Maalim
e-ISSN: 2600-6944 | p-ISSN: 2170-0052
Volume 19 Issue 1/ Juin 2026 PP. 25-54

Generative AI in Language Instruction: Personalization, Practice, and Proficiency. A Pedagogy-First Framework

Dr. Attia Youseif *
Indiana University-Bloomington, USA
ayouseif@iu.edu

DOI: 10.33705/1111-019-001-002

Received: 16/06/2026
Accepted: 21/06/2026
Published: 30/06/2026

***Corresponding Author**

Citation :
Attia,Y. (2026).
Generative AI in Language Instruction: Personalization, Practice, and Proficiency. A Pedagogy-First Framework
Maalim
I(1), 25-54

Maalim
© 2026 The Author(s).
Published by the High council of the Arabic language.
This is an open access article under the CC BY license

Abstract:

Generative artificial intelligence has entered language education with unusual speed, but its pedagogical meaning remains unsettled. This article argues that the educational value of generative AI in language instruction is best understood through three linked commitments: personalization, practice, and proficiency. Personalization refers to the capacity to generate differentiated input, scaffolds, and feedback at a scale that teacher labor alone cannot usually sustain. Practice refers to the expansion of low-stakes opportunities for conversation, writing, revision, and feedback beyond the time limits of the classroom. Proficiency refers to the need to anchor AI-supported tasks to recognized standards of communicative ability rather than to the surface fluency of machine output. The article develops these commitments through a pedagogy-first lens grounded in distributed agency, where learning emerges from the interaction among learner, teacher, and machine, but where the teacher remains the central designer, curator, and evaluator of the learning ecology. Arabic serves throughout as a stress test rather than a side example. As a less commonly taught but strategically important language, Arabic exposes the limitations of current models: uneven dialectal coverage, cultural bias, diglossic complexity, and thin evidence for long-term classroom effects. The article concludes that generative AI offers real affordances for language learning, but only under teacher-mediated conditions that protect authenticity, learner agency, assessment validity, and cultural integrity.

Keywords: generative AI; language instruction; personalization; practice; proficiency; Arabic; teacher-in-the-loop; distributed agency; ACTFL; ILR; cultural bias; less commonly taught languages.

الذكاء الاصطناعي التوليدي في تدريس اللغات:
التخصيص والممارسة و إتقان اللغة. إطار عمل يركز على التربية

د. يوسف عطيه
جامعة بلومينكتون، إنديانا، الولايات المتحدة الأمريكية

الملخّص:

دخل الذكاء الاصطناعي التوليدي مجال تعليم اللغات بسرعة غير مسبوقة، لكنّ أبعاده التربوية لا تزال غير محددة. ولفهم القيمة التعليمية للذكاء الاصطناعي التوليدي في تدريس اللغات؛ يرى المقال أن أفضل طريقة لذلك تتجلى من خلال ثلاثة عناصر مترابطة: التخصيص، والممارسة، والكفاءة. يشير التخصيص إلى القدرة على توليد مدخلات متنوعة، ودعامات تعليمية، وردود فعل على نطاق لا يستطيع عمل المعلم وحده عادةً أن يواكبه. وتشير الممارسة إلى توسيع فرص المحادثة والكتابة والمراجع. أما الكفاءة فتشير إلى الحاجة إلى ربط المهام المدعومة بالذكاء الاصطناعي بمعايير معترف بها لتحسين التواصل بدلاً من الطلاقة السطحية لمخرجات الآلة. يطور البحث هذه الالتزامات من خلال منظور يركز على التربية أولاً حيث ينشأ التعلم من التفاعل بين المتعلم والمعلم والآلة، لكن يظل المعلم هو المصمم الرئيسي والمنسق والمقيم لبيئة التعلّم. وتُستخدم اللغة العربية في هذا المقال كله كاختبار للتحمّل وليس كمثال جانبي. وباعتبارها لغة أقل شيوعًا في التدريس لكنها ذات أهمية استراتيجية، وتكشف اللغة العربية عن محدودية النماذج الحالية: التغطية اللهجية غير المتكافئة، والتحيز الثقافي، وتعقيد ثنائية اللغة، وقلة الأدلة على الآثار طويلة المدى في الفصول الدراسية.

يخلص المقال إلى أن الذكاء الاصطناعي التوليدي يوفر إمكانيات حقيقية لتعلم اللغة، ولكن فقط في ظل ظروف يتوسط فيها المعلم لحماية الأصالة، وكفاءة المتعلم، وصحة التقييم، والسلامة الثقافية.

الكلمات المفتاحيّة: الذكاء الاصطناعي التوليدي؛ تعليم اللغة؛ التخصيص؛ الممارسة؛ الكفاءة؛ اللغة العربية؛ تدخل المعلم؛ الوكالة الموزعة؛ACTFL ؛ILR ؛ التحيز الثقافي؛ اللغات الأقل شيوعًا.

Introduction: The Generative Turn in Language Education :The public release of ChatGPT in late 2022 marked a visible turn in the history of computer-assisted language learning. Earlier digital language tools had already changed classrooms in meaningful ways. They made dictionaries searchable, delivered drills automatically, tracked learner responses, and supported forms of online interaction that could not have been organized as easily in print-based instruction. Yet most of those tools still worked within bounded pathways. A learner selected from a menu, completed an exercise,

clicked through feedback, or interacted with a dialogue system whose possibilities had been anticipated in advance. The machine could process language, but it rarely produced sustained, context-sensitive language that felt responsive to the learner's communicative intention.

Generative AI alters that condition. Large language models do not merely retrieve a fixed exercise or mark an answer against an answer key. They generate new language in response to open-ended prompts, and they can adjust topic, register, length, difficulty, genre, and feedback style within the same interaction. This is why the current moment is not simply another stage in the CALL lineage. The difference is not that the tool is faster, more convenient, or more polished, although it is often all three. The difference is that the tool can participate in language production itself. It can simulate a conversation, reformulate a learner's paragraph, produce a level-adjusted text, or generate feedback on a draft in ways that earlier systems could only approximate (Kohnke, L., Moorhouse, B. L., & Zou, D. 2023; Chiu, T. K. F. 2024).

This change has produced both enthusiasm and anxiety. Some discussions present generative AI as an always-available private tutor capable of giving every learner individualized support. Others focus on academic dishonesty, fabricated citations, overreliance, and the weakening of learner autonomy. Both responses contain an element of truth, but neither is sufficient as an educational argument. The key question is not whether AI is impressive, nor whether it is dangerous in the abstract. The question is what language teachers can responsibly do with it, under what conditions, for which learners, and toward what goals.

A pedagogy-first frame is therefore necessary. Hype-first discourse begins with the tool and asks what it can do. Pedagogy-first thinking begins with language development and asks what learners need. In that second frame, the value of generative AI in language instruction can be understood through three linked commitments: personalization, practice, and proficiency. Personalization concerns differentiated input and support. Practice concerns abundant, low-stakes opportunities to use language and receive feedback. Proficiency concerns the alignment of AI-supported work with recognized descriptors of communicative ability, including ACTFL and ILR frameworks (American Council on the Teaching of Foreign Languages 2012; Interagency Language Roundtable n.d.).

These commitments depend on the teacher rather than replacing the teacher. Generative AI can produce material quickly, but speed is not judgment. It can respond fluently, but fluency is not cultural knowledge. It can evaluate learner language, but its evaluation must be read against pedagogical standards it does not itself own. The teacher remains the person who defines the learning objective, selects the appropriate task, evaluates the fit between output and goal, and protects the human dimensions of language learning that no model can supply: trust, motivation, identity, cultural interpretation, and the relational texture of classroom life (Godwin-Jones, R. 2024).

Arabic provides the recurring test case for this argument. It is not included simply to diversify the examples. It is chosen because it makes the limits of current generative AI visible. Arabic is widely taught among less commonly taught languages in the United States and has long been treated as a critical language for diplomacy, security, research, and global engagement. At the same time, it presents exactly the kinds of linguistic and cultural conditions that challenge large language models: diglossia, dialectal variation, rich morphology, script-specific processing, and cultural contexts that are often flattened or distorted in English-dominant training data. If personalization, practice, and proficiency can be implemented responsibly in Arabic, the framework has strength beyond the well-resourced cases of English and a small number of dominant world languages. If the framework breaks down in Arabic, that failure reveals something important about the limits of generative AI more generally.

The article proceeds in eight sections. Section 2 develops the conceptual frame, moving from a tool-based view of educational technology toward distributed agency and teacher-in-the-loop design. Section 3 examines personalization and the tension between differentiated materials and the cultural-linguistic limits of model coverage in Arabic. Section 4 turns to practice, especially conversational interaction and L2 writing feedback. Section 5 anchors the discussion in proficiency frameworks and assessment validity. Section 6 consolidates cautions around integrity, overreliance, fabricated output, bias, and equity. Section 7 draws out implications for teacher preparation and classroom practice. Section 8 returns to the thesis: personalization, practice, and proficiency are not

three separate uses of AI, but one integrated agenda held together by pedagogy and by the teacher's judgment.

Arabic helps clarify this shift because the teacher's expertise is needed before, during, and after the interaction. Before the activity, the teacher must specify whether learners are working in Modern Standard Arabic, a regional variety, or a controlled combination. During the activity, learners may need strategies for recognizing drift in register or variety. After the activity, the teacher must help them interpret what the model did well and where it failed. The AI may appear to be the conversational partner, but the teacher remains the designer of the conditions under which that partnership has educational value.

This uncertainty is not a reason to keep AI outside language education. It is a reason to change the unit of design. Teachers cannot design only a worksheet or a prompt. They must design the entire interactional frame: what the learner may ask, what kind of feedback the model should give, how the learner should evaluate that feedback, what evidence of independent performance will be collected, and how the activity connects to the next class meeting. The interaction with AI therefore becomes one part of a larger pedagogical sequence rather than a self-contained learning event.

One reason the pre-2022 baseline matters is that it prevents the argument from becoming technologically amnesiac. Language teachers were not waiting for AI before they individualized assignments, used authentic media, organized online exchanges, or built learner autonomy through digital tools. What changed is the relationship between learner input and machine response. In earlier systems, adaptation was usually planned in advance by the designer. In generative systems, adaptation is produced in the moment. That difference creates new pedagogical possibilities, but it also creates new uncertainty, because the teacher cannot fully know in advance what language, explanation, or cultural example the model will produce.

From Tools to Partners: A Working Conceptual Frame: Before personalization, practice, and proficiency can be examined responsibly, the role of generative AI itself must be clarified. The older language of educational technology often treated digital tools as instruments. A learner used a dictionary, completed a grammar drill, listened to a recording, or submitted answers to an

automated tutor. The tool extended access or efficiency, but it remained largely passive. Generative AI unsettles that view because it behaves less like a static instrument and more like an interactive participant in a learning ecology. It produces language, adjusts to prior turns, imitates genres, responds to prompts, and offers feedback that can influence what the learner does next.

The most productive frame for this shift is distributed agency. Godwin-Jones argues that generative AI in language learning should be understood through shared agency among human and non-human actors, rather than through a simple model in which a learner uses a neutral tool (Godwin-Jones, R. 2024). Agency is not transferred to the machine, and it is not held entirely by the learner. It is distributed across the activity system. The learner sets intentions, prompts, revises, accepts, rejects, and reflects. The AI generates language, examples, questions, and feedback. The teacher designs the activity, defines the boundaries of acceptable use, curates the output, interprets learner performance, and connects the interaction to broader curricular goals.

These reframing matters because it protects against two errors. The first error is automation fantasy, the belief that an AI tutor can replace the teacher by delivering individualized support at scale. The second is technological dismissal, the assumption that AI is merely another gadget and therefore can be handled by existing policy language without deeper pedagogical adjustment. Distributed agency avoids both. It recognizes that the model shapes the interaction, but it also insists that the teacher remains responsible for the instructional meaning of that interaction.

The teacher-in-the-loop principle follows from this account. In language education, the teacher is not only a content provider. The teacher is a designer of communicative conditions. That design work includes selecting topics, sequencing tasks, regulating difficulty, interpreting errors, responding to affective needs, and judging whether language use is appropriate for a given audience, register, and cultural setting. AI can support parts of that work, but it cannot assume the whole of it. When it generates Arabic dialogue, for example, it may produce grammatically acceptable Modern Standard Arabic where a dialectal exchange would be more authentic. When it gives writing feedback, it may correct form while missing rhetorical purpose. When it evaluates a performance, it

may reward surface fluency while ignoring whether the learner actually accomplished the communicative task.

The phrase 'AI tutor' should therefore be used carefully. A tutor, in the human sense, notices confusion, adapts to a learner's history, interprets hesitation, adjusts encouragement, and understands the cultural and interpersonal stakes of language use. A generative model can simulate parts of this role. It can ask follow-up questions, give corrections, and maintain a conversation. Yet it has no lived experience, no classroom memory unless one is supplied, no stable pedagogical conscience, and no independent understanding of the social world to which language belongs. Tseng and Warschauer's argument about AI writing tools is relevant here: rather than trying to ban or romanticize these tools, educators should teach learners to understand them, access them equitably, prompt them productively, corroborate their output, and incorporate their assistance responsibly (Tseng, W., & Warschauer, M. 2023).

For language teachers, this means that prompt design is not a narrow technical trick. It is a pedagogical act. A good classroom prompt reflects a theory of learning, a target proficiency level, an understanding of the learner, and a sense of what kind of output will support the next instructional move. Asking a model to 'write a dialogue in Arabic' is not the same as asking it to 'generate a short Novice High interpersonal exchange in Egyptian Arabic between two university students greeting each other for the first time, using familiar vocabulary, culturally appropriate politeness formulas, and no English translation.' Even then, the output must be checked. Prompting can improve quality, but it does not eliminate the need for expert reading.

This conceptual frame sets up the argument that follows. Personalization is not automatic individualization; it is teacher-mediated differentiation supported by AI generation. Practice is not simply more chatbot use; it is structured engagement that protects reflection, revision, and learner agency. Proficiency is not the model's judgment that a response is good; it is the teacher's alignment of activity and evidence to external standards. In each case, AI becomes useful only when it is located within a designed pedagogical system.

A strong personalization routine can use AI in stages. The teacher first generates draft materials with explicit constraints. The teacher then edits for accuracy, cultural fit, and level. Learners use the material and produce evidence of comprehension or performance. Finally, learners reflect on what kind of support helped them and what kind of support distracted them. In that cycle, personalization is not only the tailoring of content to a learner. It becomes the learner's growing awareness of how support works and how to use it responsibly.

For this reason, adaptive materials should be judged by fit, not only by apparent quality. A generated reading may be grammatically clean and still be wrong for the course because it contains vocabulary the class has not yet encountered, cultural assumptions the teacher has not prepared learners to interpret, or sentence structures that pull attention away from the intended objective. A dialogue may be level-appropriate but pragmatically dull. A writing model may offer sophisticated transitions that make a Novice learner's work look less like the learner's own language. The teacher's question should therefore be: what does this output make possible for the learner now?

Personalization also requires restraint in relation to learner identity. It is tempting to describe AI as capable of meeting every learner exactly where that learner is. In practice, learner needs are often more complex than the information available in a prompt. A heritage learner may need support that affirms oral competence while making literacy development explicit. A learner with anxiety may need gradual risk-taking rather than constant correction. A learner preparing for government service, study abroad, religious literacy, or community engagement may need different registers and genres. A model can respond to categories, but it cannot know the learner as a teacher can know the learner over time.

Personalization: Personalization is the most obvious promise of generative AI, and also the promise most easily overstated. Language teachers have always recognized that learners differ. They differ in prior exposure, literacy background, motivation, heritage status, processing speed, confidence, professional goals, cultural knowledge, and proficiency profile across skills. Differentiation has rarely been limited by teachers' pedagogical imagination. It has been limited by time. Preparing three versions of a reading, several tiers of grammar support, individualized

feedback on drafts, and alternative speaking prompts for different proficiency levels is labor-intensive. In many programs, especially smaller programs in less commonly taught languages, the teacher simply does not have enough hours to produce all the differentiated material learners could use.

Generative AI is genuinely suited to this problem. It can produce level-adjusted input, generate examples around topics that matter to learners, create follow-up questions, simplify or elaborate a text, build vocabulary tasks, and offer draft feedback within seconds. Earlier work on AI-assisted personalized language learning identified adaptive input, feedback, and learner-specific pathways as central affordances of AI in language education (Chen, X., Zou, D., Cheng, G., & Xie, H. 2021). Godwin-Jones extends this argument in relation to generative AI, noting that learners and teachers can specify conversational parameters such as level, register, topic, feedback type, and study goals (Godwin-Jones, R. 2024). The model can therefore extend the teacher's ability to act on what the teacher already knows about learner variation.

The strongest use of AI personalization is not the production of polished material for passive consumption. It is the creation of adjustable learning conditions. A teacher can generate a short text at Novice High, then ask for the same content at Intermediate Mid. The teacher can ask for comprehension questions that target gist, detail, inference, or cultural interpretation. A learner preparing for a professional interview can rehearse different versions of the same scenario. A heritage learner can receive scaffolding that respects oral competence while strengthening literacy. A beginning learner can practice formulaic exchanges without waiting for classroom time. These uses are meaningful because they respond to a real structural problem in language education: learners need more tailored input and feedback than teachers can usually produce by hand.

Yet personalization rests on a demanding assumption: the model must be able to generate linguistically accurate, culturally appropriate, and level-appropriate language for the learner's target context. That assumption is not equally safe across languages. It is especially unsafe in Arabic. Arabic is often described as a single language in institutional documents, but Arabic instruction lives inside a complex sociolinguistic ecology. Modern Standard Arabic is the language of formal writing, news,

education, and many official domains. Spoken life, however, is carried by regional dialects that differ substantially from one another in phonology, morphology, vocabulary, syntax, discourse markers, and pragmatic conventions. Personalization in Arabic cannot mean simply choosing 'Arabic' as the target language. It has to mean choosing which Arabic, for which communicative context, and toward which learner goal.

Current models often struggle with exactly that distinction. Evaluation work on Arabic NLP has shown that general-purpose models perform more reliably on Modern Standard Arabic than on dialectal Arabic, and that Arabic-specific or fine-tuned systems can outperform large general models on specialized Arabic tasks (Khondaker, M. T. I., Waheed, A., Nagoudi, E. M. B., & Abdul-Mageed, M. 2023). The pedagogical consequence is serious. A teacher may ask for Egyptian Arabic and receive formal Arabic with a few dialectal tokens added. A learner may ask for Levantine conversational practice and receive a hybrid that no native speaker would naturally use. The text may look fluent to a non-specialist, but fluency at the wrong register is not personalization. It is misalignment.

The problem extends beyond dialect to cultural framing. Naous, Ryan, Ritter, and Xu examined cultural bias in large language models and found systematic preference for Western over Arab entities when models operated in Arabic, including names, foods, clothing, religious references, and other culturally marked categories (Naous, T., Ryan, M. J., Ritter, A., & Xu, W. 2024). Such findings matter directly for language teachers because culture is not an optional supplement to language. A personalized Arabic lesson about daily life, family, meals, religious routines, hospitality, campus interaction, or professional communication must reflect culturally plausible worlds. If the model defaults to Western social assumptions while generating Arabic, the result may be grammatically polished but culturally displaced.

Dai, Suzuki, and Chen make a related point in their discussion of generative AI for intercultural professional communication. AI can support practice in intercultural contexts, but it must be used with care because culturally situated communication involves far more than correct wording (Dai, D. W., Suzuki, S., & Chen, G. 2025). This is highly relevant to Arabic. A greeting in Arabic is not only a greeting. It carries expectations about formality, gender, age, relationship, region, religious register,

and conversational sequencing. A refusal is not only a refusal. It often requires mitigation, relational softening, and culturally recognizable phrasing. A model may generate a literal answer, but the teacher must decide whether that answer is socially inhabitable.

Personalization without teacher curation therefore risks producing what might be called fluent thinness: language that sounds plausible but lacks the dialectal, pragmatic, and cultural density that real communication requires. In Arabic, this risk is amplified because learners may not know enough to detect the problem. A beginning student cannot easily distinguish Modern Standard Arabic from a dialectal hybrid. An intermediate learner may accept a culturally odd phrase because it is grammatically well formed. Even advanced learners may not recognize subtle pragmatic misalignment if the scenario is unfamiliar.

The solution is not to reject AI personalization. It is to design it more rigorously. Teachers should specify target variety, proficiency level, communicative mode, genre, audience, and cultural frame. They should ask models to explain variety choices, but they should not trust those explanations without checking. They should compare AI output with authentic materials, native-speaker judgments, corpus evidence, or locally validated examples. They should involve learners in noticing the limits of AI-generated language by asking them to identify where a generated dialogue feels too formal, too generic, too English-like, or culturally incomplete. In this way, AI output becomes material for learning, not merely material for use.

For less commonly taught languages, this point has an equity dimension. The languages that stand to benefit most from scalable differentiation are often the languages for which model coverage is weakest. Arabic programs may have fewer commercial resources than programs in Spanish or French, and teachers may carry heavy responsibility for curriculum design, assessment, and material creation. Generative AI can help with that burden, but only if institutions recognize that less commonly taught language teachers need time and training to vet output. Personalization is not free labor. It shifts labor from first drafting to critical evaluation, adaptation, and cultural-linguistic quality control.

Seen this way, personalization is not an automated feature of generative AI. It is a pedagogical commitment enacted through teacher design. The model can generate, vary, and respond. The teacher must decide whether the output belongs in the learner's path. For Arabic, that decision is particularly consequential, because the difference between level-appropriate personalization and culturally flattened standardization can be difficult to see on the surface.

Practice should also return to the classroom. AI-mediated homework becomes more powerful when it is not hidden as private interaction but brought back as material for shared analysis. Students can compare different model responses to the same prompt, identify better and weaker corrections, discuss cultural oddities, or revise an AI-generated text into a more authentic one. This turns practice into inquiry. Learners do not simply use AI; they study its language behavior as part of developing their own.

In Arabic writing instruction, practice must also distinguish between mechanical accuracy and discourse development. A model may correct agreement, case endings, or word order, but advanced writing requires control of rhetorical sequencing, evidence, stance, and audience. In media Arabic, learners may need to summarize without translating word for word. In academic Arabic, they may need to construct a claim, qualify it, and connect evidence. In heritage learner contexts, they may need to move from orally natural phrasing toward formal written control without treating their home variety as deficient. These are instructional decisions. AI can assist with examples and feedback, but it cannot determine the developmental priority on its own.

The same principle applies to writing. AI feedback can be immediate, but immediacy is not always pedagogically best. Sometimes learners need to sit with uncertainty long enough to notice what they do not know. If correction arrives too quickly, the learner may never formulate a hypothesis about the language. Teachers might therefore delay AI feedback until after peer review, or restrict feedback to one dimension at a time: verb tense, cohesion, audience, register, or argument structure. A narrower feedback request often produces better learning than a broad request to 'improve my writing.'

Practice also needs variety. Repeating the same chatbot pattern can quickly create a narrow interactional habit: the learner asks, the model answers, the learner follows. Human communication is less orderly. Learners need to initiate topics, repair misunderstanding, ask for clarification, negotiate meaning, and adjust to interlocutors who are not always maximally helpful. Teachers can design AI practice to include some of these demands by instructing the model to misunderstand occasionally, ask for clarification, change roles, introduce a complication, or require the learner to ask follow-up questions before the conversation can continue. Such design turns the chatbot from a smooth answer machine into a controlled site of communicative pressure.

Practice: The second commitment, practice, is where the case for generative AI is most immediately visible. Language learning requires repeated use. Learners need to try, fail, reformulate, ask, answer, narrate, describe, interpret, persuade, and revise. Classrooms, however, are limited by time. Even in a well-run class, each learner may receive only a small number of turns in extended interaction. Writing feedback also faces limits. Teachers can provide rich responses, but the time required to comment carefully on every draft makes rapid individualized feedback difficult, especially in large classes or multi-level programs.

Generative AI offers something classrooms have always needed: abundant, low-stakes, repeatable interaction. A learner can practice a dialogue at night, revise a paragraph before class, ask for clarification on a grammar point, simulate a service encounter, or rehearse a presentation several times. The absence of peer judgment can matter. Many language learners avoid risk because mistakes feel socially exposed. AI-mediated practice can lower that affective barrier. It gives learners a space to experiment before bringing language back into the classroom (Kohnke, L. et al. 2023).

Conversational practice is one obvious use. A learner can ask a model to play the role of a hotel clerk, a professor, a classmate, a host family member, or an interviewer. The learner can request slow language, simpler vocabulary, follow-up questions, or corrections after the exchange. These features can support autonomy when embedded in a course structure. They can also support equity when learners lack access to conversation partners outside class. For less commonly taught languages, where local speech communities may be small or unavailable, the appeal is easy to understand.

The risk is that simulated conversation can become too cooperative. Human interlocutors interrupt, misunderstand, shift topics, use idioms, speak at uneven speed, and bring social expectations into the exchange. A model often keeps the interaction smooth. That smoothness can be useful for beginners, but it can also hide weaknesses. A learner may feel successful because the model understands fragmented or inaccurate language that a human interlocutor would not understand as easily. Practice, in that case, produces confidence without necessarily producing transferable communicative control. Teachers need to design AI conversations that introduce appropriate difficulty, repair sequences, and follow-up tasks. Learners should not only complete the chatbot exchange. They should analyze it, summarize it, revise it, and bring evidence of their performance back into class.

Writing feedback has received even more attention. Barrot discusses the potentials and pitfalls of using ChatGPT for second language writing, noting that AI can support idea generation, organization, language accuracy, and feedback, while also raising concerns about authorship, dependence, and the quality of revision (Barrot, J. S. 2023). Yan's exploratory investigation of ChatGPT in an L2 writing practicum similarly shows that learners recognize the tool's usefulness, but also encounter questions about academic honesty, equity, and how much of the writing process should remain their own (Yan, D. 2023). These studies suggest that AI writing support is neither simply beneficial nor simply harmful. Its value depends on how the task defines learner responsibility.

A weak AI writing task asks students to submit a prompt and accept the improved output. A stronger task asks them to compare the AI feedback with peer and teacher feedback, decide which suggestions are useful, explain which suggestions they reject, and revise in a way that preserves authorial control. This difference is not minor. In the first design, AI reduces writing to output management. In the second, AI becomes a source of metalinguistic noticing and rhetorical decision-making. The teacher's assignment design determines which path the learner is likely to follow.

Overreliance is therefore not only a student problem. It is a design problem. Learners become dependent when the task rewards clean final products more than visible learning processes. If the grade is based only on a polished final paragraph, AI-generated prose becomes tempting. If the grade

includes drafts, revision notes, error analysis, oral defense, and reflection on accepted and rejected feedback, the learner must remain intellectually present. In this sense, responsible AI use in writing instruction does not require banning assistance. It requires making assistance accountable.

The aggregated evidence should be read with the same care. Li, Wang, and Yang's meta-analysis asks whether generative AI chatbots promote second language acquisition and provides a needed corrective to anecdotal claims (Li, M., Wang, Y., & Yang, X. 2025). A meta-analysis can show whether chatbot-mediated practice is associated with measurable learning gains across studies, but it cannot by itself tell teachers that any chatbot activity will be effective. The instructional conditions still matter: task type, target skill, duration, proficiency level, feedback design, learner training, and the role of the teacher. Lyu, Lai, and Guo's meta-analysis of chatbot effectiveness in language learning similarly points toward positive effects while reminding the field that design, modality, and implementation shape outcomes (Lyu, B., Lai, C., & Guo, J. 2025).

Arabic again requires caution. The current evidence base for AI in language learning remains stronger for English and other well-researched languages than for Arabic classroom contexts. Arabic-specific studies often evaluate model performance on NLP tasks, grammatical error correction, translation, or bias rather than sustained classroom learning. That evidence is valuable, but it is not the same as evidence that Arabic learners improve speaking, writing, reading, or listening proficiency through long-term AI-supported instruction. Claims about Arabic practice should therefore be measured. AI can provide more opportunities to practice Arabic, but more practice is not automatically better practice.

For Arabic speaking practice, teachers must decide whether the goal is Modern Standard Arabic, a specific dialect, or controlled movement between registers. A beginner in a program centered on Modern Standard Arabic may benefit from slow formal exchanges that reinforce classroom structures. A learner preparing for study abroad in Amman, Cairo, Rabat, or Abu Dhabi needs different pragmatic routines. If the model cannot sustain the target variety, the teacher must either adjust the activity or use AI output as a comparison exercise rather than as a model for imitation. In

writing, the teacher must likewise distinguish between feedback on formal accuracy and feedback on discourse, argument, cohesion, genre, and register.

The most defensible claim is therefore this: generative AI can expand language practice when it is used as a structured practice partner under teacher supervision. It can increase the number of turns learners take, the number of drafts they revise, and the immediacy of feedback they receive. But practice becomes pedagogically valuable only when learners remain responsible for noticing, interpreting, and revising. The classroom should not disappear into the chatbot. The chatbot should feed better classroom work.

For Arabic programs, proficiency alignment also raises curricular questions that predate AI. Should early instruction emphasize Modern Standard Arabic alone, introduce dialect from the beginning, or adopt an integrated approach? AI does not solve that debate. It makes the consequences more visible. If a program values integrated Arabic, then AI practice and assessment must reflect movement between formal and spoken registers. If a program prioritizes MSA literacy, then AI dialectal output must be constrained or treated as supplementary. Proficiency descriptors must be interpreted through the program's sociolinguistic goals, not applied mechanically.

ACTFL and ILR frameworks also remind teachers that different skills develop unevenly. AI may help a learner write beyond current speaking ability, or read with dictionary-like support beyond current listening ability. Such unevenness is normal. The teacher's task is to prevent a tool-supported strength from masking another area of need. A student who can present a polished AI-assisted essay may still need interpersonal practice in asking follow-up questions, managing hesitation, or negotiating meaning. Proficiency-aligned instruction keeps these differences visible rather than allowing a single polished product to stand for global ability.

A proficiency-oriented AI task should therefore make the support visible. The teacher might ask learners to submit a short note describing the prompt they used, the feedback they received, and the changes they made. This note is not an administrative burden. It is part of the evidence. It shows whether the learner understood the feedback, whether the learner could make a judgment, and whether the learner's revision reflects growing control. In oral work, a similar record might include

a transcript of an AI practice conversation followed by a live classroom performance on a related but not identical task.

The distinction between performance and proficiency is especially useful here. Performance describes what a learner can do in a practiced context under familiar conditions. Proficiency concerns the broader, more durable ability to use language across contexts that may not have been rehearsed. AI can improve performance quickly because it can scaffold the task heavily. That is not a weakness. Scaffolding is central to teaching. The problem begins when scaffolded performance is mistaken for independent proficiency. A learner who can produce a strong paragraph after extensive AI-guided revision may be developing, but the final paragraph alone does not show what the learner can do without that support.

Proficiency :

The third commitment, proficiency, keeps the article anchored to standards rather than novelty. Generative AI can produce language that looks advanced, but the appearance of advanced language is not the same as learner proficiency. A learner who submits a polished AI-assisted paragraph may not be able to produce similar language independently. A learner who completes a chatbot dialogue may not be able to sustain a conversation with a human interlocutor. For this reason, AI-supported instruction must be tied to recognized descriptors of what learners can do with language.

The ACTFL Proficiency Guidelines provide one of the most influential frameworks for describing communicative ability across speaking, writing, listening, and reading. They focus not on a list of grammatical structures, but on functions, contexts, text types, accuracy, and the ability to communicate in spontaneous, unrehearsed situations (American Council on the Teaching of Foreign Languages 2012). The ILR scale, widely used in federal contexts in the United States, similarly describes levels of functional language ability across skills and professional contexts (Interagency Language Roundtable n.d.). Both frameworks remind teachers that proficiency is performance-based. It concerns what a learner can do with language, not what a text produced in the learner's name looks like.

This distinction is especially important in AI-supported teaching. A model can generate a text at an ACTFL Advanced level even if the learner is at Intermediate Low. That output may be useful as input, but it cannot be treated as evidence of the learner's proficiency. A model can correct errors in a Novice learner's writing, but if the correction produces sentence patterns the learner cannot understand or reuse, it may obscure rather than support development. Proficiency-oriented AI use therefore requires careful calibration. The question is not simply whether AI improves the product. The question is whether AI helps the learner move toward the next sustainable level of independent performance.

Can-Do statements are useful here because they translate proficiency into learner-facing goals. A Can-Do orientation asks what learners can interpret, exchange, and present in meaningful contexts. AI-supported tasks should be built around those outcomes. For example, at the Novice level, AI can help learners practice formulaic exchanges, identify familiar words in short texts, and create simple messages about themselves. At the Intermediate level, AI can support question-answer exchanges, sentence-level narration, short explanations, and connected writing on familiar topics. At the Advanced level, AI can support argumentation, hypothesis, narration across time frames, and audience-sensitive presentations, but only if the learner remains the author of the performance.

Assessment design across the interpretive, interpersonal, and presentational modes must therefore be intentional. In the interpretive mode, AI might generate or adapt texts, but assessment should determine what the learner can understand without full AI mediation. In the interpersonal mode, AI can provide role-play practice, but evidence of growth should include human interaction, repair, clarification, and spontaneous response. In the presentational mode, AI can help learners revise drafts, but teachers should collect process evidence: first drafts, revision explanations, oral defenses, and reflections on feedback. These artifacts help separate learner development from machine polish.

The validity problem becomes sharper when AI both generates practice and evaluates it. If the same model creates the prompt, supplies examples, responds to the learner, and then judges the learner's performance, the task risks circularity. Learners may become skilled at satisfying the model

rather than at communicating with people. The model may reward grammatical smoothness while underweighting communicative success, pragmatic appropriateness, register, cultural fit, or discourse control. Chiu's broader analysis of generative AI in education emphasizes the need for policy and research directions that address such changes in practice rather than treating them as simple tool adoption (Chiu, T. K. F. 2024). The assessment question is not whether AI can produce a score. The question is whether the score supports a valid inference about learner ability.

Arabic makes this validity problem concrete. A learner practicing with an AI system optimized for Modern Standard Arabic may improve in formal written structures while remaining unable to handle everyday dialectal interaction. Conversely, a learner using dialectal prompts may receive inconsistent or hybrid forms that confuse the distinction between standard and spoken varieties. If the model corrects dialectal forms toward Modern Standard Arabic, it may teach learners that authentic dialectal usage is wrong. If it overgeneralizes dialectal forms, it may weaken formal control. Proficiency in Arabic must therefore be defined in relation to the learner's program goals: MSA literacy, dialectal interaction, integrated MSA-dialect performance, professional reading, media comprehension, heritage literacy, or study-abroad survival.

Teachers need rubrics that make these goals visible. A proficiency-referenced rubric should not focus only on grammar. It should identify communicative function, comprehensibility, text type, discourse organization, vocabulary range, control of time frames where appropriate, register, cultural appropriateness, and independence from support. When AI is allowed, the rubric should specify what kind of assistance is acceptable. Is AI being used for brainstorming, vocabulary review, grammar checking, reformulation, pronunciation practice, or feedback? Does the final product show what the learner can do, or what the model can do? Without these distinctions, AI integration can weaken the very evidence teachers need for proficiency-based instruction.

A standards-based approach also helps teachers resist the pressure of novelty. The purpose of a language course is not to maximize AI interaction. It is to develop learners' ability to communicate. If AI helps learners practice more, receive faster feedback, and engage in richer reflection, it serves

that purpose. If it produces polished output while hiding weak learner control, it does not. Proficiency is the anchor that allows teachers to tell the difference.

Responsible integration, then, is both ethical and pedagogical. It asks teachers to protect learners from bad output, but also to protect learners from losing the productive struggle that language development requires. A model can reduce frustration, but some difficulty is necessary. Learners need time to search for words, negotiate form, experience breakdown, repair meaning, and revise imperfectly. The goal is not frictionless language production. The goal is increasingly capable human communication.

Detection tools do not resolve the integrity problem. AI-generated text is difficult to identify reliably, and false accusations can harm students. A more constructive approach is to design assignments in which the process matters and in which appropriate AI use is disclosed. Oral follow-up, in-class writing, drafts, reflective memos, and individualized topics make it easier for teachers to see learning. They also reduce the need for surveillance. The purpose is not to catch students. The purpose is to build courses where the easiest path is also the learning path.

Another caution concerns the status of authentic materials. AI-generated texts can be useful, but they should not replace contact with real language communities, real media, and real voices. In Arabic, this matters deeply. Learners need to hear different accents, read different genres, encounter regional variation, and engage with perspectives that do not arrive pre-smoothed for textbook consumption. AI can prepare learners for authentic texts by building background knowledge or previewing vocabulary. It can help teachers create bridges. It should not become the main cultural source.

Privacy should be part of the same cautionary frame. Language learning data can be highly personal. Students may write about family, identity, religion, migration, political views, health, or personal experience. Speaking practice may involve voice data. Teachers should not assume that any available AI platform is appropriate for all student work. Institutions need clear policies about data retention, consent, approved tools, and alternatives for students who cannot or do not wish to

use a particular platform. Equity includes access, but it also includes the right not to surrender personal language data without understanding where it goes.

Cautions and Responsible Integration: A credible argument for generative AI in language education must name the risks clearly. Advocacy without caution reads as promotion. Caution without attention to affordance reads as refusal. The more responsible position recognizes both. Generative AI can expand access to input, feedback, and practice, but it also produces fluent errors, fabricated references, cultural bias, overreliance, privacy concerns, and inequitable performance across languages. These problems are not marginal. They arise from how large language models are trained and from how easily fluent output can be mistaken for reliable knowledge.

Integrity is the most visible concern in many classrooms. Students can use AI to write, translate, summarize, paraphrase, or revise with minimal effort. Traditional plagiarism policies are not always sufficient because the problem is not only copying. It is the outsourcing of thinking, composing, and language production. Barrot identifies academic integrity and overreliance as central pitfalls in ChatGPT-supported L2 writing (Barrot, J. S. 2023). Tseng and Warschauer argue that educators should respond not with simple prohibition, but with explicit instruction in responsible use, including corroboration and ethical incorporation of AI output (Tseng, W., & Warschauer, M. 2023).

Fabrication is a related but distinct problem. Generative AI can produce plausible references, explanations, examples, and cultural claims that are inaccurate or invented. In language instruction, this matters in several ways. A model might invent a grammar rule, mislabel a dialectal form, attribute a phrase to the wrong region, or provide a cultural explanation that sounds confident but has no basis in actual practice. The danger is greatest for learners who lack the expertise to check the output. Teachers must therefore treat verification as a language-learning skill. Students should learn to ask: What source supports this? Does a speaker of this variety actually say this? Does this phrase fit the context? Is the model correcting an error, or imposing a different register?

Overreliance should also be understood carefully. Learners have always used tools: dictionaries, grammar charts, tutors, spell checkers, corpora, textbooks, translation aids, and peer feedback. The goal is not tool-free learning. The goal is responsible mediation. Overreliance occurs when the tool

performs the core learning work in place of the learner. If the assignment is meant to develop presentational writing and the model writes the paragraph, the learner has bypassed the target skill. If the assignment is meant to develop interpretive reading and the model explains the text before the learner struggles with it, comprehension evidence disappears. If the assignment is meant to develop interpersonal repair and the model smooths over every misunderstanding, the learner may not practice repair at all.

Cultural bias is especially important for Arabic. Naous, Ryan, Ritter, and Xu's study of sixteen language models found systematic preference for Western over Arab entities when models operated in Arabic, with bias appearing across names, food, clothing, religion, and other culturally marked domains (Naous, T. et al. 2024). Their findings are not merely technical. They speak directly to language pedagogy. A model that generates Arabic sentences while centering Western cultural worlds may teach language detached from the communities it is meant to represent. This is not authenticity in any meaningful sense.

Bias interacts with dialect. Arabic learners often need to understand why a phrase belongs to one variety, why another belongs to a formal register, and why a third may be understood but socially inappropriate. A model trained on uneven data may blur these distinctions. The output may appear inclusive because it uses Arabic script and Arabic words, but it may erase regional specificity. It may offer a Gulf setting with generic MSA dialogue, a Moroccan market scene with Levantine expressions, or an Egyptian casual exchange that sounds like a textbook news broadcast. The teacher must identify these failures and decide whether to correct, replace, or use them as teaching moments.

Equity is the larger structural issue. Generative AI performs unevenly across languages because digital data are unevenly distributed. English and other high-resource languages benefit from larger and richer training data, more evaluation benchmarks, and stronger commercial incentives. Less commonly taught languages often receive thinner coverage, fewer specialized tools, and weaker research attention. Godwin-Jones notes that practical constraints and ethical concerns fall especially heavily on less privileged populations and languages (Godwin-Jones, R. 2024). In this sense, AI may widen rather than close gaps if institutions assume that one tool works equally well for all languages.

Responsible integration begins with policy, but it cannot end there. Teachers should define permissible and impermissible AI uses for each assignment. They should ask students to document how AI was used. They should design tasks that preserve evidence of learner thinking, including drafts, annotations, reflections, oral checks, and in-class performance. They should build verification into the learning process. In Arabic courses, they should require learners to label the target variety, reflect on register, and compare AI output with authentic examples. Programs should also develop local guidelines for AI use in assessment, especially when proficiency claims carry placement, certification, or programmatic consequences.

There is no need to present responsible integration as suspicion toward students. It is better framed as professional transparency. Learners deserve to know when AI can help them, when it can mislead them, and how to use it without surrendering their own development. Teachers deserve institutional support to evaluate tools before adopting them. Less commonly taught language programs deserve resources for language-specific vetting rather than generic AI workshops designed around English. Responsible use is not a slogan. It is a set of instructional practices, assessment safeguards, and professional judgments that must be built into the course.

Teacher education programs should also prepare future teachers to explain AI use to students in accessible language. Students do not need abstract warnings. They need concrete guidance: use AI to generate practice questions, but do not let it write your assignment; ask for feedback on one feature at a time; verify cultural explanations; record the prompts you used; be ready to explain your revisions; do not upload private information; do not assume that corrected Arabic is authentic Arabic. Such guidance respects students as developing users of technology and as developing language learners.

Professional development should include collaboration among language teachers, instructional technologists, assessment specialists, and native-speaker consultants when possible. No single teacher can evaluate every tool, dialect, task type, and policy issue alone. Shared repositories of vetted prompts and annotated model failures would be particularly useful for less commonly taught

languages. A prompt that works well for French or Spanish cannot simply be translated into Arabic and assumed to work. Arabic teachers need examples built from Arabic's own instructional realities.

Programs should also build shared evaluation checklists. For Arabic, such a checklist might ask: Does the output maintain the requested variety? Does it mix registers without explanation? Does it use culturally plausible names, places, routines, and politeness formulas? Does it introduce vocabulary beyond the target level? Does it overcorrect learner language toward a different register? Does it provide feedback in language the learner can understand? Does it encourage independent revision rather than supplying a finished answer? A checklist of this kind turns teacher intuition into a repeatable professional practice.

Teacher preparation should also include a repertoire of classroom routines. One useful routine is 'AI output as object of analysis.' The teacher presents a model-generated Arabic dialogue and asks students to identify what is appropriate, what is too formal, what belongs to a different variety, and what could be revised. Another routine is 'feedback triage.' Students classify AI feedback into three categories: accept, reject, and investigate. A third routine is 'human comparison.' Students compare AI feedback with teacher feedback or peer feedback and discuss what each source noticed. These routines make AI use visible and teachable.

Implications for Teacher Preparation and Practice : None of the three commitments described in this article can be realized without educator capacity. Personalization requires teachers who can define the learner's target, evaluate AI-generated materials, and adapt them responsibly. Practice requires teachers who can design tasks that preserve learner agency and prevent passive dependence. Proficiency requires teachers who can align AI-supported activity with communicative standards, collect valid evidence, and interpret performance across modes. AI literacy is therefore not a technical add-on to language teacher education. It is becoming part of professional language pedagogy.

The first implication is that language teachers need a working understanding of how generative AI fails. They do not need to become computer scientists, but they do need to know that model output is probabilistic, that fluency does not guarantee accuracy, that hallucinated citations can look

real, and that cultural and linguistic bias may be hidden beneath polished prose. Ng and colleagues describe AI literacy as involving knowledge of AI concepts, use, evaluation, and ethical awareness (Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. 2021). For language teachers, the evaluative dimension is especially important because errors may be subtle: a misplaced register, a culturally odd scenario, an unnatural collocation, or an overcorrection of dialectal usage.

The second implication concerns tool selection. Teachers and programs should not choose AI tools only because they are popular or powerful in English. They should test them with the tasks they will actually use. For Arabic, that means testing Modern Standard Arabic, target dialects, code-switching, register control, cultural scenarios, grammatical feedback, and learner-level adaptation. A tool that performs well in a general demonstration may fail in a Moroccan Arabic role-play, an advanced media Arabic summary, or a heritage learner writing task. Tool selection should be empirical, local, and revisited regularly as models change.

A practical vetting routine can be simple but disciplined. First, identify the target task: for example, Intermediate Mid interpersonal practice in Levantine Arabic. Second, generate model output using a carefully written prompt. Third, evaluate the output for language variety, level, cultural fit, pragmatic appropriateness, and usefulness for the learner. Fourth, revise the prompt and test again. Fifth, decide whether the output is ready for learners, needs teacher editing, or should be used only as a comparison exercise. This process takes time, but it is far less risky than assuming that an AI system understands the pedagogical and sociolinguistic conditions of the course.

The third implication is assignment design. Teachers should create AI-supported tasks that keep the teacher central and the learner active. A writing assignment, for instance, might allow AI feedback only after a first draft is completed independently. Students would then submit the first draft, the AI feedback, a revision memo, and the final version. They would explain which suggestions they accepted, which they rejected, and why. A speaking assignment might require students to complete an AI role-play, identify moments where the model's Arabic sounded too formal or culturally odd, and then perform a revised version with a classmate. In both cases, AI supports practice, but the learner remains responsible for analysis and performance.

The fourth implication concerns assessment. Programs need clear boundaries between formative and summative uses of AI. Formative use can be broad. Learners may practice with chatbots, ask for examples, receive feedback, and revise. Summative assessment requires stricter control. If the goal is to evaluate what learners can do, the task must preserve independent performance. AI may be part of preparation, but it should not generate the performance being assessed unless the assessment objective explicitly concerns AI-mediated communication. Even then, the rubric must distinguish language proficiency from tool management.

The fifth implication is professional development for less commonly taught languages. Generic AI training often assumes English or high-resource language contexts. Arabic teachers need training that addresses diglossia, dialectal authenticity, transliteration, script, morphological complexity, register, and cultural bias. They also need shared repositories of vetted prompts, sample outputs, evaluation checklists, and corrected examples. A single teacher should not have to discover every model failure alone. Departments, language centers, and professional organizations can support collective vetting and exchange.

The final implication is ethical. Teachers should talk with learners about why AI is being used and what kind of language development it is meant to support. Students should understand that using AI responsibly does not mean hiding its role. It means using it in ways that make learning more visible. They should know how to disclose assistance, verify claims, protect privacy, and avoid outsourcing the core work of learning. This ethical conversation is part of language education because language learning is not only skill acquisition. It is participation in communities of meaning.

Conclusion: Personalization, practice, and proficiency are sometimes treated as separate uses of generative AI. They are better understood as one integrated agenda. Personalization without proficiency alignment can produce attractive material that leads nowhere developmentally. Practice without teacher-designed structure can produce dependence rather than autonomy. Proficiency without personalized support and abundant practice remains a standard on paper rather than a pathway for learners. The three commitments need each other, and all three require teacher judgment.

The argument of this article has been deliberately measured. Generative AI is not merely another CALL tool, because it changes the scale and flexibility of language generation available to teachers and learners. It can produce differentiated input, simulate interaction, provide feedback, and support revision in ways that earlier tools could not. Those affordances are real. They matter for programs where teacher time is scarce, feedback demands are heavy, and learners need more opportunities to use the language than class time allows.

The risks are equally real. Model output can be fluent and wrong. It can fabricate sources, flatten cultures, blur dialects, reward surface correctness, and hide the difference between learner performance and machine assistance. The risks become sharper in less commonly taught languages, where training data are thinner and evaluation research is less developed. Arabic makes this visible. Its diglossic structure, dialectal range, cultural density, and uneven representation in model training create conditions under which uncurated AI output can easily become misleading. Arabic is therefore not a side case. It is the stress test for the entire argument.

If generative AI can be integrated responsibly in Arabic, with attention to variety, register, culture, proficiency, and assessment validity, then the three-P framework has broad value. If it cannot, then the limits exposed by Arabic must temper claims made for AI in more comfortable language contexts. The honest position is measured optimism: real affordances, real risks, and an evidence base still maturing. The teacher-in-the-loop is not a temporary safeguard until better models arrive. It is the pedagogical condition that makes AI useful in the first place.

Language teaching is ultimately about helping learners enter human worlds through language. Generative AI can assist that work, but it cannot replace the human judgment that gives the work its direction. The future of AI in language instruction should therefore be neither tool-driven nor fear-driven. It should be designed.

The article's practical claim is therefore modest but consequential. Teachers should not wait for perfect tools before experimenting, but they should not mistake experimentation for evidence. They should use AI where it clearly extends practice, scaffolding, and feedback. They should slow down where authenticity, assessment, or cultural meaning is at stake. They should document what works

and what fails, especially in less commonly taught languages where published evidence remains thin. In that disciplined middle ground, generative AI becomes neither a threat to language teaching nor a substitute for it. It becomes a powerful, imperfect resource whose educational value depends on the quality of the pedagogy surrounding it.

The next phase of research should follow the same logic. The field needs fewer general claims that AI improves language learning and more careful studies of which learners, which languages, which tools, which tasks, and which instructional designs produce durable gains. Arabic-specific classroom research is especially needed. Studies should examine not only whether students like AI or whether their drafts improve immediately, but whether AI-supported practice transfers to human interaction, whether dialectal competence develops, whether learners become better at evaluating feedback, and whether proficiency gains persist beyond short interventions.

References:

1. American Council on the Teaching of Foreign Languages. (2012). ACTFL proficiency guidelines 2012. Alexandria, VA: Author. https://www.actfl.org/educator-resources/actfl-proficiency-guidelines

2. Barrot, J. S. (2023). Using ChatGPT for second language writing: Pitfalls and potentials. Assessing Writing, 57, 100745. https://doi.org/10.1016/j.asw.2023.100745

3. Chen, X., Zou, D., Cheng, G., & Xie, H. (2021). Artificial intelligence-assisted personalized language learning: Systematic review and co-citation analysis. In 2021 IEEE International Conference on Advanced Learning Technologies (ICALT) (pp. 241-245). IEEE. https://doi.org/10.1109/ICALT52272.2021.00079

4. Chiu, T. K. F. (2024). The impact of generative AI (GenAI) on practices, policies and research direction in education: A case of ChatGPT and Midjourney. Interactive Learning Environments, 32(10), 6187-6203. https://doi.org/10.1080/10494820.2023.2253861

5. Dai, D. W., Suzuki, S., & Chen, G. (2025). Generative AI for professional communication training in intercultural contexts: Where are we now and where are we heading? Applied Linguistics Review, 16(2), 763-774. https://doi.org/10.1515/applirev-2024-0184

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

7. Godwin-Jones, R. (2024). Distributed agency in language learning and teaching through generative AI. Language Learning & Technology, 28(2), 5-31. https://doi.org/10125/73570

8. Interagency Language Roundtable. (n.d.). ILR skill level descriptions. https://www.govtilr.org/Skills/ILRscale2.htm

9. Khondaker, M. T. I., Waheed, A., Nagoudi, E. M. B., & Abdul-Mageed, M. (2023). GPTAraEval: A comprehensive evaluation of ChatGPT on Arabic NLP. In H. Bouamor, J. Pino, & K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (pp. 220-247). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.16

10. Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for language teaching and learning. RELC Journal, 54(2), 537-550. https://doi.org/10.1177/00336882231162868

11. Li, M., Wang, Y., & Yang, X. (2025). Can generative AI chatbots promote second language acquisition? A meta-analysis. Journal of Computer Assisted Learning, 41(4), e70060. https://doi.org/10.1111/jcal.70060

12. Lusin, N., Peterson, T., Sulewski, C., & Zafer, R. (2023). Enrollments in languages other than English in United States institutions of higher education, Fall 2021. Modern Language Association. https://www.mla.org/content/download/191324/file/Enrollments-in-Languages-Other-Than-English-in-US-Institutions-of-Higher-Education-Fall-2021.pdf

13. Lyu, B., Lai, C., & Guo, J. (2025). Effectiveness of chatbots in improving language learning: A meta-analysis of comparative studies. International Journal of Applied Linguistics, 35(2), 834-851. https://doi.org/10.1111/ijal.12668

14. Naous, T., Ryan, M. J., Ritter, A., & Xu, W. (2024). Having beer after prayer? Measuring cultural bias in large language models. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 16366-16393). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.862

15. Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. Computers and Education: Artificial Intelligence, 2, 100041. https://doi.org/10.1016/j.caeai.2021.100041

16. Tseng, W., & Warschauer, M. (2023). AI-writing tools in education: If you can't beat them, join them. Journal of China Computer-Assisted Language Learning, 3(2), 258-262. https://doi.org/10.1515/jccall-2023-0008

17. Yan, D. (2023). Impact of ChatGPT on learners in an L2 writing practicum: An exploratory investigation. Education and Information Technologies, 28, 13943-13967. https://doi.org/10.1007/s10639-023-11742-4