwide. Strategic investment, thoughtful governance, and collaboration between educators and AI practitioners will be essential to realize this vision. If leveraged effectively, AI tutors can become a powerful instrument for spreading the Arabic language in a smart, efficient, and globally scalable manner, reinforcing Saudi Arabia's role as a leader in both education and artificial intelligence.

The immediate priority should therefore be disciplined experimentation. Institutions can begin with targeted pilots, define measurable learning objectives, keep qualified educators in the evaluation loop, and document both successful and unsuccessful interactions. Over time, evidence from these pilots can support more ambitious forms of agentic tutoring. The objective should not be maximum autonomy as quickly as possible; it should be reliable educational value with a level of autonomy that grows only when the system has demonstrated that it can use that autonomy responsibly.

In practical terms, this requires sustained collaboration among Arabic linguists, teachers of Arabic as a second language, AI engineers, instructional designers, speech specialists, and institutional leaders. Each group sees a different part of the problem. Linguists can identify subtle language errors that may escape technical benchmarks; teachers can judge whether an intervention is useful at a particular stage; engineers can design controls and monitoring; and leaders can align investment with long-term educational goals. A successful tutor will emerge from this shared process rather than from model capability alone.

If this approach is followed, AI tutoring can extend rather than dilute the human strengths of Arabic education. It can give learners more chances to practice, give teachers better visibility into recurring difficulties, and give institutions a scalable way to reach adults whose schedules or locations make traditional study difficult. The technology is most valuable when it handles repetition, personalization, and immediate support while preserving human judgment for curriculum, culture, motivation, and complex educational decisions. In that form, agentic AI can become a practical partner in expanding high-quality Arabic learning while maintaining the linguistic depth and educational responsibility the field requires.

ISSN : 2410-1818

Ultimately, success should be judged by sustained learner progress, trustworthy instruction, and the extent to which AI expands meaningful access to high-quality Arabic education.

ISSN : 2410-1818

References

Abu Elhija, D. (2014). A new writing system? Developing orthographies for writing Arabic dialects in electronic media. *Writing Systems Research,* 6(2), 190–214.

Alayba, A. M. (2025). Arabic Natural Language Processing (NLP): A Comprehensive Review of Challenges, Techniques, and Emerging Trends. *Computers*, 14(11), 497.

Chennoufi, A., & Mazroui, A. (2017). Morphological, syntactic and diacritics rules for automatic diacritization of Arabic sentences. *Journal of King Saud University – Computer and Information Sciences,* 29(2), 156–163.

Cinta, F., Irawan, B., & Hasan, N. (2023). Semantic study: Polysemy in Arabic forms of verbs and nouns. *Aksara: Jurnal Bahasa dan Sastra,* 24, 520–530.

Hegazi, M. O., Al-Dossari, Y., Al-Yahy, A., Al-Sumari, A., & Hilal, A. (2021). Preprocessing Arabic text on social media. *Heliyon,* 7(4).

Midhwah, A. A., & Alhawary, M. T. (2020). Arabic diacritics and their role in facilitating reading speed, accuracy, and comprehension by English L2 learners of Arabic. *The Modern Language Journal,* 104(2), 418–438.

ISSN : 2410-1818

مجلـــة الأندلـــس للعلـــوم الإنسانيـــة والاجتماعيـــة
مجلـــة دوليـــة شهـريـــة علميـــة محكمـــة
الترقيـــم الدولي الإلكتروني : ISSN 2410- 521X
الترقيـــم الدولـــي الورقـــي :ISSN 2410- 1818
البريد الإلكتروني : journal@andalusuniv.net

المجلـــة مفهرســـة فـــي المواقـــع الآتيـــة :

العام	2022	2023	2024	2025	2026
معامل أرسيف	0.1954	0.3759	0.3068	0.5978	0.4135
معامل التأثير العربي	1.73	1.25	1.55	1.59	1.58